Shape Detection and Mediation Analysis using Semi-parametric Shape-Restricted Regression Spline with Applications

by

Qing Yin

Bachelor of Science, Stony Brook University, 2015

Master of Arts, Columbia University, 2017

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF BIOSTATISTICS

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Qing Yin

It was defended on

July 2nd 2021

and approved by

Jong H. Jeong, PhD, Professor and Interim Chair, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Jennifer J. Adibi, MPH, ScD, Assistant Professor, Department of Epidemiology and Department of Obstetrics/Gynecology and Reproductive Sciences, Graduate School of Public Health and School of Medicine, University of Pittsburgh

Jeanine M. Buchanich, MEd, MPH, PhD, Research Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Gong Tang, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Shape Detection and Mediation Analysis using Semi-parametric Shape-Restricted Regression Spline with Applications

Qing Yin, PhD

University of Pittsburgh, 2021

Abstract

Linear models are widely used in the field of epidemiology to model the relationship between two continuous variables, such as circulating levels of the placental hormone hCG and infant genital size. When researchers suspect curvilinear relationship exists, some non-parametric techniques can be used to model the relationship. By applying nonparametric techniques, researchers can relax the linearity assumption and capture scientifically meaningful or appropriate shapes.

In the first part of the dissertation, a shape detection method based on regression splines is developed. The proposed method can help researchers select the most suitable shape to describe their data among increasing, decreasing, convex and concave shapes. Specifically, we develop a technique based on mixed effects regression spline to analyze hormonal data, but the method is general enough to be applied to other similar problems.

Analyzing the association between two variables is usually the first step of some research project. Researchers also want to explore the causal relationship between an exposure and a potential outcome caused by the exposure. In many cases, the exposure may not directly lead to the outcome, but instead, it induces the outcome through a process. Mediation analysis is designed to explain the causal relationship between the exposure and the outcome by examining the intermediate stage, which helps researchers understand the pathway whereby the exposure affects the outcome.

In the second part of the dissertation, we develop a method to analytically estimate the direct and indirect effects when we have some prior knowledge on the relationship between the mediator and the outcome (increasing, decreasing, convex or concave). In order to make suitable inferences, the asymptotic confidence intervals of those effects are obtained via delta

method.

Public health significance: The shape detection technique can help researchers make judgements on the potential relationship between the exposure and the outcome while controlling for confounders. With such judgements, researchers can avoid the bias caused by model misspecification when building models. The regression-based mediation analysis within the shape-restricted framework offers researchers a flexible and efficient approach to perform the causal inference. The method helps researchers estimate causal effects using reasonable models.

Table of Contents

List of Tables

List of Figures

Preface

I would like to thank Dr. Shyamal D. Peddada and Dr. Jong-Hyeon Jeong for their primary guidance on my research. Dr. Peddada and Dr. Jeong always provide me with innovative ideas to solve difficult statistical problems and offer me help when I have questions. I would like to acknowledge my funding sources from Dr. Jennifer J. Adibi (3-year Graduate Student Researcher), Dr. Ajay D. Wasan (2-year Graduate Student Researcher), Dr. Howard B. Degenholtz (2-semester Graduate Student Researcher) and Department of Biostatistics (3-semester Teaching Assistant). I would like to express my special gratitude to Dr. Adibi for her help on my research in the aspects of epidemiology and biology. I also want to thank my committee members, Dr. Jeanine M. Buchanich and Dr. Gong Tang, for their suggestions on my dissertation. Finally, I would like to thank my parents, my relatives and my friends for their continued support.

1.0 Introduction

1.1 Motivation

The developmental origins of health and disease is an important research area in epidemiology. In the past, researchers used to examine the direct effect of maternal exposures and infant short-term or long-term outcomes. For example, a study showed that the infants whose mothers suffered from famine in the first two trimesters tended to have lower birth weight (Lumey, 1992); some studies showed that maternal nutrition associated with offspring's metabolic and cardiovascular functions (Dörner, 1973). The analysis of direct effect is usually intuitive and easy to interpret, but in some cases, it is not enough to exhibit the underlying molecular biology. The placenta, as an interface between the mother and the fetus, plays an important role in fetal growth. Its functions, including providing the fetus with necessary nutrients, removing various waste products and preventing the fetus from many environmental toxins, are always of interest and can be reflected by placental biomarkers, such as circulating placental-fetal hormones. Because of the special role of the placenta, the placental function is regarded as a mediator between maternal environment and fetal growth. Some studies attempted to address the effect of placenta on fetal growth during pregnancy - for example, a study modeled the associations between circulating levels of the placental hormone human chorionic gonadotropin (hCG) and infant genital size, a marker of future fertility, using linear regression (Adibi et al., 2015), and another study modeled the associations between free thyroxine concentrations and children's IQ using cubic splines (Korevaar et al., 2016), while some other studies attempted to address the effect of environmental toxins on the placental functions - for example, a study analyzed the association between maternal levels of plasticizers called phthalates and placental hormone hCG (Adibi et al., 2015), and another study analyzed the association between maternal phthalates and thyroid hormones (Huang et al., 2007). In order to systematically understand the associations among the maternal exposures, the placental mediators and the infant outcomes, we should introduce the mediation analysis to analyze the direct and indirect effects simultaneously.

Before modeling the association of maternal exposure and infant outcome by way of the hormonal mediator, it is crucial to first establish the hormonal mediator-outcome association in a reasonable manner. Researchers often find it challenging to model complex biologic relationships using observational data. Standard models, such as the linear or polynomial regression models, are often used to model the relationship between two variables while controlling for potential confounders for simplicity of implementation and interpretation, and such models might be reasonable in some cases. However, this could possibly result in model misspecification and incorrect inferences. Since *a priori* one does not know shape of the model except that it might be monotonic or convex/concave, there is a need for a nonparametric shape-restricted methodology to simultaneously evaluate multiple shapes that might capture and best represent the underlying biology.

Once researchers have prior knowledge on the association between the mediator and the outcome, they can choose a more appropriate approach to perform mediation analysis. Mediation analysis is designed to explain the causal relationship between the exposure and the outcome by examining the intermediate stage. It helps researchers understand the pathway whereby the exposure affects the outcome, investigate the specific process and refine the intervention strategies. The regression-based mediation analysis has been formulated and developed in the last decade (VanderWeele, 2015). In the circumstance where the mediator and the outcome are continuous, the classical linear regression is introduced to build the models and perform the analysis. However, the relationship between the mediator and the outcome may not be linear in many cases. For example, the relationship between free thyroxine concentrations and children's IQ is shown to be curvilinear in one study (Korevaar et al., 2016). In such cases, using linear models will result in model misspecification and introduce bias. Therefore, the approach of applying generalized additive model is discussed by Imai et al. as a remedial measure, where the direct and indirect effects along with their confidence intervals are estimated using simulations (Imai et al., 2010). Estimating effects using simulations may not be accurate, and if researchers have prior knowledge on the association between the mediator and the outcome (increasing, decreasing, convex and concave), applying shape-restricted regression spline technique may be a better choice.

1.2 M-Spline, I-Spline and C-Spline

The spline is a drawing instrument in real life, which can be used to draw irregular curves in road, architectural or other designs. When drawing the curve, draftsmen fix several stakes and let the spline pass through those stakes to obtain the desirable shape. The spline in statistics, a linear combination of piecewise-defined polynomials, serves the similar purpose. Passing through the designated points smoothly, the spline can well approximate the complex shape underlying the data. The piecewise-defined polynomials, usually called the spline basis functions, are building blocks of splines. There are many choices of spline basis functions, including truncated power series basis, B-spline basis, etc. The shape-restricted regression spline technique discussed in the entire dissertation is based on two types of spline basis functions, I-spline basis and C-spline basis, which are derived from M-spline basis.

M-spline basis functions, $M_i(x|k,t)$, were defined by Curry and Schoenberg, who also derived their properties (Curry and Schoenberg, 1966). The M-splines were expressed using the divided difference of truncated power functions in the original literature, but for computational convenience, they can be specified using the recursive form as follows:

Order $k = 1$:

$$M_i(x|1,t) = \begin{cases} \frac{1}{t_{i+1}-t_i}, & \text{if } t_i \le x < t_{i+1} \\ 0, & \text{otherwise} \end{cases},$$

Order $k > 1$:

$$M_i(x|k,t) = \begin{cases} \frac{k[(x-t_i)M_i(x|k-1,t)+(t_{i+k}-x)M_{i+1}(x|k-1,t)]}{(k-1)(t_{i+k}-t_i)}, & \text{if } t_i \le x < t_{i+k} \\ 0, & \text{otherwise} \end{cases},$$

where $t = \{t_1, t_2, ..., t_{n+k}\}$ is the knot sequence, n is the number of free parameters that specify the spline function having the specified continuity characteristics, and k is the order

of the basis functions. For example, if $k = 2$, the M-splines $M_i(x|2,t)$ will be

$$M_i(x|2,t) = \begin{cases} 0, & \text{if } x < t_i \\ \frac{2(x-t_i)}{(t_{i+2}-t_i)(t_{i+1}-t_i)}, & \text{if } t_i \leq x < t_{i+1} \\ \frac{2(t_{i+2}-x)}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})}, & \text{if } t_{i+1} \leq x < t_{i+2} \\ 0, & \text{if } x \geq t_{i+2} \end{cases}.$$

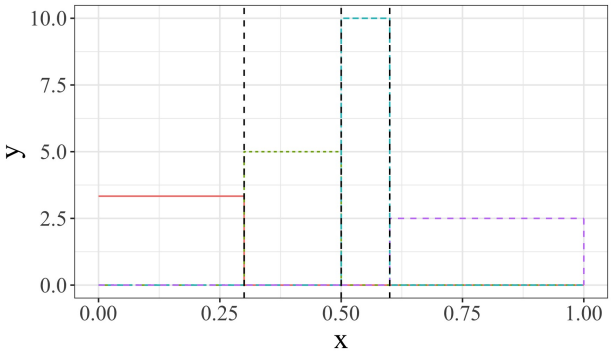
The M-splines of order $k = 1$ with knot sequence $t = \{0, 0.3, 0.5, 0.6, 1\}$, order $k = 2$ with knot sequence $t = \{0, 0, 0.3, 0.5, 0.6, 1, 1\}$, and order $k = 3$ with the knot sequence $t = \{0, 0, 0, 0.3, 0.5, 0.6, 1, 1, 1\}$ are shown in Figure 1.2.1. The M-splines are of interest in deriving the I-splines and C-splines because of the property that $M_i(x|k,t) > 0$ only when $t_i \leq x < t_{i+k}$ and $M_i(x|k,t) = 0$ otherwise.

Ramsay defined the I-spline basis functions as $I_i(x|k,t) = \int_L^x M_i(u|k,t)du$ (Ramsay, 1988). Because each M-spline basis function $M_i(x|k,t)$ is a piecewise polynomial of degree $k-1$, each I-spline basis function $I_i(x|k,t)$ will be a piecewise polynomial of degree k. We shall use the term "order k" to refer to M-splines with degree $k-1$ or the associated I-splines with degree k. The I-splines of order $k = 1$ with knot sequence $t = \{0, 0.3, 0.5, 0.6, 1\}$, order $k = 2$ with knot sequence $t = \{0, 0, 0.3, 0.5, 0.6, 1, 1\}$, and order $k = 3$ with the knot sequence $t = \{0, 0, 0, 0.3, 0.5, 0.6, 1, 1, 1\}$ are shown in Figure 1.2.2. The quadratic I-splines $I_i(x|2,t)$ are obtained by integrating the M-splines of order 2 and can be expressed as follows:

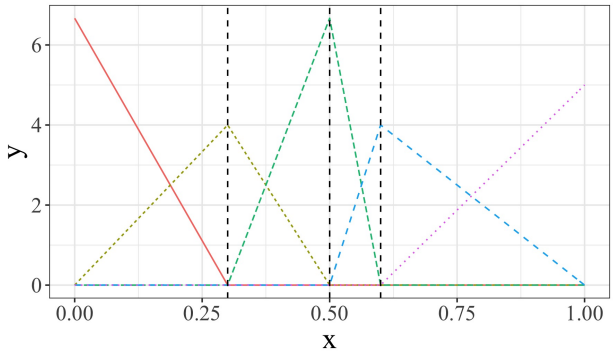
$$I_i(x|2,t) = \begin{cases} 0, & \text{if } x < t_i \\ \frac{(x-t_i)^2}{(t_{i+2}-t_i)(t_{i+1}-t_i)}, & \text{if } t_i \leq x < t_{i+1} \\ 1 - \frac{(t_{i+2}-x)^2}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})}, & \text{if } t_{i+1} \leq x < t_{i+2} \\ 1, & \text{if } x \geq t_{i+2} \end{cases}.$$

The quadratic I-splines are of interest in shape-restricted regression spline because of the following fact:

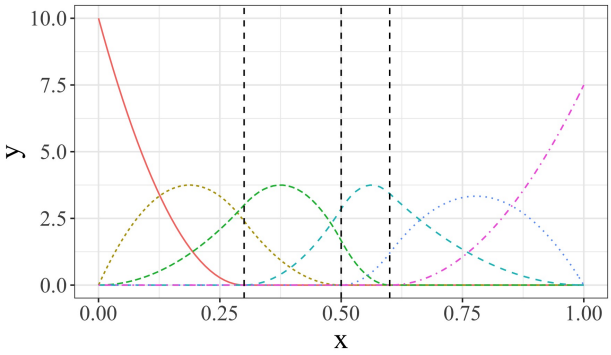
Fact 1.2.1. *A linear combination of quadratic I-spline basis functions is increasing if and only if the coefficients are positive.*



(a) M-splines of order $k = 1$



(b) M-splines of order $k = 2$



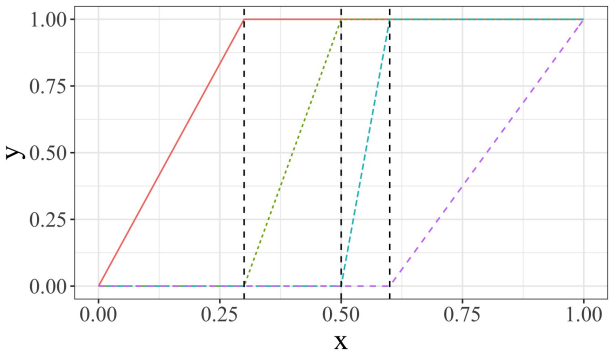
(c) M-splines of order $k = 3$

Figure 1.2.1: Plots of M-splines with inner knots placed at 0.3, 0.5 and 0.6 and orders 1, 2, and 3

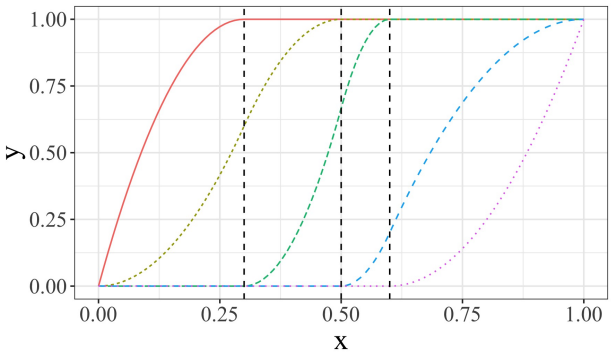
Proof. $\Rightarrow$ Suppose that a linear combination of quadratic I-spline basis functions is increasing and suppose that some coefficient of quadratic I-spline basis function is non-positive. There is a fact that at each knot, only one quadratic I-spline basis function has positive slope while other quadratic I-spline basis functions have zero slopes. Since we suppose that some coefficient of quadratic I-spline basis function is non-positive, at the knot where the quadratic I-spline basis function has the non-positive coefficient, the curve will be non-increasing, which contradicts the condition that a linear combination of quadratic I-spline basis functions is increasing.

$\Leftarrow$ Suppose that the coefficients are positive. Because I-splines are derived by integrating M-splines, which are non-negative, the derivative of the linear combination of quadratic I-spline basis functions with positive coefficients is positive. Therefore, the linear combination of quadratic I-spline basis functions with positive coefficients is increasing. $\square$

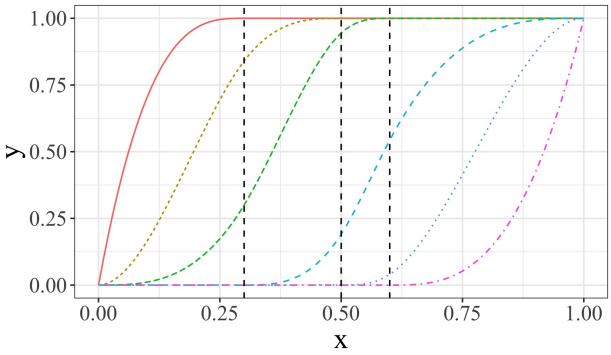
Therefore, for the curve to be non-decreasing, all coefficients of the quadratic I-spline basis functions must be non-negative, and for the curve to be non-increasing, all coefficients of the quadratic I-spline basis functions must be non-positive.



(a) I-splines of order $k = 1$



(b) I-splines of order $k = 2$



(c) I-splines of order $k = 3$

Figure 1.2.2: Plots of I-splines with inner knots placed at 0.3, 0.5 and 0.6 and orders 1, 2, and 3

Meyer defined the C-spline basis functions as $C_i(x|k,t) = \int_L^x I_i(u|k,t)du$ (Meyer, 2008). Because each I-spline basis function $I_i(x|k,t)$ is a piecewise polynomial of degree k, each

C-spline basis function $C_i(x|k,t)$ will be a piecewise polynomial of degree $k+1$. We shall use the term "order k" to refer to I-splines with degree k or the associated C-splines with degree $k+1$. The C-splines of order $k=1$ with knot sequence $t=\{0,0.3,0.5,0.6,1\}$, order $k=2$ with knot sequence $t=\{0,0,0.3,0.5,0.6,1,1\}$, and order $k=3$ with the knot sequence $t=\{0,0,0,0.3,0.5,0.6,1,1,1\}$ are shown in Figure 1.2.3. The cubic C-splines $C_i(x|2,t)$ are obtained by integrating the quadratic I-splines and can be expressed as follows:

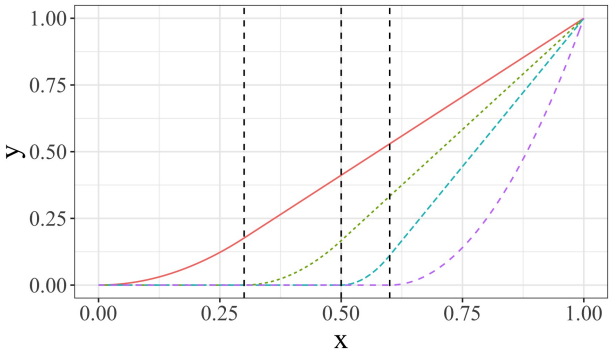
$$C_i(x|2,t)=\begin{cases}0, & \text{if } x<t_i\\ \frac{(x-t_i)^3}{3(t_{i+2}-t_i)(t_{i+1}-t_i)}, & \text{if } t_i\le x<t_{i+1}\\ x-\frac{t_i+t_{i+1}+t_{i+2}}{3}+\frac{(t_{i+2}-x)^3}{3(t_{i+2}-t_i)(t_{i+2}-t_{i+1})}, & \text{if } t_{i+1}\le x<t_{i+2}\\ x-\frac{t_i+t_{i+1}+t_{i+2}}{3}, & \text{if } x\ge t_{i+2}\end{cases}.$$

The cubic C-splines are of interest in shape-restricted regression spline because of the following fact:

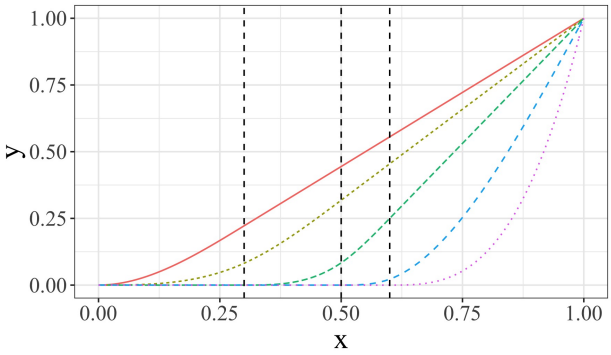
Fact 1.2.2. *A linear combination of cubic C-spline basis functions is convex if and only if the coefficients are positive.*

Proof. $\Rightarrow$ Suppose that a linear combination of cubic C-spline basis functions is convex and suppose that some coefficient of cubic C-spline basis function is non-positive. There is a fact that at each knot, only one cubic C-spline basis function is quadratic while other cubic C-spline basis functions are either linear or zero. Since we suppose that some coefficient of cubic C-spline basis function is non-positive, at the knot where the cubic C-spline basis function has the non-positive coefficient, the curve will be non-convex, which contradicts the condition that a linear combination of cubic C-spline basis functions is convex.
$\Leftarrow$ Suppose that the coefficients are positive. Because C-splines are derived by double integrating M-splines, which are non-negative, the second derivative of the linear combination of cubic C-spline basis functions with positive coefficients is positive. □

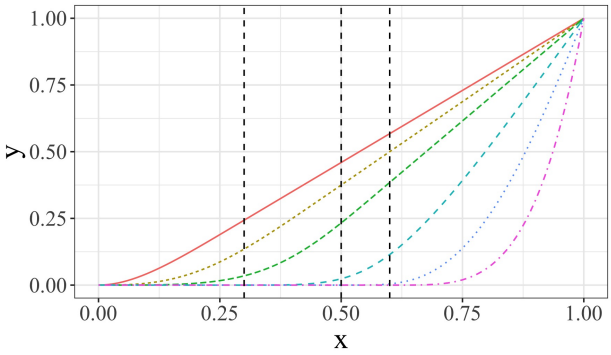
Therefore, for the curve to be convex, all coefficients of the cubic C-spline bases must be non-negative, and for the curve to be concave, all coefficients of the cubic C-spline bases must be non-positive.



(a) C-splines of order $k = 1$



(b) C-splines of order $k = 2$



(c) C-splines of order $k = 3$

Figure 1.2.3: Plots of C-splines with inner knots placed at 0.3, 0.5 and 0.6 and orders 1, 2, and 3

1.3 Constrained Statistical Inference

Constrained statistical inference refers to the procedure of parameter estimation and hypothesis testing in a subset of Euclidean space. One simple example is to test the hypotheses: $H_0 : \mu = 0$ vs. $H_a : \mu > 0$. When μ is known to be non-negative, one-sided test will be more powerful than two-sided test. Another example is to test the hypotheses: $H_0 : \mu_1 = \mu_2 = \mu_3$ vs. $H_a : \mu_1 \leq \mu_2 \leq \mu_3$ and μ_1, μ_2, μ_3 are not all equal. The fact incorporated into the above test is that the parameter space is a simple order space, i.e..

$\mu \in M = \{\mu \in R^3 : \mu_1 \leq \mu_2 \leq \mu_3\}$. Constrained statistical inference plays an important role in the entire dissertation because of Fact 1.2.1 and Fact 1.2.2 described in Section 1.2. For example, to test the null hypothesis that the curve is flat against the alternative hypothesis that the curve is increasing, we should formulate the hypotheses testing problem as: $H_0 : \beta = 0$ vs. $H_a : \beta \geq 0$, where the parameter space is a non-negative orthant, i.e., $\beta \in B = \{\beta \in R^p : \beta \geq 0\}$.

In linear model, an important procedure is to project the vector of observations y onto a linear space $C(X)$ through a projection matrix $X(X^TX)^{-1}X$. The idea of projection is also crucial in constrained statistical inference, where the unconstrained estimator is usually projected onto a convex cone. The definitions of convex set and cone are given below in Definition 1.3.1 and Definition 1.3.2.

Definition 1.3.1 (Convex set)**.** *A set $A \subset R^p$ is said to be convex if $\{\lambda x + (1-\lambda)y\} \in A$ whenever $x, y \in A$ and $0 < \lambda < 1$.*

Definition 1.3.2 (Cone)**.** *A set A is said to be a cone with vertex x_0 if $x_0 + k(x - x_0) \in A$ for every $x \in A$ and $k \geq 0$. If the vertex is the origin, then we shall simply refer to it as a cone.*

In simple words, a set A is a convex set if the line segment joining x and y is in A whenever the points x and y are in A, and a cone is a set that consists of infinite straight lines starting from the origin. The concept of convex set is important because if the space that a point is projected onto is a closed convex set, then the projection will exist uniquely and can be characterized by the angle between the projection line and the convex set. Specifically, when the projection of unconstrained estimator onto the convex set is treated as origin, the angle between the unconstrained estimator and the true value in the convex set must be obtuse. The non-negative orthant mentioned above is a special convex cone, called polyhedral cone, which is defined below in Definition 1.3.3.

Definition 1.3.3 (Polyhedral cone)**.** *Let $a_1, ..., a_q$ be q points in R^p and $P = \{x \in R^p : a_i^T x \geq 0 \forall i\}$. Then P is a closed convex cone and it is called a polyhedral cone. With the $p \times q$ matrix A defined as $[a_1, ..., a_q]$, we may express P as $\{x \in R^p : A^T x \geq 0\}$*

A polyhedral cone P is a convex set because suppose $x_1, x_2 \in P$, then given $\theta \in [0, 1]$,

$\theta x_1 + (1-\theta)x_2 \in P$. The non-negative orthant is a polyhedral cone because it can be expressed as $\{x \in R^p : A^T x \geq 0\}$, where A is an identity matrix.

1.4 Regression-Based Mediation Analysis

The regression-based mediation analysis is a parametric approach to calculate the direct and indirect effects, which is formulated within the potential outcome framework for causal inference. Applying the regression-based approach, the estimates of the effects can be obtained through combining the estimated regression coefficients and the associated confidence intervals can be calculated using delta method.

Within the potential outcome framework, in order to obtain an estimable quantity, we have to make two assumptions: the consistency assumption and the no-unmeasured-confounding assumption. We denote A as the exposure for an individual, C as the confounding variables, Y as the outcome, and Y_a as the outcome of an individual whose exposure A were set to a. The consistency assumption states that for an individual with actual exposure $A = a$, the actual outcome Y is Y_a. Thus, with a binary exposure A, the causal effect for an individual can be defined as $Y_1 - Y_0$. The no-unmeasured-confounding assumption states that given all confounding variables C, the potential outcome Y_a is independent of the exposure A, which is denoted by $Y_a \perp A|C$. This assumption makes potential outcomes comparable across the exposure groups. With these two assumptions, the average causal effect for a population given all confounding variables $E[Y_1 - Y_0|C]$ can be expressed as $E[Y|A = 1, C] - E[Y|A = 0, C]$, which is possible to be estimated using the observed data.

Before proceeding to build linear models, we need to define three quantities, the controlled direct effect, the natural direct effect and the natural indirect effect. In addition to the notations given above, we denote M as the mediator, M_a as the mediator of an individual whose exposure A were set to a, and Y_{am} as the outcome of an individual whose exposure A were set to a and mediator M were set to m. The controlled direct effect (CDE) is defined by $Y_{am} - Y_{a^*m}$, which measures the direct effect of A on Y when the mediator M is controlled at m. The natural direct effect (NDE) is defined by $Y_{aM_{a^*}} - Y_{a^*M_{a^*}}$, which measures the direct

effect of A on Y after keeping the mediator M for each individual at the level it naturally would have taken in the circumstance of $A = a^*$. The natural indirect effect (NIE) is defined by $Y_{aM_a} - Y_{aM_{a^*}}$, which compares what would have happened if the mediator M were set to what it would have been in the circumstance of $A = a$ vs. what would have happened if the mediator M were set to what it would have been in the circumstance of $A = a^*$ after setting the exposure A to some level a. In order to identify the CDE, we need the assumptions that there is no unmeasured exposure-outcome confounding, i.e., $Y_{am} \perp A|C$, and there is no unmeasured mediator-outcome confounding, i.e., $Y_{am} \perp M|A, C$. In order to identify the NDE and the NIE, besides the two assumptions above, we need two additional assumptions: there is no unmeasured exposure-mediator confounding, i.e., $M_a \perp A|C$, and there is no mediator-outcome confounding affected by exposure, i.e., $Y_{am} \perp M_{a^*}|C$.

Then, we build two linear models with continuous M and Y, where the exposure-outcome model is $Y = \theta_0+\theta_1 A+\theta_2 M+\theta_3 AM+\theta_4 C+\epsilon_1$ and the exposure-mediator model is $M = \beta_0+\beta_1 A+\beta_2 C+\epsilon_2$. If all assumptions hold and the models are correctly specified, the expected controlled direct effect, the expected natural direct effect and the expected natural indirect effect, conditioning on $C = c$, are given by $(\theta_1+\theta_3 m)(a-a^*)$, $[\theta_1+\theta_3(\beta_0+\beta_1 a^*+\beta_2 c)](a-a^*)$ and $(\theta_2\beta_1+\theta_3\beta_1 a)(a-a^*)$, respectively, and their standard errors can be obtained via delta method. The details on derivations of these effects can be found in book by VanderWeele (2015).

2.0 Shape Detection using Semi-parametric Shape-Restricted Mixed Effects Regression Spline

2.1 Introduction

The linear regression is a standard approach to model the relationship between two continuous variables and being widely used in many research fields, such as medicine and epidemiology. However, there is a strong assumption associated with linear regression, that is, the relationship between the dependent variable and the independent variable must be linear in expectation. This assumption may be true in some cases but not always. For example, one study showed that the association between maternal free thyroxine concentrations and child IQ is inverted U-shaped (Korevaar et al., 2016). Therefore, there is a need to introduce other techniques to model the curvilinear relationship, which include local polynomial regression, kernel regression, splines, etc. By applying these nonparametric techniques, researchers can relax the linearity assumption and build more flexible models.

In this chapter, we develop a shape detection method based on the regression splines technique to help researchers select the most suitable shape to describe their data among increasing, decreasing, convex and concave shapes. The development of this method is motivated by the interest of examining the associations between several placental-fetal hormones and birth weight using population-level prenatal serum screening data for the State of California, where the associations are suspected to fall into the shape categories described above. This method is not suitable in some other cases, such as the case of cyclical or rhythmic shape (Larriba et al., 2016). The splines that we adopt to build the models are I-splines and C-splines, and the proposed method is based on the properties of these two types of splines. Ramsay (1988) defined the I-splines as $I_i(x|k,t) = \int_L^x M_i(u|k,t)du$, where $M_i(x|k,t)$ is the M-splines. If we linearly combine quadratic I-splines with non-negative/non-positive coefficients, then we should obtain a non-decreasing/non-increasing curve. Meyer (2008) defined the C-splines as $C_i(x|k,t) = \int_L^x I_i(u|k,t)du$. If we apply the linear combination of cubic C-splines with non-negative/non-positive coefficients, then we should obtain a con-

vex/concave curve. Meyer (2018) also described the parameter estimation procedure for constrained partial regression splines $Y = f(T) + X^T\beta + \epsilon$. This chapter aims to extend the constrained partial regression splines to constrained partial mixed effects regression splines, derive a test statistic to test the null hypothesis of constant function against the alternative that there is an underlying shape and apply Holm-Bonferroni method (Holm, 1979) to classify the underlying shape into a reasonable category.

2.2 Methodology

2.2.1 Model setup

2.2.1.1 Linear mixed model

A linear mixed model is of the form

$$y = X\beta + Zb + \epsilon, \tag{2.2.1}$$

where $b \sim N(0, \tilde{D})$ and $\epsilon \sim N(0, R)$. In model (2.2.1), $y = (y_1^T, ..., y_c^T)^T$ is the response vector, $X = (X_1^T, ..., X_c^T)^T$ is the fixed effects covariate matrix, β is the fixed effects vector, $Z = diag(Z_1, ..., Z_c)$ is the random effects covariate matrix, $b = (b_1^T, ..., b_c^T)^T$ is the random effects vector, $\epsilon = (\epsilon_1^T, ..., \epsilon_c^T)^T$ is the measurement error vector, and $\tilde{D} = diag(D_1, ..., D_c)$ and $R = diag(R_1, ..., R_c)$ are variance component matrices. The random effects vector b is assumed to be independently distributed of the measurement error vector ϵ. There are c clusters in total and within the i^{th} cluster, there are n_i subjects.

2.2.1.2 Semi-parametric shape-restricted mixed effects regression spline model

The proposed semi-parametric shape-restricted mixed effects model is of the form

$$y = f(x_{main}) + X\beta + Zb + \epsilon, \tag{2.2.2}$$

where $b \sim N(0, \tilde{D}) \perp \epsilon \sim N(0, R)$. In model (2.2.2), y is the response vector, x_{main} is the main effect, X contains other covariates, β is the fixed effects vector, Z is the random effects

covariate matrix, b is the random effects vector, ϵ is the measurement error vector, and $\tilde{D}$ and R are variance component matrices. However, model (2.2.2) can be rewritten as a linear mixed model using spline basis functions.

Thus, the semi-parametric shape-restricted mixed effects regression spline model for monotonicity can be written as

$$y = X_{IS}\beta_{IS} + X_{IF}\beta_{IF} + Zb_I + \epsilon_I, \tag{2.2.3}$$

where $b_I \sim N(0, \tilde{D}) \perp \epsilon_I \sim N(0, R)$, and the semi-parametric shape-restricted mixed effects regression spline model for convexity can be written as

$$y = X_{CS}\beta_{CS} + X_{CF}\beta_{CF} + Zb_C + \epsilon_C, \tag{2.2.4}$$

where $b_C \sim N(0, \tilde{D}) \perp \epsilon_C \sim N(0, R)$. In models (2.2.3) and (2.2.4), $y = (y_1^T, ..., y_c^T)^T$ is the response vector, $X_{IS} = (X_{IS_1}^T, ..., X_{IS_c}^T)^T$, $X_{IF} = (X_{IF_1}^T, ..., X_{IF_c}^T)^T$, $X_{CS} = (X_{CS_1}^T, ..., X_{CS_c}^T)^T$ and $X_{CF} = (X_{CF_1}^T, ..., X_{CF_c}^T)^T$ are the fixed effects covariate matrices, where X_{IS} contains quadratic I-spline basis functions of the main effect and X_{IF} contains other potential confounders, while X_{CS} contains cubic C-spline basis functions of the main effect and X_{CF} contains main effect and other potential confounders, β_{IS}, β_{IF}, β_{CS} and β_{CF} are the fixed effects vectors, $Z = diag(Z_1, ..., Z_c)$ is the random effects covariate matrix, $b_I = (b_{I_1}^T, ..., b_{I_c}^T)$ and $b_C = (b_{C_1}^T, ..., b_{C_c}^T)$ are the random effects vectors, $\epsilon_I = (\epsilon_{I_1}^T, ..., \epsilon_{I_c}^T)$ and $\epsilon_C = (\epsilon_{C_1}^T, ..., \epsilon_{C_c}^T)$ are the measurement error vectors, and $\tilde{D} = diag(D_1, ..., D_c)$ and $R = diag(R_1, ..., R_c)$ are variance component matrices. There are c clusters in total and within the i^{th} cluster, there are n_i subjects.

2.2.2 Estimation

2.2.2.1 Linear mixed model

There are several approaches to estimate the unknown parameters and make predictions on random effects in linear mixed model, where one approach is based on Henderson's Mixed Model Equations. According to Henderson's formulation, for known $\tilde{D}$ and R, the joint density function of y and b is $f(y,b) = f(y|b)f(b)$, where $y|b \sim N_N(X\beta + Zb, R)$ and $b \sim N_{qc}(0, \tilde{D})$, $N = \sum_{i=1}^{c} n_i$ and qc is the dimension of b (if we only assume a random intercept model, then $q = 1$). From the joint density, Henderson developed a set of equations, which is known as Henderson's Mixed Model Equations (MME), to solve the Best Linear Unbiased Estimates (BLUE) and Best Linear Unbiased Predictions (BLUP) simultaneously (Searle et al., 2006). By taking the first partial derivatives of the twice negative logarithm of the joint density function with respect to β and b and equating them to zero, we will obtain

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + \tilde{D}^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ b \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}. \tag{2.2.5}$$

By solving the MME (2.2.5), we obtain $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ and $\hat{b} = \tilde{D} Z^T V^{-1} (y - X\hat{\beta})$, where $V = Z\tilde{D}Z^T + R$. The corresponding covariance matrices of $\hat{\beta}$ and $\hat{b}$ are $COV(\hat{\beta}) = (X^T V^{-1} X)^{-1}$ and $COV(\hat{b}) = \tilde{D} Z^T V^{-1} Z \tilde{D}^T - \tilde{D} Z^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} Z \tilde{D}^T$. The technical details can be found in Appendix A.1.

If the variance components $\tilde{D}$ and R are unknown, we should numerically solve for the BLUE ($\hat{\beta}$), BLUP ($\hat{b}$) and variance components ($\hat{\tilde{D}}$ and $\hat{R}$) via some iterative procedures, including iterative procedures based on MME, Expectation-Maximum (EM) algorithms, Fisher scoring algorithms, etc. The details on how to apply these algorithms can be found in relevant literature (Searle et al., 2006; Wu and Zhang, 2006; Bates et al., 2015). If the estimated variance components are indefinite, the Minimum Norm Quadratic Estimation (MINQE) can be invoked (Rao and Kleffe, 1988). The iterative procedures based on MME are also summarized in Appendix A.2. Once we obtain the point estimates of $\tilde{D}$ and R, $\hat{\tilde{D}}$ and $\hat{R}$ as denoted above, then we estimate V using $\hat{V} = Z\hat{\tilde{D}}Z^T + \hat{R}$.

2.2.2.2 Semi-parametric shape-restricted mixed effects regression spline model

In semi-parametric shape-restricted mixed effects regression spline model, we use a two-step approach to estimate parameters. In the first step, we obtain unconstrained estimates for all parameters of the linear mixed model using the algorithms described in Section 2.2.2.1. We then project the unconstrained estimates of β_{IS} or β_{CS}, denoted as $\hat{\beta}_{IS}$ or $\hat{\beta}_{CS}$ respectively, onto the suitable cone of constraints of our interest in the second step, i.e. for monotonicity,

$$\tilde{\beta}_{IS} = \arg\min_{\beta_{IS} \in O_i} (\beta_{IS} - \hat{\beta}_{IS})^T COV(\hat{\beta}_{IS})^{-1} (\beta_{IS} - \hat{\beta}_{IS}), i = 1, 2, \tag{2.2.6}$$

and for convexity,

$$\tilde{\beta}_{CS} = \arg\min_{\beta_{CS} \in O_i} (\beta_{CS} - \hat{\beta}_{CS})^T COV(\hat{\beta}_{CS})^{-1} (\beta_{CS} - \hat{\beta}_{CS}), i = 3, 4, \tag{2.2.7}$$

where O_1 and O_3 are the non-negative orthants and O_2 and O_4 are the non-positive orthants, and if $COV(\hat{\beta}_{IS})$ or $COV(\hat{\beta}_{CS})$ is unknown, which typically is the case in applications, we shall use the corresponding estimators, namely, $\widehat{COV(\hat{\beta}_{IS})}$ or $\widehat{COV(\hat{\beta}_{CS})}$ respectively. Lastly, we combine the original estimates of β_{IF} or β_{CF} and the updated estimates of β_{IS} or β_{CS} to get the final estimates of the fixed effect parameters. The projection step is developed based on Fact 1.2.1 and Fact 1.2.2 in Section 1.2 and used for deriving the asymptotic distribution of test statistics described in the following section (Section 2.2.3).

2.2.3 Testing for shape

The ultimate goal of the method is to test the null hypothesis of constant function against the alternative that there is an underlying shape and classify the underlying shape into a reasonable category. Since the shape of the functional relationship is unknown *a priori* except that it might be monotonic or convex/concave, we formulate four tests with the same form of test statistic to perform the hypothesis testing and apply Holm-Bonferroni method to implement a multiple testing procedure in order to select the most reasonable underlying pattern.

The hypotheses are

$$H_{0i} : \beta_i = 0 \text{ vs } H_{ai} : \beta_i \in O_i, i = 1, 2, 3, 4, \tag{2.2.8}$$

where β_1 and β_2 are the fixed effects vectors corresponding to the quadratic I-spline basis functions β_{IS}. Similarly, β_3 and β_4 are the fixed effects vectors corresponding to the cubic C-spline basis functions β_{CS}. Lastly, O_1 and O_3 are the non-negative orthants, i.e. $\{\beta \in R^p : \beta \geq 0\}$, and O_2 and O_4 are the non-positive orthants, i.e. $\{\beta \in R^p : \beta \leq 0\}$.

The test statistics associated with the hypotheses are

$$T_i = \frac{\hat{\beta}_i^T \Sigma_i^{-1} \hat{\beta}_i - \min_{\beta_i \in O_i} (\beta_i - \hat{\beta}_i)^T \Sigma_i^{-1} (\beta_i - \hat{\beta}_i)}{\hat{\beta}_i^T \Sigma_i^{-1} \hat{\beta}_i}, i = 1, 2, 3, 4, \tag{2.2.9}$$

where $\Sigma_i = COV(\hat{\beta}_i)$, and if $COV(\hat{\beta}_i)$ is unknown, we will use its estimate $\widehat{COV(\hat{\beta}_i)}$ and accordingly denote the above test statistic by $\hat{T}_i$. We assume that $\widehat{COV(\hat{\beta}_i)}$ is a consistent estimator of $COV(\hat{\beta}_i)$.

Theorem 2.2.1. *Suppose β_i is a p-dimensional vector. If Σ_i is known, then $P(T_i \leq c|H_{0i}) = \sum_{j=0}^{p} w_j(p, \Sigma_i, \mathbb{R}^{+p}) P(Beta(\frac{j}{2}, \frac{p-j}{2}) \leq c)$ and if Σ_i is unknown, then $P(\hat{T}_i \leq c|H_{0i}) \overset{asymp.}{=} \sum_{j=0}^{p} w_j(p, \Sigma_i, \mathbb{R}^{+p}) P(Beta(\frac{j}{2}, \frac{p-j}{2}) \leq c)$, where $w_j(p, \Sigma_i, \mathbb{R}^{+p})$ are non-negative weights and $\sum_{j=0}^{p} w_j(p, \Sigma_i, \mathbb{R}^{+p}) = 1$.*

Proof. In the proof, the semi-parametric shape-restricted mixed effects regression spline model will be expressed as $y = X_S\beta_S + X_F\beta_F + Zb + \epsilon$ for both monotonicity and convexity.

In linear mixed model, we have $\hat{\beta} \sim N(\beta, (X^T V^{-1} X)^{-1})$, so $\begin{pmatrix} \hat{\beta_S} \\ \hat{\beta_F} \end{pmatrix} \sim N(\begin{pmatrix} \beta_S \\ \beta_F \end{pmatrix}, \begin{pmatrix} \Sigma_S & \Sigma_{SF} \\ \Sigma_{SF} & \Sigma_F \end{pmatrix})$, where $\begin{pmatrix} \Sigma_S & \Sigma_{SF} \\ \Sigma_{SF} & \Sigma_F \end{pmatrix} = (X^T V^{-1} X)^{-1}$. Therefore, $\hat{\beta_S} \sim N(\beta_S, \Sigma_S)$.

Let $\tilde{\beta}_S^1 = \Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p}) = \arg\min_{\beta_S \in \mathbb{R}^{+p}} (\beta_S - \hat{\beta}_S)^T \Sigma_S^{-1} (\beta_S - \hat{\beta}_S)$ and $\tilde{\beta}_S^0 = \Pi_{\Sigma_S}(\hat{\beta}_S|0) = \arg\min_{\beta_S = 0} (\beta_S - \hat{\beta}_S)^T \Sigma_S^{-1} (\beta_S - \hat{\beta}_S)$. Then,

$$\begin{aligned} T &= \frac{\hat{\beta}_S^T \Sigma_S^{-1} \hat{\beta}_S - \min_{\beta_S \in \mathbb{R}^{+p}} (\beta_S - \hat{\beta}_S)^T \Sigma_S^{-1} (\beta_S - \hat{\beta}_S)}{\hat{\beta}_S^T \Sigma_S^{-1} \hat{\beta}_S} \\ &= \frac{||\hat{\beta}_S - \Pi_{\Sigma_S}(\hat{\beta}_S|0)||_{\Sigma_S}^2 - ||\hat{\beta}_S - \Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p})||_{\Sigma_S}^2}{||\hat{\beta}_S - \Pi_{\Sigma_S}(\hat{\beta}_S|0)||_{\Sigma_S}^2}. \end{aligned}$$

Since $\hat{\beta}_S - \Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p}) \perp \Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p}) - \Pi_{\Sigma_S}(\hat{\beta}_S|0)$ (Silvapulle and Sen, 2005, Proposition 3.6.1), $||\hat{\beta}_S - \Pi_{\Sigma_S}(\hat{\beta}_S|0)||^2_{\Sigma_S} - ||\hat{\beta}_S - \Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p})||^2_{\Sigma_S} = ||\Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p}) - \Pi_{\Sigma_S}(\hat{\beta}_S|0)||^2_{\Sigma_S}$. Thus,

$$\begin{aligned} T &= \frac{||\Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p}) - \Pi_{\Sigma_S}(\hat{\beta}_S|0)||^2_{\Sigma_S}}{||\Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p}) - \Pi_{\Sigma_S}(\hat{\beta}_S|0)||^2_{\Sigma_S} + ||\hat{\beta}_S - \Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p})||^2_{\Sigma_S}} \\ &= \frac{||\tilde{\beta}^1_S - \tilde{\beta}^0_S||^2_{\Sigma_S}}{||\tilde{\beta}^1_S - \tilde{\beta}^0_S||^2_{\Sigma_S} + ||\hat{\beta}_S - \tilde{\beta}^1_S||^2_{\Sigma_S}} \\ &= \frac{||\tilde{\beta}^1_S||^2_{\Sigma_S}}{||\tilde{\beta}^1_S||^2_{\Sigma_S} + ||\hat{\beta}_S - \tilde{\beta}^1_S||^2_{\Sigma_S}}. \end{aligned}$$

Suppose $\Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p} \cap 0^{\perp}) \in ri(F)$ for some $F \in \{F_1, ..., F_k\}$, where $ri(F)$ represents the relative interior of F and $dim(ri(F)) = j$. Then $||\tilde{\beta}^1_S||^2_{\Sigma_S} = ||\Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p})||^2_{\Sigma_S} = ||\Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p}) - \Pi_{\Sigma_S}(\hat{\beta}_S|0)||^2_{\Sigma_S} \overset{H_0}{\sim} \chi^2_j$. Since $||\Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p}) - \Pi_{\Sigma_S}(\hat{\beta}_S|0)||^2_{\Sigma_S} \overset{H_0}{\sim} \chi^2_j$, $||\hat{\beta}_S - \Pi_{\Sigma_S}(\hat{\beta}_S|0)||^2_{\Sigma_S} = ||\hat{\beta}_S||^2_{\Sigma_S} \overset{H_0}{\sim} \chi^2_p$, and $\hat{\beta}_S - \Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p}) \perp \Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p}) - \Pi_{\Sigma_S}(\hat{\beta}_S|0)$, then $||\hat{\beta}_S - \tilde{\beta}^1_S||^2_{\Sigma_S} = ||\hat{\beta}_S - \Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p})||^2_{\Sigma_S} \overset{H_0}{\sim} \chi^2_{p-j}$. Therefore,

$$T \overset{H_0}{\sim} Beta(\frac{j}{2}, \frac{p-j}{2}).$$

Finally,

$$\begin{aligned} P(T \leq c|H_0) &= \sum_i P(T \leq c|\Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p} \cap 0^{\perp}) \in ri(F_i)) \times \\ &\qquad P(\Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p} \cap 0^{\perp}) \in ri(F_i)) \\ &= \sum_{j=0}^{p} w_j(p, \Sigma_S, \mathbb{R}^{+p}) P(Beta(\frac{j}{2}, \frac{p-j}{2}) \leq c). \end{aligned}$$

For the test statistic for non-positive orthant, since $\mathbb{R}^{-p} = \{\beta \in \mathbb{R}^p : R\beta \geq 0\}$, where R is a $p \times p$ diagonal matrix with $R_{ii} = -1$ if $\beta_i \leq 0$ and $R_{ii} = 1$ if $\beta_i > 0$, $w_j(p, \Sigma_S, \mathbb{R}^{-p}) = w_j(p, R\Sigma_S R^T, \mathbb{R}^{+p}) = w_j(p, \Sigma_S, \mathbb{R}^{+p})$ (Silvapulle and Sen, 2005, Proposition 3.6.1). Thus, the null distribution of the test statistic for non-positive orthant will be the same as that of the test statistic for non-negative orthant.

If V is unknown, then we will use the maximum likelihood estimator of V, $\hat{V}$. Because $\hat{V}$ is the maximum likelihood estimator of V, under suitable regularity conditions, by the

consistency property of the maximum likelihood estimator, $\hat{V} \xrightarrow{p} V$. Therefore, according to continuous mapping theorem, $(X^T\hat{V}^{-1}X)^{-1} \xrightarrow{p} (X^TV^{-1}X)^{-1}$ and $\hat{\Sigma_S} \xrightarrow{p} \Sigma_S$. Since $\hat{\beta_S} \xrightarrow{d} N(\beta_S, \Sigma_S)$ and $\hat{\Sigma_S} \xrightarrow{p} \Sigma_S$, by continuous mapping theorem and Slutsky's theorem, $||\Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p}) - \Pi_{\Sigma_S}(\hat{\beta}_S|0)||^2_{\Sigma_S} \xrightarrow{d} \chi^2_j$ under H_0 and $||\hat{\beta}_S - \Pi_{\Sigma_S}(\hat{\beta}_S|\mathbb{R}^{+p})||^2_{\Sigma_S} \xrightarrow{d} \chi^2_{p-j}$ under H_0. Therefore,

$$\hat{T} \xrightarrow{d} Beta(\frac{j}{2}, \frac{p-j}{2}) \text{ under } H_0,$$

and

$$P(\hat{T} \leq c|H_0) \overset{asymp.}{=} \sum_{j=0}^{p} w_j(p, \Sigma_S, \mathbb{R}^{+p}) P(Beta(\frac{j}{2}, \frac{p-j}{2}) \leq c).$$

□

If Σ_i is known, then the test statistic T_i under null hypothesis H_{0i} follows a beta-bar distribution. An example of beta-bar density is shown in Figure 2.2.1, where the beta-bar is $0.2462 \times Beta(0, 2.5) + 0.4415 \times Beta(0.5, 2) + 0.2524 \times Beta(1, 1.5) + 0.0559 \times Beta(1.5, 1) + 0.0040 \times Beta(2, 0.5) + 0.0000 \times Beta(2.5, 0)$. In order to obtain the p-value from null distribution, we need to calculate the beta-bar weights. According to Silvapulle and Sen (2005), the exact computation of beta-bar weights is quite difficult. However, because of the fact that the beta-bar weight $w_j(p, \Sigma_i, \mathbb{R}^{+p})$ equals the probability of $\Pi_{\Sigma_i}(\hat{\beta}_i|\mathbb{R}^{+p})$ has exactly j positive components, where $\hat{\beta}_i \sim N(0, \Sigma_i)$ and $\Pi_{\Sigma_i}(\hat{\beta}_i|\mathbb{R}^{+p}) = \arg\min_{\beta_i \in \mathbb{R}^{+p}} (\beta_i - \hat{\beta}_i)^T \Sigma_i^{-1} (\beta_i - \hat{\beta}_i)$ (Silvapulle and Sen, 2005, Proposition 3.6.1), the weights can be approximated by simulation. The simulation steps of beta-bar weights are summarized in Appendix A.3. If Σ_i is unknown, then the test statistic $\hat{T}_i$ under null hypothesis H_{0i} follows a beta-bar distribution asymptotically. When we simulate the beta-bar weights, we can use $\hat{\Sigma}_i$, the estimate of Σ_i, instead of Σ_i. If Σ_i is unknown, we can also adopt residual bootstrap method to simulate the null distribution of the test statistic $\hat{T}_i$ and obtain the p-value (Farnan et al., 2014). As in Farnan et al. (2014), we shall bootstrap the BLUP and residuals. The details of family-wise error rate and power simulations using residual bootstrap method are described in Appendix A.4 and A.5.

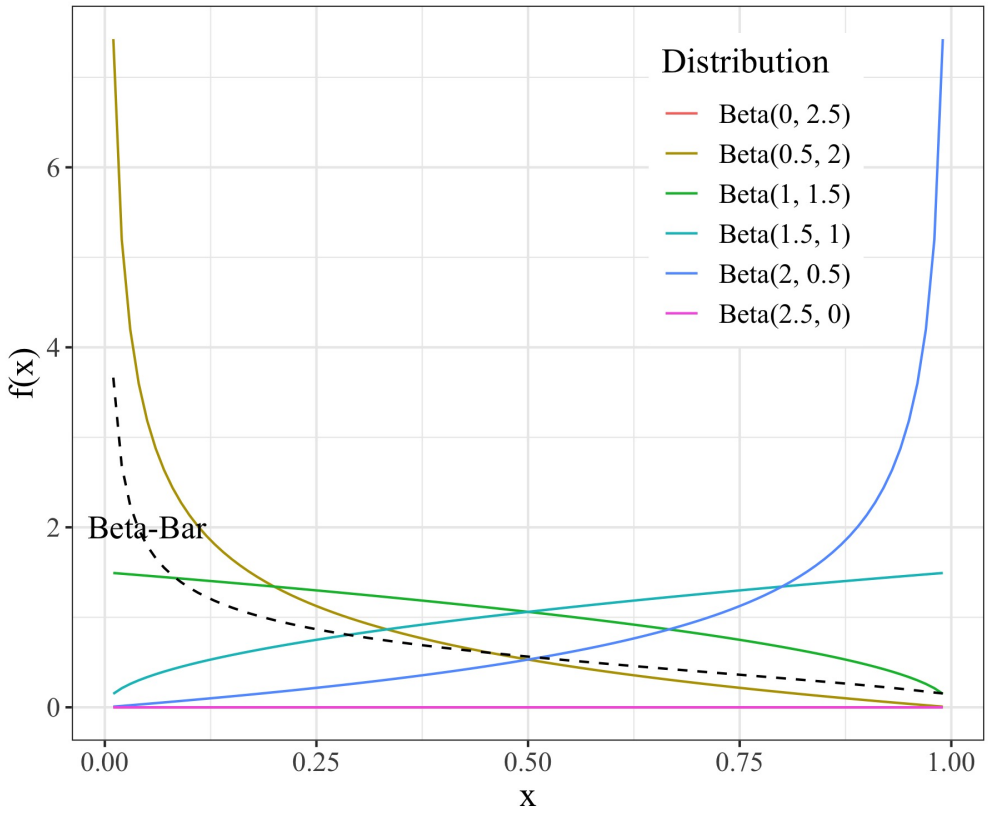


Figure 2.2.1: Probability density functions of beta and beta-bar distributions

2.2.4 Shape detection

After obtaining the p-values from those four tests, we shall apply Holm-Bonferroni method (Holm, 1979) to perform multiple testing of the four hypotheses. To begin with, we sort the p-values in ascending order, i.e. $p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq p_{(4)}$, and then compare the p-values with the corresponding significance levels, i.e. $\alpha_{(1)} = \frac{\alpha}{4} \leq \alpha_{(2)} = \frac{\alpha}{3} \leq \alpha_{(3)} = \frac{\alpha}{2} \leq \alpha_{(4)} = \alpha$. The decisions of rejections are made sequentially. If $p_{(1)} \leq \alpha_{(1)}$, then we reject the corresponding null hypothesis $H_{0(1)}$ and go to next step; otherwise, we stop and reject none of the null hypotheses. If $H_{0(1)}$ is rejected and $p_{(2)} \leq \alpha_{(2)}$, then we reject the corresponding null hypothesis $H_{0(2)}$ and go to next step; otherwise, we stop and only reject $H_{0(1)}$. If $H_{0(1)}$ and $H_{0(2)}$ are rejected and $p_{(3)} \leq \alpha_{(3)}$, then we reject the corresponding null hypothesis $H_{0(3)}$

and go to next step; otherwise, we stop and only reject $H_{0(1)}$ and $H_{0(2)}$. If $H_{0(1)}$, $H_{0(2)}$ and $H_{0(3)}$ are rejected and $p_{(4)} \leq \alpha_{(4)}$, then we reject the corresponding null hypothesis $H_{0(4)}$ and stop; otherwise, we only reject $H_{0(1)}$, $H_{0(2)}$ and $H_{0(3)}$.

Depending upon the rejections by using the above decision rules, we make decisions regarding the shape as described in Table 2.2.1. Different shapes are shown in Figure 2.2.2.

Table 2.2.1: Decision making on shape category

Null Hypothesis/Hypotheses Rejected	Shape Category	Shape Number
H_{01}	Increasing	①
H_{02}	Decreasing	②
H_{03}	Convex	③
H_{04}	Concave	④
H_{01} and H_{03}	Convex with Increasing Trend	⑤
H_{01} and H_{04}	Concave with Increasing Trend	⑥
H_{02} and H_{03}	Convex with Decreasing Trend	⑦
H_{02} and H_{04}	Concave with Decreasing Trend	⑧

2.2.5 Remarks

As noted in Hwang and Peddada (1994), since the cones of interest are polyhedral cones, the projected vector performs better than the unrestricted vector. In our method, the test statistics are based on the entire vector. Therefore, the proposed tests are expected to have higher power than tests using the unrestricted vector. However, if one is interested in any individual component, then the projected estimator may potentially have zero coverage probability as the dimension increases. In such cases, some other strategies should be introduced.

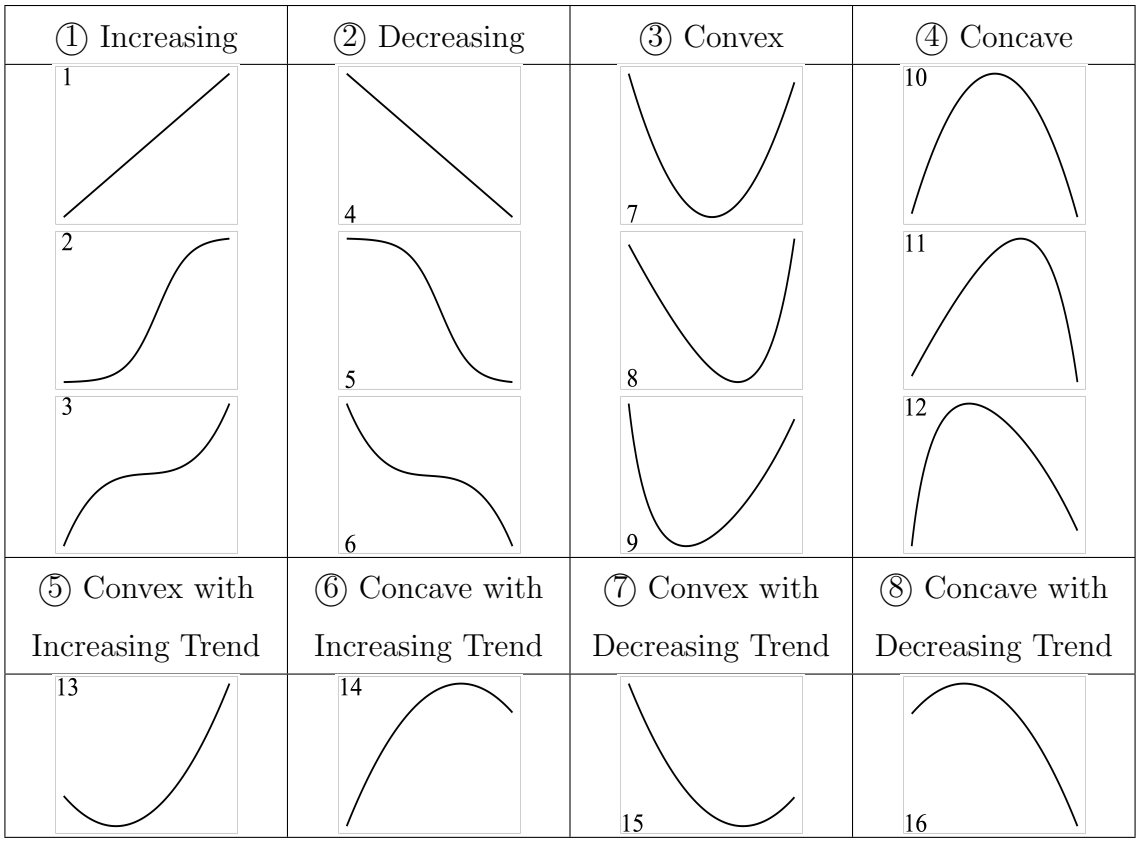


Figure 2.2.2: Figures of different shapes

2.3 Simulation

We create a data set similar to a prenatal screening program data set. The data set contained 245 observations and consisted of 8 fixed effects (X): hormone, age, inverse maternal weight, race, season of blood draw, smoking status, ovum donor status and pre-existing diabetes status, and 1 random effect, namely, geographic location as measured by zip code (Z). We created 5 geographic locations with sizes 45, 55, 50, 45 and 50 respectively. The covariates for each location were created as follows: (1) hormone (Gestational Age-Multiple of Median): randomly sampled from a normal distribution with mean 0 and variance 4, (2) age (years): randomly sampled from 18 to 40 with an increment 0.5, (3) inverse maternal weight: randomly sampled from 2 to 14.3 with an increment 0.1, (4) race: randomly sampled from

1 to 5 with probabilities 0.46, 0.28, 0.13, 0.1 and 0.03, (5) season of blood draw: randomly sampled from 1 to 4 with the same probability, (6) smoking status: randomly sampled from a binomial distribution with success probability 0.25, (7) ovum donor status: randomly sampled from a binomial distribution with success probability 0.15 and (8) pre-existing diabetes status: randomly sampled from a binomial distribution with success probability 0.2.

For monotonicity assumption, X_{IF} contains age, inverse maternal weight, race, season of blood draw, smoking status, ovum donor status and pre-existing diabetes status, and X_{IS} contains the quadratic I-spline bases generated from hormone. For convexity assumption, X_{CF} contains hormone, age, inverse maternal weight, race, season of blood draw, smoking status, ovum donor status and pre-existing diabetes status, and X_{CS} contains the cubic C-spline bases generated from hormone.

2.3.1 Family-wise error rate

For both monotonicity and convexity assumptions, b_I and b_C are drawn from $N(0, 10^2)$, $N(0, 15^2)$ and $N(0, 20^2)$, ϵ_I and ϵ_C are drawn from $N(0, 10^2)$, $N(0, 15^2)$ and $N(0, 20^2)$, and the number of internal knots is set to 4. In total, we performed three types of simulations to describe family-wise error rate under the global null. We considered (1) T_i, whose null distribution is the simulated beta-bar distribution (1000 iterations for data and 10000 iterations for the beta-bar weight), (2) $\hat{T}_i$, whose null distribution is the asymptotic beta-bar distribution (500 iterations for data and 5000 iterations for the beta-bar weight), and (3) $\hat{T}_i$, whose null distribution is simulated using residual bootstrap method (400 iterations for data and 500 iterations for the bootstrap samples). The simulated family-wise error rates are shown in Table 2.3.1.

According to the results in Table 2.3.1, under each type of simulation, our methodology controls the family-wise error rate reasonably. In reality, the variances are unknown, so we need to simulate the null distributions using either the estimated variances or residual bootstrap method.

Table 2.3.1: Family-wise error rate simulation results

Variance		Family-Wise Error Rate		
Variance of b	Variance of ϵ	Family-Wise Error Rate under Simulation Type (1)	Family-Wise Error Rate under Simulation Type (2)	Family-Wise Error Rate under Simulation Type (3) with Asymptotic 95% C.I.
10^2	10^2	0.053	0.060	0.065 (0.041, 0.089)
10^2	15^2	0.043	0.056	0.063 (0.039, 0.086)
10^2	20^2	0.048	0.052	0.048 (0.027, 0.068)
15^2	10^2	0.045	0.054	0.065 (0.041, 0.089)
15^2	15^2	0.048	0.048	0.035 (0.017, 0.053)
15^2	20^2	0.050	0.056	0.068 (0.043, 0.092)
20^2	10^2	0.050	0.050	0.058 (0.035, 0.080)
20^2	15^2	0.054	0.052	0.048 (0.027, 0.068)
20^2	20^2	0.037	0.044	0.060 (0.037, 0.083)

2.3.2 Power

For power simulation, we consider 8 cases out of 16 (see Figure 2.2.2) because other 8 cases can be treated as "complementary" cases. The functions of the main effect of the 8 cases that we consider were $f(x) = 5.5x + 70$ (shape 1), $f(x) = 50\frac{e^{1.2x}}{2+e^{1.2x}} + 50$ (shape 2), $f(x) = (\frac{2x}{3})^3 + \frac{x}{2} + 50$ (shape 3), $f(x) = \frac{-e^x - 100x^2}{50} + 100$ (shape 10), $f(x) = 300\ln(-e^{x/2} + x + 40) - 1000$ (shape 11), $f(x) = 70\ln(-e^{-x/2} - x + 10) - 60$ (shape 12), $f(x) = -1.2(x-2)^2 + 100$ (shape 14), and $f(x) = -1.2(x+1.5)^2 + 100$ (shape 16). For both monotonicity and convexity, b_I and b_C are drawn from $N(0, 10^2)$, $N(0, 15^2)$ and $N(0, 20^2)$, ϵ_I and ϵ_C are drawn from $N(0, 10^2)$, $N(0, 15^2)$ and $N(0, 20^2)$, and the number of internal knots is set to 4. We performed simulations using a variety of shapes for the response function. The null distribution of test statistic was derived using (a) simulations, and (b) residual bootstrap. The simulated powers are summarized in Table 2.3.2 to Table 2.3.9 (the power curves can be found in Appendix A.6).

According to the results in Table 2.3.2 to Table 2.3.9, the powers are relatively large in all 8 cases when the variance of ϵ is relatively small. As expected, as the variance of ϵ increases, the powers decrease, especially for shapes 3, 14 and 16. As the error variance, i.e. variance of ϵ, increases, for shape 3 the methodology is prone to detect the shape to be flat. For shape 14, the method is more likely to declare it to be either increasing or concave shape. For shape 16, the method is more likely to declare the shape as either decreasing or concave shape.

2.4 Discussion

We adopted ideas from Ramsay that the linear combination of quadratic I-spline basis functions yields monotonicity, and from Meyer that linear combination of cubic C-spline basis functions yields convexity. We applied these ideas to linear mixed models to account for random effects. Using this framework, we developed a methodology to identify different shapes of relationships between a response variable and a predictor of interest. The shapes considered were increasing, decreasing, convex and concave shapes. Our simulation study

Table 2.3.2: Power simulation results for shape 1 ($f(x) = 5.5x + 70$)

Variance		Power under Simulation Type (a)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0.976	0	0	0	0.012	0.012	0	0
10^2	15^2	0.957	0	0	0	0.013	0.02	0	0
10^2	20^2	0.87	0	0	0	0.017	0.017	0	0
15^2	10^2	0.975	0	0	0	0.015	0.01	0	0
15^2	15^2	0.949	0	0	0	0.019	0.021	0	0
15^2	20^2	0.893	0	0	0	0.018	0.023	0	0
20^2	10^2	0.961	0	0	0	0.019	0.02	0	0
20^2	15^2	0.956	0	0	0.001	0.024	0.011	0	0
20^2	20^2	0.884	0	0	0.001	0.017	0.017	0	0
Variance		Power under Simulation Type (b)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0.965	0	0	0	0.015	0.02	0	0
10^2	15^2	0.933	0	0	0	0.028	0.023	0	0
10^2	20^2	0.87	0	0	0	0.023	0.018	0	0
15^2	10^2	0.97	0	0	0	0.023	0.008	0	0
15^2	15^2	0.963	0	0	0	0.015	0.008	0	0
15^2	20^2	0.888	0	0	0.003	0.025	0.015	0	0
20^2	10^2	0.97	0	0	0	0.015	0.015	0	0
20^2	15^2	0.94	0	0	0	0.028	0.015	0	0
20^2	20^2	0.905	0	0	0	0.018	0.008	0	0

Table 2.3.3: Power simulation results for shape 2 ($f(x) = 50\frac{e^{1.2x}}{2+e^{1.2x}} + 50$)

Variance		Power under Simulation Type (a)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	1	0	0	0	0	0	0	0
10^2	15^2	0.991	0	0	0	0.004	0	0	0
10^2	20^2	0.947	0	0	0	0.011	0	0	0
15^2	10^2	1	0	0	0	0	0	0	0
15^2	15^2	0.991	0	0	0	0.005	0	0	0
15^2	20^2	0.965	0	0	0	0.011	0	0	0
20^2	10^2	0.998	0	0	0	0.002	0	0	0
20^2	15^2	0.988	0	0	0	0.008	0	0	0
20^2	20^2	0.95	0	0	0	0.01	0	0	0
Variance		Power under Simulation Type (b)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	1	0	0	0	0	0	0	0
10^2	15^2	0.98	0	0	0	0.008	0	0	0
10^2	20^2	0.955	0	0	0	0.018	0	0	0
15^2	10^2	1	0	0	0	0	0	0	0
15^2	15^2	0.98	0	0	0	0.008	0	0	0
15^2	20^2	0.953	0	0	0	0.013	0	0	0
20^2	10^2	0.998	0	0	0	0	0	0	0
20^2	15^2	0.995	0	0	0	0	0	0	0
20^2	20^2	0.96	0	0	0	0.008	0	0	0

Table 2.3.4: Power simulation results for shape 3 ($f(x) = (\frac{2x}{3})^3 + \frac{x}{2} + 50$)

Variance		Power under Simulation Type (a)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0.767	0	0	0	0.009	0	0	0
10^2	15^2	0.562	0	0	0	0.014	0	0	0
10^2	20^2	0.407	0	0.003	0	0.021	0	0	0
15^2	10^2	0.774	0	0	0	0.005	0	0	0
15^2	15^2	0.619	0	0	0	0.016	0	0	0
15^2	20^2	0.44	0	0.001	0	0.031	0.002	0	0
20^2	10^2	0.748	0	0	0	0.007	0	0	0
20^2	15^2	0.553	0	0	0	0.018	0	0	0
20^2	20^2	0.443	0	0.001	0	0.02	0	0	0
Variance		Power under Simulation Type (b)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0.76	0	0	0	0.003	0	0	0
10^2	15^2	0.62	0	0	0	0.02	0	0	0
10^2	20^2	0.46	0	0	0	0.023	0	0	0
15^2	10^2	0.823	0	0	0	0.005	0	0	0
15^2	15^2	0.618	0	0.003	0	0.023	0	0	0
15^2	20^2	0.46	0	0.003	0	0.018	0.005	0	0
20^2	10^2	0.748	0	0	0	0.008	0	0	0
20^2	15^2	0.575	0	0.005	0	0.018	0	0	0
20^2	20^2	0.505	0	0.003	0	0.028	0	0	0

Table 2.3.5: Power simulation results for shape 10 ($f(x) = \frac{-e^x - 100x^2}{50} + 100$)

Variance		Power under Simulation Type (a)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0	0	0	1	0	0	0	0
10^2	15^2	0	0	0	0.985	0	0	0	0.001
10^2	20^2	0	0	0	0.895	0	0	0	0
15^2	10^2	0	0	0	1	0	0	0	0
15^2	15^2	0	0	0	0.982	0	0	0	0
15^2	20^2	0	0	0	0.92	0	0	0	0
20^2	10^2	0	0	0	1	0	0	0	0
20^2	15^2	0	0	0	0.982	0	0	0	0
20^2	20^2	0	0	0	0.904	0	0	0	0.001
Variance		Power under Simulation Type (b)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0	0	0	0.998	0	0	0	0
10^2	15^2	0	0	0	0.97	0	0	0	0
10^2	20^2	0	0	0	0.873	0	0	0	0.003
15^2	10^2	0	0	0	0.995	0	0	0	0
15^2	15^2	0	0	0	0.963	0	0	0	0
15^2	20^2	0	0	0	0.85	0	0	0	0
20^2	10^2	0	0	0	0.993	0	0	0	0
20^2	15^2	0	0	0	0.938	0	0	0	0
20^2	20^2	0	0	0	0.82	0	0	0	0

Table 2.3.6: Power simulation results for shape 11 ($f(x) = 300\ln(-e^{x/2} + x + 40) - 1000$)

Variance		Power under Simulation Type (a)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0	0	0	1	0	0	0	0
10^2	15^2	0	0	0	0.978	0	0	0	0
10^2	20^2	0	0	0	0.883	0	0	0	0
15^2	10^2	0	0	0	1	0	0	0	0
15^2	15^2	0	0	0	0.971	0	0	0	0
15^2	20^2	0	0	0	0.899	0	0	0	0
20^2	10^2	0	0	0	1	0	0	0	0
20^2	15^2	0	0	0	0.973	0	0	0	0
20^2	20^2	0	0	0	0.891	0	0	0	0
Variance		Power under Simulation Type (b)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0	0	0	0.995	0	0	0	0
10^2	15^2	0	0	0	0.93	0	0	0	0
10^2	20^2	0	0	0	0.763	0	0	0	0
15^2	10^2	0	0	0	0.995	0	0	0	0
15^2	15^2	0	0	0	0.91	0	0	0	0
15^2	20^2	0	0	0	0.775	0	0.003	0	0
20^2	10^2	0	0	0	0.988	0	0	0	0
20^2	15^2	0	0	0	0.89	0	0	0	0
20^2	20^2	0	0	0	0.803	0	0	0	0

Table 2.3.7: Power simulation results for shape 12 ($f(x) = 70\ln(-e^{-x/2} - x + 10) - 60$)

Variance		Power under Simulation Type (a)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0	0	0	0.999	0	0	0	0
10^2	15^2	0	0	0	0.954	0	0	0	0
10^2	20^2	0	0	0	0.845	0	0	0	0
15^2	10^2	0	0	0	0.997	0	0	0	0
15^2	15^2	0	0	0	0.959	0	0	0	0
15^2	20^2	0	0	0	0.87	0	0	0	0
20^2	10^2	0	0	0	0.999	0	0	0	0
20^2	15^2	0	0	0	0.971	0	0	0	0
20^2	20^2	0	0	0	0.829	0	0	0	0
Variance		Power under Simulation Type (b)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0	0	0	0.99	0	0	0	0
10^2	15^2	0	0	0	0.93	0	0	0	0
10^2	20^2	0	0	0	0.773	0	0	0	0
15^2	10^2	0	0	0	1	0	0	0	0
15^2	15^2	0	0	0	0.945	0	0	0	0
15^2	20^2	0	0	0	0.808	0	0	0	0
20^2	10^2	0	0	0	0.993	0	0	0	0
20^2	15^2	0	0	0	0.928	0	0	0	0
20^2	20^2	0	0	0	0.788	0	0	0	0

Table 2.3.8: Power simulation results for shape 14 ($f(x) = -1.2(x-2)^2 + 100$)

Variance		Power under Simulation Type (a)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0.013	0	0	0.138	0	0.835	0	0
10^2	15^2	0.09	0	0	0.239	0	0.545	0	0
10^2	20^2	0.16	0	0	0.277	0	0.31	0	0
15^2	10^2	0.031	0	0	0.14	0	0.813	0	0
15^2	15^2	0.11	0	0	0.212	0	0.584	0	0
15^2	20^2	0.161	0	0	0.268	0	0.309	0	0
20^2	10^2	0.015	0	0	0.173	0	0.798	0	0
20^2	15^2	0.101	0	0	0.255	0	0.532	0	0
20^2	20^2	0.18	0	0	0.264	0	0.312	0	0
Variance		Power under Simulation Type (b)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0.023	0	0	0.13	0	0.84	0	0
10^2	15^2	0.103	0	0	0.208	0	0.585	0	0
10^2	20^2	0.183	0	0	0.235	0	0.345	0	0
15^2	10^2	0.033	0	0	0.155	0	0.795	0	0
15^2	15^2	0.1	0	0	0.238	0	0.575	0	0
15^2	20^2	0.163	0	0	0.263	0	0.345	0	0
20^2	10^2	0.023	0	0	0.158	0	0.798	0	0
20^2	15^2	0.115	0	0	0.233	0	0.545	0	0
20^2	20^2	0.173	0	0	0.285	0	0.328	0	0

Table 2.3.9: Power simulation results for shape 16 ($f(x) = -1.2(x + 1.5)^2 + 100$)

Variance		Power under Simulation Type (a)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0	0.018	0	0.079	0	0	0	0.896
10^2	15^2	0	0.14	0	0.189	0	0	0	0.609
10^2	20^2	0	0.2	0	0.186	0	0	0	0.412
15^2	10^2	0	0.034	0	0.079	0	0	0	0.88
15^2	15^2	0	0.137	0	0.131	0	0	0	0.675
15^2	20^2	0	0.242	0	0.189	0	0	0	0.401
20^2	10^2	0	0.021	0	0.061	0	0	0	0.907
20^2	15^2	0	0.125	0	0.159	0	0	0	0.641
20^2	20^2	0	0.216	0	0.16	0	0	0	0.433
Variance		Power under Simulation Type (b)							
Variance of b	Variance of ϵ	Shape Number							
		①	②	③	④	⑤	⑥	⑦	⑧
10^2	10^2	0	0.02	0	0.085	0	0	0	0.883
10^2	15^2	0	0.123	0	0.173	0	0	0	0.618
10^2	20^2	0	0.205	0	0.168	0	0	0	0.428
15^2	10^2	0	0.035	0	0.088	0	0	0	0.863
15^2	15^2	0	0.133	0	0.173	0	0	0	0.65
15^2	20^2	0	0.233	0	0.218	0	0	0	0.403
20^2	10^2	0	0.028	0	0.063	0	0	0	0.905
20^2	15^2	0	0.138	0	0.173	0	0	0	0.61
20^2	20^2	0	0.22	0	0.188	0	0	0	0.433

suggests that the proposed methodology controls the family-wise error rate while maintaining good power in detecting the correct shape of the response curve. By applying the method to the real data, researchers may gain insight into specific biological mechanisms.

The method aims to add shape restriction on the mean effect but not on the individual effect. Hence, the method is applicable if researchers are interested in population-level inference. However, if researchers are interested in making inferences or predictions at the individual- or cluster-level, then the methodology needs to be modified for individual- or cluster-level shape restrictions. The estimation procedure can also be improved through iteratively projecting the estimated parameters $\hat{\beta}_{IS}/\hat{\beta}_{CS}$ onto a closed orthant in the iterative algorithm and updating the variance accordingly. The simulation study in this chapter is based on small number of knots. The parameter estimation and the power of the test may be influenced by the number of knots and the placement of knots. Researchers can further explore how the number of knots and the placement of knots will influence the analysis.

3.0 Mediation Analysis using Semi-parametric Shape-Restricted Regression Spline

3.1 Introduction

Researchers often build models to analyze the relationship between an exposure and a potential outcome caused by the exposure. In some cases, the exposure has direct effect on the outcome. For example, some studies showed that the maternal obesity, characterized by high circulating glucose, fatty acids, etc., has negative impact on offspring's metabolic and cardiovascular functions (Gillman et al., 2003; Drake and Reynolds, 2010). In some other cases, the exposure may not directly lead to the outcome, but instead, it induces the outcome through a process. For example, some studies showed that phthalates, chemicals used in plastics, have impact on placental biomarkers, such as human chorionic gonadotropin (hCG), and further influence offspring's brain development (Adibi et al., 2015, 2021b). Mediation analysis is designed to explain the causal relationship between the exposure and the outcome by examining the intermediate stage, which helps researchers understand the pathway whereby the exposure affects the outcome. When performing the mediation analysis, researchers often build two models simultaneously - the exposure-mediator model and the exposure-outcome model. In regression-based mediation analysis, if the assumptions hold and the models are correctly specified, different effects, including the controlled direct effect, the natural direct effect and the natural indirect effect, can be estimated through combining the estimated parameters, and the associated variances can be obtained by applying delta method. In the circumstance that the relationship between the mediator and the outcome is curvilinear, in order to reduce the bias introduced by model misspecification, the exposure-outcome model should be modeled using nonparametric techniques.

In this chapter, we build the exposure-outcome model using semi-parametric shape-restricted regression spline, estimate the direct and indirect effects analytically, and obtain the asymptotic confidence intervals of those effects by applying delta method. The regression spline is built using quadratic I-spline and cubic C-spline because of the two facts described

in introduction (Section 1.2): (1) a linear combination of quadratic I-spline basis functions is increasing if and only if the coefficients are positive, and (2) a linear combination of cubic C-spline basis functions is convex if and only if the coefficients are positive. The estimation of the coefficients in exposure-outcome model follows the logic of method proposed by Meyer (Meyer, 2018), and since the model involves with the factor-by-curve interaction, the estimation procedure is slightly extended. Once the exposure-outcome and exposure-mediator models are established, the effects as well as their asymptotic variances can be obtained analytically.

3.2 Methodology

3.2.1 Model setup

3.2.1.1 Basic exposure-outcome and exposure-mediator models

Let Y be the continuous outcome, A be the binary exposure, M be the continuous mediator, and C be the confounding variable. Suppose the interaction between exposure and mediator exists, and the relationships between mediator and outcome in both exposure and non-exposure groups are curvilinear and known to be increasing, decreasing, convex or concave. The exposure-outcome model will be

$$Y = \beta_0 + \beta_1 A + f_1(M)A + f_2(M)(1 - A) + \beta_4 C + \epsilon_1, \tag{3.2.1}$$

where $f_1(M)$ is the curve of the mediator for the exposure group, $f_2(M)$ is the curve of the mediator for the non-exposure group, and $\epsilon_1 \sim N(0, \sigma_1^2)$, and the exposure-mediator model will be

$$M = \gamma_0 + \gamma_1 A + \gamma_2 C + \epsilon_2, \tag{3.2.2}$$

where $\epsilon_2 \sim N(0, \sigma_2^2)$.

3.2.1.2 Specified exposure-outcome model using I-splines and C-splines

In model (3.2.1), if $f_1(M)$ is fitted using the quadratic I-splines, then

$$f_1(M) = \beta_{21} I_1(M|2,t) + ... + \beta_{2k} I_k(M|2,t),$$

and if $f_1(M)$ is fitted using the cubic C-splines, then

$$f_1(M) = \beta_{20} M + \beta_{21} C_1(M|2,t) + ... + \beta_{2k} C_k(M|2,t);$$

if $f_2(M)$ is fitted using the quadratic I-splines, then

$$f_2(M) = \beta_{31} I_1(M|2,t) + ... + \beta_{3k} I_k(M|2,t),$$

and if $f_2(M)$ is fitted using the cubic C-splines, then

$$f_2(M) = \beta_{30} M + \beta_{31} C_1(M|2,t) + ... + \beta_{3k} C_k(M|2,t).$$

In specific, we can fit the following four models depending upon the assumptions on the shapes of $f_1(M)$ and $f_2(M)$.

Let $IS = [I_1(M|2,t), ..., I_k(M|2,t)]$ and $CS = [M, C_1(M|2,t), ..., C_k(M|2,t)]$. If $f_1(M)$ and $f_2(M)$ are assumed to be increasing or decreasing, then they should be fitted using I-splines, and model (3.2.1) will be expressed as

$$Y = [1, A, IS \bullet A, IS \bullet (1-A), C] \times [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4]^T + \epsilon_1; \tag{3.2.3}$$

if $f_1(M)$ and $f_2(M)$ are assumed to be convex or concave, then they should be fitted using C-splines, and model (3.2.1) will be expressed as

$$Y = [1, A, CS \bullet A, CS \bullet (1-A), C] \times [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4]^T + \epsilon_1; \tag{3.2.4}$$

if $f_1(M)$ is assumed to be increasing or decreasing and $f_2(M)$ is assumed to be convex or concave, then $f_1(M)$ should be fitted using I-splines and $f_2(M)$ should be fitted using C-splines, and model (3.2.1) will be expressed as

$$Y = [1, A, IS \bullet A, CS \bullet (1-A), C] \times [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4]^T + \epsilon_1; \tag{3.2.5}$$

if $f_1(M)$ is assumed to be convex or concave and $f_2(M)$ is assumed to be increasing or decreasing, then $f_1(M)$ should be fitted using C-splines and $f_2(M)$ should be fitted using I-splines, and model (3.2.1) will be expressed as

$$Y = [1, A, CS \bullet A, IS \bullet (1 - A), C] \times [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4]^T + \epsilon_1. \tag{3.2.6}$$

In the formulations above, $\beta_2 = [\beta_{21}, ..., \beta_{2k}]$ if the corresponding matrix is $IS \bullet A$, or $\beta_2 = [\beta_{20}, \beta_{21}, ..., \beta_{2k}]$ if the corresponding matrix is $CS \bullet A$; $\beta_3 = [\beta_{31}, ..., \beta_{3k}]$ if the corresponding matrix is $IS \bullet (1 - A)$, or $\beta_3 = [\beta_{30}, \beta_{31}, ..., \beta_{3k}]$ if the corresponding matrix is $CS \bullet (1 - A)$. The symbol "$\bullet$" denotes the face-splitting product.

3.2.2 Estimation and inference

3.2.2.1 Parameter estimation of exposure-outcome and exposure-mediator models

The parameter estimation procedure of model (3.2.1) follows the logic of method proposed by Meyer (Meyer, 2018), and we slightly extend Meyer's method by including the factor-by-curve interaction. In order to proceed with the estimation, we need to specify the matrices W_0, W and Z for each of the models (3.2.3), (3.2.4), (3.2.5), and (3.2.6).

Let $CS = [CS_1, CS_2]$, where $CS_1 = M$ and $CS_2 = [C_1(M|2, t), ..., C_k(M|2, t)]$. For model (3.2.3),

$$W_0 = [1], W = [A, C], Z_1 = [IS \bullet A], Z_0 = [IS \bullet (1 - A)] \text{ and } Z = [Z_1, Z_0]; \tag{3.2.7}$$

for model (3.2.4),

$$W_0 = [1, CS_1 \bullet A, CS_1 \bullet (1-A)], W = [A, C], Z_1 = [CS_2 \bullet A], Z_0 = [CS_2 \bullet (1-A)] \text{ and } Z = [Z_1, Z_0]; \tag{3.2.8}$$

for model (3.2.5),

$$W_0 = [1, CS_1 \bullet (1 - A)], W = [A, C], Z_1 = [IS \bullet A], Z_0 = [CS_2 \bullet (1 - A)] \text{ and } Z = [Z_1, Z_0]; \tag{3.2.9}$$

for model (3.2.6),

$$W_0 = [1, CS_1 \bullet A], W = [A, C], Z_1 = [CS_2 \bullet A], Z_0 = [IS \bullet (1-A)] \text{ and } Z = [Z_1, Z_0]. \quad (3.2.10)$$

Following Meyer's method, we let $V = [W_0, W]$, and $\Delta = (I - P_V)Z$. The projection matrix $I - P_V = I - V(V^TV)^{-1}V^T$ projects Z onto the null space of V^T, $N(V^T)$. Using hinge algorithm (see Appendix B.1) for cone projection (Meyer, 2013), a subset of columns of Δ can be determined. We then keep the corresponding columns of Z and estimate the parameters in model (3.2.1) using ordinary least squares. The parameters corresponding to the eliminated columns of Z are estimated as 0's. During the process of hinge algorithm for cone projection, if the signs of coefficients for the exposure or non-exposure group splines are assumed to be non-positive, i.e., the curve is assumed to be decreasing or concave, then we will use $[1 - I_1(M|2,t), ..., 1 - I_k(M|2,t)]$ or $[1 - C_1(M|2,t), ..., 1 - C_k(M|2,t)]$ instead of $[I_1(M|2,t), ..., I_k(M|2,t)]$ or $[C_1(M|2,t), ..., C_k(M|2,t)]$. For model (3.2.2), we use ordinary least squares to estimate the parameters.

3.2.2.2 Mediation effects estimation

Under the consistency and no-unmeasured-confounding assumptions described in introduction (Section 1.4), different mediation effects, including controlled direct effect ($Y_{am} - Y_{a^*m}$), natural direct effect ($Y_{aM_{a^*}} - Y_{a^*M_{a^*}}$) and natural indirect effect ($Y_{aM_a} - Y_{aM_{a^*}}$), are able to be identified. Given the models (3.2.1) and (3.2.2), we can find the expected CDE, NDE and NIE using the formulas $(\beta_1 + f_1(m) - f_2(m))(a - a^*)$, $(\beta_1 + E[f_1(M)|a^*, c] - E[f_2(M)|a^*, c])(a - a^*)$ and $a(E[f_1(M)|a, c] - E[f_1(M)|a^*, c]) + (1 - a)(E[f_2(M)|a, c] - E[f_2(M)|a^*, c])$ respectively, where $a = 1$ and $a^* = 0$. Once we determine the shape of $f_1(M)$ and $f_2(M)$, the functions can be parameterized using the I-splines and C-splines, and the expectations of functions can be obtained using the formula $E[g(X)] = \int_x g(x)f(x)dx$, where $f(x)$ is the probability density function of X. The technical details are shown in Proposition 3.2.1. Since different mediation effects are different combinations of the parameters in models (3.2.1) and (3.2.2), they can be estimated by plugging in the estimated parameters.

Proposition 3.2.1. *Under the consistency, no unmeasured exposure-outcome confounding, no unmeasured mediator-outcome confounding, no unmeasured exposure-mediator confounding and no mediator-outcome confounding affected by exposure assumptions, and with the models (3.2.1) and (3.2.2) being correctly specified, the expected controlled direct effect, natural direct effect and natural indirect effect, conditioning on $C = c$, are given by*

$$E[Y_{am} - Y_{a^*m}|c] = (\beta_1 + f_1(m) - f_2(m))(a - a^*), \tag{3.2.11}$$

$$E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] = (\beta_1 + E[f_1(M)|a^*, c] - E[f_2(M)|a^*, c])(a - a^*), \tag{3.2.12}$$

and

$$E[Y_{aM_a} - Y_{aM_{a^*}}|c] = a(E[f_1(M)|a, c] - E[f_1(M)|a^*, c]) + (1-a)(E[f_2(M)|a, c] - E[f_2(M)|a^*, c]), \tag{3.2.13}$$

respectively. If $f_1(M)$ is fitted using I-splines, then

$$\begin{aligned} E[f_1(M)|a, c] = \sum_{i=2}^{k} \{ \int_{t_i}^{t_{i+1}} [\beta_{21} + ... + \beta_{2,i-2} + \beta_{2,i-1}(1 - \frac{(t_{i+1} - m)^2}{(t_{i+1} - t_i)(t_{i+1} - t_{i-1})}) \\ + \beta_{2,i}(\frac{(m - t_i)^2}{(t_{i+1} - t_i)(t_{i+2} - t_i)})] f(m|a, c) dm \}, \end{aligned} \tag{3.2.14}$$

where $f(m|a, c) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{\frac{-(m-(\gamma_0+\gamma_1 a+\gamma_2 c))^2}{2\sigma_2^2}}$; if $f_1(M)$ is fitted using C-splines, then

$$\begin{aligned} E[f_1(M)|a, c] = \beta_{20}(\gamma_0 + \gamma_1 a + \gamma_2 c) \\ + \sum_{i=2}^{k} \{ \int_{t_i}^{t_{i+1}} [\beta_{21}(m - \frac{t_1 + t_2 + t_3}{3}) + ... + \beta_{2,i-2}(m - \frac{t_{i-2} + t_{i-1} + t_i}{3}) \\ + \beta_{2,i-1}(m - \frac{t_{i-1} + t_i + t_{i+1}}{3} + \frac{(t_{i+1} - m)^3}{3(t_{i+1} - t_i)(t_{i+1} - t_{i-1})}) \\ + \beta_{2,i}(\frac{(m - t_i)^3}{3(t_{i+1} - t_i)(t_{i+2} - t_i)})] f(m|a, c) dm \}, \end{aligned} \tag{3.2.15}$$

where $f(m|a, c) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{\frac{-(m-(\gamma_0+\gamma_1 a+\gamma_2 c))^2}{2\sigma_2^2}}$. The similar results hold for $E[f_2(M)|a, c]$.

Proof. Under the consistency, no unmeasured exposure-outcome confounding and no unmeasured mediator-outcome confounding, and with the models (3.2.1) and (3.2.2) being correctly specified, the expected CDE can be calculated as follows:

$$\begin{aligned}
E[Y_{am} - Y_{a^*m}|c] &= E[Y|a,m,c] - E[Y|a^*,m,c] \\
&= (\beta_0 + \beta_1 a + f_1(m)a + f_2(m)(1-a) + \beta_4 c) \\
&\quad - (\beta_0 + \beta_1 a^* + f_1(m)a^* + f_2(m)(1-a^*) + \beta_4 c \\
&= (\beta_1 + f_1(m) - f_2(m))(a - a^*).
\end{aligned}$$

Under the consistency, no unmeasured exposure-outcome confounding, no unmeasured mediator-outcome confounding, no unmeasured exposure-mediator confounding and no mediator-outcome confounding affected by exposure assumptions, and with the models (3.2.1) and (3.2.2) being correctly specified, the expected NDE and NIE can be calculated as follows:

$$\begin{aligned}
E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] &= \int_m \{E[Y|a,m,c] - E[Y|a^*,m,c]\} f(m|a^*,c) dm \\
&= \int_m \{(\beta_0 + \beta_1 a + f_1(m)a + f_2(m)(1-a) + \beta_4 c) \\
&\quad - (\beta_0 + \beta_1 a^* + f_1(m)a^* + f_2(m)(1-a^*) + \beta_4 c)\} f(m|a^*,c) dm \\
&= \int_m \{(\beta_1 + f_1(m) - f_2(m))(a - a^*)\} f(m|a^*,c) dm \\
&= (\beta_1 + E[f_1(M)|a^*,c] - E[f_2(M)|a^*,c])(a - a^*),
\end{aligned}$$

and

$$\begin{aligned}
E[Y_{aM_a} - Y_{aM_{a^*}}|c] &= \int_m E[Y|a,m,c]\{f(m|a,c) - f(m|a^*,c)\} dm \\
&= \int_m (\beta_0 + \beta_1 a + f_1(m)a + f_2(m)(1-a) + \beta_4 c)\{f(m|a,c) - f(m|a^*,c)\} dm \\
&= \{(\beta_0 + \beta_1 a + E[f_1(M)|a,c]a + E[f_2(M)|a,c](1-a) + \beta_4 c) \\
&\quad - (\beta_0 + \beta_1 a + E[f_1(M)|a^*,c]a + E[f_2(M)|a^*,c](1-a) + \beta_4 c)\} \\
&= a(E[f_1(M)|a,c] - E[f_1(M)|a^*,c]) + (1-a)(E[f_2(M)|a,c] - E[f_2(M)|a^*,c]).
\end{aligned}$$

Let the knot sequence for I-splines and C-splines be $L = t_1 = t_2 < t_3 < ... < t_k < t_{k+1} = t_{k+2} = U$.

If $f_1(M)$ is fitted using I-splines, then

$$f_1(M) = \beta_{21} I_1(M|2,t) + ... + \beta_{2k} I_k(M|2,t)$$

$$= \begin{cases} \beta_{21}(1 - \frac{(t_3-M)^2}{(t_3-t_2)(t_3-t_1)}) + \beta_{22}(\frac{(M-t_2)^2}{(t_3-t_2)(t_4-t_2)}), & \text{if } t_2 \le M < t_3 \\ \beta_{21} + \beta_{22}(1 - \frac{(t_4-M)^2}{(t_4-t_3)(t_4-t_2)}) + \beta_{23}(\frac{(M-t_3)^2}{(t_4-t_3)(t_5-t_3)}), & \text{if } t_3 \le M < t_4 \\ ... & \\ \beta_{21} + ... + \beta_{2,k-2} + \beta_{2,k-1}(1 - \dfrac{(t_{k+1}-M)^2}{(t_{k+1}-t_k)(t_{k+1}-t_{k-1})}) & \\ \quad + \beta_{2,k}(\dfrac{(M-t_k)^2}{(t_{k+1}-t_k)(t_{k+2}-t_k)}), & \text{if } t_k \le M < t_{k+1} \end{cases}.$$

The expectation of $f_1(M)$ given a and c can be calculated as

$$\begin{aligned} E[f_1(M)|a,c] &= E[\beta_{21} I_1(M|2,t) + ... + \beta_{2k} I_k(M|2,t)|a,c] \\ &= \int_m (\beta_{21} I_1(m|2,t) + ... + \beta_{2k} I_k(m|2,t)) f(m|a,c) dm \\ &= \sum_{i=2}^{k} \{ \int_{t_i}^{t_{i+1}} [\beta_{21} + ... + \beta_{2,i-2} + \beta_{2,i-1}(1 - \frac{(t_{i+1}-m)^2}{(t_{i+1}-t_i)(t_{i+1}-t_{i-1})}) \\ &\qquad + \beta_{2,i}(\frac{(m-t_i)^2}{(t_{i+1}-t_i)(t_{i+2}-t_i)})] f(m|a,c) dm \}, \end{aligned}$$

where $f(m|a,c) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{\frac{-(m-(\gamma_0+\gamma_1 a+\gamma_2 c))^2}{2\sigma_2^2}}$.

If $f_1(M)$ is fitted using C-splines, then

$$f_1(M) = \beta_{20}M + \beta_{21}C_1(M|2,t) + ... + \beta_{2k}C_k(M|2,t)$$

$$= \begin{cases} \beta_{20}M, & \text{if } M < t_2 \\ \beta_{20}M + \beta_{21}(M - \frac{t_1+t_2+t_3}{3} + \frac{(t_3-M)^3}{3(t_3-t_2)(t_3-t_1)}) + \beta_{22}(\frac{(M-t_2)^3}{3(t_3-t_2)(t_4-t_2)}), & \text{if } t_2 \le M < t_3 \\ \beta_{20}M + \beta_{21}(M - \dfrac{t_1+t_2+t_3}{3}) & \\ \quad + \beta_{22}(M - \dfrac{t_2+t_3+t_4}{3} + \dfrac{(t_4-M)^3}{3(t_4-t_3)(t_4-t_2)}) & \text{if } t_3 \le M < t_4 \\ \quad + \beta_{23}(\dfrac{(M-t_3)^3}{3(t_4-t_3)(t_5-t_3)}), & \\ ... & \\ \beta_{20}M + \beta_{21}(M - \dfrac{t_1+t_2+t_3}{3}) + ... + \beta_{2,k-2}(M - \dfrac{t_{k-2}+t_{k-1}+t_k}{3}) & \\ \quad + \beta_{2,k-1}(m - \dfrac{t_{k-1}+t_k+t_{k+1}}{3} + \dfrac{(t_{k+1}-m)^3}{3(t_{k+1}-t_k)(t_{k+1}-t_{k-1})}) & \text{if } t_k \le M < t_{k+1} \\ \quad + \beta_{2,k}(\dfrac{(m-t_k)^3}{3(t_{k+1}-t_k)(t_{k+2}-t_k)}), & \\ \beta_{20}M, & \text{if } M \ge t_{k+1} \end{cases}.$$

The expectation of $f_1(M)$ given a and c can be calculated as

$$\begin{aligned} E[f_1(M)|a,c] &= E[\beta_{20}M + \beta_{21}C_1(M|2,t) + ... + \beta_{2,k}C_k(M|2,t)|a,c] \\ &= \int_m (\beta_{20}m + \beta_{21}C_1(m|2,t) + ... + \beta_{2k}C_k(m|2,t))f(m|a,c)dm \\ &= \beta_{20}(\gamma_0 + \gamma_1 a + \gamma_2 c) \\ &\quad + \sum_{i=2}^{k}\{\int_{t_i}^{t_{i+1}}[\beta_{21}(m - \frac{t_1+t_2+t_3}{3}) + ... + \beta_{2,i-2}(m - \frac{t_{i-2}+t_{i-1}+t_i}{3}) \\ &\quad + \beta_{2,i-1}(m - \frac{t_{i-1}+t_i+t_{i+1}}{3} + \frac{(t_{i+1}-m)^3}{3(t_{i+1}-t_i)(t_{i+1}-t_{i-1})}) \\ &\quad + \beta_{2,i}(\frac{(m-t_i)^3}{3(t_{i+1}-t_i)(t_{i+2}-t_i)})]f(m|a,c)dm\}, \end{aligned}$$

where $f(m|a,c) = \frac{1}{\sqrt{2\pi\sigma_2^2}}e^{\frac{-(m-(\gamma_0+\gamma_1 a+\gamma_2 c))^2}{2\sigma_2^2}}$. □

3.2.2.3 Mediation effects inference

According to Proposition 3.2.1, the expected CDE is a function of β_1, β_2 and β_3, the expected NDE is a function of β_1, β_2, β_3, γ_0, γ_1, γ_2 and σ_2^2, and the expected NIE is a function of β_2, γ_0, γ_1, γ_2 and σ_2^2. In the parameter estimation procedure described in Section 3.2.2.1, when we estimate the parameters in model (3.2.1), we keep the necessary columns of spline bases and estimate the parameters using ordinary least squares. Therefore, we can apply delta method to obtain the asymptotic variances of different mediation effects, and the asymptotic variances can be estimated by plugging in the estimated parameters. Once we estimate the variances of different mediation effects, the 95% confidence intervals can be obtained as $(g(\hat{\theta}) - z_{0.975}\sqrt{\widehat{var(g(\hat{\theta}))}}, g(\hat{\theta}) + z_{0.975}\sqrt{\widehat{var(g(\hat{\theta}))}})$. The technical details are shown in Proposition 3.2.2.

Proposition 3.2.2. *Let $\theta_{CDE} = (\beta_1, \beta_2, \beta_3)$, $\theta_{NDE} = (\beta_1, \beta_2, \beta_3, \gamma_0, \gamma_1, \gamma_2, \sigma_2^2)$ and $\theta_{NIE} = (\beta_2, \gamma_0, \gamma_1, \gamma_2, \sigma_2^2)$, where $\beta_2 = (\beta_{21}, ..., \beta_{2k})$ if $f_1(M)$ is fitted using I-splines or $\beta_2 = (\beta_{20}, \beta_{21}, ..., \beta_{2k})$ if $f_1(M)$ is fitted using C-splines, and $\beta_3 = (\beta_{31}, ..., \beta_{3k})$ if $f_2(M)$ is fitted using I-splines or $\beta_3 = (\beta_{30}, \beta_{31}, ..., \beta_{3k})$ if $f_2(M)$ is fitted using C-splines. Denote the expected controlled direct effect as $g_{CDE}(\theta_{CDE})$, the expected natural direct effect as $g_{NDE}(\theta_{NDE})$ and the expected natural indirect effect as $g_{NIE}(\theta_{NIE})$. Then the asymptotic variances of expected CDE, NDE and NIE are $\nabla_{\theta_{CDE}} g_{CDE}(\theta_{CDE})^T \Sigma_{\theta_{CDE}} \nabla_{\theta_{CDE}} g_{CDE}(\theta_{CDE})$, $\nabla_{\theta_{NDE}} g_{NDE}(\theta_{NDE})^T \Sigma_{\theta_{NDE}} \nabla_{\theta_{NDE}} g_{NDE}(\theta_{NDE})$ and $\nabla_{\theta_{NIE}} g_{NIE}(\theta_{NIE})^T \Sigma_{\theta_{NIE}} \nabla_{\theta_{NIE}} g_{NIE}(\theta_{NIE})$ respectively, where $\Sigma_{\theta_{CDE}}$, $\Sigma_{\theta_{NDE}}$ and $\Sigma_{\theta_{NIE}}$ are the covariance matrices corresponding to θ_{CDE}, θ_{NDE} and θ_{NIE}. If $f_1(M)$ is fitted using I-splines, then*

$$\begin{aligned}
\frac{\partial E[f_1(M)|a,c]}{\partial \beta_{2i}} = &\int_{t_i}^{t_{i+1}} \frac{(m-t_i)^2}{(t_{i+1}-t_i)(t_{i+2}-t_i)} f(m|a,c)dm \\
&+ \int_{t_{i+1}}^{t_{i+2}} (1 - \frac{(t_{i+2}-m)^2}{(t_{i+2}-t_{i+1})(t_{i+2}-t_i)}) f(m|a,c)dm \\
&+ \int_{t_{i+2}}^{t_{i+3}} f(m|a,c)dm + ... \\
&+ \int_{t_k}^{t_{k+1}} f(m|a,c)dm, i = 1, ..., k,
\end{aligned} \tag{3.2.16}$$

$$\frac{\partial E[f_1(M)|a,c]}{\partial \gamma_0} = \sum_{i=2}^{k}\{\int_{t_i}^{t_{i+1}}[\beta_{21}+...+\beta_{2,i-2}+\beta_{2,i-1}(1-\frac{(t_{i+1}-m)^2}{(t_{i+1}-t_i)(t_{i+1}-t_{i-1})}) + \beta_{2,i}(\frac{(m-t_i)^2}{(t_{i+1}-t_i)(t_{i+2}-t_i)})]f(m|a,c)\frac{2(m-(\gamma_0+\gamma_1 a+\gamma_2 c)}{2\sigma_2^2}dm\}, \tag{3.2.17}$$

$$\frac{\partial E[f_1(M)|a,c]}{\partial \sigma_2^2} = \sum_{i=2}^{k}\{\int_{t_i}^{t_{i+1}}[\beta_{21}+...+\beta_{2,i-2}+\beta_{2,i-1}(1-\frac{(t_{i+1}-m)^2}{(t_{i+1}-t_i)(t_{i+1}-t_{i-1})}) + \beta_{2,i}(\frac{(m-t_i)^2}{(t_{i+1}-t_i)(t_{i+2}-t_i)})]f(m|a,c)(-\frac{1}{2\sigma_2^2}+\frac{(m-(\gamma_0+\gamma_1 a+\gamma_2 c))^2}{2(\sigma_2^2)^2})dm\}, \tag{3.2.18}$$

where $f(m|a,c)=\frac{1}{\sqrt{2\pi\sigma_2^2}}e^{\frac{-(m-(\gamma_0+\gamma_1 a+\gamma_2 c))^2}{2\sigma_2^2}}$, *and* $\frac{\partial E[f_1(M)|a,c]}{\partial \gamma_1}$ *and* $\frac{\partial E[f_1(M)|a,c]}{\partial \gamma_2}$ *have similar results as* $\frac{\partial E[f_1(M)|a,c]}{\partial \gamma_0}$. *If* $f_1(M)$ *is fitted using C-splines, then*

$$\frac{\partial E[f_1(M)|a,c]}{\partial \beta_{20}} = \gamma_0+\gamma_1 a+\gamma_2 c, \tag{3.2.19}$$

$$\begin{aligned}\frac{\partial E[f_1(M)|a,c]}{\partial \beta_{2i}} &= \int_{t_i}^{t_{i+1}}\frac{(m-t_i)^3}{3(t_{i+1}-t_i)(t_{i+2}-t_i)}f(m|a,c)dm \\ &+\int_{t_{i+1}}^{t_{i+2}}(m-\frac{t_i+t_{i+1}+t_{i+2}}{3}+\frac{(t_{i+2}-m)^3}{3(t_{i+2}-t_{i+1})(t_{i+2}-t_i)})f(m|a,c)dm \\ &+\int_{t_{i+2}}^{t_{i+3}}(m-\frac{t_i+t_{i+1}+t_{i+2}}{3})f(m|a,c)dm+... \\ &+\int_{t_k}^{t_{k+1}}(m-\frac{t_i+t_{i+1}+t_{i+2}}{3})f(m|a,c)dm, i=1,...,k,\end{aligned} \tag{3.2.20}$$

$$\begin{aligned}\frac{\partial E[f_1(M)|a,c]}{\partial \gamma_0} &= \beta_{20} \\ &+\sum_{i=2}^{k}\{\int_{t_i}^{t_{i+1}}[\beta_{21}(m-\frac{t_1+t_2+t_3}{3})+...+\beta_{2,i-2}(m-\frac{t_{i-2}+t_{i-1}+t_i}{3}) \\ &+\beta_{2,i-1}(m-\frac{t_{i-1}+t_i+t_{i+1}}{3}+\frac{(t_{i+1}-m)^3}{3(t_{i+1}-t_i)(t_{i+1}-t_{i-1})}) \\ &+\beta_{2,i}(\frac{(m-t_i)^3}{3(t_{i+1}-t_i)(t_{i+2}-t_i)})]f(m|a,c)\frac{2(m-(\gamma_0+\gamma_1 a+\gamma_2 c)}{2\sigma_2^2}dm\},\end{aligned} \tag{3.2.21}$$

$$\frac{\partial E[f_1(M)|a,c]}{\partial \sigma_2^2} = \sum_{i=2}^{k}\{\int_{t_i}^{t_{i+1}}[\beta_{21}(m - \frac{t_1+t_2+t_3}{3}) + ... + \beta_{2,i-2}(m - \frac{t_{i-2}+t_{i-1}+t_i}{3}) + \beta_{2,i-1}(m - \frac{t_{i-1}+t_i+t_{i+1}}{3} + \frac{(t_{i+1}-m)^3}{3(t_{i+1}-t_i)(t_{i+1}-t_{i-1})}) + \beta_{2,i}(\frac{(m-t_i)^3}{3(t_{i+1}-t_i)(t_{i+2}-t_i)})]f(m|a,c)(-\frac{1}{2\sigma_2^2} + \frac{(m-(\gamma_0+\gamma_1 a+\gamma_2 c))^2}{2(\sigma_2^2)^2})dm\}, \tag{3.2.22}$$

where $f(m|a,c) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{\frac{-(m-(\gamma_0+\gamma_1 a+\gamma_2 c))^2}{2\sigma_2^2}}$, *and* $\frac{\partial E[f_1(M)|a,c]}{\partial \gamma_1}$ *and* $\frac{\partial E[f_1(M)|a,c]}{\partial \gamma_2}$ *have similar results as* $\frac{\partial E[f_1(M)|a,c]}{\partial \gamma_0}$.

3.3 Simulation

We evaluate the performance of our method by measuring the coverage probability, average absolute relative bias and average mean squared error (MSE). The data set is created with the similar characteristics as a prenatal screening program data set. The simulated data set contains 500 observations and 10 variables. The confounding variables are age (randomly sampled from 18 to 40 with an increment 0.5), inverse maternal weight (randomly sampled from 2 to 14.3 with an increment of 0.1), race (randomly sampled from 1 to 5 with probabilities 0.46, 0.28, 0.13, 0.1 and 0.03 respectively), season of blood draw (randomly sampled from 1 to 4 with the same probability 0.25), smoking status (randomly sampled from a binomial distribution with success probability 0.25), ovum donor status (randomly sampled from a binomial distribution with success probability 0.15) and pre-existing diabetes status (randomly sampled from a binomial distribution with success probability 0.2). The exposure variable is pesticide exposure (randomly sampled from a binomial distribution with success probability 0.5). The mediator variable is hormone (gestational age multiple of median, calculated via exposure-mediator model). The outcome variable is birth weight (grams, calculated via exposure-outcome model).

We consider 6 different combinations of functions for exposure-outcome model, which are summarized in Table 3.3.1, and the corresponding plots are shown in Figure 3.3.1. We

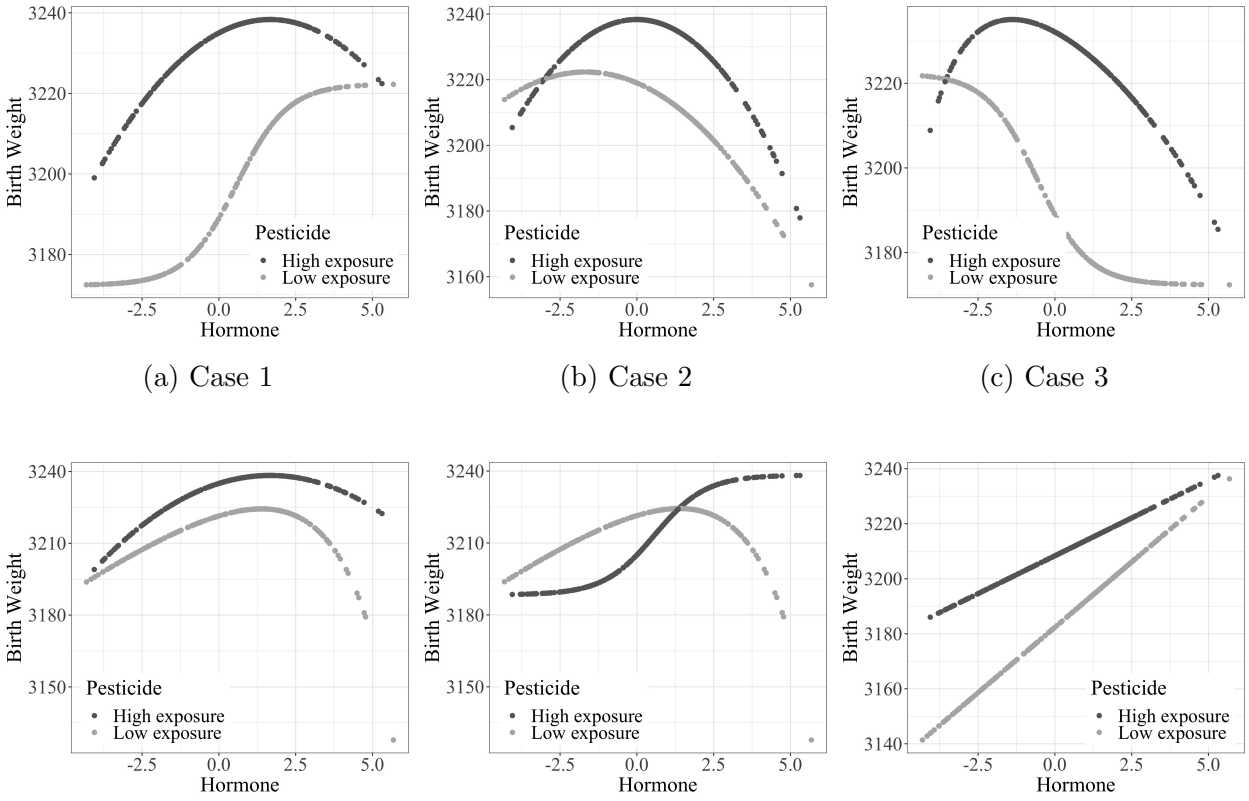
fix the variance of ϵ_2 in exposure-mediator model at 0.3^2, and ϵ_1 in exposure-outcome model is draw from $N(0, 10^2)$, $N(0, 20^2)$, $N(0, 30^2)$ and $N(0, 40^2)$. The number of bases is set to 5. For CDE, the mediator is set to its mean value. Since the parameters β_1, γ_0, γ_1, γ_2 and σ_2^2 and the functions $f_1(m)$ and $f_2(m)$ are known, the true effects can be calculated using the formulas in proposition 3.2.1. The number of simulations is 500, and the results of coverage probability, average absolute relative bias and average MSE are shown in Tables 3.3.2 - 3.3.7 (the plots can be found in Appendix B.2).

Table 3.3.1: Different combinations of functions for exposure-outcome model

Case No.	$f_1(M)$	$f_2(M)$
1	$\frac{-6(M-5/3)^2}{5} + 100$	$\frac{50e^{6M/5}}{2+e^{6M/5}} + 50$
2	$\frac{-e^M-100M^2}{50} + 100$	$\frac{-6(M+5/3)^2}{5} + 100$
3	$70\ln(-e^{-M/2} - M + 10) - 60$	$\frac{50e^{-6M/5}}{2+e^{-6M/5}} + 50$
4	$\frac{-6(M-5/3)^2}{5} + 100$	$300\ln(-e^{M/2} + M + 40) - 1000$
5	$\frac{50e^{6M/5}}{2+e^{6M/5}} + 50$	$300\ln(-e^{M/2} + M + 40) - 1000$
6	$5.5M + 70$	$9.5M + 60$

If $f_1(M)$ and $f_2(M)$ are not linear, semi-parametric shape-restricted regression spline outperforms linear regression in general. Using semi-parametric shape-restricted regression spline, the coverage probability always keeps around 95%, but using linear regression, the coverage probability tends to be 0 when the variance of ϵ_1 is small and the effect size is large (see Proposition 3.3.1). Both average absolute relative bias and average MSE of the estimated effects from semi-parametric shape-restricted regression spline become large when the variance of ϵ_1 increases, but they are much smaller than those from linear regression, especially for small variance of ϵ_1 and large effect size. If $f_1(M)$ and $f_2(M)$ are linear, although linear regression performs better, the metrics from semi-parametric shape-restricted regression spline are still acceptable.

Proposition 3.3.1. *If the mediation effect from linear regression deviates from the true mediation effect, then as the variance of ϵ_1 decreases, the coverage probability decreases.*



(a) Case 1 (b) Case 2 (c) Case 3

(d) Case 4 (e) Case 5 (f) Case 6

Figure 3.3.1: Plots of hormone vs. birth weight varying by pesticide under different cases

Proof. Using linear regression, the exposure-outcome model becomes

$$Y = \beta_0 + \beta_1 A + \beta_2 M + \beta_3 AM + \beta_4 C + \epsilon_1 \tag{3.3.1}$$

where $\epsilon_1 \sim N(0, \sigma_1^2)$, while the exposure-mediator model keeps the same as model (3.2.2).

The expected CDE, NDE and NIE, conditioning on $C = c$, are given by

$$E[Y_{am} - Y_{a^*m}|c] = (\beta_1 + \beta_3 m)(a - a^*), \tag{3.3.2}$$

$$E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}|c] = (\beta_1 + \beta_3(\gamma_0 + \gamma_1 a^* + \gamma_2 c))(a - a^*), \tag{3.3.3}$$

Table 3.3.2: Simulation results of coverage probability, average absolute relative bias and average MSE for case 1 (true CDE: ∼43.99, true NDE: 44.85, true NIE: 1.030)

Semi-parametric shape-restricted regression spline									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0.936	0.954	0.964	0.031	0.030	0.407	3.453	2.719	0.283
20^2	0.942	0.938	0.954	0.058	0.055	0.720	10.83	9.597	0.906
30^2	0.946	0.950	0.958	0.083	0.080	1.029	21.80	20.02	1.862
40^2	0.946	0.958	0.956	0.107	0.103	1.335	36.07	33.86	3.151
Linear regression									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0	0	0.916	0.186	0.193	0.432	74.61	76.36	0.311
20^2	0.032	0.024	0.942	0.184	0.192	0.746	76.58	78.37	0.957
30^2	0.268	0.226	0.938	0.183	0.190	1.090	80.74	82.59	2.045
40^2	0.512	0.458	0.938	0.183	0.190	1.443	87.09	89.01	3.575

Table 3.3.3: Simulation results of coverage probability, average absolute relative bias and average MSE for case 2 (true CDE: ∼21.10, true NDE: 19.58, true NIE: -0.292)

Semi-parametric shape-restricted regression spline									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0.952	0.976	0.980	0.072	0.071	1.375	3.328	2.963	0.265
20^2	0.946	0.962	0.970	0.131	0.131	2.694	10.97	10.15	1.008
30^2	0.952	0.964	0.966	0.185	0.186	3.995	21.97	20.68	2.209
40^2	0.942	0.960	0.966	0.235	0.238	5.283	35.57	33.92	3.859
Linear regression									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0.552	0.638	0.928	0.124	0.116	1.804	6.798	6.376	0.439
20^2	0.802	0.830	0.940	0.133	0.129	2.827	9.748	9.376	1.112
30^2	0.884	0.896	0.940	0.157	0.157	3.999	14.89	14.58	2.228
40^2	0.912	0.926	0.940	0.189	0.191	5.216	22.21	21.99	3.786

Table 3.3.4: Simulation results of coverage probability, average absolute relative bias and average MSE for case 3 (true CDE: ∼46.84, true NDE: 43.58, true NIE: -1.308)

Semi-parametric shape-restricted regression spline									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0.902	0.926	0.980	0.033	0.034	0.300	3.060	3.501	0.254
20^2	0.940	0.940	0.972	0.059	0.061	0.539	10.56	11.04	0.828
30^2	0.938	0.946	0.962	0.084	0.087	0.780	22.33	22.88	1.738
40^2	0.938	0.944	0.956	0.109	0.113	1.012	37.37	38.00	2.947
Linear regression									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0	0	0.792	0.235	0.222	0.501	107.1	94.50	0.637
20^2	0.002	0.008	0.886	0.233	0.220	0.711	108.8	96.36	1.324
30^2	0.082	0.132	0.930	0.232	0.218	0.961	112.6	100.1	2.453
40^2	0.288	0.366	0.940	0.230	0.217	1.226	118.7	106.7	4.023

Table 3.3.5: Simulation results of coverage probability, average absolute relative bias and average MSE for case 4 (true CDE: ∼13.64, true NDE: 13.60, true NIE: 1.030)

Semi-parametric shape-restricted regression spline									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0.936	0.954	0.968	0.104	0.100	0.390	3.090	2.864	0.261
20^2	0.942	0.956	0.964	0.187	0.182	0.755	10.13	9.578	0.978
30^2	0.956	0.962	0.962	0.264	0.258	1.119	20.19	19.27	2.141
40^2	0.954	0.962	0.956	0.339	0.331	1.482	33.10	31.79	3.744
Linear regression									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0.742	0.802	0.930	0.140	0.125	0.494	4.794	3.95	0.416
20^2	0.858	0.876	0.936	0.175	0.165	0.792	8.388	7.53	1.085
30^2	0.896	0.902	0.934	0.224	0.216	1.124	14.17	13.30	2.195
40^2	0.908	0.926	0.940	0.277	0.271	1.469	22.14	21.28	3.748

Table 3.3.6: Simulation results of coverage probability, average absolute relative bias and average MSE for case 5 (true CDE: ∼-14.37, true NDE: -15.26, true NIE: 4.191)

Semi-parametric shape-restricted regression spline									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0.928	0.944	0.906	0.100	0.088	0.186	3.484	2.893	0.923
20^2	0.938	0.948	0.912	0.185	0.163	0.278	10.76	9.848	2.132
30^2	0.950	0.952	0.932	0.265	0.235	0.381	21.52	20.45	4.067
40^2	0.950	0.954	0.936	0.342	0.305	0.488	35.49	34.52	6.706
Linear regression									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0	0	0.128	0.655	0.615	0.467	92.92	89.39	4.225
20^2	0.010	0.008	0.496	0.660	0.620	0.469	97.65	94.08	4.809
30^2	0.166	0.148	0.720	0.666	0.625	0.491	104.6	101.0	5.835
40^2	0.386	0.358	0.804	0.674	0.632	0.532	113.7	110.1	7.303

Table 3.3.7: Simulation results of coverage probability, average absolute relative bias and average MSE for case 6 (true CDE: ∼25.20, true NDE: 25.65, true NIE: 1.65)

Semi-parametric shape-restricted regression spline									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0.938	0.936	0.956	0.056	0.054	0.299	3.043	3.027	0.378
20^2	0.936	0.930	0.902	0.108	0.105	0.541	11.39	11.40	1.262
30^2	0.944	0.936	0.866	0.158	0.153	0.773	24.31	24.27	2.691
40^2	0.942	0.940	0.854	0.206	0.200	1.019	41.78	41.60	4.819
Linear regression									
Variance of ϵ_1	Coverage Probability			Average |Relative Bias|			Average MSE		
	CDE	NDE	NIE	CDE	NDE	NIE	CDE	NDE	NIE
10^2	0.936	0.934	0.938	0.033	0.032	0.235	1.101	1.104	0.239
20^2	0.936	0.936	0.942	0.066	0.065	0.453	4.357	4.405	0.897
30^2	0.936	0.934	0.942	0.099	0.097	0.676	9.801	9.908	1.996
40^2	0.936	0.934	0.944	0.132	0.130	0.900	17.43	17.62	3.538

and

$$E[Y_{aM_a} - Y_{aM_{a^*}}|c] = (\beta_2\gamma_1 + \beta_3\gamma_1 a)(a - a^*), \tag{3.3.4}$$

respectively.

Let $X = [1, A, M, AM, C]$, then $\hat{\beta} \sim N(\beta, \sigma_1^2(X^TX)^{-1})$. The expected CDE, NDE and NIE can all be expressed as a linear combination of β. Let $\hat{\theta}_{LR} = a\hat{\beta} \sim N(a\beta, \sigma_1^2 a(X^TX)^{-1}a^T)$. Then,

$$\begin{aligned}
&P(|\hat{\theta}_{LR} - \theta_{true}| \leq z_{\alpha/2}\sqrt{Var(\hat{\theta}_{LR})}) \\
&= P(-z_{\alpha/2}\sqrt{Var(\hat{\theta}_{LR})} \leq \hat{\theta}_{LR} - \theta_{true} \leq z_{\alpha/2}\sqrt{Var(\hat{\theta}_{LR})}) \\
&= P(-z_{\alpha/2} \leq \frac{\hat{\theta}_{LR} - \theta_{LR} + \theta_{LR} - \theta_{true}}{\sqrt{Var(\hat{\theta}_{LR})}} \leq z_{\alpha/2}) \\
&= P(-z_{\alpha/2} + \frac{\theta_{true} - \theta_{LR}}{\sqrt{Var(\hat{\theta}_{LR})}} \leq \frac{\hat{\theta}_{LR} - \theta_{LR}}{\sqrt{Var(\hat{\theta}_{LR})}} \leq z_{\alpha/2} + \frac{\theta_{true} - \theta_{LR}}{\sqrt{Var(\hat{\theta}_{LR})}}) \\
&= \phi(z_{\alpha/2} + \frac{\theta_{true} - \theta_{LR}}{\sigma_1\sqrt{a(X^TX)^{-1}a^T}}) - \phi(-z_{\alpha/2} + \frac{\theta_{true} - \theta_{LR}}{\sigma_1\sqrt{a(X^TX)^{-1}a^T}}).
\end{aligned}$$

Therefore, as σ_1 decreases, the coverage probability decreases if $\theta_{true} - \theta_{LR}$ is large. □

3.4 Discussion

If researchers have evidence that the relationship between the mediator and the outcome is curvilinear, then using linear regression to build exposure-outcome model will result in biased estimator. Our method is designed to relax the linearity assumption when building the exposure-outcome model to reduce the bias introduced by model misspecification. In our method, the exposure-outcome model is specified using quadratic I-spline basis functions and/or cubic C-spline basis functions depending upon the prior knowledge on the specific shapes between the mediator and the outcome in both exposure and non-exposure groups. The parameter estimation procedure follows the logic of method proposed by Meyer (2018), and since the model involves the factor-by-curve interaction, the estimation procedure is slightly extended. The core algorithm used in the estimation procedure is hinge algorithm (Meyer, 2013), which selects the necessary columns of spline bases and forces the coefficients

of unselected columns to be 0. Once the parameters in both exposure-mediator and exposure-outcome models are estimated, the mediation effects as well as their asymptotic variances can be obtained analytically.

We extended the regression-based mediation analysis into the shape-restricted framework, where the relationship between the mediator and the outcome is not limited to linear. Our simulation study suggested that the proposed method performs well in terms of coverage probability, average absolute relative bias and average mean squared error. If researchers have prior knowledge on the specific shapes between the mediator and the outcome (increasing, decreasing, convex or concave), then they can apply our method. Although monotonic curves do not allow peaks and valleys and convex or concave curves do not allow any sort of wiggling, in order to make more precise predictions, the number of knots and the knots placement may still need to be considered. The proposed method is not suitable if shapes other than monotonic, convex and concave are considered. In such cases, researchers may apply the simulation-based method developed by Imai et al. (2010).

4.0 Illustrations

4.1 Shape Detection using Semi-parametric Shape-Restricted Mixed Effects Regression Spline

Relationships between several serum placental-fetal biomarkers and birth weight are well-studied in the literature. For example, maternal serum levels of PAPP-A were reported to be significantly lower in SGA (small for gestational age) newborns than in controls and significantly higher in LGA (large for gestational age) newborns than in controls (Tul et al., 2003; Canini et al., 2008), and maternal serum estriol levels in the 29^{th} week and at delivery were significantly positively correlated with birth weight (Nagata et al., 2006). However, the relationships are usually examined on specific birth cohort rather than at the population-level and the overall trends are rarely reported.

Using a population-level data set from a prenatal screening program, we model the relationships between 1st and 2nd trimester maternal serum placental-fetal biomarkers (GA-MoM) and neonatal birth weight (gram) (N = 810,812, N_{female} = 397,820, N_{male} = 409,653). We consider serum placental-fetal biomarkers, hCG, PAPP-A, estriol, AFP and inhibin-A in our analysis, and all analyses are stratified by fetal sex and adjusted for maternal race, year of blood draw, month of blood draw, smoking status, ovum donor status, pre-existing diabetes status, maternal age and inverse maternal weight. To adjust for confounding by geographic variability across the state and all the unmeasured confounders that determine where a woman lives, we treated zip code of maternal residence as a random effect.

Based on scientific literature, we believe that the relationships between serum placental-fetal biomarkers and neonatal birth weight have a specific shape (increasing, decreasing, convex or concave) but not always linear. Therefore, we apply our method to the data to detect the underlying shape in each relationship. Because we have a large sample size, we set the number of internal knots to 4 and let the knots be placed at the 20^{th}, 40^{th}, 60^{th} and 80^{th} percentiles of the data. The results are shown in Table 4.1.1.

According to the results in Table 4.1.1, we can make some relevant inferences on the

Table 4.1.1: Shape detection results on population-level prenatal screening program data

Female Infant						
	1^{st}-trimester hCG	PAPP-A	2^{nd}-trimester hCG	Estriol	AFP	Inhibin-A
T_1	0.9985	0.9982	0.8594	0.9994	0	0.3330
(p-value)	(0.0034)	(0.0031)	(0.0567)	(0.0039)	(1)	(0.3840)
T_2	0	0	0.0189	0	1	0.5628
(p-value)	(1)	(1)	(0.7535)	(1)	(<0.0001)	(0.2210)
T_3	0	0	0	0	0	0
(p-value)	(1)	(1)	(1)	(1)	(1)	(1)
T_4	0.9733	1	1	0.9975	0.7996	0.9830
(p-value)	(0.0056)	(<0.0001)	(<0.0001)	(0.0014)	(0.0658)	(0.0045)
Shape Number	⑥	⑥	④	⑥	②	④
Male Infant						
	1^{st}-trimester hCG	PAPP-A	2^{nd}-trimester hCG	Estriol	AFP	Inhibin-A
T_1	0.9952	0.9994	0.6048	0.9968	0	0.1546
(p-value)	(0.0043)	(0.0024)	(0.1852)	(0.0046)	(1)	(0.5660)
T_2	0	0	0.1611	0	0.9974	0.7640
(p-value)	(1)	(1)	(0.5401)	(1)	(0.0056)	(0.1075)
T_3	0	0	0	0	0	0
(p-value)	(1)	(1)	(1)	(1)	(1)	(1)
T_4	1	0.9993	0.9976	0.7940	0.9019	0.9879
(p-value)	(<0.0001)	(0.0003)	(0.0009)	(0.0697)	(0.0276)	(0.0035)
Shape Number	⑥	⑥	④	①	②	④

relationships between placental-fetal hormones and birth weight. The relationships between 1^{st}-trimester hCG and birth weight and between PAPP-A and birth weight are categorized as concave with increasing trend in both female and male infants. The relationships between 2^{nd}-trimester hCG and birth weight and between Inhibin-A and birth weight are categorized as concave in both female and male infants. The relationships between AFP and birth weight are categorized as decreasing in both female and male infants. The relationship between Estriol and birth weight is categorized as concave with increasing trend in female infants while it is categorized as increasing in male infants.

Results from our methodology will help researchers to make judgements on the potential relationships between maternal serum placental-fetal biomarkers and neonatal birth weight. With such judgements, researchers can correctly choose corresponding prediction methods to make relevant predictions.

4.2 Mediation Analysis using Semi-parametric Shape-Restricted Regression Spline

Pesticide is used in agriculture to control pests and improve yields. However, it usually has negative influences on organisms and ecosystem and may also negatively impact birth outcomes of human beings (Larsen et al., 2017). Larsen et al. (2017) reported that for individuals in high exposure group in 1^{st}-trimester pregnancies, the birth weight is about 13 grams lower, and being in high exposure group reduces gestational age and increases the probability of preterm birth and the probability of birth abnormality. Chemicals often do not directly affect the fetuses but instead they alter the levels of placental biomarkers and further influence the birth outcomes, which is regarded as a placentally-mediated effect (Adibi et al., 2021a). In order to examining the mediation effects of placental biomarkers, we perform the mediation analysis using semi-parametric shape-restricted regression spline.

The pesticide data is from California's pesticide use reporting (PUR) program, where pesticide use is reported monthly. The data is at the county level, so for each zip code, we have a pesticide exposure data point. We merge the pesticide data with the population-

level prenatal screening data, dichotomize the pesticide exposure variable using its median, and model the relationship among dichotomized pesticide exposure, 1st trimester hCG (GA-MoM) and neonatal birth weight (gram). The analyses are stratified by fetal sex and adjusted for maternal race, year of blood draw, month of blood draw, smoking status, ovum donor status, pre-existing diabetes status, maternal age and inverse maternal weight.

We apply our method to several subsets of data with specific types of pesticide. For permethrin, within male infants cohort, there are 21,433 observations, the relationship between 1st trimester hCG and neonatal birth weight is inferred as increasing in above-median exposure group and as concave with increasing trend in below-median exposure group; within female infants cohort, there are 20,895 observations, the relationship between 1st trimester hCG and neonatal birth weight is inferred as concave with increasing trend in above-median exposure group and as increasing in below-median exposure group. For glyphosate isopropylamine salt, within male infants cohort, there are 56,299 observations, the relationship between 1st trimester hCG and neonatal birth weight is inferred as concave with increasing trend in above-median exposure group and as increasing in below-median exposure group; within female infants cohort, there are 55,052 observations, the relationship between 1st trimester hCG and neonatal birth weight is inferred as increasing in above-median exposure group and as concave with increasing trend in below-median exposure group. To perform mediation analysis using semi-parametric shape-restricted regression spline, the number of bases is set to 5 and the confounding variables are controlled at their mean values; for CDE, the mediator is set to its mean value. The results are summarized in Table 4.2.1.

For permethrin exposure, when the confounding variables are controlled at their mean values, within female infants cohort, the CDE with the mediator being set to its mean value is 13.27 (95% C.I.: -4.09 - 30.63), the NDE is 6.70 (95% C.I.: -5.25 - 18.65) and the NIE is -1.19 (95% C.I.: -2.06 - -0.31); within male infants cohort, the CDE with the mediator being set to its mean value is 12.45 (95% C.I.: -3.55 - 28.46), the NDE is 15.67 (95% C.I.: 3.53 - 27.80) and the NIE is -2.42 (95% C.I.: -3.55 - -1.29). For glyphosate isopropylamine salt exposure, when the confounding variables are controlled at their mean values, within female infants cohort, the CDE with the mediator being set to its mean value is 4.95 (95% C.I.: -6.96 - 16.87), the NDE is 8.89 (95% C.I.: -11.68 - 29.46) and the NIE is -1.63 (95% C.I.:

Table 4.2.1: Mediation analysis results on population-level prenatal screening program data

Female Infant			
	CDE with 95% C.I.	NDE with 95% C.I.	NIE with 95% C.I.
permethrin	13.270 (-4.0927, 30.633)	6.7042 (-5.2455, 18.654)	-1.1889 (-2.0636, -0.3142)
glyphosate isopropylamine salt	4.9524 (-6.9646, 16.869)	8.8868 (-11.683, 29.457)	-1.6303 (-2.2427, -1.0178)
Male Infant			
	CDE with 95% C.I.	NDE with 95% C.I.	NIE with 95% C.I.
permethrin	12.454 (-3.5468, 28.455)	15.665 (3.5325, 27.797)	-2.4189 (-3.5479, -1.2899)
glyphosate isopropylamine salt	8.0234 (-3.0360, 19.083)	5.8625 (-1.6837, 13.409)	-2.4338 (-3.1993, -1.6683)

-2.24 - -1.02); within male infants cohort, the CDE with the mediator being set to its mean value is 8.02 (95% C.I.: -3.04 - 19.08), the NDE is 5.86 (95% C.I.: -1.68 - 13.41) and the NIE is -2.43 (95% C.I.: -3.20 - -1.67). The NIE is significant in each case, indicating that 1st trimester hCG has significant impact on the relationship between permethrin/glyphosate isopropylamine salt exposure and neonatal birth weight within male infants cohort or female infants cohort.

5.0 Discussion and Future Work

5.1 Discussion and Future Work for Chapter 2

The linear regression is a widely used statistical model in numerous fields because it is easy to be applied and the results are easy to be interpreted. However, if the underlying pattern between two continuous variables is curvilinear, using linear regression will result in model misspecification and introduce bias. In such a scenario, nonparametric techniques, such as splines, should be introduced to build a flexible model. In the first part of the dissertation, we develop a shape detection method to help researchers identify the most probable shape of relationship between two continuous variables among increasing, decreasing, convex and concave shapes while controlling for confounders and accounting for random effects. In specific, we build the mixed effects models, derive a test statistic to test the null hypothesis of constant function against the alternative that there is an underlying shape, and apply Holm-Bonferroni method to classify the underlying shape into a reasonable category. The proposed method is based on the properties of I-splines and C-splines, i.e., a linear combination of quadratic I-splines is increasing if and only if the coefficients are positive, and a linear combination of cubic C-splines is convex if and only if the coefficients are positive.

5.1.1 Group-level shape detection using semi-parametric shape-restricted regression splines

In practice, the functional form of the covariate effect may vary across different groups. For example, the association between circulating levels of the placental hormone human chorionic gonadotropin (hCG) and infant anogenital distance may be different in different maternal stressful life event groups. Coull et al. (2001) proposed a method to incorporate factor-by-curve interactions into generalized additive models, where they used penalized spline method and truncated power bases. We can adopt the idea from Coull et al. (2001) to extend the original shape-restricted regression spline by incorporating the factor-by-curve

interaction. In order to determine the group-level shape of the curve, the constraints should be applied to both main and interaction effects.

If the interaction between two groups is absent, we will fit a semi-parametric shape-restricted regression spline model of the form

$$y_i = \beta_0 + \beta_1 a_i + f(x_i) + \beta_2 c_{1i} + ... + \beta_p c_{pi} + \epsilon_i, \tag{5.1.1}$$

where x_i is a continuous predictor, y_i is a continuous outcome variable, a_i is a binary group variable (exposure vs. non-exposure, male vs. female, etc.), $c_{1i}, ..., c_{pi}$ are potential confounding variables, and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. In this model, $f(x_i)$ is the curve of the predictor, and it can be fitted using the combination of quadratic I-spline bases, i.e., $f(x_i) = \theta_1 I_1(x_i|2,t) + \theta_2 I_2(x_i|2,t) + ... + \theta_k I_k(x_i|2,t)$, or using the combination of cubic C-spline bases, i.e., $f(x_i) = \theta_0 x_i + \theta_1 C_1(x_i|2,t) + \theta_2 C_2(x_i|2,t) + ... + \theta_k C_k(x_i|2,t)$. We estimate all parameters using ordinary least-squares, and project the unconstrained estimates of $\theta = (\theta_1, ..., \theta_k)^T$, denoted as $\hat{\theta}$, onto a suitable polyhedral cone with respect to $\Sigma_\theta = Cov(\hat{\theta})$ to obtain the constrained estimates of θ, $\tilde{\theta} = \arg\min_{\theta \in O} (\theta - \hat{\theta})^T \Sigma_\theta^{-1} (\theta - \hat{\theta})$. In practice, Σ_θ is usually unknown, and thus the estimate of Σ_θ, $\hat{\Sigma}_\theta = \widehat{Cov(\hat{\theta})}$, is used. We test the hypotheses $H_0 : \theta = 0$ vs. $H_a : \theta \in O$ using the test statistic $T = \frac{\hat{\theta}^T \Sigma_\theta^{-1} \hat{\theta} - \min_{\theta \in O} (\theta - \hat{\theta})^T \Sigma_\theta^{-1} (\theta - \hat{\theta})}{\hat{\theta}^T \Sigma_\theta^{-1} \hat{\theta}}$. Once we obtain four test statistics, we will apply Holm-Bonferroni procedure to classify the shape of the curve into the category increasing, decreasing, convex, concave, convex with increasing trend, concave with increasing trend, convex with decreasing trend, or concave with decreasing trend (see Chapter 2 for details).

If the interaction between two groups is considered, we will fit a semi-parametric shape-restricted regression spline model incorporating the factor-by-curve interaction of the form

$$y_i = \beta_0 + \beta_1 a_i + f_1(x_i) + f_2(x_i) a_i + \beta_2 c_{1i} + ... + \beta_p c_{pi} + \epsilon_i. \tag{5.1.2}$$

In this model, $f_1(x_i)$ is the curve of the predictor for the non-exposure group and $f_2(x_i)$ is the difference of the curves for the exposure group and non-exposure group, and they can be fitted using the combination of quadratic I-spline bases simultaneously, i.e., $f_1(x_i) = \theta_1 I_1(x_i|2,t) + \theta_2 I_2(x_i|2,t) + ... + \theta_k I_k(x_i|2,t)$ and $f_2(x_i) a_i = \gamma_1 I_1(x_i|2,t) a_i + \gamma_2 I_2(x_i|2,t) a_i + ... + \gamma_k I_k(x_i|2,t) a_i$, or using the combination of cubic C-spline bases simultaneously, i.e.,

$f_1(x_i) = \theta_0 x_i + \theta_1 C_1(x_i|2,t) + \theta_2 C_2(x_i|2,t) + ... + \theta_k C_k(x_i|2,t)$ and $f_2(x_i)a_i = \gamma_0 x_i a_i + \gamma_1 C_1(x_i|2,t)a_i + \gamma_2 C_2(x_i|2,t)a_i + ... + \gamma_k C_k(x_i|2,t)a_i$. Model (5.1.2) can be written in matrix notation as

$$y = X\beta + \epsilon, \tag{5.1.3}$$

where

$$X = \begin{bmatrix} 1 & a_1 & I_1(x_1) & ... & I_k(x_1) & I_1(x_1)a_1 & ... & I_k(x_1)a_1 & c_{11} & ... & c_{p1} \\ 1 & a_2 & I_1(x_2) & ... & I_k(x_2) & I_1(x_2)a_2 & ... & I_k(x_2)a_2 & c_{12} & ... & c_{p2} \\ ... & ... & ... & & ... & ... & & ... & ... & & ... \\ 1 & a_n & I_1(x_n) & ... & I_k(x_n) & I_1(x_n)a_n & ... & I_k(x_n)a_n & c_{1n} & ... & c_{pn} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_0 & \beta_1 & \theta_1 & ... & \theta_k & \gamma_1 & ... & \gamma_k & \beta_2 & ... & \beta_p \end{bmatrix}^T$$

and $\epsilon \sim N(0, \sigma^2 I)$ for model fitted using the combination of quadratic I-spline bases, or

$$X = \begin{bmatrix} 1 & a_1 & x_1 & C_1(x_1) & ... & C_k(x_1) & x_1 a_1 & C_1(x_1)a_1 & ... & C_k(x_1)a_1 & c_{11} & ... & c_{p1} \\ 1 & a_2 & x_2 & C_1(x_2) & ... & C_k(x_2) & x_2 a_2 & C_1(x_2)a_2 & ... & C_k(x_2)a_2 & c_{12} & ... & c_{p2} \\ ... & ... & ... & & ... & ... & & ... & ... & & ... & & \\ 1 & a_n & x_n & C_1(x_n) & ... & C_k(x_n) & x_n a_n & C_1(x_n)a_n & ... & C_k(x_n)a_n & c_{1n} & ... & c_{pn} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_0 & \beta_1 & \theta_0 & \theta_1 & ... & \theta_k & \gamma_0 & \gamma_1 & ... & \gamma_k & \beta_2 & ... & \beta_p \end{bmatrix}^T$$

and $\epsilon \sim N(0, \sigma^2 I)$ for model fitted using the combination of cubic C-spline bases. Note that we use $I_j(x_i)$ and $C_j(x_i)$ as shorthand notations of $I_j(x_i|2,t)$ and $C_j(x_i|2,t)$.

To determine the shape of the curve for non-exposure group, i.e., $f_1(x_i)$, we should follow the steps described in Chapter 2. To determine the shape of the curve for exposure group, i.e., $f_1(x_i) + f_2(x_i)$, we will follow the steps described below.

The unconstrained estimates of all parameters are obtained using ordinary least-squares, which are denoted as $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\theta}_1, ..., \hat{\theta}_k, \hat{\gamma}_1, ..., \hat{\gamma}_k, \hat{\beta}_2, ..., \hat{\beta}_p)^T$ for model fitted using the combination of quadratic I-spline bases and $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\theta}_0, \hat{\theta}_1, ..., \hat{\theta}_k, \hat{\gamma}_0, \hat{\gamma}_1, ..., \hat{\gamma}_k, \hat{\beta}_2, ..., \hat{\beta}_p)^T$ for model fitted using the combination of cubic C-spline bases. Once we obtain the unconstrained estimates, we will project $\hat{\eta} = (I_k, I_k)(\hat{\theta}, \hat{\gamma})^T = (\hat{\theta}_1 + \hat{\gamma}_1, ..., \hat{\theta}_k + \hat{\gamma}_k)^T$ onto the suitable polyhedral cone with respect to $\Sigma_\eta = \Sigma_{\theta+\gamma} = (I_k, I_k)\Sigma_{(\theta,\gamma)}(I_k, I_k)^T$, where $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_k)$,

$\hat{\gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_k)$, I_k is a $k \times k$ identity matrix and $\Sigma_{(\theta,\gamma)} = Cov((\hat{\theta}, \hat{\gamma})^T)$. The constrained estimates are

$$\tilde{\eta}_i = \arg\min_{\eta_i \in O_i} (\eta_i - \hat{\eta}_i)^T \Sigma_{\eta_i}^{-1} (\eta_i - \hat{\eta}_i),\ \text{i} = 1,\ 2,\ 3,\ 4 \tag{5.1.4}$$

where η_1 and η_2 involve θ and γ from the model fitted using the combination of quadratic I-spline bases and η_3 and η_4 involve θ and γ from the model fitted using the combination of cubic C-spline bases, and O_1 and O_3 are the non-negative orthants, i.e., $\{\eta | \eta \geq 0\}$, and O_2 and O_4 are the non-positive orthants, i.e., $\{\eta | \eta \leq 0\}$. In practice, Σ_{η_i} is usually unknown, and thus we will use its estimate $\hat{\Sigma}_{\eta_i} = \widehat{Cov((\hat{\theta}, \hat{\gamma})^T)}$. We then formulate four tests with the same form of test statistics to test the hypotheses

$$H_{0i} : \eta_i = 0 \text{ vs. } H_{ai} : \eta_i \in O_i, \tag{5.1.5}$$

and the test statistics are

$$T_i = \frac{\hat{\eta}_i^T \Sigma_{\eta_i}^{-1} \hat{\eta}_i - (\tilde{\eta}_i - \hat{\eta}_i)^T \Sigma_{\eta_i}^{-1} (\tilde{\eta}_i - \hat{\eta}_i)}{\hat{\eta}_i^T \Sigma_{\eta_i}^{-1} \hat{\eta}_i}. \tag{5.1.6}$$

Corollary 5.1.1. *Suppose η_i is a k-dimensional vector. If Σ_{η_i} is known, then $P(T_i \leq c | H_{0i}) = \sum_{j=0}^{k} w_j(k, \Sigma_{\eta_i}, \mathbb{R}^{+p}) P(Beta(\frac{j}{2}, \frac{k-j}{2}) \leq c)$ and if Σ_{η_i} is unknown, then $P(\hat{T}_i \leq c | H_{0i}) \overset{asymp.}{=} \sum_{j=0}^{k} w_j(k, \Sigma_{\eta_i}, \mathbb{R}^{+p}) P(Beta(\frac{j}{2}, \frac{k-j}{2}) \leq c)$, where $w_j(k, \Sigma_{\eta_i}, \mathbb{R}^{+p})$ are non-negative weights and $\sum_{j=0}^{k} w_j(k, \Sigma_{\eta_i}, \mathbb{R}^{+p}) = 1$.*

Proof. The unconstrained estimate of β in model 5.1.3, $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$, and thus $R\beta \sim N(R\beta, \sigma^2 R(X^T X) R^T)$. For the model fitted using the combination of quadratic I-spline bases, $R = (0, 0, I_k, I_k, 0, ..., 0)$, and for the model fitted using the combination of cubic C-spline bases, $R = (0, 0, 0, I_k, 0, I_k, 0, ..., 0)$, where I_k is a $k \times k$ identity matrix and 0 is a $k \times 1$ vector. The remaining proof is the same as the proof of Theorem 2.2.1 in Chapter 2. □

If Σ_{η_i} is known, then the test statistic T_i under null hypothesis H_{0i} follows a beta-bar distribution. If Σ_{η_i} is unknown, then the test statistic $\hat{T}_i$ under null hypothesis H_{0i} follows a beta-bar distribution asymptotically.

After obtaining the p-values from those four tests, we will apply Holm-Bonferroni method to perform multiple testing of the four sets of hypotheses. According to Holm-Bonferroni method, we sort the p-values in ascending order, i.e. $p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq p_{(4)}$, and then compare the p-values with the corresponding significance levels, i.e. $\alpha_{(1)} = \frac{\alpha}{4} \leq \alpha_{(2)} = \frac{\alpha}{3} \leq \alpha_{(3)} = \frac{\alpha}{2} \leq \alpha_{(4)} = \alpha$. The decisions of rejections are made sequentially. Depending upon the rejections by using the decision rules, we make decisions regarding the shape of the curve for exposure group as described in Table 2.2.1 in Chapter 2.

The simulation study should be performed in the future to show the family-wise error rate as well as power of the group-level shape detection method.

5.1.2 Knots consideration

The spline technique that we use in the entire dissertation is regression spline. The estimation procedure of a model using regression splines is relatively straightforward comparing with other spline techniques. Once the spline basis functions are determined, we can use least squares to estimate the parameters. However, in order to determine the spline basis functions, we need to ascertain the degree of the functions, the number of knots and the locations of knots (Wegman and Wright, 1983). From Fact 1.2.1 and Fact 1.2.2 in Chapter 1, the quadratic I-splines and the cubic C-splines are the building blocks of our regression spline models. Nevertheless, the number of knots and the locations of knots are still undetermined. There are several popular methods of knots placement, including "equally spaced" method, "equally spaced sample quantiles as knots" method and "model selection based" method (Wu and Zhang, 2006). In the simulation study in Chapter 2 and the real data analysis in Chapter 4, we adopt the "equally spaced sample quantiles as knots" method. In specific, we set the number of internal knots to 4 and let the knots be placed at the 20^{th}, 40^{th}, 60^{th} and 80^{th} percentiles of the data. The choice of the number of knots is arbitrary in our analysis, so in order to systematically understand the influence of the number of knots, we refine the

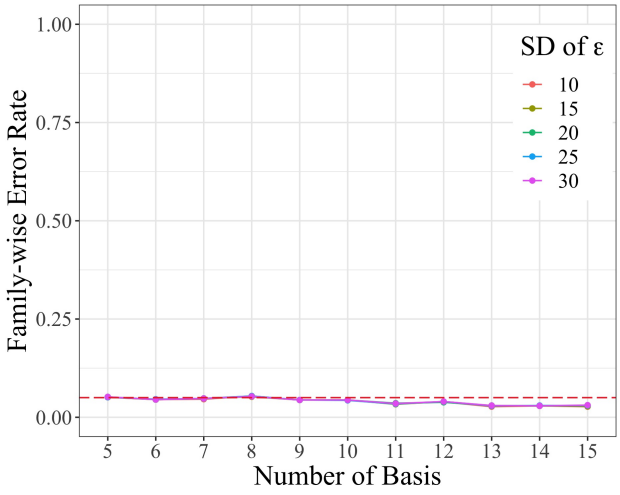
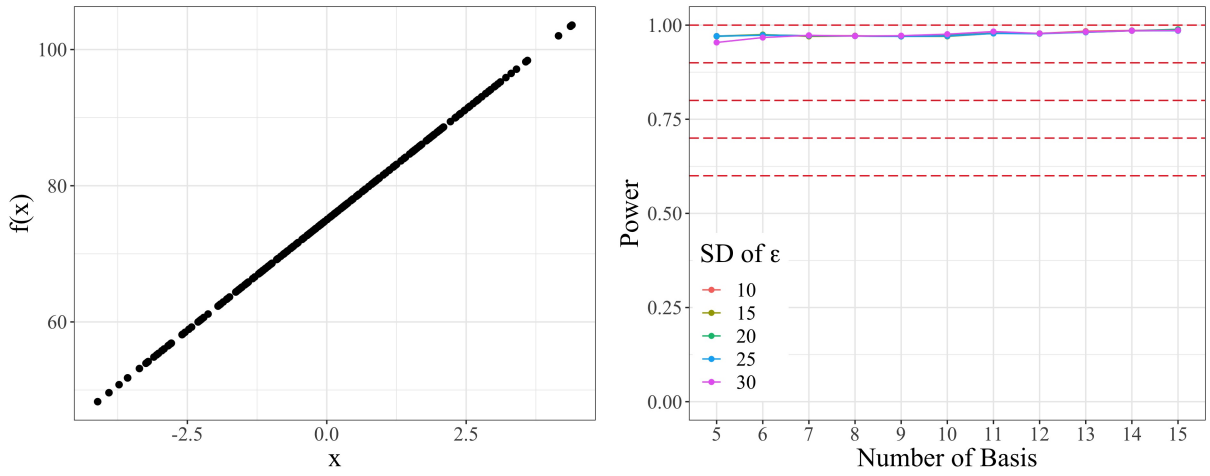


Figure 5.1.1: Family-wise error rate curve

functions of the main effect and examine the family-wise error rates and the powers under different number of bases in a fixed effect regression spline setting. The random error ϵ is drawn from $N(0, 10^2)$, $N(0, 15^2)$, $N(0, 20^2)$, $N(0, 25^2)$ and $N(0, 30^2)$, and the number of bases is from 5 to 15. The simulated family-wise error rates and the simulated powers are shown in Figure 5.1.1 to Figure 5.1.9.

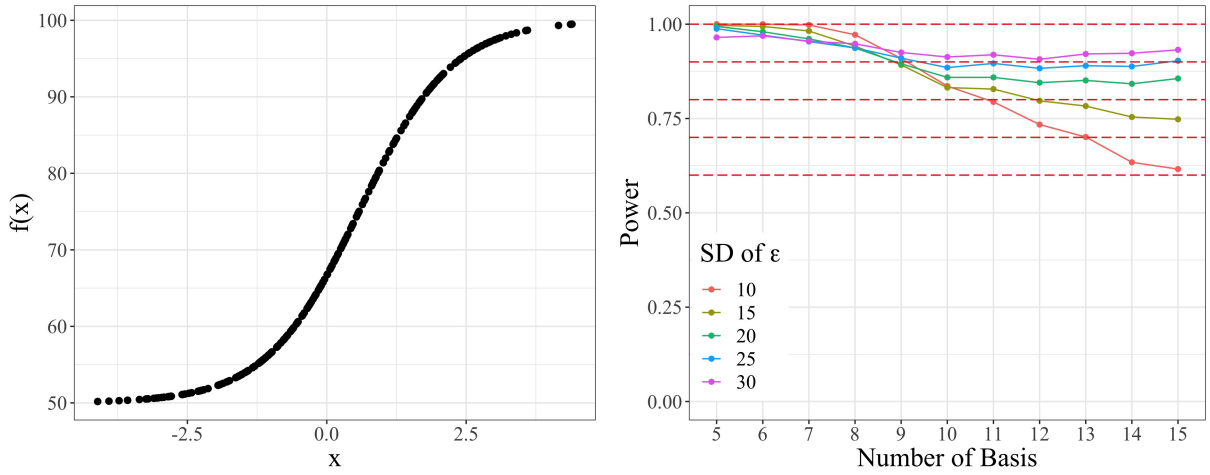
According to Figure 5.1.1 to Figure 5.1.9, the family-wise error rates are controlled reasonably overall. For shape 1 and shape 10, the powers are large under each number of bases and each standard deviation of ϵ. For shape 2, shape 11 and shape 12, the powers are large when the number of bases is relatively small (< 10); as the number of bases increases, the powers decrease, especially when the standard deviation of ϵ is small, where the method is prone to detect the shape as convex with increasing trend for shape 2, concave with increasing trend for shape 11, and concave with decreasing trend for shape 12. For shape 3, the powers are large overall, but when the number of bases is small, the method will have chance to report the shape as flat. For shape 14 and shape 16, powers are large when the standard deviation of ϵ is relatively small; as the standard deviation of ϵ increases, the powers decrease, especially when the number of bases is small; with large standard deviation of ϵ and small number of bases, the method is prone to detect the shape as increasing or concave



(a) Scatter plot of simulated curve

(b) Power Curve

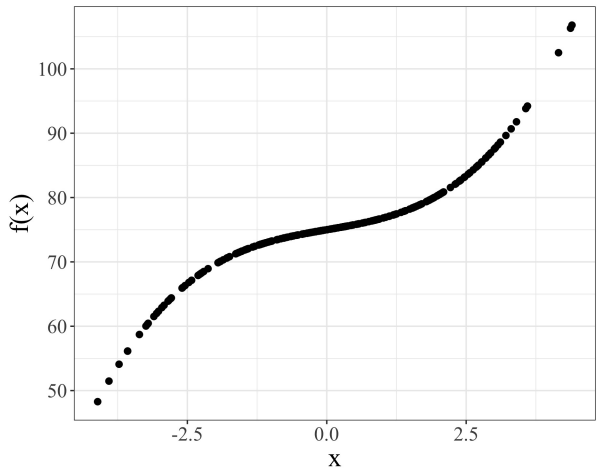
Figure 5.1.2: Power curve for shape 1 ($f(x) = 6.5x + 75$)



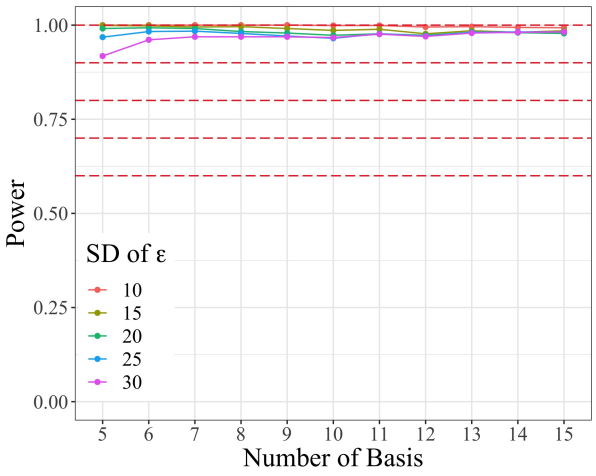
(a) Scatter plot of simulated curve

(b) Power Curve

Figure 5.1.3: Power curve for shape 2 ($f(x) = 50\frac{e^{1.2x}}{2+e^{1.2x}} + 50$)

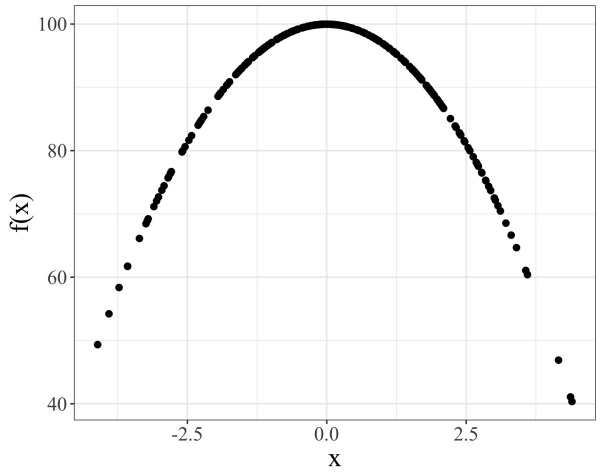


(a) Scatter plot of simulated curve

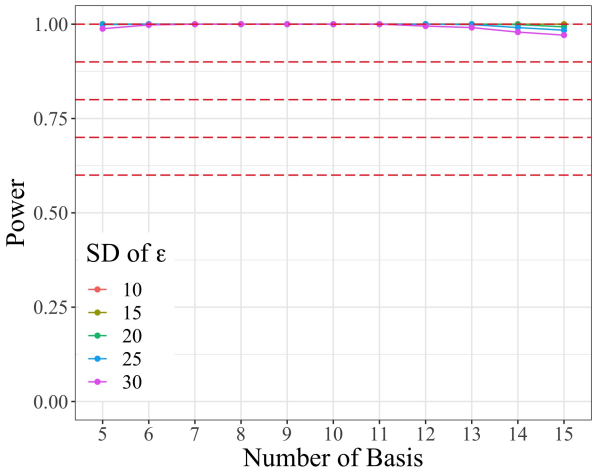


(b) Power Curve

Figure 5.1.4: Power curve for shape 3 ($f(x) = (\frac{2x}{3})^3 + \frac{3x}{2} + 75$)

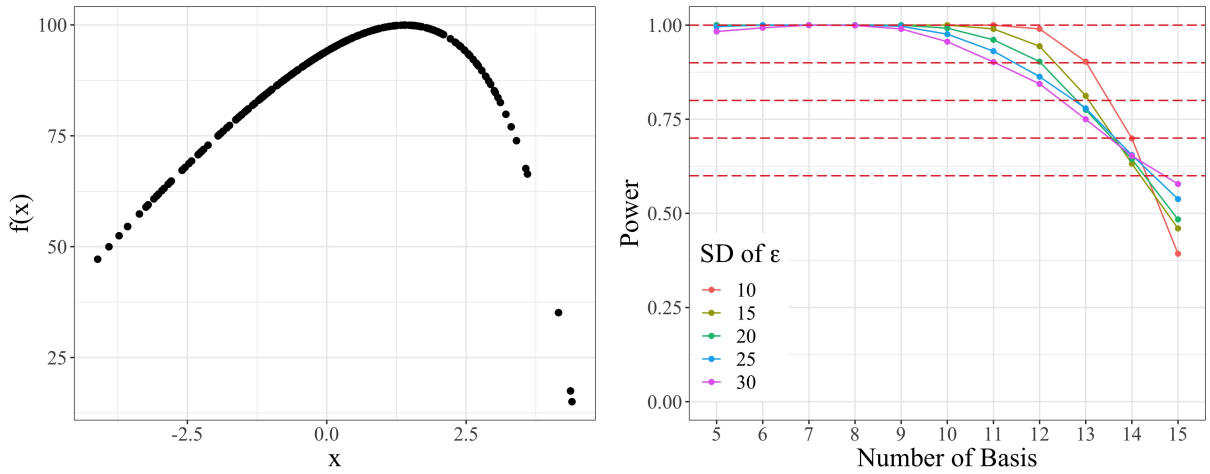


(a) Scatter plot of simulated curve



(b) Power Curve

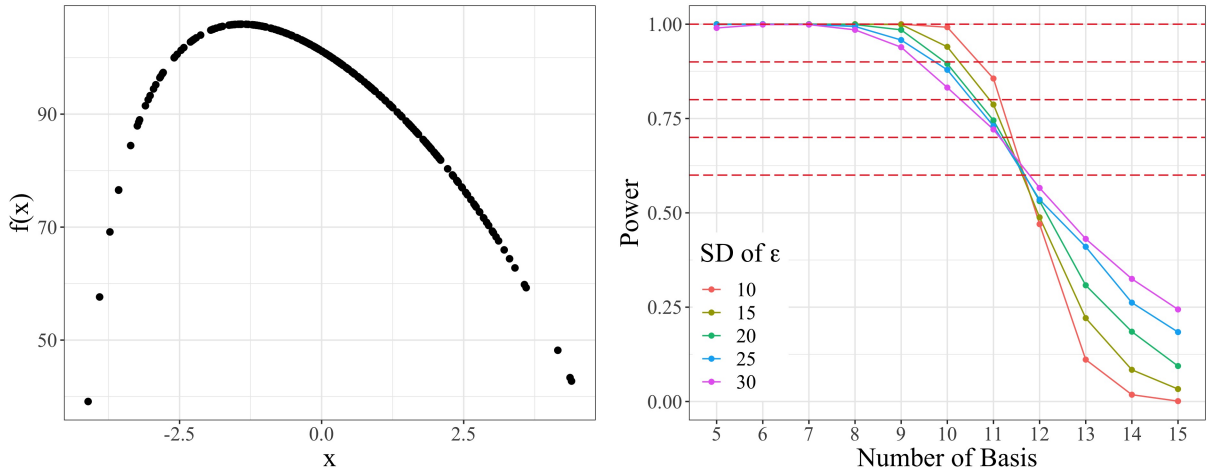
Figure 5.1.5: Power curve for shape 10 ($f(x) = \frac{-e^x - 150x^2}{50} + 100$)



(a) Scatter plot of simulated curve

(b) Power Curve

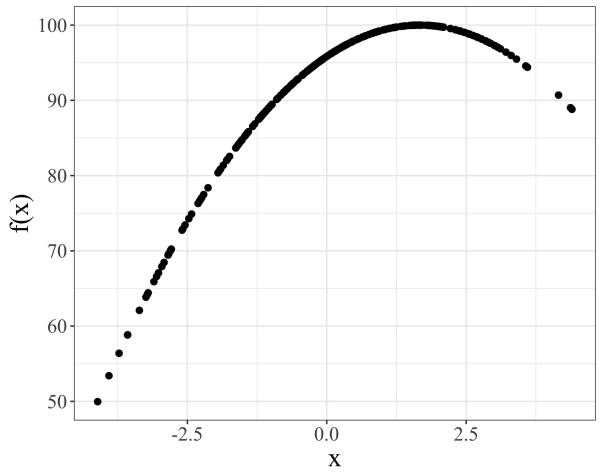
Figure 5.1.6: Power curve for shape 11 ($f(x) = 300\ln(-e^{x/1.55} + 1.6x + 40) - 1005$)



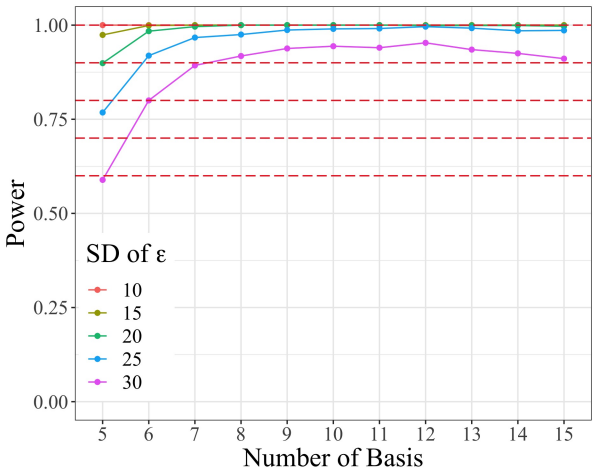
(a) Scatter plot of simulated curve

(b) Power Curve

Figure 5.1.7: Power curve for shape 12 ($f(x) = 70\ln(-e^{-x/1.6} - 1.5x + 11) - 60$)

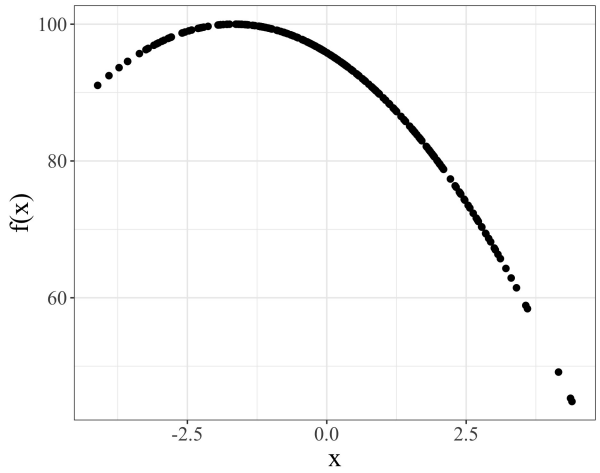


(a) Scatter plot of simulated curve

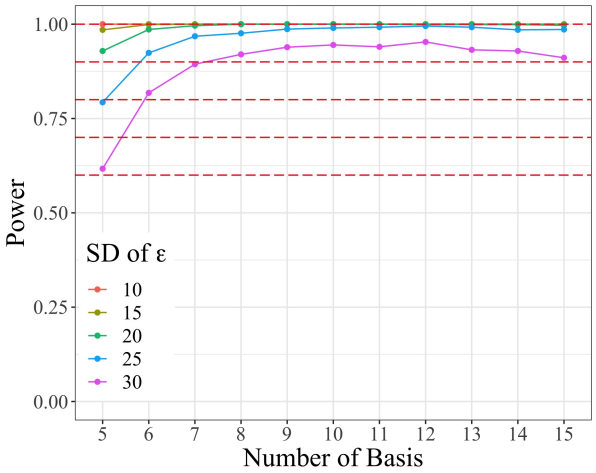


(b) Power Curve

Figure 5.1.8: Power curve for shape 14 ($f(x) = -1.5(x - 5/3)^2 + 100$)



(a) Scatter plot of simulated curve



(b) Power Curve

Figure 5.1.9: Power curve for shape 16 ($f(x) = -1.5(x + 5/3)^2 + 100$)

for shape 14 and decreasing or concave for shape 16, while with large standard deviation of ϵ and large number of bases, the method is prone to detect the shape as increasing for shape 14 and decreasing for shape 16. Overall, 8 or 9 is a better choice for the number of bases.

Some knot selection techniques for shape-restricted regression spline are proposed (Meyer, 2012, 2013; Choi et al., 2019), whose aim is to minimize the distance between the observed and the estimated responses. However, the goal of our method is to test the null hypothesis of constant function against the alternative that there is an underlying shape and classify the underlying shape into a reasonable category. Therefore, the proposed knot selection techniques do not seem to be suitable in our setting, and thus we need further explorations on knots placement. There are some recommendations on knots placement that can be followed, for example, more knots should be placed in the regions where the shape changes rapidly (Choi et al., 2019), and extrema should be centered between knots and inflection points should be placed near knots (Wegman and Wright, 1983).

5.2 Discussion and Future Work for Chapter 3

The mediation analysis is developed to interpret the causal relationship between an exposure and a potential outcome directly or indirectly caused by the exposure through examining the intermediate stage. The regression-based mediation analysis developed by VanderWeele (2015) is based on linear regression. If the relationship between the mediator and the outcome is curvilinear, applying method based on linear regression may lead to biased estimates of mediation effects. In such a case, the exposure-outcome model should be built using nonparametric techniques. In the second part of the dissertation, we develop a method to help researchers analytically estimate the mediation effects and make inferences on them when the shape of relationship between the mediator and the outcome is previously inferred or known to be increasing, decreasing, convex or concave. In specific, we build the exposure-mediator and exposure-outcome models, estimate the exposure-mediator model parameters through least squares and the exposure-outcome model parameters through a cone projection method, combine the model parameters through a specific process to estimate the mediation

effects, and apply delta method to obtain the asymptotic variances of different mediation effects.

5.2.1 Mediation analysis with continuous exposure, continuous mediator and continuous outcome

The method that we develop in Chapter 3 is based on binary exposure, continuous mediator and continuous outcome. However, in some scenarios, the exposure could also be measured in continuous scale. If the exposure, the mediator and the outcome are all continuous, the relationships between the exposure and the mediator, between the exposure and the outcome and between the mediator and the outcome are curvilinear, and the interaction between the exposure and the mediator exist, then the exposure-outcome model will be

$$Y = \beta_0 + f_1(A) + f_2(M) + f_3(A, M) + \beta_4 C + \epsilon_1, \tag{5.2.1}$$

where $f_1(A)$ is the curve of the exposure, $f_2(M)$ is the curve of the mediator, $f_3(A, M)$ is the surface of the interaction between the exposure and the mediator and $\epsilon_1 \sim N(0, \sigma_1^2)$, and the exposure-mediator model will be

$$M = \gamma_0 + g_1(A) + \gamma_2 C + \epsilon_2, \tag{5.2.2}$$

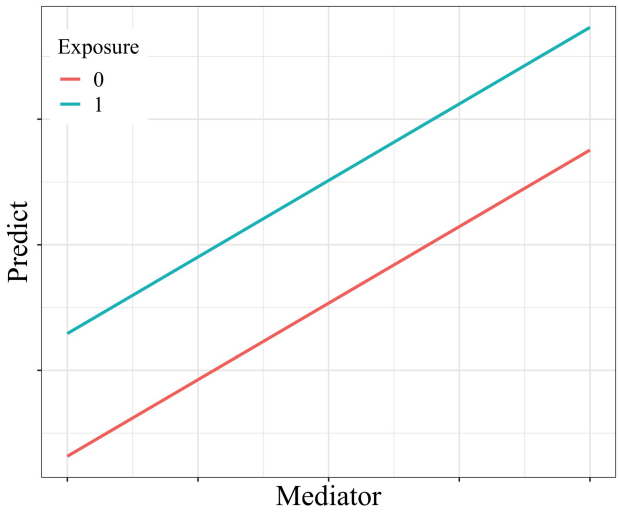
where $g_1(A)$ is the curve of the exposure and $\epsilon_2 \sim N(0, \sigma_2^2)$.

In model 5.2.1, $f_1(A)$ and $f_2(M)$ can be fitted using one-dimensional smoothers, such as cubic splines and P-splines, and $f_3(A, M)$ can be fitted using tensor product splines (Wood, 2017). In model 5.2.2, $g_1(A)$ can be fitted using one-dimensional smoother. To facilitate readers' understanding of the exposure-outcome model, we generate several figures (see Figure 5.2.1 and Figure 5.2.2). Figure 5.2.1a to Figure 5.2.1c are drawn in the circumstance of binary exposure, continuous mediator and continuous outcome. With binary exposure, we are able to visualize the relationship among the exposure, the mediator and the predicted outcome after controlling for confounders in a 2D plot. Figure 5.2.2a to Figure 5.2.2c are drawn in the circumstance of continuous exposure, continuous mediator and continuous outcome. When the exposure is continuous, we need to draw a 3D plot to visualize the

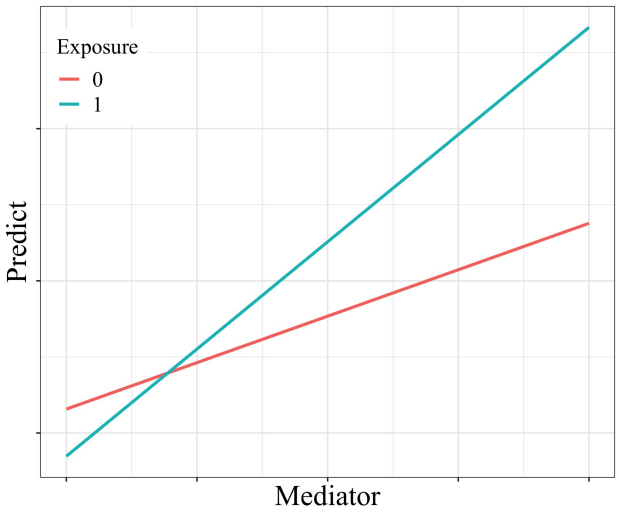
relationship among the exposure, the mediator and the predicted outcome after controlling for confounders. Figure 5.2.2c is an example of model 5.2.1.

To estimate the mediation effects, we need to adopt the ideas from simulation-based mediation analysis (Imai et al., 2010). The estimation procedure is slightly modified and summarized below:

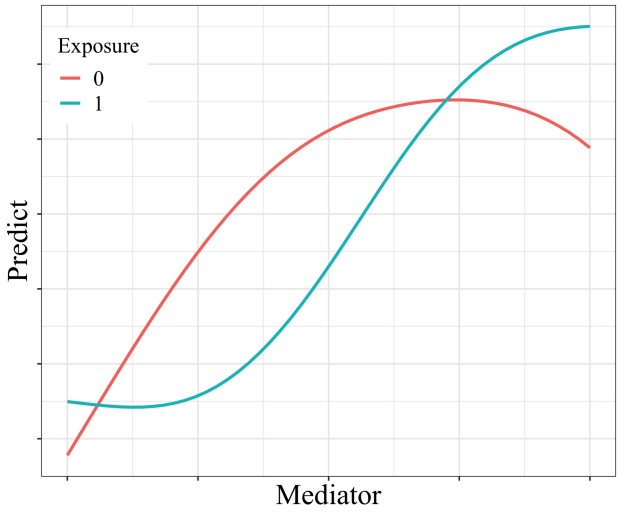
- Step 1: Draw a random sample with replacement of size n from the original data.
- Step 2: Fit models 5.2.1 and 5.2.2.
- Step 3: Keep A at some level a_1 (say 25^{th} percentile) and calculate M_{a_1} for each individual using fitted model 5.2.2 in Step 2; keep A at some other level a_2 (say 75^{th} percentile) and calculate M_{a_2} for each individual using fitted model 5.2.2 in Step 2.
- Step 4: Keep A at some level a_1 (say 25^{th} percentile) and calculate $Y_{a_1 M_{a_1}}$ or $Y_{a_1 M_{a_2}}$ for each individual using fitted model 5.2.1 in Step 2 and calculated M_{a_1} or M_{a_2} in Step 3; keep A at some other level a_2 (say 75^{th} percentile) and calculate $Y_{a_2 M_{a_2}}$ or $Y_{a_2 M_{a_1}}$ for each individual using fitted model 5.2.1 in Step 2 and calculated M_{a_2} or M_{a_1} in Step 3.
- Step 5: Calculate the NDE for the random sample as $\frac{1}{n}\sum(Y_{a_2 M_{a_1}} - Y_{a_1 M_{a_1}})$ and the NIE for the random sample as $\frac{1}{n}\sum(Y_{a_2 M_{a_2}} - Y_{a_2 M_{a_1}})$.
- Step 6: Repeat Step 1 to Step 5 m times and take the median of all calculated sample NDEs/NIEs as the point estimate of NDE/NIE, the standard deviation of all calculated sample NDEs/NIEs as the standard error of NDE/NIE and the percentiles of all calculated sample NDEs/NIEs to construct the confidence interval of NDE/NIE.



(a) Linear, no interaction

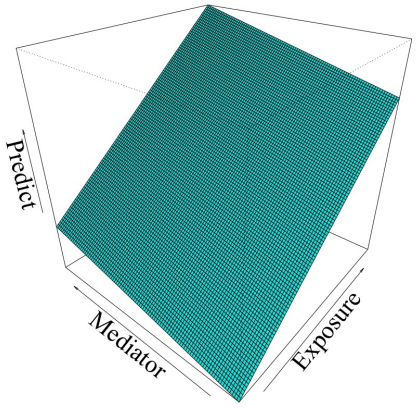


(b) Linear, with interaction

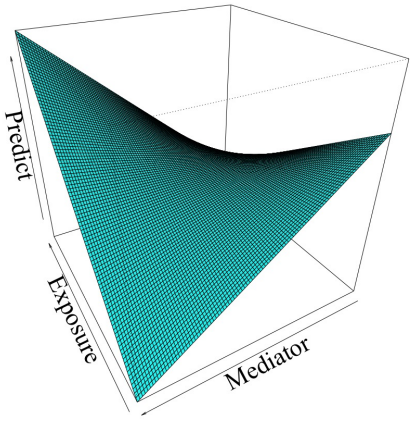


(c) Curvilinear, with interaction

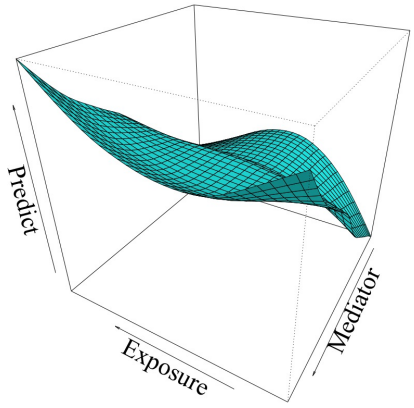
Figure 5.2.1: Plots of mediator and predicted outcome varying by binary exposure



(a) Linear, no interaction



(b) Linear, with interaction



(c) Curvilinear, with interaction

Figure 5.2.2: Plots of mediator and predicted outcome varying by continuous exposure

Appendix A Appendix for Chapter 2

A.1 Technical details of Henderson's Mixed Model Equations

Since in model (2.2.1), $b \sim N(0, \tilde{D})$ and $\epsilon \sim N(0, R)$, then $y|b \sim N(X\beta + Zb, R)$, and thus the density function of b is $f(b) = \frac{e^{-\frac{1}{2}[b^T \tilde{D}^{-1} b]}}{(2\pi)^{\frac{1}{2}(qc)}|\tilde{D}|^{1/2}}$ and the conditional density function of y given b is $f(y|b) = \frac{e^{-\frac{1}{2}[(y-X\beta-Zb)^T R^{-1}(y-X\beta-Zb)]}}{(2\pi)^{\frac{1}{2}(N)}|R|^{1/2}}$. Therefore, the joint density function of y and b is $f(y,b) = f(y|b)f(b) = \frac{e^{-\frac{1}{2}[(y-X\beta-Zb)^T R^{-1}(y-X\beta-Zb)+b^T \tilde{D}^{-1} b]}}{(2\pi)^{\frac{1}{2}(N+qc)}|R|^{1/2}|\tilde{D}|^{1/2}}$. The twice negative logarithm of the joint density function of y and b is $l(\beta, b|y) = (y - X\beta - Zb)^T R^{-1}(y - X\beta - Zb) + b^T \tilde{D}^{-1} b + (N + qc)\log(2\pi) + \log|\tilde{D}| + \log|R|$. The partial derivative of $l(\beta, b|y)$ with respect to β is $\frac{\partial l(\beta,b|y)}{\partial \beta} = -2X^T R^{-1}(y - X\beta - Zb)$, and by setting it to 0, we obtain $X^T R^{-1} X\beta + X^T R^{-1} Zb = X^T R^{-1} y$ (A.1.1). The partial derivative of $l(\beta, b|y)$ with respect to b is $\frac{\partial l(\beta,b|y)}{\partial b} = -2Z^T R^{-1}(y - X\beta - Zb) + 2\tilde{D}^{-1} b$, and by setting it to 0, we obtain $Z^T R^{-1} X\beta + Z^T R^{-1} Zb + \tilde{D}^{-1} b = Z^T R^{-1} y$ (A.1.2). Equations (A.1.1) and (A.1.2) are Henderson's Mixed Model Equations and can be rewritten into a matrix form as

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + \tilde{D}^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ b \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}.$$

From equation (A.1.2), we obtain $\hat{b} = (Z^T R^{-1} Z + \tilde{D}^{-1})^{-1}(Z^T R^{-1} y - Z^T R^{-1} X\hat{\beta})$, and by plugging $\hat{b}$ into equation (A.1.1), we obtain $X^T R^{-1} X\beta + X^T R^{-1} Z(Z^T R^{-1} Z + \tilde{D}^{-1})^{-1} Z^T R^{-1}(y - X\hat{\beta}) = X^T R^{-1} y$ (A.1.3). Let $Z\tilde{D}Z^T + R = V$, then according to the results on Schur complements, $V^{-1} = R^{-1} - R^{-1} Z(Z^T R^{-1} Z + \tilde{D}^{-1})^{-1} Z^T R^{-1}$, and thus equation (A.1.3) can be written as $X^T V^{-1} X\hat{\beta} = X^T V^{-1} y$. Therefore, $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$. According to the results on Schur complements, $(Z^T R^{-1} Z + \tilde{D}^{-1})^{-1} Z^T R^{-1} = (\tilde{D} - \tilde{D} Z^T V^{-1} Z\tilde{D}) Z^T R^{-1} = \tilde{D} Z^T (R^{-1} - V^{-1} Z\tilde{D} Z^T R^{-1}) = \tilde{D} Z^T (R^{-1} - V^{-1}(V - R)R^{-1}) = \tilde{D} Z^T V^{-1}$, and then $\hat{b} =$

$\tilde{D}Z^TV^{-1}(y - X\beta)$. The covariance matrix of $\hat{\beta}$ is

$$\begin{aligned} COV(\hat{\beta}) &= ((X^TV^{-1}X)^{-1}X^TV^{-1})COV(y)((X^TV^{-1}X)^{-1}X^TV^{-1})^T \\ &= (X^TV^{-1}X)^{-1}X^TV^{-1}V(V^{-1})^TX((X^TV^{-1}X)^{-1})^T \\ &= (X^TV^{-1}X)^{-1}((X^TV^{-1}X)^{-1}X^TV^{-1}X)^T \\ &= (X^TV^{-1}X)^{-1}, \end{aligned}$$

and the covariance matrix of $\hat{b}$ is

$$\begin{aligned} COV(b) &= (\tilde{D}Z^TV^{-1})COV(y - X\beta)(\tilde{D}Z^TV^{-1})^T \\ &= (\tilde{D}Z^TV^{-1})(I - X(X^TV^{-1}X)^{-1}X^TV^{-1})V(I - X(X^TV^{-1}X)^{-1}X^TV^{-1})^T(\tilde{D}Z^TV^{-1})^T \\ &= (\tilde{D}Z^TV^{-1})(V - X(X^TV^{-1}X)^{-1}X^T)(\tilde{D}Z^TV^{-1})^T \\ &= \tilde{D}Z^TV^{-1}Z\tilde{D} - \tilde{D}Z^TV^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}Z\tilde{D}. \end{aligned}$$

A.2 Iterative procedures based on Henderson's Mixed Model Equations

We assume $\tilde{D} = diag(\sigma_1^2 I_{n_1}, ..., \sigma_c^2 I_{n_c})$ and $R = \sigma_e^2 I_N$, then the iterative procedure based on MME for ML has the following steps:

- Step 0: Set $r = 0$, and set the starting values of σ_i^2 and σ_e^2 as $\sigma_{i(0)}^2$ and $\sigma_{e(0)}^2$.
- Step 1: Calculate $W_{(r)} = (\sigma_{e(r)}^2 I + Z^TZ\tilde{D}_{(r)})^{-1}\sigma_{e(r)}^2$, set $r = r + 1$, and update $\beta_{(r)}$ and $b_{(r)}$ using
$$\begin{bmatrix} X^TX & X^TZ\tilde{D}_{(r-1)} \\ Z^TX & W_{(r-1)} \end{bmatrix} \begin{bmatrix} \beta_{(r)} \\ v_{(r)} \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \end{bmatrix},$$
and $b_{(r)} = \tilde{D}_{(r-1)}v_{(r)}$.
- Step 2: Update $\sigma_{e(r)}^2$ and $\sigma_{i(r)}^2$ using $\sigma_{e(r)}^2 = \frac{y^T(y - X\beta_{(r)} - Zb_{(r)})}{N}$ and $\sigma_{i(r)}^2 = \frac{b_{i(r)}^T b_{i(r)}}{n_i - tr(W_{ii(r-1)})}$.
- Step 3: Repeat Steps 1 and 2 until convergence.

and the the iterative procedure based on MME for REML has the following steps:

- Step 0: Set $r = 0$, and set the starting values of σ_i^2 and σ_e^2 as $\sigma_{i(0)}^2$ and $\sigma_{e(0)}^2$.
- Step 1: Calculate $W_{(r)} = (\sigma_{e(r)}^2 I + Z^T Z \tilde{D}_{(r)})^{-1} \sigma_{e(r)}^2$, $T_{(r)} = (I + \frac{Z^T[I - X(X^TX)^- X^T]Z\tilde{D}_{(r)}}{\sigma_{e(r)}^2})^{-1}$ set $r = r + 1$, and update $\beta_{(r)}$ and $b_{(r)}$ using

$$\begin{bmatrix} X^TX & X^TZ\tilde{D}_{(r-1)} \\ Z^TX & W_{(r-1)} \end{bmatrix} \begin{bmatrix} \beta_{(r)} \\ v_{(r)} \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \end{bmatrix},$$

  and $b_{(r)} = \tilde{D}_{(r-1)} v_{(r)}$.
- Step 2: Update $\sigma_{e(r)}^2$ and $\sigma_{i(r)}^2$ using $\sigma_{e(r)}^2 = \frac{y^T(y - X\beta_{(r)} - Zb_{(r)})}{N - rank(X)}$ and $\sigma_{i(r)}^2 = \frac{b_{i(r)}^T b_{i(r)}}{n_i - tr(T_{ii(r-1)})}$.
- Step 3: Repeat Steps 1 and 2 until convergence.

A.3 Simulation of beta-bar weights

- Step 1: Generate $\hat{\beta}_i$ from $N(0, \Sigma_i)$
- Step 2: Calculate $\tilde{\beta}_i = \Pi_{\Sigma_i}(\hat{\beta}_i | \mathbb{R}^{+p}) = \arg\min_{\beta_i \in \mathbb{R}^{+p}} (\beta_i - \hat{\beta}_i)^T \Sigma_i^{-1} (\beta_i - \hat{\beta}_i)$
- Step 3: Count the number of positive components of $\tilde{\beta}_i$
- Step 4: Repeat Steps 1 to 3 N times and $w_j(p, \Sigma_i, \mathbb{R}^{+p})$ can be calculated as

$$\frac{\text{the total times of } j \text{ positive components of } \tilde{\beta}_i}{N}$$

A.4 Family-wise error rate simulation using residual bootstrap

- Step 1: Generate the response vector y_{null} under the null hypothesis
  - $y_{null} = X_F \beta_F + Zb + \epsilon$, where X_F and Z are from the real data, β_F are assumed to be the true parameters, b is drawn from $N(0, \tilde{D})$ and ϵ is drawn from $N(0, R)$
- Step 2: Fit the model $y = X_S \beta_S + X_F \beta_F + Zb + \epsilon$ under constraints on β_S using y_{null} and calculate the test statistics T_1, T_2, T_3, T_4
- Step 3: Fit the model $y = X_F \beta_F + Zb + \epsilon$ using y_{null}
- Step 4: Bootstrap the null distribution
  - Bootstrap the random effects and noise from the model in Step 2

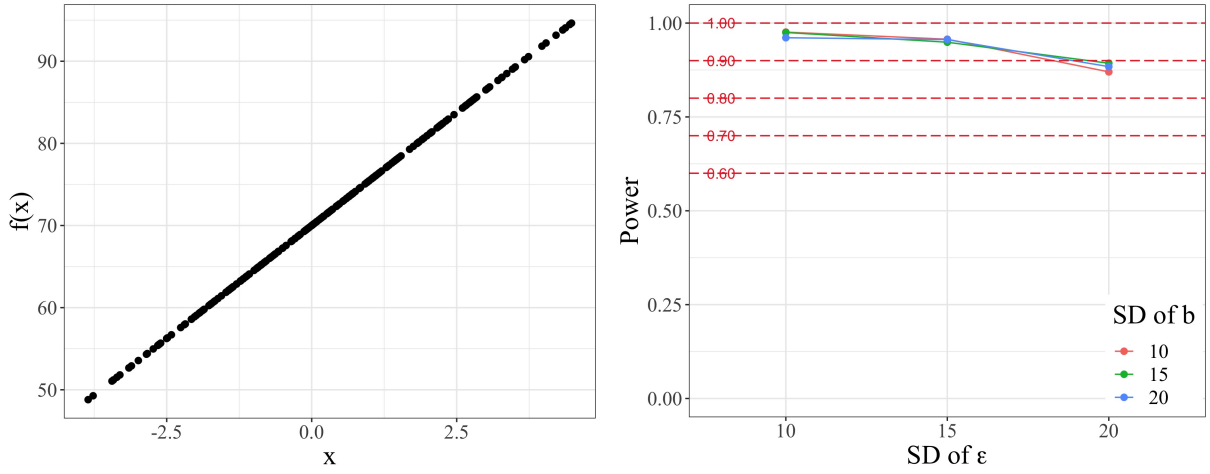
  - Adopt the fixed effects estimates $\hat{\beta}_F$ in Step 3
  - Generate the response vector y_{boot} under the null hypothesis using bootstrapped random effects and noise
  - Fit the model $y = X_S\beta_S + X_F\beta_F + Zb + \epsilon$ under constraints on β_S using y_{boot} and calculate the test statistics T_1^*, T_2^*, T_3^*, T_4^*
  - Compare T_i with T_i^* (if $T_i \leq T_i^*$ then $p_{ij} = 1$, else $p_{ij} = 0$), where $i = 1, 2, 3, 4$
- Step 5: Repeat Step 4 N times and calculate p_i as $\frac{\sum_{j=1}^{N} p_{ij}}{N}$, where $i = 1, 2, 3, 4$; calculate $p = min(p_1, p_2, p_3, p_4)$ (if $p < 0.0125$, then $U_k = 1$, else $U_k = 0$)
- Step 6: Repeat Step 1 to Step 5 M times and calculate family-wise error rate as $\frac{\sum_{k=1}^{M} U_k}{M}$

A.5 Power simulation using residual bootstrap

- Step 1: Generate the response vector y_{null} under the null hypothesis
  - $y_{null} = X_F\beta_F + Zb + \epsilon$, where X_F and Z are from the real data, β_F are assumed to be the true parameters, b is drawn from $N(0, \tilde{D})$ and ϵ is drawn from $N(0, R)$
- Step 2: Generate the response vector $y_{alternative}$ under a specific shape by using some function of the main effect
  - $y_{alternative} = f(X_{main}) + X_F\beta_F + Zb + \epsilon$, where $f(X_{main})$ is a function of the main effect, X_F and Z are from the real data, β_F are assumed to be the true parameters, b is drawn from $N(0, \tilde{D})$ and ϵ is drawn from $N(0, R)$
- Step 3: Fit the model $y = X_S\beta_S + X_F\beta_F + Zb + \epsilon$ under constraints on β_S using $y_{alternative}$ and calculate the test statistics T_1, T_2, T_3, T_4
- Step 4: Fit the model $y = X_F\beta_F + Zb + \epsilon$ using y_{null}
- Step 5: Bootstrap the null distribution
  - Bootstrap the random effects and noise from the model in Step 3
  - Adopt the fixed effects estimates $\hat{\beta}_F$ in Step 4
  - Generate the response vector y_{boot} under the null hypothesis using bootstrapped random effects and noise

  - Fit the model $y = X_S\beta_S + X_F\beta_F + Zb + \epsilon$ under constraints on β_S using y_{boot} and calculate the test statistics T_1^*, T_2^*, T_3^*, T_4^*
  - Compare T_i with T_i^* (if $T_i \leq T_i^*$ then $p_{ij} = 1$, else $p_{ij} = 0$), where $i = 1, 2, 3, 4$
- Step 6: Repeat Step 5 N times and calculate p_i as $\frac{\sum_{j=1}^{N} p_{ij}}{N}$, where $i = 1, 2, 3, 4$; sort p_1, p_2, p_3, p_4 in ascending order and follow the Holm-Bonferroni strategy to calculate V_k (for example, for increasing shape, if $p_{(1)} < 0.0125$ and $p_{(2)} \geq 0.05/3$, then $V_k = 1$, else $V_k = 0$)
- Step 7: Repeat Step 1 to Step 6 M times and calculate power as $\frac{\sum_{k=1}^{M} V_k}{M}$

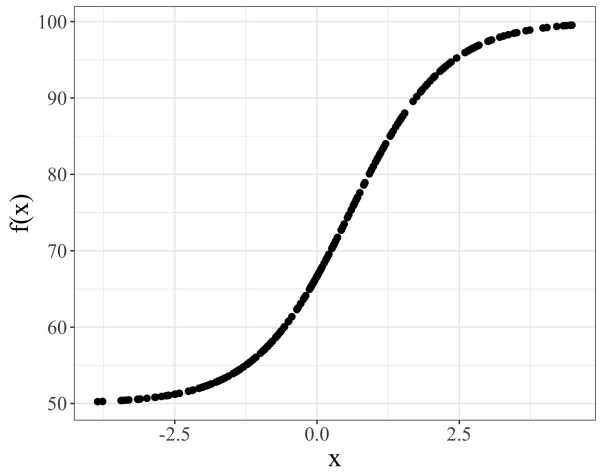
A.6 Plots of power curve under simulation type (a)



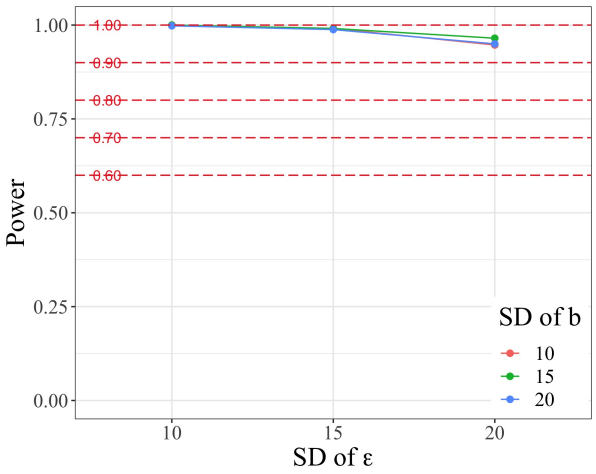
(a) Scatter plot of simulated curve

(b) Power Curve

Figure A.6.1: Power curve for shape 1 ($f(x) = 5.5x + 70$)

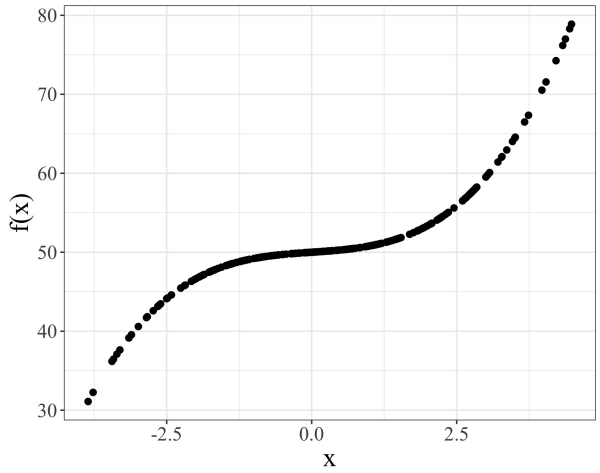


(a) Scatter plot of simulated curve

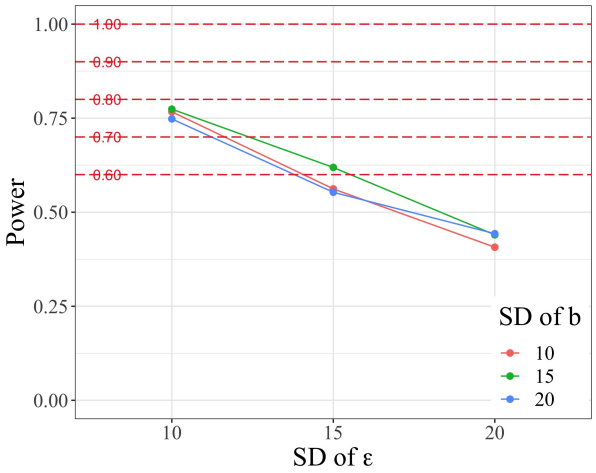


(b) Power Curve

Figure A.6.2: Power curve for shape 2 ($f(x) = 50\frac{e^{1.2x}}{2+e^{1.2x}} + 50$)

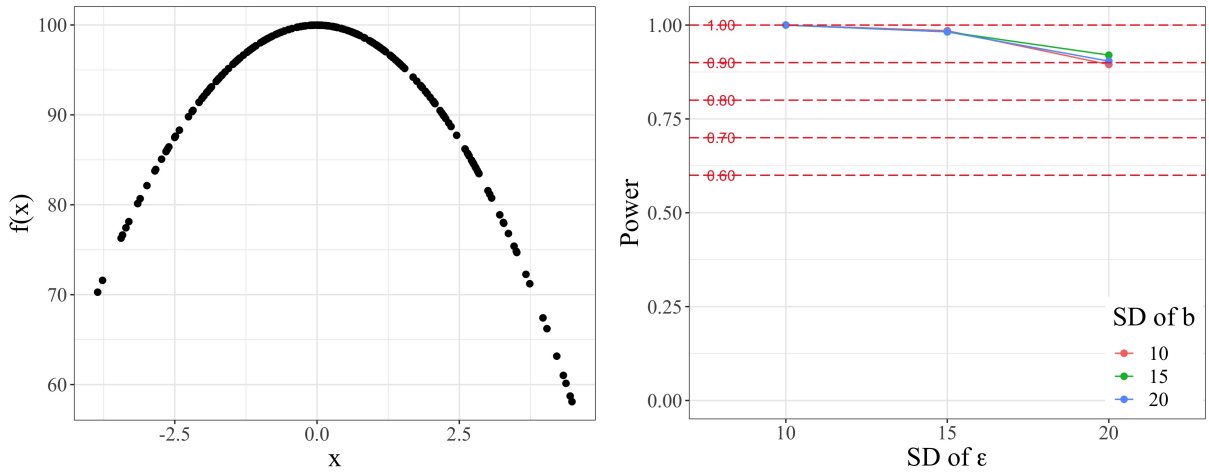


(a) Scatter plot of simulated curve



(b) Power Curve

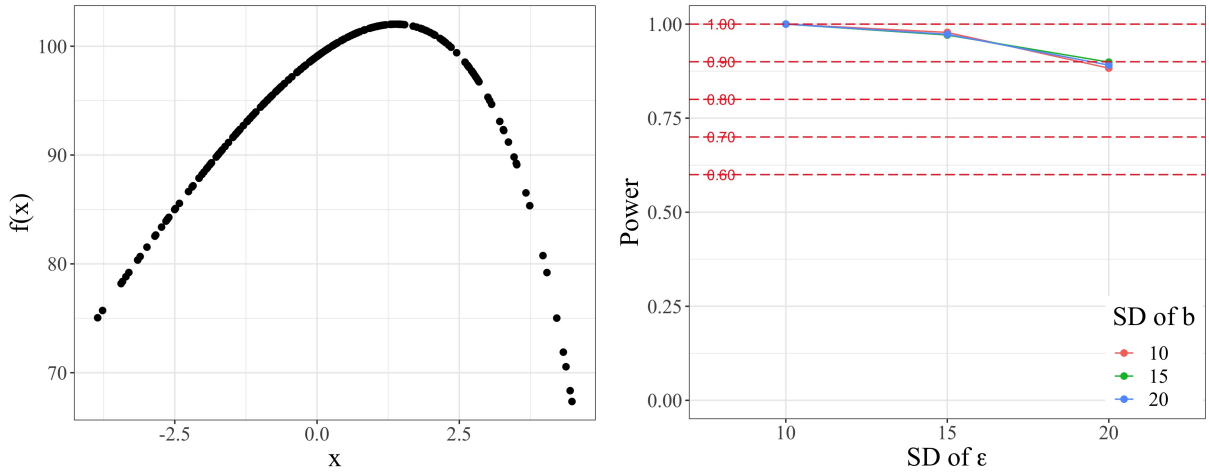
Figure A.6.3: Power curve for shape 3 ($f(x) = (\frac{2x}{3})^3 + \frac{x}{2} + 50$)



(a) Scatter plot of simulated curve

(b) Power Curve

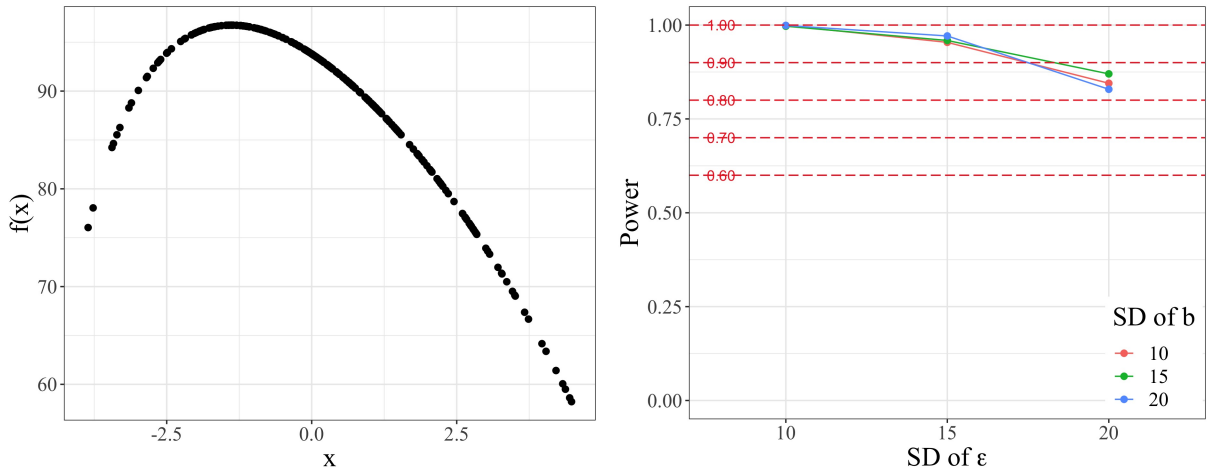
Figure A.6.4: Power curve for shape 10 ($f(x) = \frac{-e^x - 100x^2}{50} + 100$)



(a) Scatter plot of simulated curve

(b) Power Curve

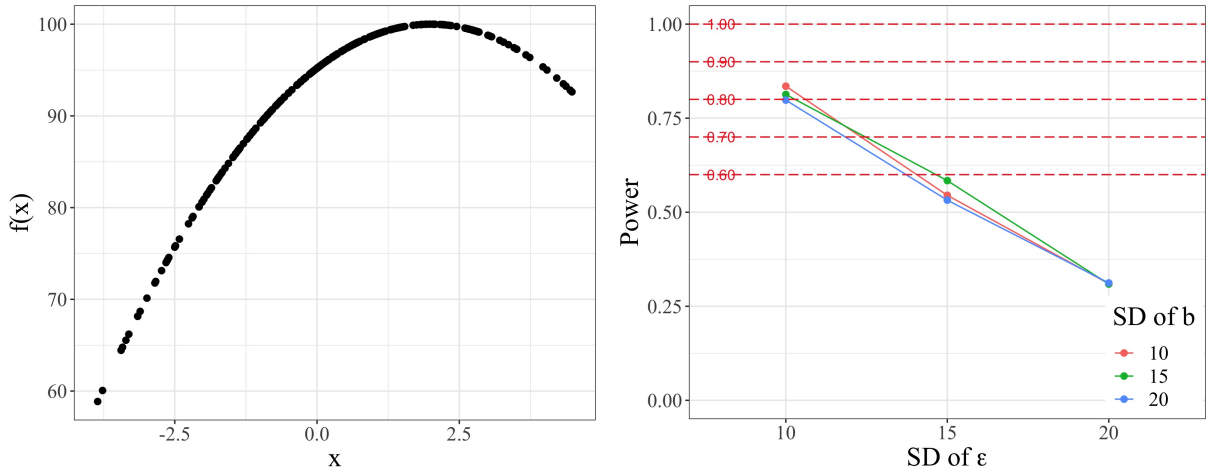
Figure A.6.5: Power curve for shape 11 ($f(x) = 300\ln(-e^{x/2} + x + 40) - 1000$)



(a) Scatter plot of simulated curve

(b) Power Curve

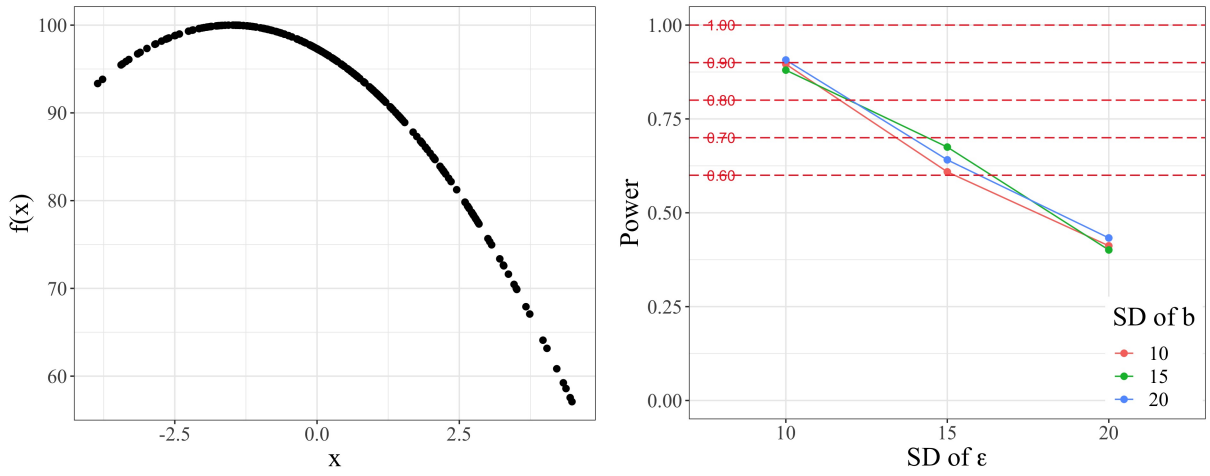
Figure A.6.6: Power curve for shape 12 ($f(x) = 70\ln(-e^{-x/2} - x + 10) - 60$)



(a) Scatter plot of simulated curve

(b) Power Curve

Figure A.6.7: Power curve for shape 14 ($f(x) = -1.2(x-2)^2 + 100$)



(a) Scatter plot of simulated curve

(b) Power Curve

Figure A.6.8: Power curve for shape 16 ($f(x) = -1.2(x + 1.5)^2 + 100$)

Appendix B Appendix for Chapter 3

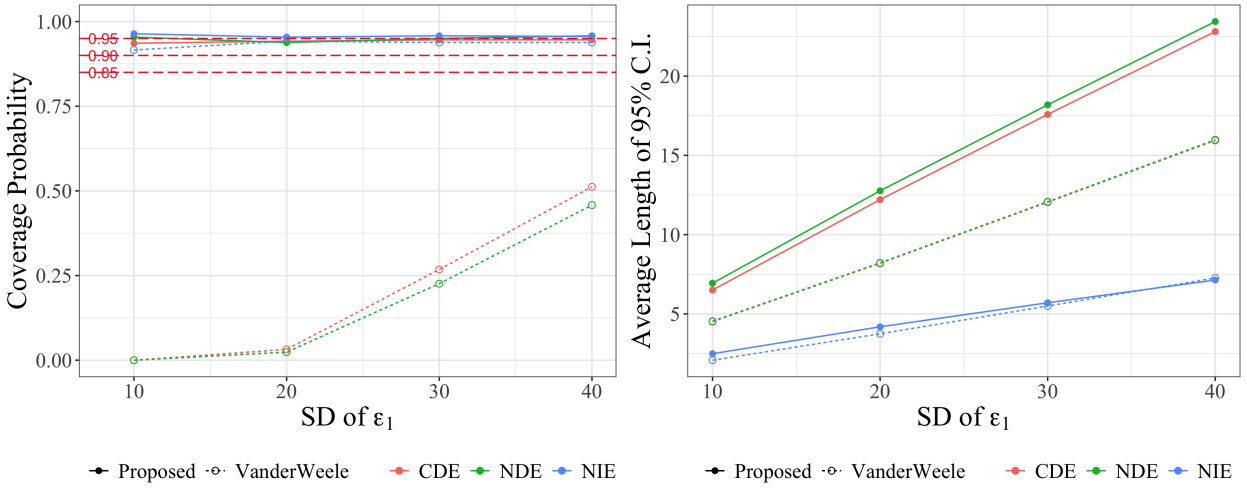
B.1 Hinge algorithm

Let $C = \{\phi \in R^n : \phi = v + \sum_{j\in J} b_j\delta^j$, where $b_j \geq 0$ and $v \in V\}$. Define $\Omega = C \cap V^{\perp}$, where $V^{\perp}$ is the linear space orthogonal to V. Then $\Omega = \{\phi \in R^n : \phi = \sum_{j\in J} b_j\delta^j$, where $b_j \geq 0\}$. The optimization problem is $\min_{\phi\in C}||y-\phi||^2$, and the necessary and sufficient conditions for the optimization problem are $< y-\hat{\phi}, \hat{\phi} >= 0$ and $< y-\hat{\phi}, \phi > \leq 0 \ \forall \ \phi \in C$.

The hinge algorithm has the following steps:

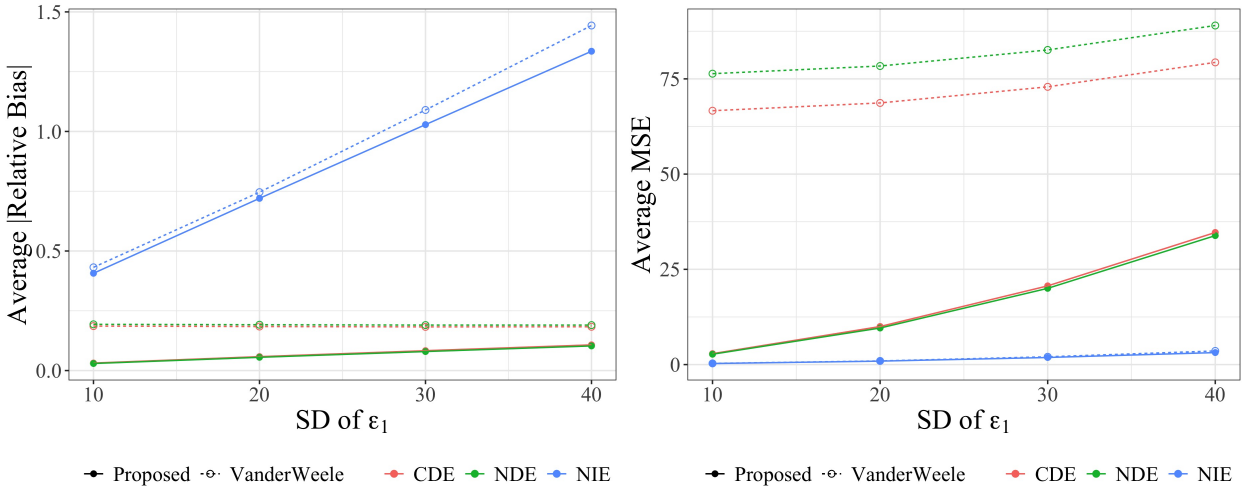
- The initial guess J_0 can be any subset of J for which the corresponding δ^j form a linearly independent set.
- At the k^{th} iteration:
  - Step 1: Project z onto the linear space spanned by $\{\delta^j, j \in J_k\}$, to get $\phi^k = \sum_{j\in J_k} b_j^{(k)}\delta_j$.
  - Step 2: Check to see if ϕ^k satisfies the constraints, that is, if all $b_j^{(k)}$ are nonnegative: if yes, go to Step 3; if no, choose j for which $b_j^{(k)}$ is smallest, remove it from the set, and go to Step 1.
  - Step 3: Compute $< y-\phi^k, \delta^j >$ for each $j \notin J_k$. If these are all nonpositive, then stop. If not, choose j for which this inner product is largest, add it to the set, and go to Step 1.

B.2 Plots of simulation results of coverage probability, average length of 95% C.I., average absolute relative bias and average MSE



(a) Coverage probability

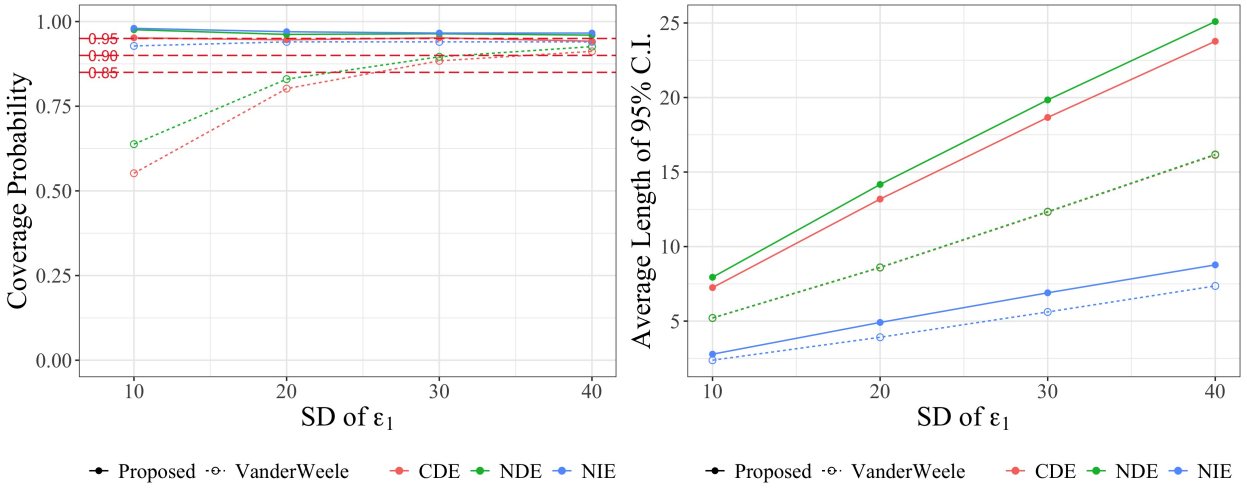
(b) Average length of 95% C.I.



(c) Average |Relative Bias|

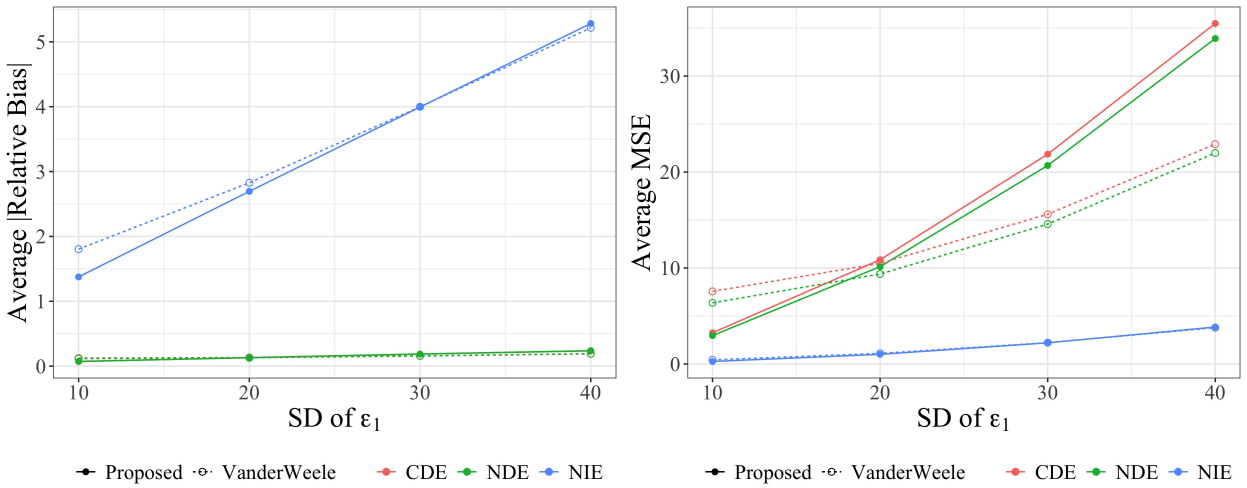
(d) Average MSE

Figure B.2.1: Plots of simulation results for case 1



(a) Coverage probability

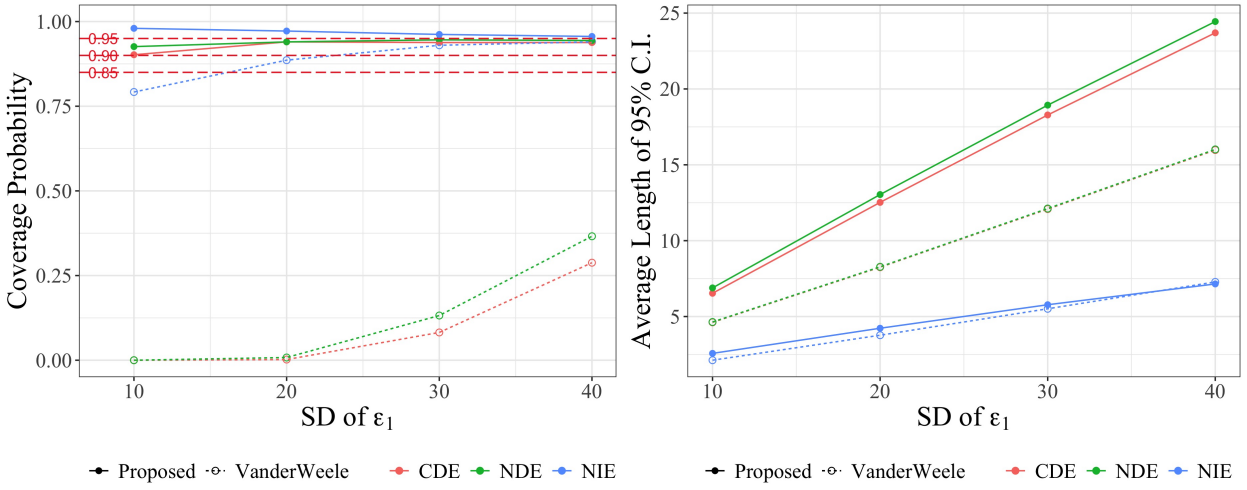
(b) Average length of 95% C.I.



(c) Average |Relative Bias|

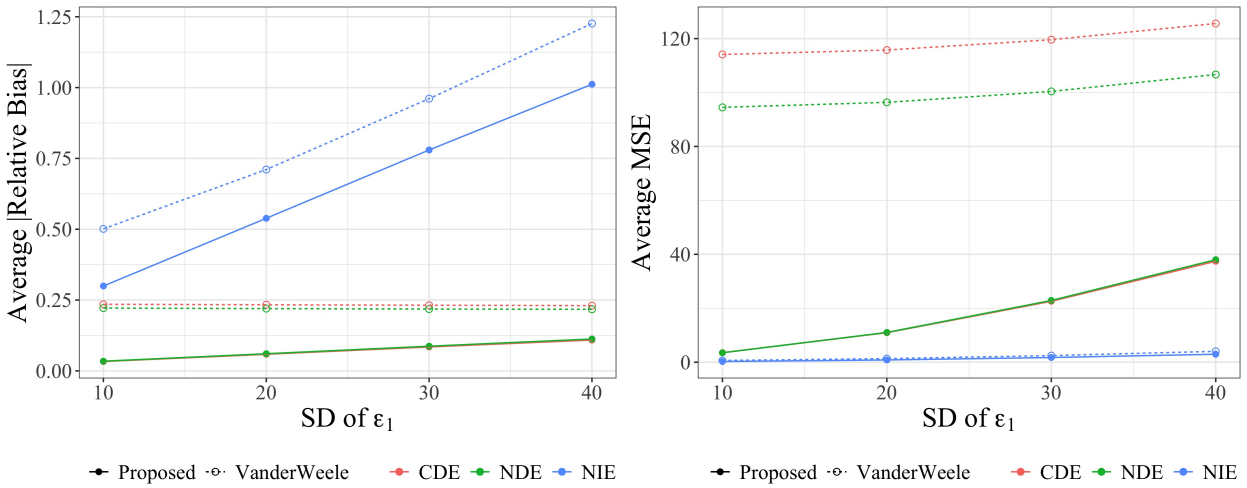
(d) Average MSE

Figure B.2.2: Plots of simulation results for case 2



(a) Coverage probability

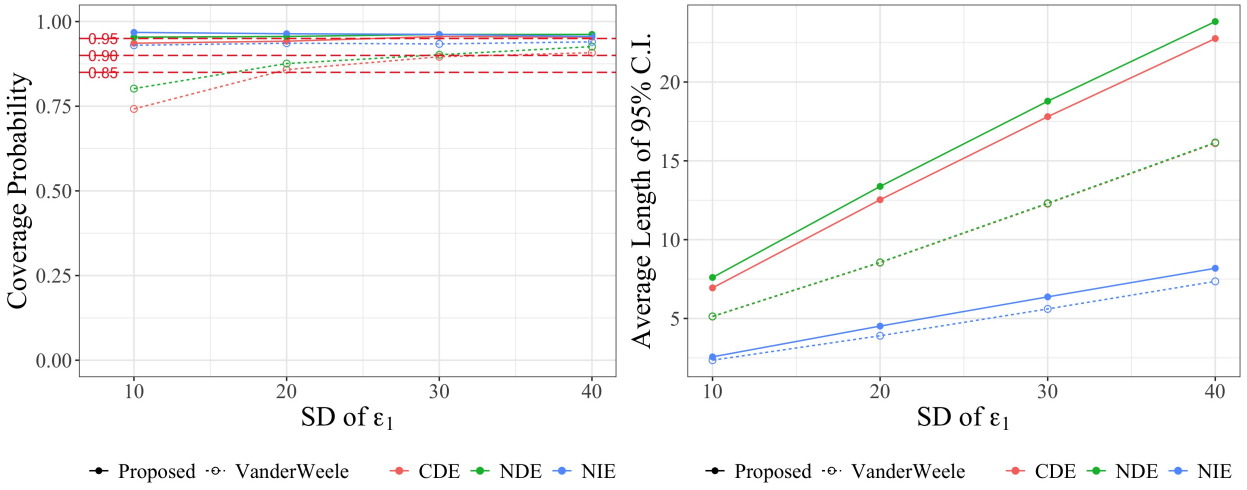
(b) Average length of 95% C.I.



(c) Average |Relative Bias|

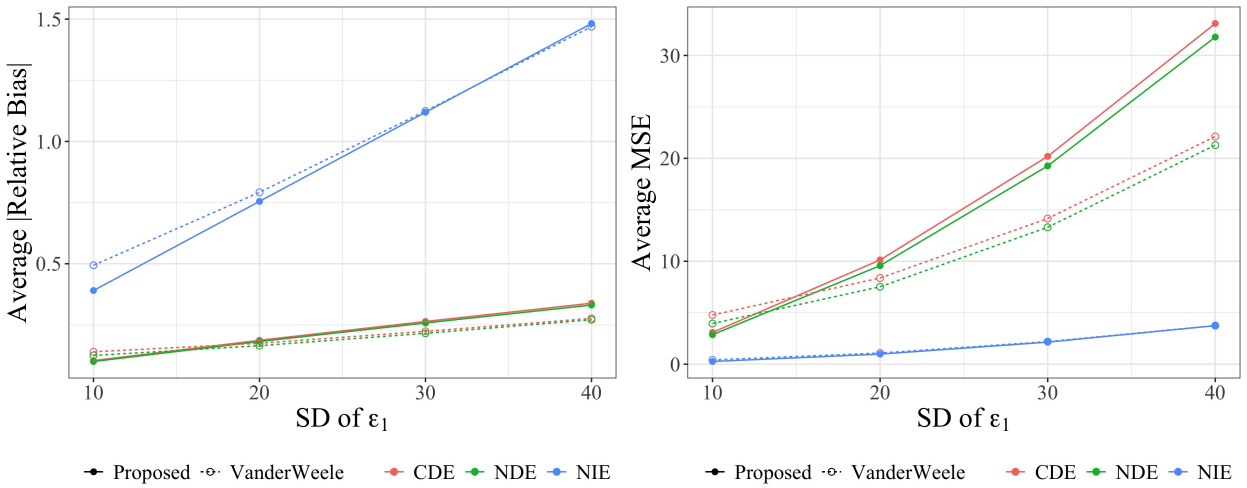
(d) Average MSE

Figure B.2.3: Plots of simulation results for case 3



(a) Coverage probability

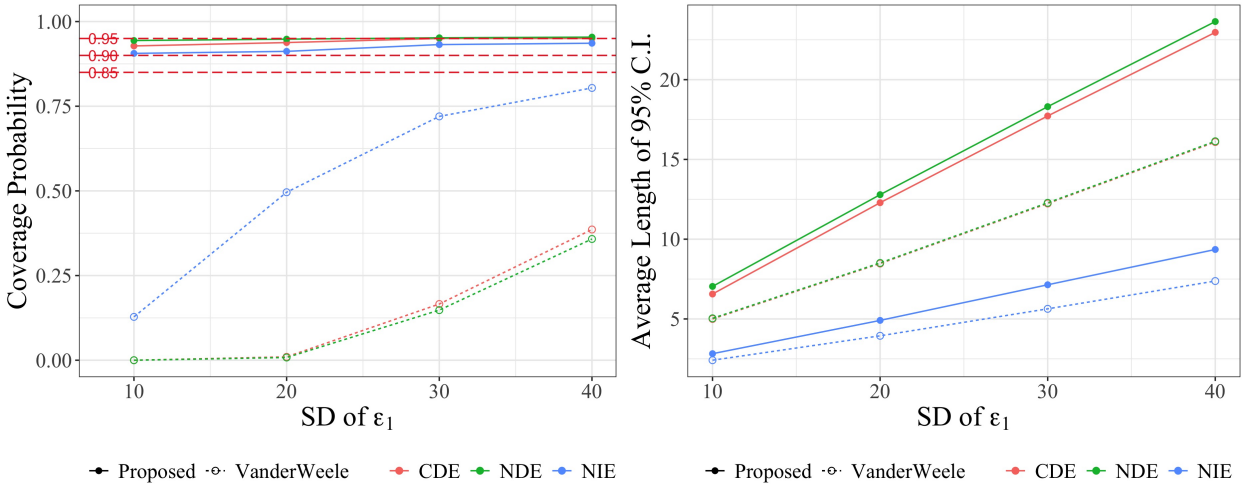
(b) Average length of 95% C.I.



(c) Average |Relative Bias|

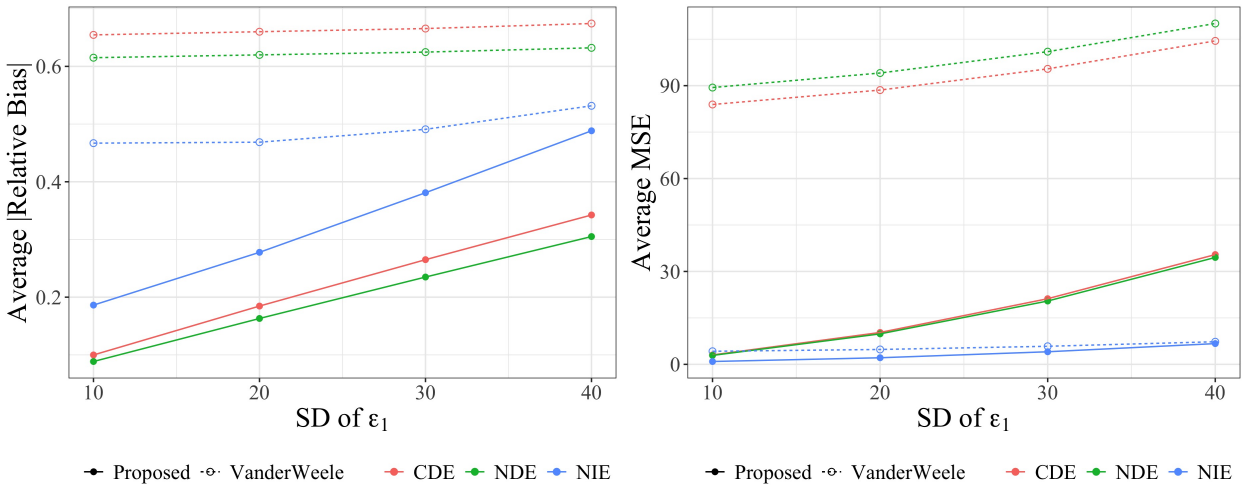
(d) Average MSE

Figure B.2.4: Plots of simulation results for case 4



(a) Coverage probability

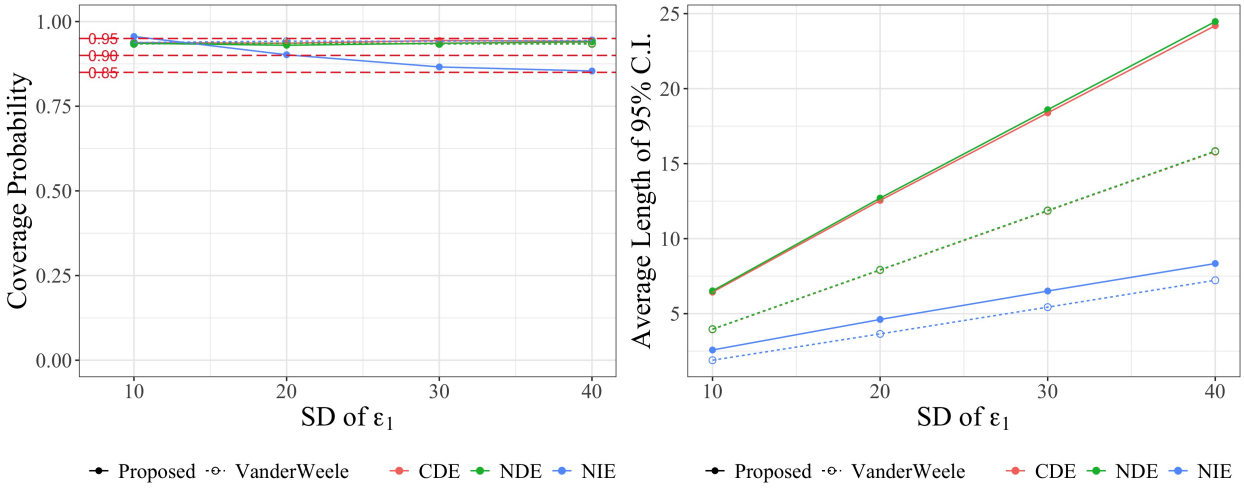
(b) Average length of 95% C.I.



(c) Average |Relative Bias|

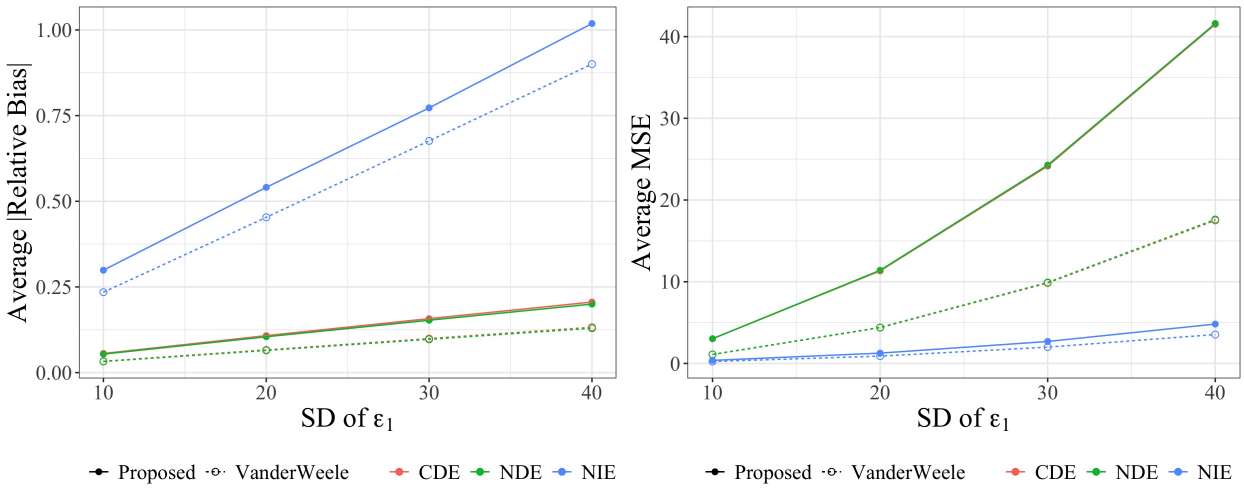
(d) Average MSE

Figure B.2.5: Plots of simulation results for case 5



(a) Coverage probability

(b) Average length of 95% C.I.



(c) Average |Relative Bias|

(d) Average MSE

Figure B.2.6: Plots of simulation results for case 6

Bibliography

J. Adibi, M. Lee, A. Naimi, E. Barrett, R. Nguyen, S. Sathyanarayana, Y. Zhao, M. Thiet, J. Redmon, and S. Swan. Human chorionic gonadotropin partially mediates phthalate association with male and female anogenital distance. *The Journal of Clinical Endocrinology and Metabolism*, 100(9):E1216–E1224, 2015.

J. Adibi, A. Layden, R. Birru, A. Miragaia, X. Xun, M. Smith, Q. Yin, M. Millenson, T. O'Connor, E. Barrett, N. Snyder, S. Peddada, and R. Mitchell. First trimester mechanisms of gestational sac placental and foetal teratogenicity: a framework for birth cohort studies. *Human reproduction update*, 27(4):747–770, 2021a. doi: https://doi.org/10.1093/humupd/dmaa063.

J. Adibi, X. Xun, Y. Zhao, Q. Yin, K. LeWinn, N. Bush, A. Panigrahy, S. Peddada, H. Alfthan, U. Stenman, F. Tylavsky, and H. Koistinen. Second trimester placental and thyroid hormones are associated with cognitive development from 1 to 3 years of age. *Journal of the Endocrine Society*, 5(5):bvab027, 2021b. doi: https://doi.org/10.1210/jendso/bvab027.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

S. Canini, F. Prefumo, D. Pastorino, L. Crocetti, C. Afflitto, P. Venturini, and P. De Biasio. Association between birth weight and first-trimester free beta-human chorionic gonadotropin and pregnancy-associated plasma protein A. *Fertility and Sterility*, 89(1): 174–178, 2008.

J. Choi, J. Lee, J. Jhong, and J. Koo. Penalized I-spline monotone regression estimation. *Communications in Statistics - Simulation and Computation*, 2019. doi: https://doi.org/10.1080/03610918.2019.1630433.

B. Coull, D. Ruppert, and M. Wand. Simple incorporation of interactions into additive models. *Biometrics*, 57(2):539–545, 2001.

H. Curry and I. Schoenberg. On pólya frequency functions IV: The fundamental spline functions and their limits. *Journal d'Analyse Mathematique*, 17:71–107, 1966.

G. Dörner. Die mögliche bedeutung der prä- und/oder perinatalen ernährung für die pathogenese der obesitas. *Acta Biologica et Medica Germanica*, 30:19–22, 1973.

A. Drake and R. Reynolds. Impact of maternal obesity on offspring obesity and cardiometabolic disease risk. *Reproduction*, 140(3):387–398, 2010.

L. Farnan, A. Ivanova, and S. Peddada. Linear mixed effects models under inequality constraints with applications. *PLoS One*, 9(1):e84778, 2014.

M. Gillman, S. Rifas-Shiman, C. Berkey, A. Field, and G. Colditz. Maternal gestational diabetes, birth weight, and adolescent obesity. *Pediatrics*, 111(3):e221–e226, 2003.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

P. Huang, P. Kuo, Y. Guo, P. Liao, and C. Lee. Associations between urinary phthalate monoesters and thyroid hormones in pregnant women. *Human reproduction*, 22:2715–2722, 2007.

J. Hwang and S. Peddada. Confidence interval estimation subject to order restrictions. *The Annals of Statistics*, 22(1):67–93, 1994.

K. Imai, L. Keele, and D. Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334, 2010.

T. Korevaar, R. Muetzel, M. Medici, L. Chaker, V. Jaddoe, Y. de Rijke, E. Steegers, T. Visser, T. White, H. Tiemeier, and R. Peeters. Association of maternal thyroid function during early pregnancy with offspring IQ and brain morphology in childhood: a

population-based prospective cohort study. *The Lancet: Diabetes and Endocrinology*, 4 (1):35–43, 2016.

Y. Larriba, C. Rueda, M. A. Fernández, and S. D. Peddada. Order restricted inference for oscillatory systems for detecting rhythmic signals. *Nucleic Acids Research*, 44(22):e163, 2016.

A. Larsen, S. Gaines, and O. Deschênes. Agricultural pesticide use and adverse birth outcomes in the san joaquin valley of California. *Nature Communications*, 8(1):302, 2017.

L. Lumey. Decreased birthweights in infants after maternal in utero exposure to the Dutch famine of 1944–1945. *Paediatric and Perinatal Epidemiology*, 6(2):240–253, 1992.

M. Meyer. Inference using shape-restricted regression splines. *Annals of Applied Statistics*, 2(3):1013–1033, 2008.

M. Meyer. Constrained penalized splines. *The Canadian Journal of Statistics*, 40(1):190–206, 2012.

M. Meyer. A simple new algorithm for quadratic programming with applications in statistics. *Communications in Statistics - Simulation and Computation*, 42(5):1126–1139, 2013.

M. Meyer. Constrained partial linear regression splines. *Statistica Sinica*, 28(1):277–292, 2018.

C. Nagata, S. Iwasa, M. Shiraki, and H. Shimizu. Estrogen and alpha-fetoprotein levels in maternal and umbilical cord blood samples in relation to birth weight. *Cancer Epidemiology, Biomarkers & Prevention*, 15(8):1469–1472, 2006.

J. Ramsay. Monotone regression splines in action. *Statistical Science*, 3(4):425–461, 1988.

C. Rao and J. Kleffe. *Estimation of variance components and applications*. North-Holland, Amsterdam, 1988.

S. Searle, G. Casella, and C. McCulloch. *Variance Components*. John Wiley and Sons, Inc., Hoboken, New Jersey, 2006.

M. Silvapulle and P. Sen. *Constrained Statistical Inference: Inequality, Order and Shape Restrictions*. John Wiley and Sons, Inc., Hoboken, New Jersey, 2005.

N. Tul, S. Pusenjak, J. Osredkar, K. Spencer, and Z. Novak-Antolic. Predicting complications of pregnancy with first-trimester maternal serum free-beta hCG, PAPP-A and inhibin-A. *Prenatal Diagnosis*, 23:990–996, 2003.

T. VanderWeele. *Explanation In Causal Inference : Methods for Mediation and Interaction*. Oxford University Press, New York, 2015.

E. Wegman and I. Wright. Splines in statistics. *Journal of the American Statistical Association*, 78(382):351–365, 1983.

S. Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2017.

H. Wu and J. Zhang. *Nonparametric Regression Methods for Longitudinal Data Analysis*. John Wiley and Sons, Inc., Hoboken, New Jersey, 2006.