ABSTRACT

Title of document: SEMIPARAMETRIC AND NONPARAMETRIC ANALYSIS FOR LONGITUDINAL DATA ON THE RELATIONSHIP BETWEEN CHILDHOOD EXTERNALIZING BEHAVIOR AND BODY MASS INDEX

Kejia Wang, MPH, 2011

Directed By: Xin He, Ph.D.
Assistant Professor
Department of Epidemiology and Biostatistics

This thesis is an extension of the longitudinal data analysis of the association between externalizing behavior in early childhood and body mass index (BMI) from age 2 to 12 years conducted in Anderson et al. (2010). Externalizing behaviors problems are characterized by aggressive, oppositional, disruptive, or inattentive behaviors beyond those that would be expected given a child’s age and development. The aim of the thesis is to estimate the children's BMI trajectory and to evaluate to what extent the externalizing behavior is related to BMI using semiparametric and nonparametric time-varying coefficient models. Some valuable insights into how the externalizing behavior and BMI are associated will be provided.

SEMIPARAMETRIC AND NONPARAMETRIC ANALYSIS FOR LONGITUDINAL DATA ON THE RELATIONSHIP BETWEEN CHILDHOOD EXTERNALIZING BEHAVIOR AND BODY MASS INDEX

By

Kejia Wang

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Public Health
2011

Advisory Committee:
Dr. Xin He, Chair
Dr. Guangyu Zhang
Dr. Brit Saksvig

Acknowledgements

First and foremost I want to express my sincere gratitude to my advisor Dr. Xin He. He is an excellent teacher and advisor. His ideas and guidance have helped me during the research and writing of this thesis. I appreciate all his contributions of time and ideas to make my master experience productive and stimulating. I also appreciate his patience, enthusiasm, immense knowledge, as well as his helpful career guidance for me. I could not imagine having a better advisor and mentor in my master study.

I also want to thank Dr. Deborah Young, for all her great ideas, suggestions and insightful comments for this thesis.

Besides, I also would like to thank the rest of my thesis committee members: Dr. Guangyu Zhang and Dr. Brit Saksvig for their valuable comments.

I am thankful for my classmates Lu Chen and Xia Li for their helps to me and all the fun we have in the same GA office during these two years.

Last but not the lease, I would like to thank my parents for supporting me spiritually throughout my life.

Table of Contents

Acknowledgements ii

Table of Contents iii

List of Tables iv

List of Figures v

Chapter 1: Introduction 1

1.1 Background 2

1.1.1 Study Population 3

1.1.2 Childhood Externalizing and Internalizing Behavior 5

1.1.3 Body Mass Index 6

1.1.4 Covariates 7

1.1.5 Key Analyses and Findings 7

1.2 Research Questions 9

Chapter 2 Semiparametric and Nonparametric Models for Longitudinal Data 10

2.1 Overview 11

2.2 Semiparametric Models 15

2.2.1 Semiparametric Population Mean Model 15

2.2.2 Semiparametric Mixed Effects Model 16

2.3 Nonparametric Time-varying Coefficient Models 17

2.3.1 Nonparametric Time-varying Coefficient Population Mean Model 18

2.3.2 Nonparametric Time-varying Coefficient Mixed Effects Model 19

Chapter 3 Results 21

Chapter 4 Discussion 26

4.1 Strengths 27

4.2 Limitations 28

4.3 Conclusion Remarks 28

Appendices 30

Bibliography 38

List of Tables

Table 1 Summary of variables in Anderson et al. (2010) 30

Table 2 Semiparametric population mean model fitted among boys: estimated covariate effects, standard deviations, approximate z-test values, P values for the parametric component 30

Table 3 Semiparametric population mean model fitted among girls: estimated covariate effects, standard deviations, approximate z-test values, P values for the parametric component 31

Table 4 Semiparametric mixed effects model fitted to dataset among boys: estimated covariate effects, standard deviations, approximate z-test values, P values for the parametric component 31

Table 5 Semiparametric mixed effects model fitted to dataset among girls: estimated covariate effects, standard deviations, approximate z-test values, P values for the parametric component 32

Table 6 Comparison of the four proposed models 33

List of Figures

Figure 1 Semiparametric population mean model fitted to boys and girls: estimated BMI trajectory with 95% pointwise SD bands 34

Figure 2 Semiparametric mixedeffects model fitted to boys and girls: fitted coefficient functions and their 95% pointwise SD bands 34

Figure 3 Time-varying coefficient nonparametric population mean model for boys: estimated coefficient functions and their 95% pointwise SD bands 35

Figure 4 Time-varying coefficient nonparametric population mean model girls: estimated coefficient functions and their 95% pointwise SD bands 35

Figure 5 Time-varying coefficient nonparametric mixed effects model for boys: fitted coefficient functions and their 95% pointwise SD bands 36

Figure 6 Time-varying coefficient nonparametric mixed effects model for girls: fitted coefficient functions and their 95% pointwise SD bands 37

Chapter 1: Introduction

Longitudinal data, which involve repeated measurements that are recorded on the same subject over a certain time period, are frequently encountered in epidemiological studies. Parametric generalized estimation equation (GEE) based marginal models (Liang & Zeger, 1986; Zeger et al., 1988) and general linear mixed effects models (LME) (Harville, 1976, 1977; Laird & Ware, 1982) are the predominant approaches for analyzing longitudinal data. Parametric models are a powerful tool for modeling the association between the outcome and covariates, but fully parametric models may also be too restrictive or limited to be adequate, and subject to model misspecification (Hoover et al., 1998; Wu & Yu, 2002). Nonparametric regression, which is well known to be more data adaptive and less restrictive than parametric approaches, thus emerges as a promising alternative to handle longitudinal data. Because of the flexibility in the form of regression models, nonparametric modeling approaches can play an important role in exploring longitudinal data similarly as their applications in cross-sectional studies. Nonparametric models are more robust against the model misspecification, but they are more complex and less efficient than parametric models. Semiparametric models, which include both parametric and nonparametric components, retain advantages of both parametric and nonparametric models (Wu & Zhang, 2006).

Nonparametric models make no assumption on the functional form of the model, but it may fail to incorporate some prior information, thus the resulting estimator of the unknown function tends to incur a greater variance. In addition, the standard nonparametric method is practically impotent when the dimension of the covariates is high. Varying-coefficient models, a class of structural nonparametric models, can effectively relax the conditions imposed on traditional parametric models and explore the hidden structure. They help one explore the dynamic feature that may exist in the dataset as well as to make the model fit the data better (Fan & Zhang, 2008).

1.1 Background

Behavior problems and obesity are very important factors that affect the health of children. Some evidence suggests that obesity is associated with childhood externalizing behavioral problems (Lumeng et al., 2003; Datar & Sturm, 2004; Mamun et al., 2009). But this association has not been observed in all studies or in children of both sexes (Lawlor et al., 2005; Datar & Sturm, 2006; Drukker et al., 2008; Bradley et al., 2008). There is controversy regarding the direction of the association (Lawlor et al., 2005; Datar & Sturm, 2006; Mamun et al., 2009). Previous studies show that children who experience maltreatment or neglect, bullying, social marginalization, or academic difficulties are more likely to have externalizing behavior problems (Strauss & Pollack, 2003; Deater-Deckard et al., 1998). These factors are also related to obesity (Lissau & Sorensen, 1993, 1994; Strauss & Pollack, 2003; Janssen et al., 2004), and this association appears to be stronger among African American females (Anderson et al., 2006). Accumulating evidence indicates that the pathways in the brain governing appetite and emotion regulation are interrelated and may be impacted by stress (McEwen, 2008). Also, it is suggested that alterations in developmental weight trajectories are associated with indicators of psychopathology. In clinical samples of adolescents and adults, both overweight and underweight are associated with significantly increased levels of psychopathology (Maddi et al, 1997; Vila et al., 2004; Kaye et al., 2004; Bollen & Wojciechowski, 2004). The findings from previous analyses suggest the need to study the associations between externalizing behavior and weight status in a diverse cohort observed from an early age, in which height and weight are measured repeatedly throughout childhood (Anderson et al., 2010).

In Anderson et al. (2010), the main aim is to examine the extent to which externalizing behavior in early childhood is related to body mass index (BMI) and to their BMI trajectory through 12 years of age, and to evaluate whether these associations differ by sex and race. In this thesis, we will consider similar research questions but use more general statistical models.

1.1.1 Study Population

The study data were collected by the National Institute of Child Health and Human Development (NICHD) Study of Early Child Care and Youth Development (SECCYD). NICHD, which is part of the National Institutes of Health (NIH) within the U.S. Department of Health and Human Services, began a study in 1991 to collect information about different non-maternal child care arrangements, and about children and families who use child care as well as those who do not. The major goal of the NICHD study is to examine how differences in child care experiences relate to children's social, emotional, intellectual, and language development, and to their physical growth and health (National Institute of Health, 2006).

The NICHD Study of Early Child Care and Youth Development (SECCYD) is a large-scale prospective longitudinal study. Since 1991, the study has followed the development of children from the time they were 1 month of age until 15½ years. Participants in the study were recruited from 24 hospitals in the vicinity of 10 data collection sites around the country (e.g., Charlottesville, VA; Irvine, CA; Lawrence, KS; Little Rock, AR; Madison, WI; Morganton, NC; Philadelphia, PA; Pittsburgh, PA; Seattle, WA; and Wellesley, MA). Researchers have used multiple methods to assess children's development (e.g., trained observers, interviewers, questionnaires, and testing), and measured many aspects of children's development (e.g., social, emotional, intellectual, and language development; behavioral problems and adjustment; and physical health).

During the selected sampling periods, study personnel visited new mothers in hospital. To be eligible, the mother had to be 18 years of age or older, healthy, and conversant in English, and the infant had to be a singleton. The collection of data from parents, children, and other adults in home, laboratory playroom, child care, and school visits proceeded in several stages from birth and is ongoing (Bradley et al., 2008). In its initial phase (1991–1994), the NICHD Study of Early Child Care followed the development of 1,364 families with healthy newborns at 10 sites with approximately equal numbers of families at each site in United States from birth through age 3. Phase II of the study (1995–1999) followed the same children's development through first grade (with 1,226 children participated in the study).

Phase III of the study (2000–2004) followed the same children through sixth grade (with 1,100 children participated in the study). Phase IV of the study (2005-2007) followed the same children through ninth grade (with 1,056 children participated in the study). The detailed recruitment and selection procedures can be found in a previously published NICHD Early Child Care Research Network study (NICHD Early Child Care Research Network, 2001). The study procedure is also available at http://secc.rti.org.

1.1.2 Childhood Externalizing and Internalizing Behavior

Externalizing and internalizing behaviors comprise the most common children's reactions to the experience of stress (Achenbach & Edelbrock, 1981; Rutter & Garmezy, 1983). While externalizing behaviors are reactions that are directed toward others, including delinquent behavior and aggressive behaviors; internalizing behaviors, such as anxiety, withdrawal, somatic complaints, are mainly directed toward the self. The Child Behavior Checklist (CBCL) is a widely used standardized form that parents fill out to describe their children's behavioral and emotional problems (Achenbach, 2000), and the reliability and validity of the instrument are well-established (Achenbach & Edelbrock 1983; Achenbach, 1992; Achenbach, 2000). It contains 100 items to evaluate whether the child's exhibited behaviors are consistent with emotional or behavioral difficulties currently or in the past two months (Achenbach, 1992). Age-normed scores (T scores) for externalizing behaviors and internalizing behaviors are calculated from the aggressive and destructive behavior scales (externalizing) and the anxious/depressed and withdrawn scales (internalizing) (Achenbach, 1992). In Anderson et al. (2010), the Child Behavior Checklist (CBCL-2/3) was completed by mothers at their child's laboratory study visits. Mothers rated the extent to which each behavior described their child using the following three-level scale: "not true" (coded as 0), "somewhat true" (coded as 1), or "very true" (coded as 2). Six scales were derived from these rating: aggressive behavior (15 items), destructive behavior (11 items), anxious/depressed (11 items), withdrawn (14 items), sleeping problems (7 items), and somatic problems (14 items).

Although the externalizing behavior T score is a continuous measurement, but to determine whether covariates were associated with high levels of externalizing problems and to assess confounding, children were categorized as having high levels of externalizing behavior if their CBCL T score was ≥65 (which is the 95th percentile of externalizing behavior at the measurement time within the cohort, and is a cut point above which children's symptoms would be considered in the clinical range) (Achenbach, 1992).

1.1.3 Body Mass Index

Children's heights and weights have been measured during laboratory visits at 24, 36, and 54 months, and when children were in the 1st, 3rd, 5th, and 6th grades. The standardized protocol is used at all time periods and all sites. Body mass index (BMI) is calculated from height and weight (BMI = weight (kg)/height (m)^2). Obesity is defined as BMI-for-age above the 95th percentile of the Centers for Disease Control and Prevention (CDC) sex-specific BMI-for-age growth charts

(Kuczmarski et al., 2002).

1.1.4 Covariates

As described in Anderson et al. (2010), the covariates included in the study were children's race, sex, maternal depression, household poverty status, and maternal education attainment. Race was reported by mothers and was categorized as White and non-White. Maternal depression defined as CES-D score>16 (Radloff, 1977). Household poverty status is an indicator for whether or not household income-to-needs ratio was less than the federal poverty threshold. Mothers reported their educational attainment at the time of their child's birth, which was categorized as "<= high school graduate", "some college" and ">= college degree" (reference group).

1.1.5 Key Analyses and Findings

In Anderson et al. (2010), logistic regression was conducted to estimate odds ratios and 95% confidence intervals for the cross-sectional association between obesity and high levels of externalizing behavior at 24 months overall and stratified by sex and race ($\alpha<0.05$ as the significance level). Linear mixed effects models were used to estimate the average BMI trajectory and to test the extent to which externalizing behaviors at 24 months were related to children's BMI trajectory. To determine the BMI trajectory pattern relative to age, models were fitted with age as linear, quadratic, and cubic terms. The study also tested whether externalizing behaviors at 24 months were associated with the average BMI and the linear change in BMI with age by including the corresponding interaction terms in the model. The summary table of all the variables is given by Table 1.

The results in Anderson et al. (2010) suggested that cross-sectionally at 24 months, children with high levels of externalizing behavior had odds of obesity that were 2.9 (95% CI: 1.3, 6.5) times as high as children with lower levels of externalizing behavior. Race, maternal education, household poverty status, and maternal depression were all associated with differences in prevalence of high externalizing behavior at 24 months. Children's BMI trajectory from age 2 to 12 years was modeled as a cubic function of age. Externalizing behavior at 24 months was associated with the BMI trajectory (p =0.02), but there was little evidence overall that this association differed by age (p =0.38). Among two-year-old children, irrespective of race, and the result predicted an average difference of three-quarters of a BMI unit between children with high levels of externalizing behavior and children with low levels of externalizing behavior.

Anderson et al. (2010) applied linear mixed effects models to estimate the average BMI trajectory and to test the extent to which externalizing behaviors at 24 months were related to the differences in children's BMI. The linear mixed effects models used were parametric models under the assumption that the effects of externalizing behaviors on BMI are constant over time, which might not be true. Thus the parametric models might not be the most appropriate choice for that specific research question. Semiparametric models, which retain the advantages of both parametric and nonparametric models, will be considered as an alternative in this thesis.

1.2 Research Questions

The objective of this research is to model the trajectory of BMI using semiparametric models, and to evaluate the time-varying effects of childhood externalizing behaviors on body mass index (BMI) using nonparametric varying-coefficient models.

In chapter 2, we give an overview of semiparametric and nonparametric regression methods for longitudinal data, and briefly review the penalized spline method used in the estimation procedure. Two semiparametric models and two nonparametric time-varying coefficient models will be introduced. In chapter 3, we apply the proposed method to the NICHD SECCYD dataset. The results will be compared between boys and girls. In chapter 4, some concluding remarks will be given with respect to semiparametric models and nonparametric time-varying coefficient models. The strengths and weaknesses of this research will also be discussed.

Chapter 2 Semiparametric and Nonparametric Models for Longitudinal Data

In biomedical and epidemiological studies, interests are often focused on evaluating the effects of treatment, dosage, risk factors or other covariates on the outcomes, such as disease progression and change of health status of a population, over time (Wu & Yu, 2002). The key difference between longitudinal and cross-sectional data is that longitudinal data are usually correlated within an individual and independent between subjects, while cross-sectional data are often independent (Wu & Zhang, 2006). Thus, how to take into account for the within-subject correlation becomes a challenge for longitudinal data analysis.

Parametric models, such as linear mixed effects models and nonlinear models are frequently used in analyzing longitudinal data, by including random effects to take into account for the within-subject correlation. However, the model specification is often hard to verify, which may lead to biased results. Therefore, there is a need to release the parametric assumption on the functional form of the model to get a more precise estimation. Various nonparametric models, such as nonparametric population mean models and nonparametric mixed effects models have been proposed for longitudinal data (Hoover et al., 1998; Wang, 1998a, b; Wu, Chiang; Fan & Zhang, 2000; Chiang, Rice & Wu, 2001).

Although nonparametric models are more robust against the model assumptions, they are usually more complex and less efficient than parametric models. semiparametric models are often used to compromise and retain the advantages of both parametric and nonparametric models,. The most commonly used semiparametric models include semiparametric population mean models (Martinussen & Scheike, 1999; Cheng & Wei, 2000; Lin & Carroll 2001a, b; Lin & Ying, 2001; He et al., 2002; Fan & Li, 2004, Hu et al., 2004; Lin & Carroll, 2005) and semiparametric mixed effects models (Zeger & Diggle, 1994; Zhang et al., 1998; Tao et al., 1999; Jacqmin-Gadda et al., 2002; Ruppert, Wand & Carroll, 2003; Durban et al., 2005).

Semiparametric mixed effects models (SPMEs) extend LMEs by modeling a covariate effect (e.g., time effect), using a nonparametric function (Zeger & Diggle, 1994; Zhang et al., 1998) and modeling other covariate effects parametrically. Also, a random effect component will be added to the LMEs to account for the within-subject correlation.

2.1 Overview

Semiparametric and nonparametric regression methods for longitudinal data have been well developed during the last two decades. Nonparametric regression methods can be broadly classified as kernel methods (Gasser & Muller, 1979; Gasser & Rosenblatt, 1984; Hart & Wehrly, 1986; Wand & Jones, 1995; Fan & Gijbels, 1996; Lin & Carroll, 2000; Chiou et al., 2002), spline methods, which consist of smoothing splines (Wahba, 1990; Green & Silverman, 1994), penalized splines (Eilers & Marx, 1996; Ruppert et al., 2003), and regression splines (Stone et al., 1997). Smoothing splines and penalized splines are based on penalized likelihoods. The traditional local likelihood based kernel methods are not able to account for the within-subject correlation effectively (Lin & Carroll, 2000).

- **Spline Methods**

Spline, which is a special function defined piecewisely by polynomials, play a central role in semiparametric and nonparametric modeling (Wand & Ormerod, 2008). Splines are often preferred than polynomials when doing interpolation, since they yield simpler results. There are several varieties of spline approaches, including smoothing spline (Wahba, 1990; Green & Silverman, 1994), regression spline (Eubank 1988, 1999), B-spine (De Boor, 1978; Dierckx, 1993), and penalized spline (P-spline) (Eilers & Marx, 1996). B-spline stands for basis spline, which is constructed from polynomial pieces and joined at the certain value of the knots (Eilers & Marx, 1996).

In regression spline smoothing, local neighborhoods are specified by a group of locations: $\tau_{0,}\tau_{1,}\tau_{2,...,}\tau_{K,}\tau_{K+1}$, where K is the number of knots. Then in the interval $[a,b]$, where $a=\tau_0<\tau_1<...<\tau_K<\tau_{K+1}=b$ **,** $\tau_{1,}\tau_{2,...,}\tau_K$ are called interior knots. These knots divide the interval into K subintervals (local neighborhoods). Within each subinterval, Taylor's expansion up to some degree can be applied. A regression spline can also be described as a piecewise polynomial within any two neighboring knots τ_r and τ_{r+1} for $r=0,1,...,K$. A regression spline can be constructed using the k -th degree truncated power basis with K knots $\tau_{1,}\tau_{2,...,}\tau_K$:

$1,t,...,t^k,(t-\tau_1)_+^k,...,(t-\tau_K)_+^k$, where $w_+^k=[w_+]^k$ denotes the power k of the positive part of w with $w_+=\max(0,w)$. For convenience, the truncated power basis is often denoted as

$$\Phi_p(t) = [1, t, ..., t^k, (t - \tau_1)_+^k, ..., (t - \tau_K)_+^k]^T, \qquad (2.1)$$

where $p = K + k + 1$ (Wu & Zhang, 2006).

- **Penalized Spline Method**

Good performance of regression splines strongly relies on the location of knots and the number of knots. To overcome this drawback, smoothing splines take all of the distinct time points as knots, and use a roughness penalty to control the smoothness of the estimated curve. However, smoothing splines are expensive to compute. When the number of distinct time points is large, the number of the parameters to be estimated is large. Moreover, we need to compute the roughness matrix for cubic smoothing splines or the roughness matrix based on a pre-specified basis for a general degree.

Eilers & Marx (1996) introduced the penalized spline (a combination of B-spline and difference penalty on the estimated coefficients), which overcomes these drawbacks by using the truncated power basis of some degree with a pre-specified number of knots that is much smaller than the total number of distinct design time points (Wu & Zhang, 2006). Also, a penalized spline avoids direct calculation of the roughness matrix by specifying a simple roughness matrix (a diagonal matrix), indicating that some coefficients of the basis functions are penalized and some are not (Ruppert et al., 2003).

Penalized spline is one of the most powerful smoothing techniques for uncorrelated or independent data, which has gained much popularity in the last decade. It has many attractive properties, including no boundary effects and straightforward extension of generalized linear regression models, and is relatively inexpensive in computation and cross validation compared to smoothing spline. Similar to the regression spline method, penalized spline uses the truncated power basis $\Phi_p(t)$ described in (2.1). Consider the following nonparametric population mean model:

$$y_{ij} = \eta(t_{ij}) + e_{ij}, \qquad (2.2)$$

Given the longitudinal dataset (y_{ij}, t_{ij}), $j = 1,2,...,n_i; i = 1,2,...,n$, n_i is the number of observations for the i-th subject, and n is the number of subjects. $\eta(t)$ is the smooth function over time, and e_{ij} is the error term. Using the penalized spline method, $\eta(t)$ can be approximately expressed as

$$\eta(t) = \Phi_p(t)^T \beta = \sum_{r=0}^{k} \beta_r t^\tau \sum_{l=1}^{K} \beta_{k+l}(t-\tau_l)_+^k \qquad (2.3)$$

where β is the coefficient vector that can be estimated by the following penalized least squares (PLS) criteria:

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij} - x_{ij}{}^T\beta)^2 + \lambda\beta^T G\beta,$$

where $x_{ij} = \Phi_p(t_{ij})$, λ represents the smoothing parameter, and G is the roughness matrix based on $\Phi_p(t)$ (Wu & Zhang, 2006).

In this thesis, penalized spline method will be used to estimate the BMI trajectory using semiparametric population mean model and semiparametric mixed effects model, and to evaluate the time-varying effect of externalizing behavior via nonparametric varying-coefficient models.

2.2 Semiparametric Models

To estimate the BMI trajectory in this thesis, both semiparametric population mean model and semiparametric mixed effects model will be considered. First, the time effects on BMI will be estimated using nonparametric methods, while the effects of externalizing behavior and other covariates will be modeled parametrically.

2.2.1 Semiparametric Population Mean Model

A semiparametric population mean model is given by

$$y_{ij} = f(t_{ij}) + w_{ij}{}^{T}\beta + \varepsilon_{ij}, \qquad j = 1, 2, ..., n_i, \; i = 1, 2, ..., n, \tag{2.4}$$

where y_{ij} is the BMI measured for subject i at t_{ij}, the effect of time on BMI is modeled by the smooth function $f(t_{ij})$, $w_{ij} = [w_{1ij}, ..., w_{rij}]^T$ is a vector of children's externalizing behavior T score and the other $r-1$ covariates (i.e., sex, race, maternal education, household poverty status, and maternal depression) for subject i observed at t_{ij}, $\beta = [\beta_1, ...\beta_r]^T$ represents the corresponding coefficient vector, and ε_{ij} is the error term.

Model (2.4) consists of two parts:

- Parametric component $w_{ij}{}^{T}\beta$
- Nonparametric component $f(t_{ij})$

Using the penalized spline method, $f(t)$ can be expressed as a regression spline $\Phi_p(t)^T\alpha$, $p = K + k + 1$, where $\alpha = [\alpha_1, ..., \alpha_p]^T$ is the associated coefficient vector. $\Phi_p(t)$ is a k-th degrees truncated power basis with K knots $\tau_{1,...,}\tau_K$ (described in (2.1)). Then the penalized spline smoother $\hat{\beta}$ and $\hat{\alpha}$ can be obtained by using the penalized least square criterion (Wu & Zhang, 2006):

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}(y_{ij}-w_{ij}{}^T\beta-x_{ij}{}^T\alpha)^2+\lambda\alpha^T G\alpha\ ,$$

where $x_{ij}=\Phi_p(t_{ij})$, λ is the smoothing parameter, and G is the roughness matrix for the penalized spline based on $\Phi_p(t)$. G is defined as $\begin{bmatrix} 0_{(k+1)\times(k+1)} & 0_{(k+1)\times K} \\ 0_{K\times(k+1)} & I_K \end{bmatrix}$, where I_K denotes the $K\times 1$ vector of one.

2.2.2 Semiparametric Mixed Effects Model

In the following model, we incorporate the subject specific effects by adding the random effect function of time

$$y_{ij}=w_{ij}{}^T\gamma+v_i(t_{ij})+\epsilon_{ij}\ ,\qquad j=1,2,...,n_i\ ,\ i=1,2,...,n\ ,\qquad (2.5)$$

where $w_{ij}=[w_{1ij},...,w_{rij}]^T$ is the same as that in (2.4), $\gamma=[\gamma_1,...,\gamma_r]^T$ represents the corresponding coefficients, $v_i(t_{ij})$ denotes a smooth process over time, and ϵ_{ij} is the error term.

Model (2.5) consists of two parts:

- Parametric component $w_{ij}{}^T\gamma$
- Nonparametric component $v_i(t_{ij})$

Let $z_{ij}=\Psi_q(t)^T$, $R_i=\sigma^2 I_N$, where $\Psi_q(t)$ is the a k_v-th degree truncated power basis with K_v knots $\delta_{1,...,}\delta_{K_v}$:

$$\Psi_q(t)=[1,t,...,t^{k_v},(t-\delta_1)_+^{k_v},...,(t-\delta_{K_v})_+^{k_v}]^T \qquad (2.6)$$

v_i(t)=$\Psi_q(t)^T b_i$, $b_i=[b_{i1},...,b_{iq}]^T$, $i=1,2,...,n$, $q=K_v+k_v+1$, $b_i\sim N(0,D_b)$, or in other words, the coefficient b_i is independently and identically distributed with mean 0 and covariance matrix D_b. Model (2.5) can be approximately expressed as a standard linear mixed effects model: $y_{ij}=w_{ij}{}^T\gamma+z_{ij}b_i+\epsilon_{ij}$, and the vector of measurement errors ϵ_i is assumed to be normal with covariance matrix R_i, then b_i and γ can be obtained by minimizing the following penalized generalized log-likelihood (PGLL) (Wu & Zhang, 2006):

$$\sum_{i=1}^{n}\{\Delta_i^T R_i^{-1}\Delta_i + b_i^T D_b^{-1} b_i\} + \lambda_v \sum_{i=1}^{n} b_i^T G_v b_i \, ,$$

where $\Delta_i = y_i - w_i^T\gamma - z_i^T b_i$, $w_i = [w_{i1,\dots,}w_{in_i}]^T$, $z_i = [z_{i1,\dots,}z_{in_i}]^T$, and the roughness matrix G_v is defined as: $G_v = \begin{bmatrix} 0_{(k_v+1)\times(k_v+1)} & 0_{(k_v+1)\times K_v} \\ 0_{K_v\times(k_v+1)} & I_{K_v} \end{bmatrix}$.

2.3 Nonparametric Time-varying Coefficient Models

Time-varying coefficient models are a class of structural nonparametric models which are particularly useful in longitudinal analyses. West et al. (1985) first introduced a “dynamic generalized linear” model, known as the time-varying coefficient (TVC) model. Time-varying coefficient models are a special case of the general varying coefficient models introduced by Hastie & Tibshirani (1993). Research on TVC models to longitudinal data analysis have been conducted by Brumback & Rice (1998), Hoover et al. (1998), Wu et al. (1998), Fan & Zhang (1998), and among others. The nonparametric time-varying coefficient model can be expressed as:

$$y_i(t) = x_i(t)^T \beta(t) + e_i(t)$$

$j = 1,2,\dots,n_i$, $i = 1,2,\dots,n$, with n_i denoting the number of measurements of the i-th subject, and n denoting the number of subject. Let $y_{ij} = y_i(t_{ij}), e_{ij} = e_i(t_{ij})$, where t_{ij} denotes the time of the j-th measurement of the i-th subject, then the above model can be re-written as a discrete version:

$$y_{ij} = x_i(t_{ij})^T \beta(t_{ij}) + e_{ij}$$

for $j = 1, 2, ..., n_i$, $i = 1, 2, ..., n$.

To account for the within-subject correlation, one can apply parametric or nonparametric techniques. When the parametric models do not fit well, nonparametric techniques can be used. Research has been done to model the within-subject correlation nonparametrically (Shi et al., 1996; Rice & Wu, 2001; Wu & Zhang, 2002; Liang, Wu & Carroll, 2003).

In this thesis, nonparametric time-varying coefficient population mean model and mixed effects model will be proposed to estimate the effects of childhood externalizing behavior on BMI.

2.3.1 Nonparametric Time-varying Coefficient Population Mean Model

The following model estimates the effects of predictors by including the time-varying coefficient function

$$y_{ij} = x_i(t_{ij})^T \lambda(t_{ij}) + e_{ij} \qquad j = 1, 2, ..., n_i, \; i = 1, 2, ..., n, \tag{2.7}$$

where t_{ij} denote the time when the j-th measurement of the i-th subject was recorded, $j = 1, 2, ..., n_i$ with n_i denoting the number of measurements of the i-th subject, $\lambda(t_{ij})$ denotes the coefficient function at time t_{ij}, and $x_i(t_{ij})$ denotes the time dependent covariate vector.

Penalized spline method will be used to express the coefficient function $\lambda_r(t)$, $r = 0, 1, ..., d$, where d is the number of basis function, with regression splines using the same truncated power basis $\Phi_p(t)^T$ (as described in (2.1)) and penalize the highest order derivative jumps of the regression splines, $p = K + k + 1$. Then $\lambda_r(t)$ can be approximately expressed as regression spline $\lambda_r(t) = \Phi_p(t)^T a_r, r = 0,1,...,d$, $a_r = [a_{r1,...,}a_{rp}]^T$.

Let G be the roughness matrix that associated with $\Phi_p(t)$ to penalize the k -th time derivative jumps, $G = \begin{bmatrix} 0_{(k+1)\times(k+1)} & 0_{(k+1)\times K} \\ 0_{K\times(k+1)} & I_K \end{bmatrix}$. α_r can be obtained by using the penalized weighted least square (PWLS) criterion (Wu & Zhang, 2006):

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_i(y_{ij} - \sum_{r=0}^{d} x_{rij}^T\alpha_r)^2 + \sum_{r=0}^{d}\lambda_r\alpha_r^T G\alpha_r ,$$

where $x_{rij} = x_{ri(t_{ij})}\Phi_p(t_{ij})$, $x_{ri(t)}$ denotes the r -th entry of $x_i(t)$, and λ_r , $r = 0,1,...,d$, is the smoothing parameter.

2.3.2 Nonparametric Time-varying Coefficient Mixed Effects Model

In the following model, we incorporate the subject specific effects by adding the random effect function of time

$$y_{ij} = x_i(t_{ij})^T \varphi(t_{ij}) + z_i(t_{ij})^T r_i(t_{ij}) + \epsilon_i(t_{ij}) \qquad j = 1,2,...,n_i, i = 1,2,...,n \quad (2.8)$$

Let $x_i(t_{ij}) = [x_{0i}(t_{ij}), x_{1i}(t_{ij})...x_{di}(t_{ij})]^T$ denote the fixed effect covariate vector. $\varphi(t_{ij}) = [\varphi_0(t_{ij}),...,\varphi_d(t_{ij})]^T$ is the fixed effect coefficient function, $z_i(t_{ij})^T r_i(t_{ij})$ represents the deviation from the population mean function, or random effect component, where $z_i(t_{ij}) = [z_{0i}(t_{ij}), z_{1i}(t_{ij})...z_{d*i}(t_{ij})]^T$ is a d^*-dimensional covariate vector, and $r_i(t_{ij}) = [r_{0i}(t_{ij}), r_{1i}(t_{ij})...r_{d*i}(t_{ij})]^T$ denotes the random effect coefficient function. When $d^* = d$, the fixed effect covariates $x_i(t_{ij})$ and random effect covariates $z_i(t)$ are the same. Let $\Phi_p(t)$ be the truncated power basis of degree k with K knots $\tau_{1,...,}\tau_K$, $p = K + k + 1$ and $\Psi_q(t)$ be the truncated power basis of degree k_v with K_v knots $\delta_{1,...,}\delta_{K_v}$, $q = K_v + k_v + 1$ (as described in (2.6)). Then the fixed effect coefficient function $\varphi(t)$ and the random effect coefficient function $r_i(t)$ can be expressed as regression splines:

$\varphi_m(t) = \Phi_p(t)^T \alpha_m$, where $m = 0,1,...,d$, $\alpha_m = [\alpha_{m1},...,\alpha_{mp}]^T$;

$v_{si}(t) = \Psi_q(t)^T b_{si}$, where $s = 0,1,...,d^*$, $b_{si} = [b_{si1,...,} b_{siq}]^T$.

Let $\Phi(t)$ and $\Psi(t)$ be two block diagonal matrices with $(d+1)$ blocks of $\Phi_p(t)$ and (d^*+1) blocks of $\Psi_q(t)$, respectively. Let $\alpha = [\alpha_0^T, \alpha_1^T,...,\alpha_d^T]^T$, $b_i = [b_{0i}^T, b_{1i}^T,...,b_{d^*i}^T]^T$, then the coefficient vector can be denoted as $\varphi(t) = \Phi(t)^T \alpha$ and $v_i(t) = \Psi(t)^T b_i$, $i = 1,2,...,n$, and the model in (2.8) can be approximately expressed as:

$$y_{ij} = x_{ij}^T \alpha + z_{ij}^T b_i + \epsilon_{ij}, \quad j = 1,2,...,n_i, \quad i = 1,2,...,n. \tag{2.9}$$

Let $X_i = [x_{i1},...,x_{in_i}]^T$ and $Z_i = [z_{i1},...,z_{in_i}]^T$, then the model in (2.9) can be further written as:

$$y_i = X_i \alpha + Z_i b_i + \epsilon_i, \quad i = 1,2,...,n. \tag{2.10}$$

The estimate of α and b_i can be obtained by using the following penalized generalized log-likelihood criterion (Wu & Zhang, 2006):

$$\sum_{i=1}^{n} \{(y_i - X_i\alpha - Z_i b_i)^T R_i^{-1} (y_i - X_i\alpha - Z_i b_i) + b_i^T D^{-1} b_i\} +$$
$$\sum_{i=1}^{n} \sum_{s=0}^{d^*} \lambda_{sv} b_{si}^T G_v b_{si} + \sum_{r=0}^{d} \lambda_m \alpha_m^T G \alpha_m,$$

where G and G_v are two P-spline roughness matrices associated with $\Phi_p(t)^T$ and $\Psi_q(t)^T$, $G = diag(0,...,0,1,...,1)$, a $p \times p$ diagonal matrix with the last K diagonal entries being 1, and $G_v = diag(0,...,0,1,...,1)$, a $q \times q$ diagonal matrix with the last K_v K diagonal entries being 1. The smoothing parameter λ_m, $m = 0,1,...,d$ and λ_{sv}, $s = 0,1,...,d^*$ can be used to trade off the goodness of fit with the roughness of the regression splines.

Chapter 3 Results

The four proposed models (2.4), (2.5), (2.7), and (2.8) are fitted to the dataset of the NICHD SECCYD using Matlab R2007b. We fit these four models using the penalized spline method described in Section 2.1, and also stratify each model by gender.

There are 1,364 subjects included in the dataset for analysis. Due to missing data, there are 1,103 subjects with the full observed information of baseline covariates (e.g., race, sex, maternal education, maternal depression, and household poverty level at 24 months), partial or complete records of BMI, externalizing behavior, and internalizing behavior at seven time points. In total, 563 boys and 540 girls have been considered in the analysis.

We first fit the model (2.4) among boys and girls using semiparametric population mean model.

$$\begin{aligned} BMI_{ij} = f(t_{ij}) + \beta_1 Ex_{i24} + \beta_2 In_{i24} + \beta_3 White_i + \beta_4 Depre_i + \\ \beta_5 Poverty_i + \beta_6 Edu_{1i} + \beta_7 Edu_{2i} + \epsilon_{ij} \end{aligned} \qquad (3.1)$$

$j = 1, 2, ..., n_i$, $i = 1, 2, ..., n$, where BMI_{ij} denotes the BMI for subject i at time t_{ij}. The baseline covariates for subject i included in the model are externalizing behavior, internalizing behavior, race, maternal depression, household poverty status, and maternal education measured at 24 months. The covariates Edu_1 and Edu_2 are indicators for whether or not maternal education level is "<=high school graduate" or "some college". The coefficients $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ and β_7 are used to model the constant effects of these covariates on the BMI trajectory. $f(t_{ij})$ is a nonparametric function which is used to model the age effect on BMI trajectory.

The error e_{ij} follows an independent and identical normal distribution $N(0,\sigma^2)$. The number of knots for the quadratic truncated power basis is taken as K=3, and the knots are scattered using the "equally spaced sample quintiles as knots" method. Under the generalized cross-validation rule (Wahba, 1977; Craven & Wahba, 1979), the smoothing parameter λ is 13.269 and 13.307 for boys and girls, respectively.

The covariate estimates for model (3.1) fitted among boys and girls are described in Table 2 and Table 3. Except for household poverty status, all the other covariates are significantly associated with BMI trajectory for boys. Externalizing behavior is significantly associated with BMI trajectories for both boys and girls. But the direction of association is different. For boys, BMI decreases 0.02 kg/m^2 per unit increase in externalizing behavior T score controlling for other covariates. For girls, BMI increases 0.03 kg/m^2 per unit increase in externalizing behavior T score controlling for other covariates. The BMI for White boys is 0.7 kg/m^2 lower than that among non-White boys controlling for other covariates. The BMI of children whose mothers have depression are 0.67 kg/m^2 higher than that of children whose mothers do not after controlling for other covariates. Among girls, internalizing behavior, race, maternal depression are not significantly associated with BMI trajectory. High level of maternal education is positively associated with BMI trajectory for both girls and boys. On average, boys have higher BMI trajectory than girls (Figure 1), and the BMI trajectories are quite similar to that in Anderson et al. (2010), which were estimated by applying cubic function of age.

We fit the model (2.5) using semiparametric mixed effects model among boys and girls. A random effect function has been included in the model to incorporate the time-varying subject specific effects in the following model:

$$\begin{aligned} BMI_{ij} &= \beta_0(t_{ij}) + \beta_1 Ex_{24i} + \beta_2 In_{24i} + \beta_3 White_i + \beta_4 Depre_i + \\ &\beta_5 Poverty_i + \beta_6 Edu_{1i} + \beta_7 Edu_{2i} + v_i(t_{ij}) + \epsilon_{ij} \end{aligned} \tag{3.2}$$

$j = 1, 2, ..., n_i$, $i = 1, 2, ..., n$.

The nonparametric random component $v_i(t_{ij})$ denotes a Gaussian process with mean zero and covariate function $\gamma(s,t)$. The smoothing parameter λ_v is used to control the roughness of the random effect functions $v_i(t_{ij})$.

Table 4 and Table 5 give the parametric component estimation results of model (3.2) for boys and girls separately. Still, boys have a higher average BMI trajectory than girls (Figure 2). The quadratic polynomial model for the intercept BMI trajectory among boys is fitted as $\beta_0(t) = 18.393 - 1.576t + 0.312t^2$. Externalizing behavior is not significantly associated with BMI trajectory among boys. Race and Maternal depression are significantly associated with BMI trajectory among boys, but are not significant among girls. The quadratic polynomial model for the intercept BMI trajectory among girls is fitted as $\beta_0(t) = 17.305 - 1.274t + 0.271t^2$. Among girls, externalizing behavior, household poverty status, and higher level of education are significantly associated with BMI trajectory, which are consistent with model (3.1).

Comparing models (3.1) to (3.2), we can notice that the significant predictors for BMI trajectory are quite consistent for girls, but are not that consistent for boys. Model (3.1) identifies more significant predictors than model (3.2) for boys. In models (3.1) and (3.2), we assume that the effects of all the covariates are constant over time, which may not provide the best fit. Thus, in models (3.3) and (3.4) the covariate effects will be treated as time-varying functions.

We fit model (2.7) using nonparametric time-varying coefficient population mean model among boys and girls. Externalizing behavior and internalizing behavior are treated as time-dependent covariates, which are repeatedly measured at seven time points. All the covariate effects are modeled as nonparametric time-varying coefficient functions in the following model:

$$\begin{aligned} BMI_{ij} &= f(t_{ij}) + \beta_1(t_{ij}) Ex_i(t_{ij}) + \beta_2(t_{ij}) In_i(t_{ij}) + \beta_3(t_{ij}) White_i + \\ &\beta_4(t_{ij}) Poverty_i + \beta_5(t_{ij}) Depre_i + \beta_6(t_{ij}) Edu_{1i} + \beta_7(t_{ij}) Edu_{2i} + \epsilon_{ij} \end{aligned} \tag{3.3}$$

for $j = 1,2,...,n_i$, $i = 1,2,...,n$.

Figure 3 and Figure 4 show the fitted coefficient functions (solid curves) among boys and girls, together with the 95% pointwise SD bands (dashed curves). The BMI trajectories for boys and girls are consistent with those in models (3.1) and (3.2). Externalizing behavior is significantly associated with BMI trajectory of both boys and girls. However, the coefficient function plot of externalizing behavior is constant over time for girls, which means that the effects of externalizing behavior on BMI trajectory for girls are not changing as they grow older. Therefore, it is reasonable to replace the coefficient function by a constant coefficient parameter. Furthermore, the effects of internalizing behavior, household poverty and higher level of education are significant for both genders, and they tend to be stronger overtime.

Finally, we fit model (2.8) using time-varying coefficient mixed effects model among boys and girls to incorporate the within subject correlation.

$$\begin{aligned} &BMI_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij})Ex_i(t_{ij}) + \beta_2(t_{ij})In_i(t_{ij}) + \beta_3(t_{ij})White_i \\ &+\beta_4(t_{ij})Poverty_i + \beta_5(t_{ij})Depre_i + \beta_6(t_{ij})Edu_{1i} + \beta_7(t_{ij})Edu_{2i} + v_i(t_{ij}) + \epsilon_i(t_{ij}) \end{aligned} \tag{3.4}$$

for $j = 1, 2, ..., n_i$, $i = 1, 2, ..., n$.

In the above model, $v_i(t_{ij})$ denotes a nonparametric random effect function, which is similar as that in model (3.2). The error $\epsilon_i = [\epsilon_{i1}, ..., \epsilon_{in_i}]^T$ follows the normal distribution *N* (0, **Ri**). Figure 5 and Figure 6 give the fitted coefficient functions (solid curves), together with the 95% pointwise SD bands (dashed curves) for boys and girls. It is clear that the 95% pointwise SD bands under model (3.4) are generally wider than those 95% pointwise SD bands under model (3.3), since model (3.4) accounts for the within-subject correlations while model (3.3) does not. The effects of internalizing behavior on BMI trajectory are no longer significant for girls. Similarly to model (3.3), the effects of race on BMI trajectory become weaker over time for both boys and girls, and the effects of maternal depression become stronger for boys as they grow older. Furthermore, the effects of high level education become stronger over time on BMI trajectory for girls.

Chapter 4 Discussion

Very few studies of non-clinical populations have examined the relationship between obesity and externalizing behavior in preadolescent children. The main aim of this research is to re-examine this relationship by applying advanced semiparametric and nonparametric methods. In this thesis, two semiparametric and two nonparametric time-varying coefficient models have been fitted among boys and girls to analyze the effect of childhood externalizing behavior on BMI and other relevant covariates on BMI trajectory, and to evaluate how those effects are changing over time, which are not able to be evaluated using the parametric logistical regression or linear mixed-effects models proposed by Anderson et al. (2010).

Compared to the population mean models, mixed effects models are most useful when the research objective is to make inferences about individuals rather than the study population (Fitzmaurice et al., 2004). For both semiparametric population mean models and mixed effects models (models (3.1) and (3.2)), the significant covariates were consistent for girls, which included externalizing behavior, maternal depression, and high level of maternal education. However, there exists inconsistency for the significant predictors among boys, while only race was significantly associated with BMI trajectory for boys in both models.

In models (3.1) and (3.2), it was assumed that all the covariate effects are polynomials of lower degrees and was constant over time, which might not provide the best fit to the data. The BMI trajectories from models (3.1) and (3.2) are similar to those in Anderson et al. (2010), which validated that the linear mixed effects model used in that study was an appropriate method to estimate BMI trajectory. In models (3.3) and (3.4), we removed the polynomial model assumption, and considered nonparametric time-dependent covariate effects. Compared to models (3.1) and (3.2), models (3.3) and (3.4) not only identified consistent significant predictors, but also provided more information on how the effects are changing over time. The effects of externalizing behavior on BMI for boys were decreasing, and were relatively constant for girls, while the effects of internalizing behavior were increasing over time for both boys and girls in model (3.3). Our results also illustrated that in models (3.3) and (3.4), the effects of race on BMI trajectory were declining overtime for both boys and girls, the effects of maternal depression were increasing overtime for boys, and the effects of household poverty on BMI trajectory were increasing overtime for girls. From this point of view, models (3.3) and (3.4) provide us with more insights about the changing effects of those covariates on BMI trajectory than models (3.1) and (3.2).

4.1 Strengths

Nonparametric time-varying coefficient models are very helpful to evaluate how the strength of covariate effects is changing over time, which cannot be done in parametric models. By using time-varying coefficient models, we are able to identify the decreasing effects of externalizing behavior and race on BMI trajectory among boys, and the increasing effects of household poverty status and high level of education on BMI trajectory among girls.

By applying nonparametric function on age, we are able to estimate the BMI trajectory nonparametrically, whose functional form is not pre-specified, but is determined from the data. By doing that, we can avoid model misspecification.

This thesis considered semiparametric and nonparametric mixed effects models, which account for the time-varying subject specific effect by adding a random effect function, while in the generalized linear mixed effects models, the subject specific effect is assumed to be constant over time. We compare the aims and components of these four models in Table 6.

4.2 Limitations

There are also some limitations for this research. Firstly, the study relied upon mother's reports of their child's behavior. The CBCL is not a diagnostic instrument and the study does not take into account the extent to which externalizing behaviors were problematic for the mother or the child. It is possible that the ratings of children's behavior are influenced by the child's weight status. Secondly, the study is not able to adjust for parental obesity, which is strongly related to child obesity and may be related to child behavior problems or maternal reports of child behavior problems. The third limitation is the lack of model checking techniques to determine which model provides the best fit to the data.

4.3 Conclusion Remarks

Nonparametric varying coefficient models are very helpful to evaluate how the strength of covariate effects is changing over time, which cannot be done in parametric models. Sometimes, it will be misleading to model the covariate effects as constant, especially for factors that in nature may change over time.

Nonparametric models are more flexible than parametric models in terms of the functional form, and it is useful to apply nonparametric models first to identify the pattern of the covariate effects of interest.

Appendices

Table 1 Summary of variables in Anderson et al. (2010)

Variables	**Scale**
Externalizing behavior	Continuous T score
Internalizing behavior	Continuous T score
Race	1-White, 0-non White
Sex	1-boys, 0-girls
Maternal depression	1-depression, 0-don't have depression
Household poverty status	1-below poverty threshold, 0-above poverty threshold
Maternal education1 (<= high school)	1-yes, 0-no (whether mother has lower or equal to high school education)
Maternal education2 (some college)	1-yes, 0-no (whether mother has some college education)

Table 2 Semiparametric population mean model (3.1) fitted among boys: estimated covariate effects, standard deviations, approximate z-test values, P values for the parametric component

(Quadratic truncated power basis with K = 3 knots was used. The smoothing parameter selected by GCV is 13.269.)

Covariate	**Effect**	**SE**	**z-test value**	**P value**
Ex. behavior	-0.0218	0.0077	-2.8281	0.0047
In. behavior	0.0316	0.0074	4.2691	0.0001
Race (White vs Non-White)	-0.6991	0.1664	-4.201	0.0001
Maternal depression	0.6687	0.1619	4.1309	0.0001
Household poverty	0.2770	0.1954	1.4178	0.1562
Maternal education (High school)	0.5673	0.1567	3.6219	0.0003
Maternal education (Some college)	0.4192	0.1421	2.951	0.0032

Table 3 Semiparametric population mean model (3.1) fitted among girls: estimated covariate effects, standard deviations, approximate z-test values, P values for the parametric component

(Quadratic truncated power basis with K = 3 knots was used. The smoothing parameter selected by GCV is 13.307.)

Covariate	Effect	SE	z-test value	P value
Ex. behavior	0.0283	0.0072	3.9207	0.0001
In. behavior	0.0098	0.0074	1.3284	0.184
Race (White vs Non-White)	-0.1848	0.1518	-1.217	0.2236
Maternal depression	-0.2096	0.1526	-1.3741	0.1694
Household poverty	0.9011	0.1771	5.0871	0.0001
Maternal education (High school)	0.0148	0.1546	0.0958	0.9237
Maternal education (Some college)	0.3196	0.1234	2.591	0.0096

Table 4 Semiparametric mixed effects model (3.2) fitted to dataset among boys: estimated covariate effects, standard deviations, approximate z-test values, P values for the parametric component

(Quadratic truncated power basis with K = 3 knots was used. The smoothing parameter selected by GCV is 1.576.)

Covariate	Coef.	Effect	SE	z-test value	P value
Intercept	β_{00}	18.393	0.1677	109.7	0.0001
	β_{01}	-1.576	0.099	-15.926	0.0001
	β_{02}	0.3116	0.0156	19.962	0.0001
Ex. behavior	β_1	0.0072	0.0053	1.356	0.1751
In. behavior	β_2	0.0008	0.005	0.1556	0.8763
White (1-white, 0-black and other)	β_3	-0.4181	0.1037	-4.0299	0.0001
Maternal depression	β_4	0.2422	0.102	2.3755	0.0175
Poverty status	β_5	0.1684	0.1208	1.3936	0.1634
Maternal education (High school)	β_{61}	0.1608	0.0966	1.6642	0.0961
Maternal education (Some college)	β_{62}	0.1263	0.0882	1.4322	0.1521

Table 5 Semiparametric mixed effects model (3.2) fitted to dataset among girls: estimated covariate effects, standard deviations, approximate z-test values, P values for the parametric component

(Quadratic truncated power basis with K = 3 knots was used. The smoothing parameter selected by GCV is 1694.7.)

Covariate	**Coef.**	**Effect**	**SE**	**z-test value**	**P value**
Intercept	β_{00}	17.305	0.1558	111.06	0.0001
	β_{01}	-1.2744	0.089	-14.315	0.0001
	β_{02}	0.2705	0.0137	19.762	0.0001
Ex. behavior	β_{1}	0.0227	0.0051	4.4952	0.0001
In. behavior	β_{2}	-0.0042	0.0049	-0.8441	0.3986
White (1-white, 0-black and other)	β_{3}	0.0618	0.0983	0.6285	0.5297
Maternal depression	β_{4}	-0.0344	0.0994	-0.3457	0.7296
Poverty status	β_{5}	0.4406	0.116	3.7969	0.0001
Maternal education (High school)	β_{61}	0.0218	0.1011	0.2158	0.8291
Maternal education (Some college)	β_{62}	0.1708	0.0795	2.148	0.0317

Table 6 Comparison of the four proposed models

Models		Aims	Components
3.1	**Semiparametric population mean model** $BMI_{ij} = f(t_{ij}) + \beta_1 Ex_{i24} + \beta_2 In_{i24} + \beta_3 White_i + \beta_4 Depre_i + \beta_5 Poverty_i + \beta_6 Edu_{1i} + \beta_7 Edu_{2i} + \epsilon_{ij}$ $j = 1,2,...,n_i$, $i = 1,2,...,n$	Model the BMI trajectory from 2-12 years of age nonparametrically; Evaluate the cross-sectional association of externalizing behavior at 24 months and BMI trajectory.	**Parametric**: Externalizing, internalizing behavior at 24 months and all the baseline covariates. **Nonparametric:** Effects of age on BMI is modeled as: $f(t_{ij})$
3.2	**Semiparametric mixed effects model** $BMI_{ij} = \beta_0(t_{ij}) + \beta_1 Ex_{24i} + \beta_2 In_{24i} + \beta_3 White_i + \beta_4 Depre_i + \beta_5 Poverty_i + \beta_6 Edu_{1i} + \beta_7 Edu_{2i} + v_i(t_{ij}) + \epsilon_{ij}$ $j = 1,2,...,n_i$, $i = 1,2,...,n$	Same with model 3.1, but also able to evaluate the changing subject-specific effects nonparametrically by adding the nonparametric random effect function.	**Parametric**: Externalizing, internalizing behavior at 24 months and all the baseline covariates; Effects of age on BMI is modeled as pre-specified form. **Nonparametric**: Random effect is modeled as: $v_i(t_{ij})$
3.3	**Nonparametric time-varying coefficient population mean model** $BMI_{ij} = f(t_{ij}) + \beta_1(t_{ij}) Ex_i(t_{ij}) + \beta_2(t_{ij}) In_i(t_{ij}) + \beta_3(t_{ij}) White_i + \beta_4(t_{ij}) Poverty_i + \beta_5(t_{ij}) Depre_i + \beta_6(t_{ij}) Edu_{1i} + \beta_7(t_{ij}) Edu_{2i} + \epsilon_{ij}$ $j = 1,2,...,n_i$, $i = 1,2,...,n$	Assess the time-varying effects of externalizing, internalizing behavior and other covariates on BMI trajectory.	**Parametric**: The effects of all predictors are modeled as time-varying coefficient functions with pre-specified forms. **Nonparametric:** Effects of age on BMI is modeled as: $f(t_{ij})$
3.4	**Nonparametric time-varying coefficient mixed effects model** $BMI_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij}) Ex_i(t_{ij}) + \beta_2(t_{ij}) In_i(t_{ij}) + \beta_3(t_{ij}) White_i + \beta_4(t_{ij}) Poverty_i + \beta_5(t_{ij}) Depre_i + \beta_6(t_{ij}) Edu_{1i} + \beta_7(t_{ij}) Edu_{2i} + v_i(t_{ij}) + \epsilon_i(t_{ij})$ $j = 1,2,...,n_i$, $i = 1,2,...,n$	Same with model 3.3, but also able to evaluate the changing subject-specific effects nonparametrically by adding the nonparametric random effect function.	**Parametric**: The effects of all predictors are modeled as nonparametric time-varying coefficient functions with pre-specified forms. **Nonparametric:** Random effect function is modeled as: $v_i(t_{ij})$

Figure 1 Semiparametric population mean model (3.1) fitted to boys and girls: estimated BMI trajectory (solid curves) with 95% pointwise SD bands (dashed curves)

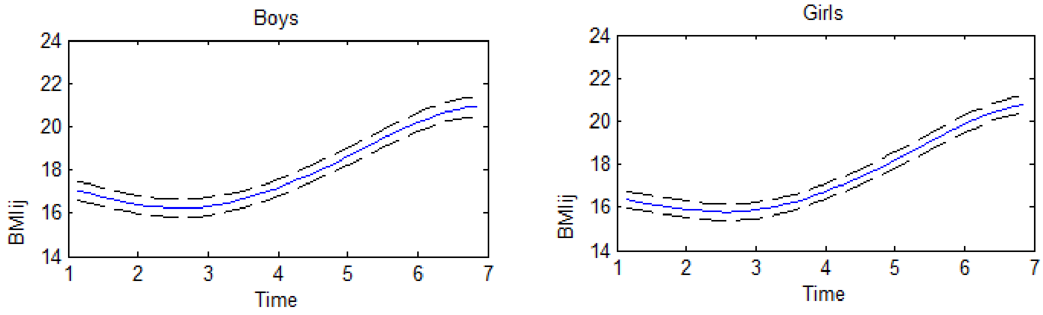


Figure 2 Semiparametric mixedeffects model (3.2) fitted to boys and girls: fitted coefficient functions (solid curves) and their 95% pointwise SD bands (dashed curves)

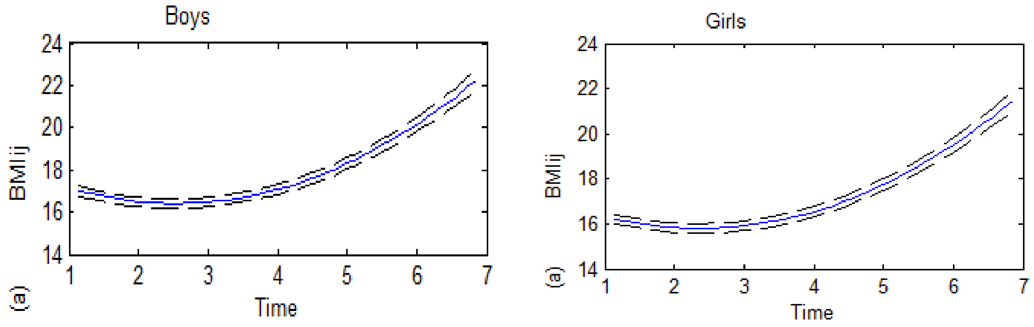


Figure 3 Time-varying coefficient nonparametric population mean model (3.3) for boys: estimated coefficient functions (solid curves) and their 95% pointwise SD bands (dashed curves)

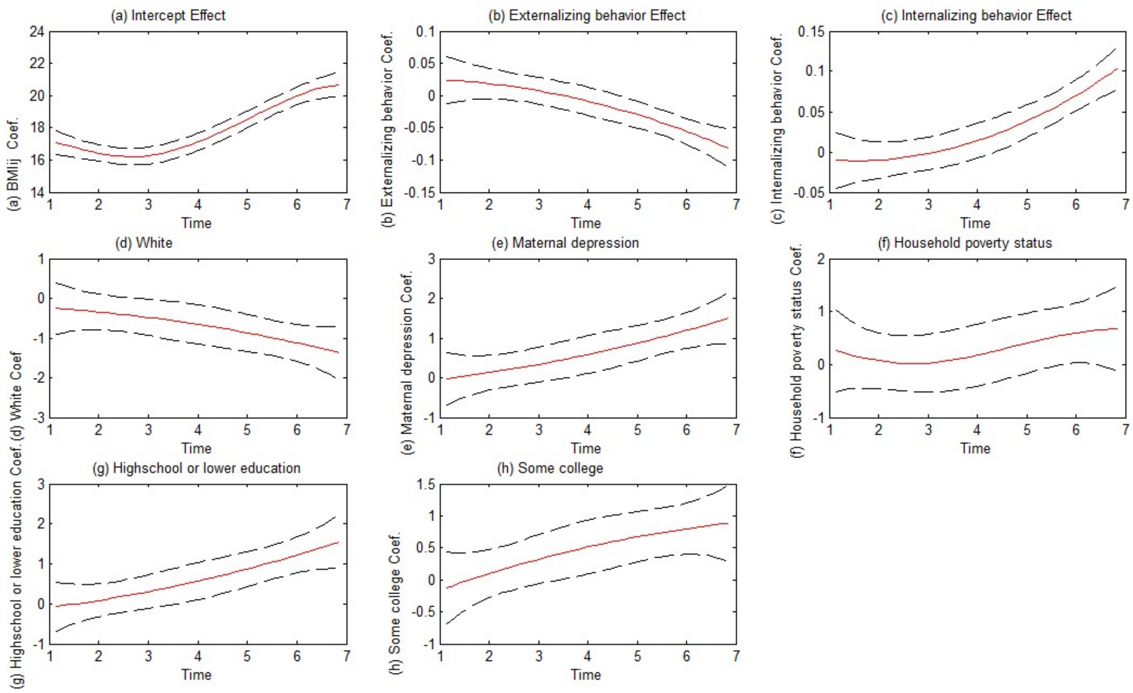


Figure 4 Time-varying coefficient nonparametric population mean model (3.3) girls: estimated coefficient functions (solid curves) and their 95% pointwise SD bands (dashed curves)

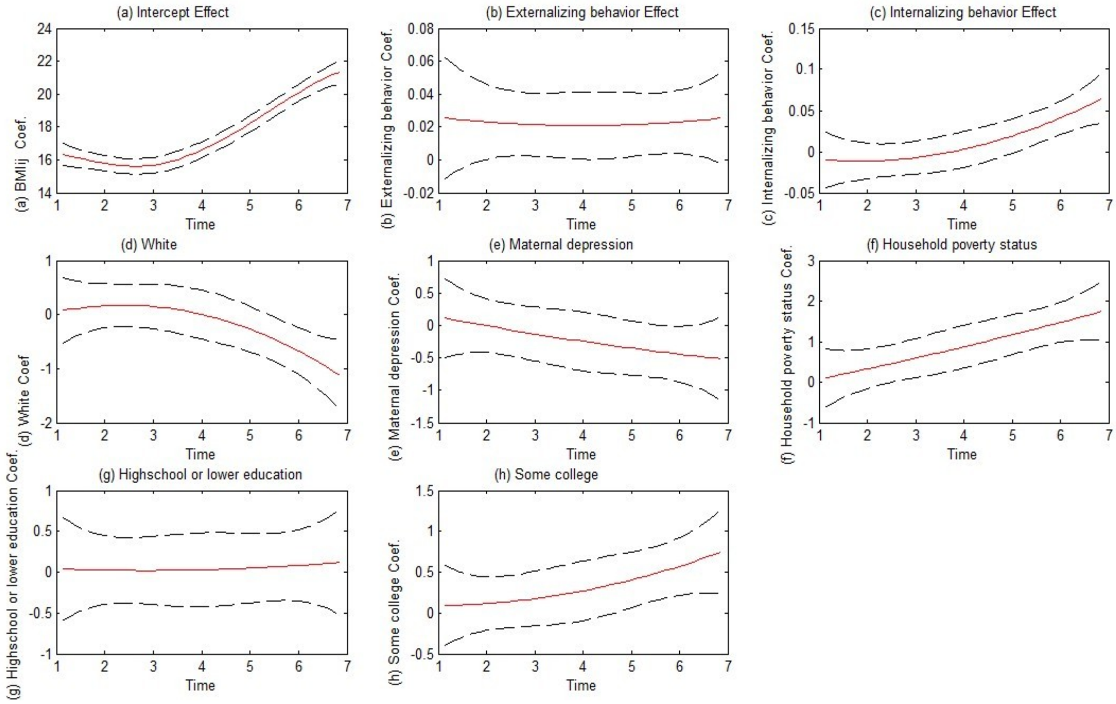


Figure 5 Time-varying coefficient nonparametric mixed effects model (3.4) for boys: fitted coefficient functions (solid curves) and their 95% pointwise SD bands (dashed curves)

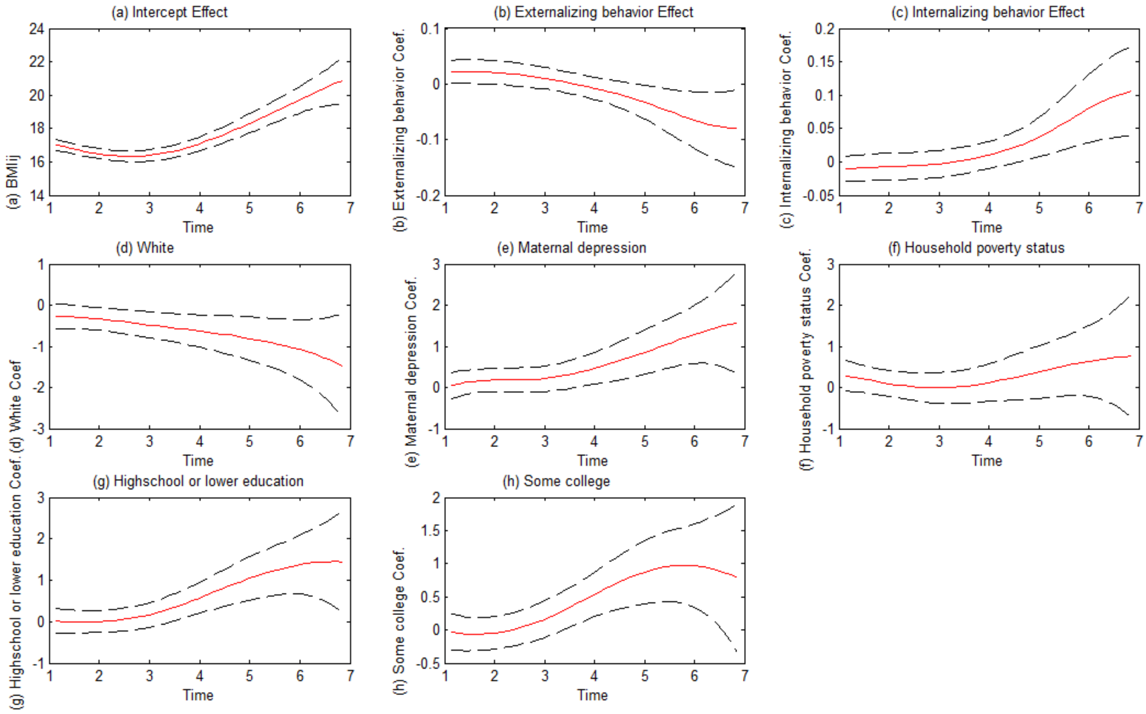
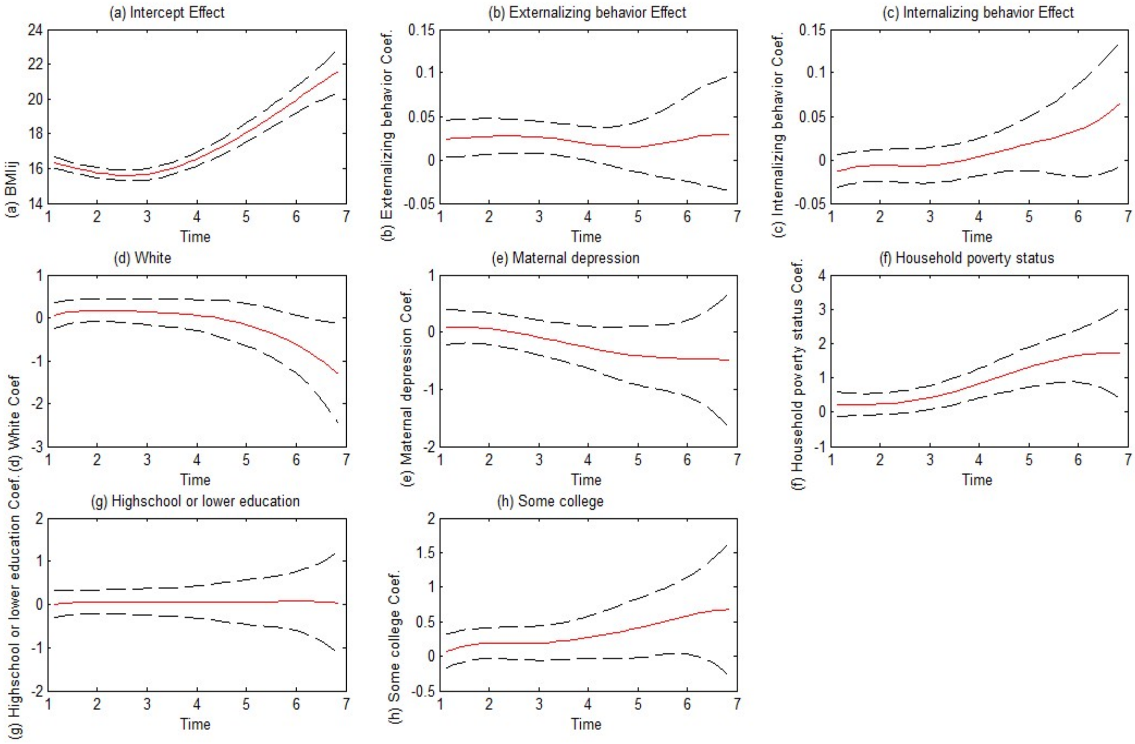


Figure 6 Time-varying coefficient nonparametric mixed effects model (3.4) for girls: fitted coefficient functions (solid curves) and their 95% pointwise SD bands (dashed curves)



Bibliography

Achenbach T.M. & Edelbrock C. (1981). Behavioral problems and competencies reported by parents of normal and disturbed children aged four through sixteen. Monographs of the Society for Research in Child Development 46. Serial no.188.

Achenbach T.M. & Edelbrock C. (1983). Manual for the Child Behavior Checklist and Revised Child Behavior Profile. Burlington: University of Vermont, Department of Psychiatry.

Achenbach T.M. (1992). Manual for the Child Behavior Checklist/2-3 and 1992 Profile. Burlington: University of Vermont, Department of Psychiatry.

Achenbach T.M. (2000). Assessment of Psychopathology. In Handbook of Developmental Psychopathology Second edition. New York: Kluwer Academic/Plenum Publishers.

Achenbach T.M. (2000). The Child Behavior Checklist and Related Forms for Assessing Behavioral/Emotional Problems and Competencies, *Pediatrics in Review,* Vol. 21, No. 1.

Akaike H. (1973). Information Theory and an Extension of the Entropy Maximization Principle. In Petrov B.N. & Csak F. (Eds.), *2nd International Symposium on Information Theory*, pp. 267-281.

Anderson S.E., Cohen P., Naumova E.N. & Must A. (2006). Relationship of childhood behavior disorders to weight gain from childhood into adulthood**.** *Ambulatory Pediatrics,* 6, 297-301.

Anderson S.E., He X., Schoppe-Sullivan S. & Must A. (2010). Externalizing behavior in early childhood and body mass index from age 2 to 12 years: longitudinal analyses of a prospective cohort study. *BMC Pediatrics*, 10, 49.

Bradley R.H., Houts R., Nader P.R., O'Brien M., Belsky J. & Crosnoe R. (2008). The relationship between body mass index and behavior in children**.** *Journal of Pediatric,* 153, 629-634.

Bollen E. & Wojciechowski F.L. (2004). Anorexia nervosa subtypes and the big five personality factors. *European Eating Disorders Review*, 12, 117-21.

Brumback B. & Rice J.A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of American Statistical Association*, 93, 961-994.

Chiou J.M., Muller H.G. & Wang J.L. (2002). Functional quasi-likelihood regression models with smooth random effects. *Journal of Royal Statistical Society*, Series B, 65, 405-423.

Chiang C.T., Rice J.A., & Wu C.O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of American Statistical Association,* 96, 605-619.

Cheng S.C. & Wei L.J. (2000). lnferences for semiparametric model with panel data. *Biometrika,* 87, 89-97.

Craven P. & Wahba G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik,* 31, 377-390.

Datar A. & Sturm R. (2004). Childhood overweight and parent and teacher reported behavior problems: evidence from a prospective study of kindergarteners. *Archives of Pediatrics & Adolescent Medicin,* 158, 804-810.

Datar A. & Sturm R. (2006). Childhood overweight and elementary school outcomes**.** *International Journal of Obesity,* 30, 1449-1460.

Deater-Deckard K., Dodge K.A., Bates J.E. & Pettit G.S. (1998). Multiple risk factors in the development of externalizing behavior problems: group and individual differences. *Development and Psychopathology*, 10, 469-493.

De Boor C. (1978). A Practical Guide to Splines. New York: Springer-Verlag.

Dierckx P. (1993). Curve and Surface Fitting With Splines. Oxford: Clarendon Press.

Durban M., Harezlak J., Wand M. P. & Carroll L. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine,* 24, 1153-1167.

Drukker M., Wojciechowski F., Feron J.M., Mengelers R. & Van Os. J. (2009). A community study of psychosocial functioning and weight in young children and adolescents**.** *International Journal of Pediatric Obesity*, 4, 91-97.

Eilers P.H. & Marx B.D. (1996). Flexible smoothing with B-splines and penalities (with discussion). *Statistical Science,* 11, 89-121.

Eubank R.L. (1988). Nonparametric Regression and Spline Smoothing and. New

York: Marcel Dekker, Inc.

Eubank R.L., Huang C., Maldonado Y., Wang N., Wang S. & Buchanan R. J. (2004). Smoothing spline estimation in varying-coefficient models. *Journal of Royal Statistical Society,* Series B, 6, 653- 667.

Fan J. & Gijbels I. (1996). Local polynomial modeling and its application. London: Chapman and Hall.

Fan J. & Zhang J.T. (1998). Comments on "Smoothing spline models for the analysis of nested and crossed samples of curves" by Brumback and Rice (1998). *Journal of American Statistical Association*, 93, 980-983.

Fan J. & Zhang W. (2008). Statistical Methods with Varying Coefficient Models. *Stat Interface*, 1, 179-195.

Fitzmaurice G.M., Larid N.M. & Ware J.H. (2004). Applied Longitudinal Analysis. New Jersey: John Wiley & Sons.

Gasser T.H. & Muller H.G. (1979). Smoothing Techniques for Curve Estimation. In Gasser T.H. & M. Rosenblatt (Ed.), pp. 23-68. Heidel-berg: Springer.

Gasser T.H. & Rosenblatt M. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics,* 11, 171-185.

Green P.J. & Silverman B.W. (1994). Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. London: Chapman & Hall.

Hart J.P. & Wehrly T.E. (1986). Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association.,* 81, 1080-1088.

Hastie T.J. & Tibshirani R.J. (1993). Varying-coefficient models. *Journal of Royal Statististical Society,* 55, 757-796.

Harville D.A. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *Annals of Statistics,* 4, 384-395.

Harville D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of American Statistical Association,* 72, 320-340.

Hoover D.R., Rice J.A., Wu, C.O. & Yang L.P. (1998). Nonparametnc smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85, 809-822.

Janssen I., Craig W., Boyce W.F. & Pickett W. (2004). Associations between overweight and obesity with bullying behaviors in school-aged children**.** *Pediatrics*, 113, 1187-1194.

Jacqmin-Gadda H., Joly P., Commenges D., Binquet C. & Chene *G.* (2002). Penalized likelihood approach to estimate a smooth mean curve on longitudinal data. *Statistics in Medicine,* 21, 2391-2402.

Kaye W.H., Bulik C.M., Thornton L., Barbarich N. & Masters K. (2004). Comorbidity of anxiety disorders with anorexia and bulimia nervosa. *American Journal of Psychiatry*, 161, 2215-2221.

Kupersmidt J.B. & Coie J.D. (1990). Preadolescent peer status, aggression, and school adjustment as predictors of externalizing problems in adolescence. *Child Development*, 61, 1350-1362.

Kuczmarski R.J., Ogden C.L., Guo S.S., Grummer-Strawn L.M., Flegal K.M., Mei Z., Wei R., Curtin L.R., Roche A.F. & Johnson C.L. (2002). 2000 CDC Growth Charts for the United States: Methods and Development. *Vital Health Statistics,* 11, 1-190.

Laird N.M. & Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics,* 38, 963-974.

Lawlor D.A., Mamun A.A., O'Callaghan M.J., Bor W., Williams G.M. & Najman J.M. (2005). Is being overweight associated with behavioral problems in childhood and adolescence? *Archives of Disease in Childhood*, 90, 692-697.

Liang H., Wu H. & Carroll R.J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient semiparametric models with measurement error. *Biostatistics*, 4, 297-312.

Liang K.Y. & Zeger S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika,* 73, 13-22.

Lin X. & Carroll R.J. (2000). Nonap召metric function estimation for clustered data when the predictor is measured without/with error. *Journal of American Statistical Association,* 95, 520-534.

Lin X. & Carroll R.J. (2001a). Semiparametric regression for clustered data using generalized estimating equations. *Journal of American Statistical Association*, 96, 1045-1056.

Lin X. & Carroll R.J. (2001b). Semiparametric regression for clustered data. *Biometrika*, 88, 1179-1185.

Lin X. & Carroll R.J. (2005). Semiparametric estimation in general repeated measures problems. *Journal of Royal Statistical Society,* B, 68, 69-88.

Lin X., Wang N., Welsh A. & Carroll R.J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered data. *Biometrika*, 91, 177-193.

Lissau I. & Sorensen T.I. (1993). School difficulties in childhood and risk of overweight and obesity in young adulthood: a ten year prospective population study**.** *International Journal of Obesity,* 17, 169-175.

Lissau I. & Sorensen T.I. (1994). Parental neglect during childhood and increased risk of obesity in young adulthood**.** *Lancet,* 343, 324-327.

Lumeng J.C., Gannon K., Cabral H.J., Frank D.A. & Zuckerman B. (2003). Association between clinically meaningful behavior problems and overweight in children. *Pediatrics,* 112, 1138-1145.

Maddi S.R., Khoshaba D.M., Persico M., Bleecker F. & VanArsdall G. (1997). Psychosocial correlates of psychopathology in a national sample of the morbidly obese. *Obese surgery*, 7, 397-404.

Mamun A.A., O'Callaghan M.J., Cramb S.M., Najman J.M., Williams G.M. & Bor W. (2009). Childhood behavioral problems predict young adults' BMI and obesity: evidence from a birth cohort study. *Obesity,* 17, 761-766.

Martinussen T. & Scheike T.H. (1999). A semiparametric additive regression model for longitudinal data. *Biometrika,* 86, 691-702.

McEwen B.S. (2008). Understanding the potency of stressful early life experiences on brain and body function**.** *Metabolism-Clinical and Experimental*, 57, S11-S15.

National Institute of Health. (2006). The NICHD Study of Early Child Care and Youth Development: Findings for children up to age 4½ years. *NIH Pub,* No. 05-4318.

NICHD Early Child Care Research Network. (2001). Nonmaternal care and family factors in early development: an overview of the NICHD study of early child care. *Journal Applied Developmental Psychology,* 22, 457-492.

O'Sullivan F. (1986). A statistical perspective on ill-posed in verse problems. *Statistics Science*, 1, 505-527.

Radloff L.S. (1977). The CES-D Scale: a self-report depression scale for research in the general population**.** *Applied Psychological Measurement*, 1, 385-401.

Rice J.A. & Wu C. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics,* 57, 253-259.

Ruppert D., Wand M.P. & Carroll R.J. (2003). Semiparametric Regression. New York: Cambridge University Press.

Rutter M. & Garmezy N. (1983). Handbook of Child Psychology: Socialization, Personality, and Social Development. In Hetherington E. M. (Eds.), *Developmental psychopathology* (pp. 775-911). New York: Wiley.

Schwarz G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6, 461-464.

Shi M., Weiss R.E. & Taylor J.M. (1996). An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics,* 45, 151-163.

Stone C.J., Hansen M., Kooperberg C. & Truong Y.K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics, 25,* 1371-1470.

Strauss R.S. & Pollack H.A. (2003). Social marginalization of overweight adolescents**.** *Archives of Pediatrics & Adolescent Medicine,* 157, 746-752.

Tao H., Palta M., Yandell B.S. & Newton, M. A. (2005). An estimation method for the semiparametric mixed-effects model. *Biometrics,* 55, 102-110.

Verberk G. & Lessaffre E. (1996). A linear mixed effect model with heterogeneity in random effects population, *Journal of American Statistical Association,* 91, 217-221.

Vila G., Zipper E., Dabbas M., Bertrand C., Robert J.J. & Ricour C. (2004). Mental disorders in obese children and adolescents. *Psychosom Med*, 66, 387-394.

Wahba G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In Krishnaiah P.R. (Ed), Applications of Statistics, pp. 507-523. Amsterdam, North Holland.

Wahba G. (1990). Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.

Wand M. P. & Jones M.C. (1995). Kernel Smoothing. London: Chapman & Hall.

Wang Y. (1998a). Mixed-effects smoothing spline ANOVA. *Journal of Royal Statistical Society,* Series B, 60, 159-174.

Wang Y. (1998b). Smoothing spline models with correlated random errors. *Journal of American Statistical Association,* 93, 341-348.

Wang Y. & Taylor J.M. (1995). Inference for smooth curves in longitudinal data with application to an AIDS clinical trial. *Statistics in Medicine,* 14, 1205-1218.

West M., Harrison P.J. & Migon H.S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of American Statistical Association,* 80, 73-83.

Wu C.O., Chiang C.T. & Hoover D.R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of American Statistical Association,* 93, 1388-1402.

Wu C.O. & Yu K.F. (2002). Nonparametric varying-coefficient models for the analysis of longitudinal data. *International Statistical Review*, 70, 373-393.

Wu H. & Zhang, J.T. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of American Statistical Association,* 97, 883-897.

Wu H. & Zhang, J.T. (2006). Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches. New York: John Wiley & Sons.

Zeger S.L., Liang K.Y. & Albert P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049-1060.

Zeger S.L. & Diggle P.J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics,* 50, 689-699.

Zhang D., Lin X., Raz J. & Sowers M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of American Statistical Association*, 93, 710-719.