2017

Semi-Parametric Generalized Estimating Equations with Kernel Smoother: A Longitudinal Study in Financial Data Analysis

Liu Yang

FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

SEMI-PARAMETRIC GENERALIZED ESTIMATING EQUATIONS WITH KERNEL SMOOTHER:

A LONGITUDINAL STUDY IN FINANCIAL DATA ANALYSIS

By

LIU YANG

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2017

Liu Yang defended this dissertation on November 15, 2017.
The members of the supervisory committee were:

Xufeng Niu

Professor Directing Dissertation

Yingmei Cheng

University Representative

Fred Huffer

Committee Member

Minjing Tao

Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Longitudinal studies are widely used in various fields, such as public health, clinic trials and financial data analysis. A major challenge for longitudinal studies is repeated measurements from each subject, which cause time dependent correlation within subjects. Generalized Estimating Equations can deal with correlated outcomes for longitudinal data through marginal effect. My model will base on Generalized Estimating Equations with semi-parametric approach, providing a flexible structure for regression models: coefficients for parametric covariates will be estimated and nuisance covariates will be fitted in kernel smoothers for non-parametric part. Profile kernel estimator and the seemingly unrelated kernel estimator (SUR) will be used to deliver consistent and efficient semi-parametric estimators comparing to parametric models. We provide simulation results for estimating semi-parametric models with one or multiple non-parametric terms. In application part, we would like to focus on financial market: a credit card loan data will be used with the payment information for each customer across 6 months, investigating whether gender, income, age or other factors will influence payment status significantly. Furthermore, we propose model comparisons to evaluate whether our model should be fitted based on different levels of factors, such as male and female or based on different types of estimating methods, such as parametric estimation or semi-parametric estimation.

CHAPTER 1

INTRODUCTION

1.1 Background

For statistical scientific studies, different experiment designs depend on different types of system under study and different goals for research. Cross-sectional studies and longitudinal studies are two types of important statistical experiment widely used in various fields, such as public health, clinical trials, and economics. Cross-sectional studies involve the analysis of data observed from a population given dependent variables of interest at one specific time point whereas longitudinal studies involve a period over time, allowing the repeated observations with the same variables for each subject.

Cross-sectional studies have the advantages that allow us comparing different variables at a same time point and collecting routine data efficiently. For instance, consider a financial study in credit card loan data, we can collect the payment information from each client for one month across different gender groups and education groups. However, this study may not include past or future payment information. Therefore, cross-sectional studies generally do not provide enough evidence for accurate relationship over different time periods.

Longitudinal studies allow the investigation of change over different time points and the effects of different factors on the change. One distinctive feature of longitudinal studies is the repeated measurements at different time points within each subject (or cluster), which take the time series correlation into account. If we apply longitudinal studies in the previous credit card loan data, we may collect monthly payment information on a client over many years and obtain more reliable relationship between payment status and many factors through different time period.

Several challenges raised from longitudinal studies include correlated outcomes, time-varying covariates and missing data from incomplete experiments. One important assumption for classic linear models is that observations from the response variable are independent. However, in a longitudinal study, repeated measurements from a subject are dependent, which may be characterized by a time series model instead of a classic linear model. Another feature in a longitudinal study is

that covariates related to the response variable may change with time, called time-varying covariates. When the relationship between response and covariates is investigated, correlations among observations of a time-varying covariate need to be considered too. Missing data from incomplete experiments in a longitudinal study need to be also dealt with. For instance, subjects may suspend their treatments during an experiment, leaving "drop-out" patterns whether at random or not. The complicated missing structure for unbalanced data cannot apply in classic linear models.

The critical challenge of dealing with longitudinal data is the correlation within subjects. When classic linear models fail to analyze repeated measurements, several alternative models may be applied to longitudinal analysis, which include marginal models, random effect models and transition models. The marginal approach fits a model for data from all subjects at each time point, indicating the mean response at each time point depends only on covariates, instead of any other source from specific subjects or previous response. Random effect models, or subject-specific models also refers to generalized linear mixed models (GLMMs) in longitudinal studies, providing source for within-subject association through a vector of random effects on mean response, allowing mean response change depend not only on covariates, but also on randomness among different subjects. Transition models handle longitudinal data with sequential nature, by modeling the conditional distribution of each mean response with explicit function of previous responses or covariates.

Marginal models only state regression on mean response, without requiring full distributions for repeated measurements. Some likelihood-based approaches had been proposed for marginal models: Gumbel (1961) proposed a latent-variable model for multivariate binary data, requiring high-dimensional integration over the joint distribution of the latent variables; GSK method (1969) provided a multinomial distribution for the vector of repeated responses within each subject with many restrictions, including categorical covariates and large sample sizes. In fact, all the likelihood-based approaches share the issues that lack of an appropriate joint distribution for multivariate responses, no closed form for the joint probabilities, sensitive incorrect estimation for β and its higher-order moments. If a binary response measured at 10 time points, the total number of multinomial probabilities to be estimated should be $2^{10} - 1$, which presents a lot of difficulty in practice.

Generalized Estimating Equation method, extended by quasi-likelihood method (Wedderburn, 1974) is an approach for marginal models without likelihood-based issues discussed above. Zeger

and Liang (1986) considered the marginal distribution of response variables, providing parametric GEE approach through extension of Generalized Linear Models. Instead of specifying joint multivariate distribution of the repeated measurements, they introduce estimating equations with consistent estimates of regression parameters. GEE methods have some decent properties, such as more precision comparing with maximum likelihood estimation and robustness with consistent estimator even if the within-subject correlation of repeated measurements has been misspecified. (Zeger and Liang 1986) One major limitation for parametric GEE is that association between mean response and covariates is fully parametric. Sometimes parametric models cannot capture the complicated relationship between response and covariates while non-parametric models or semi-parametric models may present an adequately flexible approach to explore the relationship between longitudinal outcomes and covariates.

A variety methods can be used for estimating non-parametric and semi-parametric regression models for independent data. For non-parametric regression models, we may use kernel estimation methods based on local likelihoods and splines based on penalized likelihoods. For semi-parametric regression models, we may use partial linear models, which specify the mean of outcome variable as parametric function respect to some covariates and non-parametric functions respect to other covariates. More specifically, local polynomial kernels, smoothing splines, regression spline and penalized splines have been introduced for non-parametric and semi-parametric regression estimation methods. Local polynomial kernels provide different weight for neighborhood observations. Smoothing splines fit the non-parametric function by a spline function with a set of covariates. Regression splines model the non-parametric regression part with spline basis functions with a small number of knots and penalized splines present put penalty of smoothing splines on regression splines.

When refer to longitudinal data analysis, non-parametric and semi-parametric regression should be able to deal with within subject correlation for repeated measurements. Estimating equation based methods and likelihood based methods can be used on non-parametric regression and semi-parametric regression with kernel and spline smoothing methods. Lin and Carroll (2000) proposed kernel GEE estimator through local polynomial kernel estimating equations by the extension of generalized linear model. Unlike the parametric GEE developed by Zeger and Liang (1986), kernel GEE have limited conditions for consistent estimator and cannot reach efficiency bound if account

for within-subject association. Wang (2003) provided the seemingly unrelated kernel (SUR) estimator which fulfill both consistency and efficiency if we consider within-subject association. For likelihood based settings, spline smoothing includes the generalized smoothing spline estimator, P-splines and regression splines and smoothing spline estimator has close relationship with linear mixed models.

Semi-parametric regression can be applied in marginal models and linear mixed models depends on different goals. If we focused on semiparametric regression in marginal models, several estimation method have been developed to deal with the within-subject correlations. Lin and Carroll (2001) developed profile-kernel estimating equations which estimate parametric part by profile method and non-parametric part by kernel GEE with local polynomial kernels we mentioned above. The estimator from profile-kernel methods is consistent only when ignoring within-subject correlation and is not semi-parametric efficient even without the within-subject correlation for non-parametric part. Wang, Carroll and Lin (2004) used SUR kernel model for non-parametric part and remained estimating the parametric part with profile method, providing an estimator with consistency and semi-parametric efficiency. For semi-parametric linear mixed models, we can also use profile SUR kernel methods to fit the model and spline method as well.

1.2 Motivation

Financial market plays an important role in daily life. Major types of financial market include capital markets, commodity markets, money markets and derivative markets. Capital markets involve stock markets and bond markets, providing long-term investment; commodity markets provide trading for primary economic section products, such as agricultural products or mined products; money markets treat money as commodity, providing short-term trading such as borrowing, lending, buying and selling and financial instruments derived from assets are trading in derivative markets such as option and future contracts.

Financial institutes such as commercial banks, investment banks, insurance companies and brokerages are major players for trading in financial market. Financial data analysis raises when commercial banks issue loans, investment issue securities, insurance company decide premium and brokerages settle bid price. Most financial data analysis involves time series because in financial market, time is valuable and we would like to track temporal tendency on subjects. Therefore,

once we have time changing measurements for each subject as well as covariates, we can conduct longitudinal studies for financial data analysis.

Longitudinal studies display not only relationship between covariates and response variable, but also showing the change of response variable over time. In financial data analysis, longitudinal properties can be applied to cope subjects and time tendency. Commercial banks can track customers across different time points to perform risk management for asset default; investment banks can follow the trend for stocks and bonds, evaluating returns through not only traditional time series model but also other interests such as macro-economic index or other factors that will affect pricing in financial market. Insurance companies can focus on products with time-varying properties such as health insurance premium, investigating health record over years and assigning proper quota for keeping customers and reducing risk. Longitudinal study is popular for researches on corporate firms, since it can assess performance of corporate firms through different interests with time changing effect.

A variety of longitudinal models can be applied in financial analysis. Petersen (2009) pointed out that previous researches focus on mainly three major methods: Fama-MacBeth procedure (Fama and MacBeth, 1973) estimates, dummy variables in each cluster such as fixed effect model and adjust within cluster correlation such as generalized estimating equations. Different methods should be applied depending on different interests. For subject specified effect, Generalized Linear Mixed Model (GLMM) will provide nice estimator for individual subjects and for exploring population average effect. When covariates are involved in general factor or policy, Generalized Estimating Equations can be applied to investigate relationship between response and covariates.

In order to capture the complex relationship in longitudinal data analysis, semi-parametric and non-parametric models have been devepoled for financial data analysis in longitudinal studies. Sam and Jiang (2009) propose a non-parametric estimator for short rate diffusion process with yields in longitudinal structure. To remove the bias from parametric models, they construct estimator for U.S short rate process, showing that it has economic impact on the pricing of financial assets such as bonds and interest rate derivatives.

1.3 Structure

Chapter 2 will focus on literatures for Generalized Estimating Equations and semi-parametric models. Traditional Generalized Estimating Equations methods (Zeger and Liang 1986) will be proposed first, with details about the model and some properties for the estimator. Kernel smoothers will be discussed then, showing the properties of kernel estimators and non-parametric generalized estimating equation will be described as well, especially for kernel generalized estimating equations(Lin and Carroll 2001). Semi-parametric models with kernel smoother will be described then, with several different approaches: Profile-kernel estimating equations (Lin and Carroll 2001) will be proposed first and profile SUR kernel methods (Wang, Carroll and Lin 2005) will be presented as well.

Chapter 3 will display mathematical details for semi-parametric model, kernel smoothers and semi-parametric kernel estimating equations. Different estimators with different approaches will be fully developed with closed form solutions, such as kernel average estimator(Lin and Carroll 2004) and The SUR kernel estimator(Wang, Lin and Carroll 2004).

Chapter 4 will show a simulation study follow the methods in Chapter 3. Results with estimated coefficients and overall fitting mean square errors for parametric estimators and semi-parametric estimators will be provided, showing the difference between parametric models and semi-parametric models. For each model, we display two setups and different time periods with separated training and testing datasets. Detailed table titles for different models with different setups and different time periods will be provided in page 2.

Chapter 5 would be data application. Data description will be provided first, showing details of predictors and responses variables in credit card loan dataset. We conduct an overall model first, and based on criterion of model selection, we provided results for analysis when fitting model separately based on different level of factors.

Chapter 6 shows future work, focus on develop a more reliable criterion on model selection for semi-parametric approach. We would like to extend our model for high dimensional data structure and find out how to construct the semi-parametric approach.

CHAPTER 2

LITERATURE REVIEW

In the Literature Review part, we first discuss traditional Generalized Estimating Equation that extends the Quasi-likelihood Approach in section 2.1. In section 2.2, kernel smoothers with different densities will be provided for estimating non-parametric models and semi-parametric models and bandwidth selection method will be described for the tuning bandwidth parameter. In section 2.3, non-parametric GEE approach will be discussed for estimating non-parametric models with longitudinal data. In section 2.4, semi-parametric GEE models with the profile kernel method and profile SUR method will be presented, and the second method provided an efficient semi-parametric estimator.

2.1 Generalized Estimating Equations

2.1.1 Quasi-Likelihood Approach

Wedderburn (1974) first proposed quasi-likelihood approach for estimating Generalized Linear Models and McDullagh (1983) provided more details for quasi-likelihood estimators. Instead of fully specifying the distribution density function for the response variable, quasi-likelihood approach only requires the relationship between mean function of outcomes and covariates as wells as the variance structure of the response variable as a function of mean.

Suppose $\mathbf{Y} = (Y_1, \cdots, Y_n)$ is a vector of observations and Y_i is the value observed at the ith subject. Let $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})$ be a vector of covariates for the ith subject and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ be a p-dimensional parameter vector. Let $\boldsymbol{\mu} = E(\mathbf{Y}) = (\mu_1, \cdots, \mu_n)$ be a mean vector for the response variable with the structure $\mu_i = X_i^T\boldsymbol{\beta}$. The covariance matrix of $\mathbf{Y}$ can be specified as $\boldsymbol{V}(\boldsymbol{\mu}) = \phi cov(Y)$, where $\mathbf{V}_{ij}(\mu)$ is a variance function of mean $\boldsymbol{\mu}$ and ϕ is a unknown scalar denoting a scale parameter.

McDullagh (1983) proposed the quasi-likelihood approach to estimate β by setting the score function $u(\beta) = \mathbf{D}^T\mathbf{V}^{-1}\{\mathbf{Y} - \boldsymbol{\mu}\} = 0$, where $\mathbf{D} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}$ is a $n \times p$ matrix and $\mathbf{V}^{-1}$ is the inverse matrix of $\mathbf{V}(\boldsymbol{\mu})$. McDullagh (1983) claimed that this score function $u(\beta)$ has zero mean

and finite variance and the quasi-likelihood estimator $\hat{\beta}$ is consistent. Agrasti (2002) pointed out a key property of the quasi-likelihood estimator $\hat{\beta}$: if the mean function and linear relationship are correctly specified, $\hat{\beta}$ will be consistent even if the scale parameter ϕ is misspecified.

2.1.2 Generalized Estimating Equation Approach

Zeger and Liang (1986) extended the quasi-likelihood equations to a multivariate case and provided the Generalized Estimating Equation method. Instead of specifying the joint distribution for repeated measurements, they introduced working correlation matrix to show correlations within each subjects. They stated 'independence' estimating equation first, assuming there was no association within subjects. With this "independence" estimating equations for longitudinal data, they derived a consistent estimator if mean structure with covariates is correctly specified. They defined efficiency based on the estimated variance of estimators, with high efficiency corresponding to low estimated variance of the estimators. The authors then introduced a working correlation matrix with a parameter α to capture the within subject association and obtain higher efficiency. They proved that when the number of subjects increases, the new GEE estimator follows a multivariate Gaussian distribution with the true mean values and a specified covariance matrix . Therefore, similar to the independence case, this new GEE estimator is consistent if the mean function and covariates are correctly specified.

The efficiency of a GEE estimator depends on how the working correlation matrix is approximated: the closer to the true correlation matrix for the within subject association, the more efficiency of this GEE estimator we can obtain. Zeger and Liang (1986) claimed that GEE estimators get the most efficiency when working correlation matrix equals to the true correlation matrix. If working correlation matrix are identity, GEE estimators will equal to the estimators with maximum likelihood approach.

For model selection, Pan (2001a) extended Akaike's information criterion (AIC, 1973) to the quasi-likelihood approach as a information criterion (QIC), which provided a criterion for the diagnosis of Generalized Estimating Equation method. Instead of treating numbers of parameters as penalty term, QIC used the trace of working correlation matrix as a penalty. Simulation study by Pan (2001) shows that QIC favors less restrictive structures, which means if only consider QIC as the criterion to evaluate the model, unstructured working correlation matrix dominates other structures.

2.2 Kernel Smoother

Kernel smoothers are widely used in estimating non-parametric model or the non-parametric part in a semi-parametric model. Hastie, Tibshirani and Friedman (2008, chapter 6) pointed out that kernel smoother provides a method to define local weights K based on kernel densities. For example, supposed we want to estimate a function $Y(x)$ at a target point x_0, kernel weights are assigned according to the distance between this target point and its neighborhood point x_i. A bandwidth parameter h indicates the width of the neighborhood.

2.2.1 Kernel Smoothers

Kernel smoothers are constructed from symmetric mean zero kernel density functions. Usually we use either the Epanechnikov density, or the Tri-cube density or the Gaussian density. Tri-cube density is compact, Epanechnikov density has no continuous derivatives at the boundary, and Gaussian density is continuously differentiable with infinite support. The choice of densities generally has little impact for the estimation results (Hastie, Tibshirani and Friedman 2008).

Different types of kernel-weighted estimators can be constructed by the different forms of estimation function $\hat{Y}(x)$. Nadaraya-Watson kernel-weighted average estimator (Nadaraya and Watson 1964) is widely used to evaluate the observed value on the target point x_0 when $\hat{Y}(x)$ is locally constant at neighborhood point x of x_0. N-W estimator is convenient to calculate but it may raise observation bias on the boundaries of domain. Local linear kernel estimator uses a linear form of estimation function $\hat{Y}(x)$, providing a first order correction for N-W estimator to remove boundary bias, but it may have issues on hills or valleys for the curvature of the fitted kernel smoothers. We can step forward to try local polynomial estimators with a polynomial form of $\hat{Y}(x)$ to deal with the issues of curvature and some results show that local polynomial estimator with odd degrees are better than even degrees. However, local polynomial estimator may increase variance while local linear estimator will keep little cost on variance when removing bias. There is a bias and variance trade off for different kernel estimators and if we are more interested in boundaries, local linear estimator will be more reliable.

For computation, kernel weighted average estimator is more straightforward while local linear kernel estimator has explicit form. Local polynomial kernel estimator is complicated and we may apply some other computation paths instead of a closed form solution. When compiling kernel

smoothers with Generalized Estimating Equations in longitudinal study, it may raise more difficulties.

2.2.2 Bandwidth Selection for Kernel Smoothers

Bandwidth parameter h shows the width of neighborhood of the target point. For Epanechnikov kernel smoother, bandwidth is the radius of the support region and for gaussian kernel smoother, bandwidth is the standard deviation. Bandwidth parameter is critical for kernel smoothers and there is a bias-variance trade off for estimated h: with a small bandwidth, the bias of a estimator will be smaller but variance will increase. If we have a relatively large bandwidth, which indicates we use observations x_i further away from target point x_0, the bias will increase but the variance will decrease.

Cross validation is a common method used to tune bandwidth parameter h. Least square cross-validation was first proposed by Rudemo (1982), Stone (1984) and Bowman (1984) and other cross validation methods such as likelihood cross validation (Duin 1976) can also be used in the computation of h.

2.3 Non-Parametric Generalized Estimating Equations with Kernel Smoother

2.3.1 Local Polynomial Kernel GEE

Lin and Carroll (2000) proposed local polynomial kernels GEE approach in longitudinal analysis, constructing a model which explore the relationship between response variable and covariates with non-parametric forms through a link function. They followed the estimating equations of parametric GEE estimator provided by Zeger and Liang (1986) and derived two estimating equations for local polynomial kernel GEE estimators, one for symmetric kernel density and the other for asymmetric kernel density. The computation of the estimator β can be obtained by iteratively re-weighted least squares and the initial value can be set as Generalized Linear Model estimator(Nelder and Wedderburn 1972) to start the iteration.

Lin and Carroll (2000) showed properties for local polynomial kernel estimators estimated by kernel GEE with different orders of the polynomial term for non-parametric covariates. For the local average kernel estimator, which estimates $Y(x)$ by the local polynomial kernel GEE with a zero

order polynomial term, the estimator is consistent under any specified working correlation matrix and the variance is minimized under the assumption of independent working correlation matrix. For the linear kernel estimator, which is estimated by the local polynomial kernel GEE with order one in the polynomial term, the asymptotic properties of the estimator are hard to derive for non-gaussian case because a complicated mean function or link function may raise difficulties. Therefore, they only provided Gaussian case for local linear kernel estimator. Local linear kernel estimator has the same properties as the local average kernel estimators, saying that the estimator is consistent with any working correlation matrix and obtain minimum variance when using the independent working correlation matrix. They proceeded a pth order polynomial term in kernel GEE and obtained local polynomial kernel estimator. The authors claimed that this local polynomial estimators share the same properties as local average kernel estimator and local linear kernel estimator. For bandwidth selection, they used EBBS by Ruppert (1997) instead of cross validation method we presented in Section 2.2.2 because of less computation cost.

Lin and Carroll (2000) pointed out that for any estimated kernel smoothers, when using independent working correlation matrix, the most efficient kernel GEE estimator does not take the true within subject association into account. This result conflicts the major result of traditional GEE estimator proposed by Zeger and Liang (1986), which said the most efficient estimator is obtained by using working correlation matrix which specified true within subject association. One reason for this conflicts comes from how kernel smoother is estimated: kernel methods use observations from all subjects at all different time points as the neighborhood of the target point, without concerning the cluster effect produced by association within each subject. Therefore, the most efficient kernel estimator is calculated based on the situation when ignoring within subject association.

2.3.2 The Seemingly Unrelated Kernel Estimator

Wang (2003) proposed the seemingly unrelated kernel estimator, which takes within subject associations into account. The author considered the cluster effect and pointed out that estimators ignoring within subject association may suffer loss of efficiency. When the number of subjects increases, for each subject(cluster), there was one data point which contributes to the estimators. When solving the estimating equations, working correlation matrix has a diagonal element entry which corresponds to that data point and estimators will be calculated based on the kernel weights evaluated at that data point.

Wang (2003) derived a score function to calculate the SUR estimator with an iterative algorithm. The initial set up of the algorithm can be obtained by the kernel GEE estimator with the independence working correlation matrix. Wang (2003) showed that the SUR estimator are consistent and achieve minimized variance when using true working covariance matrix, which is different from the results of local polynomial kernel GEE estimator derived by Lin and Carroll (2000), saying that the most efficient estimator comes from the estimator using independence working correlation matrix. Lin et al. (2004) provided a closed form solution to compute the SUR estimator.

2.4 Semi-Parametric Generalized Estimation Equations with Kernel Smoother

Semi-parametric model is a combination for parametric model and non-parametric model. The parameters in semi-parametric model have linear coefficients and non-parametric smoothers. Hastie and Tibshirani (1990) provided a backfitting algorithm for calculating semi-parametric estimators, showing a closed form solution for the estimated linear coefficients and the estimated non-parametric smoothers. Zerger and Diggle (1994) used the backfitting algorithm for longitudinal analysis, estimating parameters for linear part and non-parametric part by ignoring within subject association in non-parametric part. We focus on two semi-parametric estimators provided by Lin and Carroll (2004) and Wang et al. (2005), showing two semi-parametric kernel Generalized Estimating Equation methods which use different working correlation matrices in parametric part and non-parametric part.

2.4.1 Profile-Kernel Estimator

Lin and Carroll (2004) proposed a semi-parametric approach through profile-kernel estimating equation methods. They provided a profile method to estimate the parametric coefficients with a given non-parametric smoother and a kernel method to estimate the non-parametric smoother with the estimated parametric coefficients. They used this profile-kernel method in Generalized Estimating Equations and took within subject association into account through a specified working correlation matrix in both parametric part and non-parametric part.

Lin and Carroll (2004) obtained unexpected asymptotic properties and empirical simulation results for this profile-kernel estimator: the coefficients for parametric part will be consistent only

when independent working correlation matrix is used in both parametric part and non-parametric part; even when the true working correlation matrix is assumed, coefficients are not consistent unless kernel smoother is undersmoothed with a smaller bandwidth. The reason why profile-kernel method fails is because when assigning kernel weights in non-parametric part, kernel methods do not consider the time association within each subject as we discussed in section 2.3.1. Lin and Carroll (2004) also pointed out that if the non-parametric kernel smoother is constructed with a subject level covariate which ignores time changing issues, profile-kernel estimator will be consistent and obtain semi-parametric efficiency when true working correlation matrix is assigned.

2.4.2 Profile SUR Method

Wang et al. (2005) proposed profile SUR method, using non-parametric SUR estimator derived by Wang (2003) to estimate the kernel smoother in non-parametric part. They proved that profile SUR estimator is consistent without constraint of undersmoothed kernel smoother and bandwidth selection is not so critical because the estimator is insensitive to bandwidth. More important, this estimator will achieve semi-parametric efficiency when true working correlation matrix is assumed. This result shows that the authors coped time changing effect within each subject in longitudinal data and agreed the key properties of traditional Generalized Estimating Equations first proposed by Liang and Zeger (1986) as we described in section 2.2. A closed form solution for the profile SUR estimator can be calculated by the backfitting methods (Hastie and Tibshirani 1990) and Lin et al. (2006) provided such a closed form solution for the profile SUR estimator.

The kernel smoother for this profile SUR estimator uses Ebbs (Ruppert 1997) to select bandwidth and a kernel GEE polynomial term with zero order (section 2.3.1) to assign the kernel weights. Wang et al. (2005) claimed that when using kernel GEE with pth order polynomial term, the properties for the new profile SUR estimator will be the same as when a zero order polynomial term is assumed in the kernel estimating equations.

CHAPTER 3

MODELS AND METHODS

In this chapter, we propose semi-parametric models for Generalized Estimating Equations. Various kernel smoothers will be used to capture non-parametric pattern between response variable and covariates. Specifically, provide local polynomial kernel GEE estimator and the seemingly unrelated kernel estimator are two main tools for model fitting. The difference of consistency and efficiency between those estimators will be displayed when accounting to association within subjects.

3.1 Semi-Parametric Model

Hastie and Tibshirani (1990) proposed Generalized Additive Model with the form:

$$E(Y_i|X_i) = f_1(X_i) + f_2(Z_i) \quad i = 1, 2, \cdots, n \tag{3.1}$$

where i denotes the ith subject and Y_i denotes the response variable. In the model, $f_1(X_i)$ denotes a linear pattern with $f_1(X_i) = X_i^{'}\beta$ where $X_i^{'} = (X_{i1}, ..., X_{ip})$ denotes p covariates for the linear part and β is a $p \times 1$ coefficient vector for the covariates. In the second part, Z_i denotes covariates for a non-parametric pattern and $f_2(\cdot)$ denotes a smoother function.

In general, we can use a backfitting algorithm to fit Generalized Additive Model:

Algorithm 1 Fitting Generalized Additive Model

1: $f_1 = f_1^0$ ▷ Initialize f_1
2: $f_2(Z_i) = S_2(Y_i - f_1(X_i))$
3: update $f_1(X_i) = S_1(Y_i - f_2(X_i)) = X_i\hat{\beta}$
4: back to step2 until convergence

In semi-parametric model, $S_1 = X(X^{'}X)^{-1}X^{'}$ denotes a projection for the least-square estimate and S_2 denotes a non-parametric smoother to estimate f_2. Hastie and Tibshirani(1990, Chapter 5) pointed out in this case β and $\hat{f}_2$ could be solved explicitly with form:

$$\beta = \{X^{'}(I - S_2)X\}^{-1}X^{'}(I - S_2)Y \tag{3.2}$$

$$\hat{f}_2 = S_2(y - X\beta) \tag{3.3}$$

Therefore, as long as $X^{'}(I - S_2)X$ is invertible, Hastie and Tibshirani(1990) displays that this solution is consistent and unique for semi-parametric model.

3.2 Kernel Smoothers

In this study, we would like to use a kernel smoother for non-parametric part in a semi-parametric model. For example we may use Epanechnikov quadratic kernel (Hastie and Tibshirani 2009):

$$K_h(x_0, x) = D(\frac{|x - x_0|}{h(x_0)}) \tag{3.4}$$

where

$$D(t) = \begin{cases} \frac{3}{4}(1 - t^2), & \text{if} |t| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

Let $h(x_0)$ denote the bandwidth parameter for each evaluated point. For Gaussian kernel densities, $h(x_0)$ is the standard deviation σ .

There are many different ways to use a Kernel function in non-parametric estimation. For instance, we may use the Nadaraya-Watson kernel-weighted average (ref) to fit the value at an evaluated point x_0:

$$\hat{f}(x_0) = \frac{\sum_{i=1}^{N} K_h(x_0, x_i) y_i}{\sum_{i=1}^{N} K_h(x_0, x_i)} \tag{3.6}$$

In order to remove potential bias on boundary, we may use local linear estimator, at each evaluated x_0 with the form:

$$\min_{\alpha(x_0), \beta(x_0)} \sum_{i=1}^{N} K_h(x_0, x_i)[y_i - \alpha(x_0) - \beta(x_0) x_i]^2 \tag{3.7}$$

after we solved for $\hat{\alpha}(x_0)$ and $\hat{\beta}(x_0)$, the fitted value at each evaluated point x_0 is $\hat{f}(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0) x_0$.

Suppose $b(x)^{'} = (1, X)$ let B be a $N \times 2$ regression matrix with ith row $b(X_i)^{'}$, and $W(x_0)$ the $N \times N$ diagonal matrix with ith diagonal element $K_\lambda(x_0, x_i)$, we have:

$$\hat{f}(x_0) = b(x_0)^{'}(B^{'} W(x_0) B)^{-1} B^{'} W(x_0) y \tag{3.8}$$

which is an explicit form for the local linear regression estimate. If we extend local linear estimation to local polynomial regression fitting:

$$\min_{\alpha(x_0),\beta(x_0),j=1,\ldots,d} \sum_{i=1}^{N} K_h(x_0, x_i)[y_i - \alpha(x_0) - \sum_{j=1}^{d} \beta(x_0) x_i^j]^2 \tag{3.9}$$

3.3 Working Correlation Structure

In longitudinal data analysis, estimating time trend for each subject is an important and interesting topic. We would like to introduce three major types of working correlation matrices: AR(1) working correlation matrix, exchangeable working correlation matrix, and unstructured working correlation matrix. Those working correlation matrices will be used in the part of simulation and application. Let Y_t be the observations for one subject, we may consider an autoregressive model with order p for the observations:

$$\begin{aligned} Y_t &= \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \epsilon_t, \\ E(\epsilon_t) &= 0, \quad E(\epsilon_t^2) = \sigma^2, \quad E(\epsilon_t \epsilon_s) = 0, \quad \forall t \neq s \end{aligned} \tag{3.10}$$

AR(1) working correlation matrix

For an AR(1) structure with $t = 1, \cdots m$, it is well known the correlation matrix for $Y = \{Y_1, \cdots, Y_m\}$ has the form:

$$corr(Y_i) = \begin{bmatrix} 1 & \rho^1 & \rho^2 & \ldots & \rho^m \\ \rho^1 & 1 & \rho^1 & \ldots & \rho^{m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^m & \rho^{m-1} & \ldots & \rho^1 & 1_{m\times m} \end{bmatrix}$$

Exchangeable working correlation matrix

Another popular time structure for longitudinal data is the exchangeable structure with correlation matrix is:

$$corr(Y_i) = \begin{bmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \ldots & 1_{m\times m} \end{bmatrix}$$

Exchangeable working correlation matrix shows the observations of repeated measurements have only one common correlated parameter ρ. Exchangeable working correlation matrix is widely used in health study, assuming the the repeated measurements from patients of clinics have no time dependence.

Unstructured working correlation matrix

For an Unstructured working correlation matrix, with $t = 1, \cdots m$, the correlation matrix for $Y = \{Y_1, \cdots, Y_m\}$ has the form:

$$corr(Y_i) = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1m} \\ \rho_{12} & 1 & \rho_{22} & \cdots & \rho_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1m} & \rho_{2m} & \rho_{3m} & \cdots & 1_{m\times m} \end{bmatrix}$$

The Unstructured working correlation structure is the most general structure for fitting a GEE model. It has no fixed structure for estimating correlation matrix. It may not be available when we have unbalanced datasets because of the singular issues of matrix. We can use the QIC value for evaluation when comparing GEE models with different working correlation matrices.

3.4 Semi-Parametric GEE with One Kernel Smoother

Suppose that Y_{ij} and Z_{ij} are outcome and covariate scalars for subject i at time period j, $(i = 1, \cdots, N)$, $(j = 1, \cdots, n_i)$. Given some other covariates $\mathbf{X}_{ij}$, where $\mathbf{X}_{ij}^{'} = (X_{ij1}, \cdots, X_{ijp})$ is a $p \times 1$ vector. For semi-parametric regression, our model setup will be:

$$g(\mu_{ij}) = \mathbf{X}_{ij}^{'}\beta + \theta(Z_{ij}) \tag{3.11}$$

$g(.)$ is a known monotonic link function and $\theta(.)$ is a smooth function. For normal distributed responses, we use a identity link function and $g(\mu_{ij}) = \mu_{ij}$. $X_{ij}^{'}$ is the covariates for parametric part, indicating that we assume those covariates have linear relationship with response variable while Z_{ij} is the covariates for non-parametric part, denoting that we assume covariates shows pattern non-parametrically and β is a $p \times 1$ coefficient vector in linear part. We would like to show two methods for estimating this semi-parametric regression model with kernel smoother through generalized estimating equations.

3.4.1 Estimating Equations for Parametric GEE and Non-Parametric GEE

Zeger and Liang (1986) proposed generalized estimating equaitons (GEE) following concepts from the quasi-likelihood approach. For non-parametric model:

$$g(\mu_{ij}) = X_{ij}^{T}\beta, \tag{3.12}$$

They defined generalized estimating equations as:

$$\sum_{i=1}^{N}(\partial\mu_i/\partial\beta)^T V_i^{-1}[Y_i-\mu_i(\beta)]=0 \tag{3.13}$$

where Y_i and μ_i are vectors: $Y_i^{'}=(Y_{i1},\cdots,Y_{in_i})$, $\mu_i^{'}=(\mu_{i1},\cdots,\mu_{in_i})$,$\mu_{ij}=\mu_{ij}(\beta)=g^{-1}(\mathbf{X}_{ij}^{'}\beta)$. We would like to introduce a working correlation matrix to capture the within subject association:

$$V_i=\phi A_i^{1/2}R_i(\alpha)A_i^{1/2}, \tag{3.14}$$

where R_i is a working correlation matrix with parameter α. If R_i is the true correlation matrix for Y_i, V_i is the true covariance matrix for Y_i with $A_i=diag\{V_i\}$. If R_i is the identity matrix then V_i is a diagonal matrix that implies the response $(Y_{i1},\cdots,Y_{in_i})$ are independent when $(Y_{i1},\cdots,Y_{in_i})$ are normally distributed such as the Generalized Linear Models.

Lin and Carroll (2001) stated a local polynomial kernel smoother approach. Suppose that Y_{ij} is outcome, $(i=1,\cdots,N)$, $(j=1,\cdots,n_i)$. Let h denote the bandwidth parameter and $K(\cdot)$ is a symmetric kernel density function, either Epankochiv or Gaussian. Then the model and kernel estimating equations are:

$$g(\mu_{ij})=\theta(Z_{ij}),\quad i=1,\cdots,N,\quad j=1,\cdots,n_i \tag{3.15}$$

$$\sum_{i=1}^{N}Z_i^{'}\delta_i(z)K_{ih}^{1/2}(z)V_i^{-1}(z)K_{ih}^{1/2}(z)\{Y_i-\mu_i(z)\}=0, \tag{3.16}$$

respectively, where $K_{ih}(z)=diag\{K_h(Z_{ij}-z)\}$, $\mu_i(z)=\{\mu_{i1}(z),\cdots,\mu_{in_i}(z)\}^{'}$ with $\mu_{ij}(z)=g^{-1}\{Z_{ij}(z)^{'}\alpha\}$, $\delta_i=diag\{1/g^{(1)}\{\mu_{ij}(z)\}$, $V_i=S_i^{1/2}R_i(\gamma)S_i^{1/2}$, $S_i=diag[\phi^{-1}\nu\{\mu_{ij}(z)\}]$, R_i is working correlation matrix depend on γ. For non-symmetric local polynomial kernel GEEs, the estimating equation should be:

$$\sum_{i=1}^{N}Z_i^{'}\delta_i(z)V_i^{-1}(z)K_{ih}(z)\{Y_i-\mu_i(z)\}=0 \tag{3.17}$$

Through the estimating equations in 1.16, the local average kernel GEE estimator has a closed form solution:

$$\hat{\theta}_K(z)=\frac{\sum_{i=1}^{N}1_i^{'}K_{ih}^{1/2}(z)V_i^{-1}K_{ih}^{1/2}(z)Y_i}{\sum_{i=1}^{N}1_i^{'}K_{ih}^{1/2}(z)V_i^{-1}K_{ih}^{1/2}(z)1_i^{'}} \tag{3.18}$$

The SUR estimator is introduced by Wang (2003) has the estimating equation:

$$\sum_{N}^{i=1}\sum_{j=1}^{n_i}K_h(Z_{ij}-z)Z_{ij}(z)^{'}V_i^{-1}\{Y_i-\mu_{*j}(z;\alpha)\}=0, \tag{3.19}$$

Supposed that θ_K^* is the solution of estimating equation in 1.15, where V_i is working covariance matrix, $Z_{ij}(z)$ is an $n_i \times (d+1)$ matrix of zeros except the j_{th} row is $\{1, (Z_{ij}-z), ..., (Z_{ij}-z)^d\}'$. The l_{th} element of $\mu_{*j}(z;\alpha)$ is $\mu\{\hat{\alpha}_0 + \hat{\alpha}_1(Z_{il}-z)/h\}$ when $j = l$ and is $\hat{\theta}_K^{*l}(z)$ when $l \neq j$. $\hat{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_d)'$ is the solution from previous estimating equations.

Although an iteration path for the SUR estimator was proposed by Wang (2003), a closed form solution can be obtained (Lin, 2006) if we have an identity link. Suppose $\hat{\theta}_K^*(z)$ is the estimator:

$$\hat{\theta}_K^*(z) = K_{wh}^{'}(z)\{I + (\tilde{V}^{-1} - V^d)K_w\}^{-1}\tilde{V}^{-1}Y, \tag{3.20}$$

Where

$$K_{wh}(z) = \{\sum_{i=1}^{N}\sum_{j=1}^{n_i} K_h(Z_{ij}-z)v^{jj_i}\}^{-1}\{K_h(Z_{11}-z), ..., K_h(Z_{Nn_N}-z)\}' \tag{3.21}$$

Here $K_{wh}(z)$ is an $N \times 1$ vector and N^* is the total number of observations, $K_w = \{K_{wh}(Z_{11}), ..., K_{wh}(Z_{n_in_i})\}'$ is an $N^* \times N^*$ matrix, $\tilde{V} = diag(V_1, ..., V_N)$, $V_i^d = diag(v_i^{jj}) = diag(V_i^{-1})$, $\tilde{V}^d = diag(V_1^d, ..., V_1^n)$, and $Y = (Y_1^{'}, ..., Y_N^{'})'$.

3.4.2 Profile-Kernel Estimating Equations

Lin and Carroll (2001) proposed the profile kernel estimating methods for semi-parametric GEE model in 3.11, which have two steps: for a given β, we can estimate the non-parametric part using non-parametric estimating equations. After we get the non-parametric estimation result, traditional generalized estimating equation can be used to obtain β estimator.

Given β, the estimating equation for $\theta_K(t)$ estimator with symmetric kernel density function is:

$$\sum_{i=1}^{N} Z_i(z)' \triangle_i (X_i, z)K_{ih}^{1/2}(t)V_{1i}^{-1}(X_i, z)K_{ih}^{1/2}(z) \times \{Y_i - \mu_i(X_i, z)\} = 0 \tag{3.22}$$

For asymmetric kernel density function, the estimating equation is:

$$\sum_{i=1}^{N} Z_i(z)' \triangle_i (X_i, z)V_{1i}^{-1}(X_i, z)K_{ih}(t)\{Y_i - \mu_i(X_i, z)\} = 0 \tag{3.23}$$

where $K_{ih}(t) = diag\{K_h(Z_{ij}-z)\}$ is a kernel density function for i^{th} subject and $\mu_i(z) = E(Z_i)$. $\triangle_i(X_i, z) = diag\{\mu_{ij}^{(1)}\}$ and $\mu^{(1)}(.)$ is the first derivative of $\mu(.)$. $V_{1i}(X_i, Z_i) = S_i^{1/2}(X_i, Z_i)R_{1i}S_i^{1/2}(X_i, Z_i)$ and $S_i(X_i, Z_i) = diag\{\phi\omega_{ij}^{-1}V_i\}$, in which ϕ is a scale parameter and ω is weight. R_{1i} is an invertible working correlation matrix where we construct some structures such as AR(1) or exchangeable correlation forms.

After estimating the non-parametric part in 1.11, we can proceed to estimate β through solving the adjusted generalized estimating equations:

$$\sum_{i=1}^{N} \frac{\partial \mu\{X_i\beta + \hat{\theta}(Z_i;\beta)^T\}}{\partial \beta} \times V_{2i}^{-1}(X_i, Z_i) \times [Y_i - \mu\{X_i\beta + \hat{\theta}(Z_i;\beta)\}] = 0 \tag{3.24}$$

where $\hat{\theta}(Z_i;\beta) = \{\hat{\theta}(Z_{i1};\beta), \hat{\theta}(Z_{in_i};\beta)\}$, $V_{2i}(X_i, Z_i) = S_i^{1/2}(X_i, Z_i)R_{2i}S_i^{1/2}(X_i, Z_i)$, $S_i(X_i, Z_i) = diag\{\phi\omega_{ij}^{-1}V[\mu\{X_{ij}'\beta + \hat{\theta}(Z_{ij};\beta)\}]\}$, R_{2i} is a working correlation matrix for this equation.

For semi-parametric regression, the profile kernel estimator by Lin and Carroll (2001) can be obtained by backfitting steps in Algorithm 1. However, Fan and Li (2004) showed that this estimator has a closed form solution.

$$\theta_K = A_k(Y - X\beta) \tag{3.25}$$

$$\hat{\beta}_K = \{X'(I - A_k)\tilde{V}^{-1}(I - A_k)X\}^{-1}X'(I - A_k)\tilde{V}^{-1}(I - A_k)Y \tag{3.26}$$

where X is covariates matrix and Y is response variable. A_K is the coefficient for non-parametric regression estimator and $\tilde{V} = diag(V)$. If we write $\hat{\beta}_K = H_K Y$, then $Cov(\hat{\beta}_K) = H_K\tilde{\Sigma}H_K'$, $\tilde{\Sigma} = diag(\Sigma_i)$ and Σ_i is the true correlation matrix for Y.

As we discussed in Literature review part, Lin and Carroll (2001) claimed that the asymptotic properties for profile kernel estimator has properties that when $R_{1i} = R_{2i} = I$, the estimator $\hat{\beta}_K$ will be consistent and efficient. However using independent working correlation matrix will ignore association within subjects, which conflicts the properties for parametric GEE estimator (Zeger and Liang 1986) and violates the clustering properties in longitudinal data studies. We need another estimation method to take within subject association into account and place different weights according to different clusters.

3.4.3 Profile SUR Kernel Estimator

Wang, Carroll and Lin (2004) presented an estimation method based on the seemingly unrelated kernel estimator (Wang 2003), which fulfill both consistency and efficiency for within-subject association. Specifically, the estimation method in Wang, Carroll and Lin (2004) also requires backfitting steps iteration:

Step 1 : Let $\tilde{\theta}(\cdot)$ be the current estimator of $\theta(\cdot)$. Given β, let $\hat{\alpha} = \hat{\alpha}(z,\beta) = \{\hat{\alpha_0}(z,\beta), \hat{\alpha_1}(z,\beta)\}^{'}$ be the solution to the kernel equation

$$\sum_{i=1}^{N}\sum_{j=1}^{n_i} K_h(z - Z_{ij})\mu_{ij}^{(1)}(\beta,\hat{\alpha})G_{ij}^{'}(z)V_i^{-1} \times [Y_i - \mu^{\star}\{z, X_i, Z_i, \beta, \hat{\alpha}, \tilde{\theta}(Z_i;\beta)\}] = 0 \tag{3.27}$$

where the k_{th} element of $\mu^{*}\{z, X_i, Z_i, \beta, \hat{\alpha}, \tilde{\theta}(Z_i;\beta)\}$ is

$$\mu[X_{ik}^{t}\beta + I(k=j)\{\hat{\alpha_0} + \hat{\alpha_1}(z - Z_{ij})/h\} + I(k \neq j)\tilde{\theta}(Z_{ik},\beta)] \tag{3.28}$$

$\mu_{ij}^{(1)}$ is the first derivative of the function $\mu(\cdot) = g^{-1}(\cdot)$ evaluated at $X_{ij}^{'}\beta + \hat{\alpha_0} + \hat{\alpha_1}(z - Z_{ij})/h$. The updated estimator of $\theta(z)$ is $\hat{\theta}(z,\beta) = \hat{\alpha_0}(z,\beta)$ and $G_{ij}(z)$ is an $n_i \times 2$ matrix with the kth column $e_j \times \{(z - Z_{ij})/h\}^{k-1}(k = 1, 2)$, where e_j is an $n_i \times 1$ vector of zeros except with the k_{th} entry being 1 and h denotes bandwidth parameter.

Step 2: Find $\hat{\beta}$ by solving the profile estimating equation:

$$\sum_{i=1}^{n} \frac{\partial \mu\{X_i\beta + \hat{\theta}(Z_i;\beta)^T\}}{\partial \beta} V_{1i}^{-1}(X_i, Z_i) \times [Y_i - \mu\{X_i\beta + \hat{\theta}(Z_i;\beta)\}] = 0 \tag{3.29}$$

Then we can run a full iteration through those backfitting steps.

Follow the properties of semi-parametric regression, still, we use a closed form provided by Fan and Li (2004):

$$\theta_K^{*} = A_k^{*}(Y - X\beta^{*}) \tag{3.30}$$

$$\hat{\beta}_K^{*} = \{X^{'}(I - A_k^{*})\tilde{V}^{-1}(I - A_K^{*})X\}^{-1}X^{'}(I - A_k^{*})\tilde{V}^{-1}(I - A_k^{*})Y \tag{3.31}$$

Where X and Y are same as last section and A_K^{*} is the coefficient for non-parametric regression estimator and $\tilde{V} = diag(V)$. Still, if we write $\beta_K^{*} = H_K^{*}Y$, then $Cov(\hat{\beta^{*}}_K) = H_K^{*}\tilde{\Sigma}H_K^{*'}$. Wang, Carroll and Lin (2004) displayed the asymptotic properties that $\hat{\beta}_K^{*}$ is consistent with any correlation matrix and when correlation and covariance matrix show the true association within subjects, $\hat{\beta}_K^{*}$ achieves semi-parametric efficiency.

3.5 Semi-Parametric GEE with Multiple Kernel Smoothers

Suppose that Y_{ij} is a outcome scalars for subject i at time period j, $(i = 1, \cdots, N)$, $(j = 1, \cdots, n_i)$. Given some other covariates $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$, where $\mathbf{X}_{ij}^T = (X_{ij1}, \cdots, X_{ijp})$ is a $p \times 1$ vector; $\mathbf{Z}_{ij}^T = (Z_{ij1}, \cdots, Z_{ijq})$ is a $q \times 1$ vector. For semi-parametric regression, our model setup will be:

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T \beta + \sum_{d=1}^{q} \theta(\mathbf{Z}_{ijd}) \tag{3.32}$$

where $g(.)$ is a known monotonic link function and $\theta(.)$ are kernel smooth functions. For normal distributed responses, we use a identity link function and $g(\mu_{ij}) = \mu_{ij}$; for binary response, we use a logit link function and $g(\mu_{ij}) = \frac{\pi_{ij}}{1-\pi_{ij}}$ with π_{ij} is the probability when $Y_{ij} = 1$. $\mathbf{X}_{ij}^T$ is the covariates for parametric part while $\mathbf{Z}_{ij}^T$ is the covariates for non-parametric part and β is a $p \times 1$ coefficient vector in parametric part. Still, we would like to provide profile-kernel estimator and profile SUR estimator for this semi-parametric regression model with multiple kernel smoothers.

3.5.1 Generalized Additive Models

Hastie and Tibshirani (1990) proposed the local scoring procedure ACE algorithm for fitting Generalized Additive model. Suppose Y is response variable with a distribution of an exponential family and $(X_1, X_2, \cdots, X_p)$ are the predictors associated with additive terms. The mean $\mu = E(Y|X_1, X_2, \cdots, X_p)$ and the predictors can be connected by a link function $g(.)$:

$$g(\mu) = \alpha + \sum_{j=1}^{p} f_j(X_j) \tag{3.33}$$

where $f_j(X_j)$ denotes every smoother term in additive models.

Algorithm 2 Fitting Generalized Additive Model

1: $\alpha = g(\sum_1^n y_i/n)$; $f_1^0 = f_2^0, \cdots, = f_p^0$ ▷ Initialize
2: Given the rest terms $f_2^0, \cdots, f_p^0$, we can estimate and update the first term $f_1(X_1)$. ▷ update one term
3: We can proceed to estimate and update the new term f_l by fix other terms $f_1, \cdots, f_{l-1}, f_{l+1}, \cdots, f_p$ ▷ update the other terms
4: A criterion can be calculate for convergence: $\frac{\sum_{j=1}^{p} ||f_j^{new} - f_j^{old}||}{\sum_{j=1}^{p} ||f_j^{old}||}$ ▷ Convergence for some small threshold

3.5.2 Profile-Kernel Estimating Equations

We follow Lin and Carroll's method (2001), using back-fitting algorithm to calculate the profile-kernel estimator, which have three steps in general: for a given β and other kernel smoother terms, we can estimate one of the non-parametric term using non-parametric estimating equations. After we estimate that non-parametric term, we can estimate the rest kernel smoother terms and after we finished the estimator of all non-parametric terms, traditional generalized estimating equation can be used to obtain β estimator.

Suppose we have a semi-parametric model with two kernel smoother terms:

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T\beta + \theta_1(Z_{ij1}) + \theta_2(Z_{ij2}) \tag{3.34}$$

where we define $\mathbf{Z}_{ij}^T$ is covariates for non-parametric part and $\mathbf{Z}_{ij}^T = (Z_{ij1}, Z_{ij2})$.

Step 1: Given β and $\theta_1(Z_{ij1})$, the estimating equation for $\theta_2(z)$ is:

$$\sum_{i=1}^{N} \mathbf{Z}_{i2}(z)^T \triangle_i (X_i, z, \mathbf{Z}_{i1}) K_{ih}^{1/2}(z) V_{1i}^{-1}(X_i, z, \mathbf{Z}_{i1}) K_{ih}^{1/2}(z) \times \{\mathbf{Y}_i - \mu_i(X_i, z, \mathbf{Z}_{i1})\} = 0 \tag{3.35}$$

where $Z_{i2}(z)$ is an $n_i \times (r+1)$ matrix with the j^{th} row is $\{1, (Z_{ij2} - z), ..., (Z_{ij2} - z)^r\}^T$. For kernel average estimator, $r = 0$. $\mathbf{Y}_i$ and $\boldsymbol{\mu}_i$ are vectors: $\mathbf{Y}_i^T = (Y_{i1}, \cdots, Y_{in_i})$, $\boldsymbol{\mu}_i^T = (\mu_{i1}, \cdots, \mu_{in_i})$, $\mu_{ij} = E(Y_{ij}) = \mu_{ij}(\beta) = g^{-1}(\mathbf{X}_{ij}^T\beta + \theta_1(Z_{ij1}) + \theta_2(z))$ and we use identity link function. $K_{ih}(z) = diag\{K_h(Z_{ij2} - z)\}$ are kernel weighs of the target value for i^{th} subject. $\triangle_i(X_i, \mathbf{Z}_{i1}, z) = diag\{\mu_{ij}^{(1)}\}$ and $\mu^{(1)}(.)$ is the first derivative of $\mu(.)$. $V_{1i}(X_i, z, \mathbf{Z}_{i1}) = S_i^{1/2}(X_i, \mathbf{Z}_{i1}, z) R_{1i} S_i^{1/2}(X_i, \mathbf{Z}_{i1}, z)$ and $S_i(X_i, \mathbf{Z}_{i1}, z) = diag\{\phi\omega_{ij}^{-1} V_i\}$, in which ϕ is a scale parameter and ω is known weight. R_{1i} is an invertible working correlation matrix for $\theta_2(z)$ where we construct some structures such as AR(1) or exchangeable correlation forms.

Through the estimating equations in 1.4, the local average kernel GEE estimator has a closed form solution:

$$\hat{\theta}_{2K}(z) = \frac{\sum_{i=1}^{N} 1_i^T K_{ih}^{1/2}(z) V_{1i}^{-1} K_{ih}^{1/2}(z)(Y_i - \mathbf{X}_{ij}^T\beta - \theta_1(Z_{ij1}))}{\sum_{i=1}^{N} 1_i^T K_{ih}^{1/2}(z) V_i^{-1} K_{ih}^{1/2}(z) 1_i^T} \tag{3.36}$$

Step 2: After we obtain $\hat{\theta}_2(Z_{ij2})$ and given β, we can proceed to calculate $\theta_1(z)$:

$$\sum_{i=1}^{N} \mathbf{Z}_{i1}(z)^T \triangle_i (\mathbf{X}_i, z, \mathbf{Z}_{i2}) K_{ih}^{1/2}(z) V_{2i}^{-1}(\mathbf{X}_i, z, \mathbf{Z}_{i2}) K_{ih}^{1/2}(z) \times \{\mathbf{Y}_i - \mu_i(\mathbf{X}_i, z, \mathbf{Z}_{i2})\} = 0 \tag{3.37}$$

where $Z_{i1}(z)$ is an $n_i \times (r+1)$ matrix with the j^{th} row is $\{1, (Z_{ij1}-z), ..., (Z_{ij1}-z)^r\}^T$ and $\mu_{ij} = \mu_{ij}(\boldsymbol{\beta}) = g^{-1}(\mathbf{X}_{ij}^T\beta + \hat{\theta}_2(Z_{ij2}) + \theta_1(z))$. $K_{ih}(z) = diag\{K_h(Z_{ij1}-z)\}$ is the kernel weights assigned for each target value point of Z_{ij1}. $\triangle_i(X_i, \mathbf{Z}_{i2}, z) = diag\{\mu_{ij}^{(1)}\}$ and $\mu^{(1)}(.)$ is the first derivative of $\mu(.)$. $V_{2i}(X_i, z, \mathbf{Z}_{i2}) = S_i^{1/2}(X_i, \mathbf{Z}_{i2}, z)R_{2i}S_i^{1/2}(X_i, \mathbf{Z}_{i2}, z)$ and $S_i(X_i, \mathbf{Z}_{i2}, z) = diag\{\phi\omega_{ij}^{-1}V_i\}$, in which R_{2i} is an invertible working correlation matrix for $\theta_1(z)$.

Still, through the estimating equations, the local average kernel GEE estimator has a closed form solution:

$$\hat{\theta}_{1K}(z) = \frac{\sum_{i=1}^N 1_i^T K_{ih}^{1/2}(z)V_{2i}^{-1}K_{ih}^{1/2}(z)(Y_i - \mathbf{X}_{ij}^T\beta - \theta_2(Z_{ij2}))}{\sum_{i=1}^N 1_i^T K_{ih}^{1/2}(z)V_{2i}^{-1}K_{ih}^{1/2}(z)1_i^T} \tag{3.38}$$

Step 3: After estimating the non-parametric parts $\hat{\theta}_1(Z_{ij1})$ and $\hat{\theta}_2(Z_{ij2})$, we can proceed to estimate $\boldsymbol{\beta}$ through solving the adjusted generalized estimating equations:

$$\sum_{i=1}^N \frac{\partial\mu\{\mathbf{X}_i\beta + \hat{\theta}_1(\mathbf{Z}_{1i};\beta) + \hat{\theta}_2(\mathbf{Z}_{2i};\beta)\}}{\partial\beta} \times V_{3i}^{-1}(\mathbf{X}_i, \mathbf{Z}_{i1}, \mathbf{Z}_{i2}) \times [\mathbf{Y}_i - \mu\{\mathbf{X}_i\beta + \hat{\theta}_1(Z_{1i};\beta) + \hat{\theta}_2(Z_{2i};\beta)\}] = 0 \tag{3.39}$$

where $\hat{\theta}_1(\mathbf{Z}_{1i};\beta) = \{\hat{\theta}(Z_{i11};\beta), \cdots, \hat{\theta}(Z_{in_i1};\beta)\}$, $\hat{\theta}_2(\mathbf{Z}_{2i};\beta) = \{\hat{\theta}(Z_{i12};\beta), \cdots, \hat{\theta}(Z_{in_i2};\beta)\}$, $V_{3i}(X_i,$ $\mathbf{Z}_{i1}, \mathbf{Z}_{i2}) = S_i^{1/2}(X_i, \mathbf{Z}_{i1}, \mathbf{Z}_{i2})R_{3i}S_i^{1/2}(X_i, \mathbf{Z}_{i1}, \mathbf{Z}_{i2})$, $S_i(\mathbf{X}_i, \mathbf{Z}_{i1}, \mathbf{Z}_{i2}) = diag\{\phi\omega_{ij}^{-1}V[\mu\{X_{ij}^T\beta + \hat{\theta}(Z_{i1};\beta) + \hat{\theta}(Z_{i2};\beta)\}]\}$ and R_{3i} is a working correlation matrix.

The estimating equation has a closed form solution for β:

$$\hat{\beta}_K = \{X'(I - A_{1k} - A_{2k})\tilde{V}^{-1}(I - A_{1k} - A_{2k})X\}^{-1}X'(I - A_{1k} - A_{2k})\tilde{V}^{-1}(I - A_{1k} - A_{2k})Y \tag{3.40}$$

where X is covariates matrix and Y is response variable. A_{1K} is the coefficient for non-parametric regression estimator $\theta_1(Z_{ij1})$, A_{2K} is the coefficient for non-parametric regression estimator $\theta_2(Z_{ij2})$ and $\tilde{V} = diag(V)$. If we write $\hat{\beta}_K = H_K Y$, then $Cov(\hat{\beta}_K) = H_K\tilde{\Sigma}H_K'$, $\tilde{\Sigma} = diag(\Sigma_i)$ and Σ_i is the true correlation matrix for Y.

Once we obtain $\hat{\beta}$, we can update $\theta_2(z)$ and $\theta_1(z)$ until convergence.

3.5.3 Profile SUR Kernel Estimator

Following Wang, Carroll and Lin's method (2004), we propose the SUR kernel estimator for semi-parametric model with two kernel smoother terms. Still, a back-fitting 3 steps iteration can be used for the estimation.

Step 1 : Let $\tilde{\theta}_2(\cdot)$ be the current estimator of $\theta_2(\cdot)$. Given β and $\theta_1(Z_{ij1})$, let $\hat{\alpha} = \hat{\alpha}(z, \beta, Z_{ij2}) = \{\hat{\alpha_0}(z, \beta, Z_{ij2}), \hat{\alpha_2}(z, \beta, Z_{ij2}), \cdots, \hat{\alpha_r}(z, \beta, Z_{ij2})\}^T$ be the solution to the kernel equation

$$\sum_{i=1}^{N}\sum_{j=1}^{n_i} K_h(Z_{ij2} - z)\mu_{ij}^{(1)}(\beta, \hat{\alpha})G_{ij}^{'}(z)V_i^{-1} \times [Y_i - \mu^{\star}\{z, X_i, \boldsymbol{Z}_{i1}, \beta, \hat{\alpha}, \tilde{\theta}_2(\boldsymbol{Z}_{i2}; \beta)\}] = 0 \qquad (3.41)$$

where the k^{th} element of $\mu^{\star}\{z, X_i, \boldsymbol{Z}_{i1}, \beta, \hat{\alpha}, \tilde{\theta}_2(\boldsymbol{Z}_{i2}; \beta)\}$ is

$$\mu[X_{ik}^T\beta + I(k = j)\{\hat{\alpha}_0 + \hat{\alpha}_1(Z_{ij2} - z)/h + \cdots + \hat{\alpha}_d(Z_{ij2} - z)/h\} + I(k \neq j)\tilde{\theta}(Z_{i2k}, \beta)] \qquad (3.42)$$

and $\mu_{ij}^{(1)}$ is the first derivative of the function $\mu(\cdot) = g^{-1}(\cdot)$ evaluated at $(X_{ij}^T\beta + \hat{\alpha}_0 + \hat{\alpha}_1(Z_{ij2} - z)/h + \cdots + \hat{\alpha}_d(Z_{ij2} - z)/h)$. The updated estimator of $\theta_2(z)$ is $\hat{\theta}_2(z, \beta, Z_{ij1}) = \hat{\alpha_0}(z, \beta, Z_{ij1})$ and $G_{ij}(z)$ is an $n_i \times (r+1)$ matrix of zeros except the j^{th} column is $e_{ij} \times (Z_{ij2} - z)^r\}^T$, where e_j is an $n_i \times 1$ vector of zeros except with the k^{th} entry being 1 and h denotes the bandwidth parameter.

A closed form solution with identity link can be obtained by:

$$\hat{\theta}_{2K}^{*}(z) = K_{wh}^{'}(z)\{I + (\tilde{V}^{-1} - V^d)K_w\}^{-1}\tilde{V}^{-1}(Y - X\beta - \theta_1(Z_{ij1})), \qquad (3.43)$$

Where

$$K_{wh}(z) = \{\sum_{i=1}^{N}\sum_{j=1}^{n_i} K_h(Z_{ij2} - z)v_i^{jj}\}^{-1}\{K_h(Z_{112} - z), ..., K_h(Z_{Nn_{N2}} - z)\}^T \qquad (3.44)$$

Here $K_{wh}(z)$ is an $N \times 1$ vector and N^* is the total number of observations, $K_w = \{K_{wh}(Z_{112}), ..., K_{wh}(Z_{n_i n_i 2})\}^{'}$ is an $N^* \times N^*$ matrix, $\tilde{V} = diag(V_1, ..., V_N)$, $V_i^d = diag(v_i^{jj}) = diag(V_i^{-1})$, $\tilde{V}^d = diag(V_1^d, ..., V_1^n)$, and $Y = (Y_1^{'}, ..., Y_N^{'})^{'}$.

Step 2: Given β and $\hat{\theta}_2(Z_{ij2})$ we obtained from last step, $\theta_1(z)$ can be calculated by: Let $\tilde{\theta}_1(\cdot)$ be the current estimator of $\theta_1(\cdot)$ and let $\hat{\alpha} = \hat{\alpha}(z, \beta, Z_{ij1}) = \{\hat{\alpha_0}(z, \beta, Z_{ij1}), \hat{\alpha_2}(z, \beta, Z_{ij2}), \cdots, \hat{\alpha}_d(z, \beta, Z_{ij1})\}^T$ be the solution to the kernel equation

$$\sum_{i=1}^{N}\sum_{j=1}^{n_i} K_h(Z_{ij1} - z)\mu_{ij}^{(1)}(\beta, \hat{\alpha})G_{ij}^{'}(z)V_i^{-1} \times [Y_i - \mu^{\star}\{z, X_i, \boldsymbol{Z}_{i1}, \beta, \hat{\alpha}, \tilde{\theta}_1(\boldsymbol{Z}_{i1}; \beta)\}] = 0 \qquad (3.45)$$

where the k^{th} element of $\mu^{\star}\{z, X_i, \boldsymbol{Z}_{i2}, \beta, \hat{\alpha}, \tilde{\theta}_1(\boldsymbol{Z}_{i1}; \beta)\}$ is

$$\mu[X_{ik}^T\beta + I(k = j)\{\hat{\alpha}_0 + \hat{\alpha}_1(Z_{ij1} - z)/h + \cdots + \hat{\alpha}_d(Z_{ij2} - z)/h\} + I(k \neq j)\tilde{\theta}(Z_{i1k}, \beta)] \qquad (3.46)$$

and $\mu_{ij}^{(1)}$ is the first derivative of the function $\mu(\cdot) = g^{-1}(\cdot)$ evaluated at $(X_{ij}^T\beta + \hat{\alpha}_0 + \hat{\alpha}_1(Z_{ij1} - z)/h + \cdots + \hat{\alpha}_d(Z_{ij1} - z)/h)$. The updated estimator of $\theta_1(z)$ is $\hat{\theta}_1(z, \beta, Z_{ij2}) = \hat{\alpha_0}(z, \beta, Z_{ij2})$ and

$G_{ij}(z)$ is an $n_i \times (r+1)$ matrix of zeros except the j^{th} column is $e_{ij} \times (Z_{ij1} - z)^r\}^T$, where e_j is an $n_i \times 1$ vector of zeros except with the k^{th} entry being 1 and h denotes bandwidth parameter.

A closed form solution with identity link can be obtained by:

$$\hat{\theta}^*_{1K}(z) = K'_{wh}(z)\{I + (\tilde{V}^{-1} - V^d)K_w\}^{-1}\tilde{V}^{-1}(Y - X\beta - \theta_2(Z_{ij2})), \tag{3.47}$$

where

$$K_{wh}(z) = \{\sum_{i=1}^{N}\sum_{j=1}^{n_i} K_h(Z_{ij1} - z)v_i^{jj}\}^{-1}\{K_h(Z_{111} - z), ..., K_h(Z_{Nn_{N1}} - z)\}^T \tag{3.48}$$

Here $K_{wh}(z)$ is an $N \times 1$ vector and N^* is the total number of observations, $K_w = \{K_{wh}(Z_{111}), ...,$ $K_{wh}(Z_{n_i n_i 1})\}'$ is an $N^* \times N^*$ matrix, $\tilde{V} = diag(V_1, ..., V_N)$, $V_i^d = diag(v_i^{jj}) = diag(V_i^{-1})$, $\tilde{V}^d = diag(V_1^d, ..., V_1^n)$, and $Y = (Y_1', ..., Y_N')'$.

Step 3: After we obtain the estimators for two kernel smoothers, we can calculate $\hat{\beta}$ by solving the adjusted estimating equation:

$$\sum_{i=1}^{n} \frac{\partial \mu\{X_i\beta + \hat{\theta}_1(Z_{1i};\beta)^T + \hat{\theta}_2(Z_{2i};\beta)^T\}}{\partial \beta} V_{1i}^{-1}(X_i, Z_i) \times [Y_i - \mu\{X_i\beta + \hat{\theta}_1(Z_{1i};\beta)\} + \hat{\theta}_2(Z_{2i};\beta)^T] = 0 \tag{3.49}$$

and still, we can update β by:

$$\theta^*_{1K} = A^*_{1k}(Y - X\beta^* - \theta^*_{2K}) \tag{3.50}$$

$$\theta^*_{2K} = A^*_{2k}(Y - X\beta^* - \theta^*_{1K}) \tag{3.51}$$

$$\hat{\beta}^*_K = \{X'(I - A^*_{1k} - A^*_{2k})\tilde{V}^{-1}(I - A^*_{1k} - A^*_{2k})X\}^{-1}X'(I - A^*_{1k} - A^*_{2k})\tilde{V}^{-1}(I - A^*_{1k} - A^*_{2k})Y \tag{3.52}$$

Where X and Y are same as last section and A^*_{1K} is the coefficient for non-parametric regression estimator $\theta_1(Z_{ij1})$, A^*_{2K} is the coefficient for non-parametric regression estimator $\theta_2(Z_{ij2})$ and $\tilde{V} = diag(V)$. Still, if we write $\beta^*_K = H^*_K Y$, then $Cov(\hat{\beta^*}_K) = H^*_K \tilde{\Sigma} H^{*'}_K$.

Then we can run a full iteration through those backfitting steps until convergence.

CHAPTER 4

SIMULATION

In this Chapter, simulations are conducted for comparing different estimation methods. Bias, standard deviation, and mean square error for estimators will be used to evaluate the performance of different approaches in parametric and semi-parametric models. Different scenarios based on local polynomial kernel GEE estimator and the SUR estimator will be used to display when and which unbiased estimator will achieve least standard deviation under given conditions. Simulations in this part are performed by using R language, and the package "kedd" is used to compute kernel densities and cross validation results.

4.1 Simulation Set-Up

4.1.1 Semi-Parametric Model with One Kernel Smoother

Consider a model with non-parametric part and linear part in the form:

$$Y_{ij} = X_{ij}^{'}\beta + \theta(Z_{ij}) + \epsilon_{ij}, \quad i = 1, ..., n \quad and \quad j = 1, ..., m \tag{4.1}$$

where i denotes the i^{th} subject and j denotes the j^{th} time point. In the equation, $\theta(\cdot)$ is a kernel smooth function, Z_{ij} denotes covariate in the non-parametric part, X_{ij} denotes covariates in the parametric part and β is the coefficient vector. For estimating the non-parametric part $\theta(Z_{ij})$, epanchikov and gaussian density kernels will be used to construct kernel weights in non-parametric smoother and least square cross validation method (Silverman 1986) will be used to select band-width parameter h which is critical for kernel regression models.

In this simulation, data is generated with the following set-up:

- Each run with 100 subjects, each subject with and 4 or 10 time points and 200 replicates.
- $\theta(Z_{ij}) = sin(4 \times Z_{ij})$ in the first setup and $\theta(Z_{ij}) = exp(sin(2/Z_{ij}))$ in the second setup.
- X_{ij} and Z_{ij} are both scalars and time-varying covariates with $X_{ij} = b_{ij} + e_{1ji}$, $Z_{ij} = b_{ij} + e_{2ji}$. $b_{ij} \sim U[0, 1]$, where e_{1ij} and e_{2ij} are independent to each other and follow uniform distribution $U[-2, 2]$. The model in equation 4.1 becomes:

$$Y_{ij} = X_{ij}\beta + \theta(Z_{ij}) + \epsilon_{ij}, \quad i = 1, ..., n \quad and \quad j = 1, ..., m \tag{4.2}$$

- $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{in})^{'}$ is a vector that follows multivariate normal distribution with mean zero and correlation coefficient matrix R_i, which is an AR(1) working correlation matrix with a lower entry ρ=0.3 and upper entry ρ=0.7 respectively.

For estimating the semi-parametric model in (4.2), semi-WI estimator with independent working correlation matrix $R_i = I$ and semi-True estimator with true working correlation matrix R_i will be used in different scenarios. Parametric estimating approaches, such as estimators based on the following three parametric models and one generalized additive model, will also be used in this simulation:

$$para1: \quad Y_{ij} = Z_{ij} + X_{ij}\beta + \epsilon_{ij}, \tag{4.3}$$

$$para2: \quad Y_{ij} = exp(Z_{ij}) + X_{ij}\beta + \epsilon_{ij}, \tag{4.4}$$

$$para3: \quad Y_{ij} = Z_{ij} + Z_{ij}^2 + Z_{ij}^3 + X_{ij}\beta + \epsilon_{ij}, \tag{4.5}$$

and

$$gam: \quad Y_{ij} = spline(Z_{ij}) + X_{ij}\beta + \epsilon_{ij} \tag{4.6}$$

In this simulation part, we will focus on the estimation of β and overall fitting of different estimators. Mean and standard deviation of the estimated β will be displayed. Overall fitting performance of different approaches will be examined based on the mean square error. Training dataset and test dataset will be used to evaluate the performance of semi-parametric models, parametric models, and GAM model.

4.1.2 Semi-Parametric Model with Multiple Kernel Smoothers

Consider another model with two kernel smoothers in non-parametric part:

$$Y_{ij} = X_{ij}^{'}\beta + \theta_1(Z_{ij1}) + \theta_2(Z_{ij2}) + \epsilon_{ij}, \quad i = 1, ..., n \quad and \quad j = 1, ..., m \tag{4.7}$$

Still, i denotes the i^{th} subject and j denotes the j^{th} time point. In the equation, $\theta_1(\cdot)$ and $\theta_2(\cdot)$ are kernel smooth functions, Z_{ij1} and Z_{ij1} denote the covariates in the non-parametric part, X_{ij} denotes the covariates in the parametric part and β is the coefficient vector. Like the section(), epanchikov and gaussian density kernels will be used to construct kernel weights in non-parametric smoothers and least square cross validation method will be used to select bandwidth parameter h which is critical for kernel regression models.

In this simulation, data is generated with the following set-up:

- Each run with 100 subjects, each subject with and 4 or 10 time points and 200 replicates.
- The first setup is $\theta_1(Z_{ij1}) = sin(4 \times Z_{ij1})$ in the first non-parametric term and $\theta_2(Z_{ij2}) = sin(4 \times Z_{ij2})$ in the second non-parametric term; the second setup is $\theta_1(Z_{ij1}) = exp(2/Z_{ij2})$ in the first non-parametric term and $\theta_2(Z_{ij2}) = exp(2/Z_{ij2})$ in the second non-parametric term.
- X_{ij}, Z_{ij1} and Z_{ij2} are all scalars and time-varying covariates with $X_{ij} = b_{ij} + e_{1ji}$, $Z_{ij1} = b_{ij} + e_{2ji}$ and $Z_{ij2} = b_{ij} + e_{3ji}$. We set $b_{ij} \sim U[0, 1]$, where e_{1ij}, e_{2ij} and e_{3ij} are independent to each other and follow uniform distribution $U[-2, 2]$. The model in equation 4.1.2 becomes:

$$Y_{ij} = X_{ij}\beta + \theta_1(Z_{ij1}) + \theta_2(Z_{ij2}) + \epsilon_{ij}, \quad i = 1, ..., n \quad and \quad j = 1, ..., m \tag{4.8}$$

- The error term $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{in})'$ is a vector that follows multivariate normal distribution with mean zero and correlation coefficient matrix R_i, which is an AR(1) working correlation matrix with a lower entry ρ=0.3 and upper entry ρ=0.7 respectively.

We use the same setting for working correlation matrix and parametric models as section (): semi-WI estimator with independent working correlation matrix $R_i = I$ and semi-True estimator with true working correlation matrix R_i will be used in estimating non-parametric part. Estimators based on the following three parametric models and one generalized additive model will also be used in this simulation:

$$para1: \quad Y_{ij} = Z_{ij1} + Z_{ij2} + X_{ij}\beta + \epsilon_{ij}, \tag{4.9}$$

$$para2: \quad Y_{ij} = exp(Z_{ij1}) + exp(Z_{ij2}) + X_{ij}\beta + \epsilon_{ij}, \tag{4.10}$$

$$para3: \quad Y_{ij} = Z_{ij1} + Z_{ij1}^2 + Z_{ij1}^3 + Z_{ij2} + Z_{ij2}^2 + Z_{ij2}^3 + X_{ij}\beta + \epsilon_{ij}, \tag{4.11}$$

and

$$gam: \quad Y_{ij} = spline(Z_{ij1}) + spline(Z_{ij2}) + X_{ij}\beta + \epsilon_{ij} \tag{4.12}$$

In this simulation, we will focus on the estimation of β and overall fitting of different estimators. Mean and standard deviation of the estimated β will be displayed. Overall fitting performance of different approaches will be examined based on the mean square error. Training dataset and test dataset will be used to evaluate the performance of semi-parametric models, parametric models, and GAM model.

Table 4.1 β estimates using the Epanechnikov kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta_{ij} = sin(4 \times Z_{ij})$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	0.995	0.021	1.000	0.017
semi-True	0.991	0.023	0.995	0.017
para1	0.996	0.063	1.006	0.063
para2	1.000	0.062	1.007	0.063
para3	0.983	0.063	1.000	0.063

4.2 Local Kernel Estimator with One Kernel Smoother

In this section, we first show the results of local polynomial kernel GEE estimator with one kernel smoother for semi-parametric regression and other estimators for parametric regression with various scenarios such as different kernel densities, correlation entries and time periods as we discussed in Section 4.1.

Table 4.1 shows β estimates in the semi-parametric model (4.2) and parametric models in (4.3-4.6) and local kernel average estimators with Epanechnikov kernel density and least square cross validation methods are used in the calculations. From the table, we can see that the semi-WI estimates have the smallest standard error 0.021, which is slightly better than the standard error of the semi-True estimates. More over, the standard errors of the β estimates based on semi-WI and semi-True are at least 3 times less than the standard errors of the estimators from the three parametric models (para1-para3). In addition, we found that the standard errors for $\rho = 0.7$ are slightly lower than those for $\rho = 0.3$ in the semi-parametric estimators.

Table 4.2 shows overall fitting mean square errors in the semi-parametric model (4.2) and parametric models (4.3-4.6). For the training dataset, the mean square errors in the parametric estimators and GAM estimator are higher than the mean square errors in semi-parametric estimators. For the test dataset, the mean square errors in parametric estimators and GAM estimator are larger than those for the semi-parametric estimators. Among the parametric models, polynomial model (para3 in equation (4.5)) has the smallest mean square error. Mean square errors for training data are slightly higher than testing data for all models.

Table 4.3 shows β estimates in the semi-parametric model (4.2) and parametric models in (4.3-4.6). The results in Table 4.3 based on the Gaussian Kernel density are very similar to the results

Table 4.2 Overall MSE using the Epanechnikov kernel and 200 replications. Each of the 100 subjects with 4 time points and $\theta_{ij} = sin(4 \times Z_{ij})$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.229	1.236	1.476	1.479	1.445	1.466
$\rho = 0.7$	1.227	1.243	1.484	1.487	1.455	1.470
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.305	1.310	1.507	1.507	1.506	1.507
$\rho = 0.7$	1.300	1.310	1.519	1.517	1.515	1.516

Table 4.3 β estimates using the Gaussian kernel and 200 replications. Each of the 100 subjects with 4 time points and $\theta_{ij} = sin(4 \times Z_{ij})$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	0.998	0.019	0.994	0.014
semi-True	0.993	0.021	0.997	0.016
para1	1.005	0.068	1.003	0.065
para2	1.009	0.068	1.005	0.065
para3	0.992	0.069	0.997	0.066

in Table 4.1 using the Epanechnikov kernel density: the standard errors of the β estimates based on semi-parametric estimators are at least 3 times less than the standard errors of the estimators from the three parametric models (para1-para3). Similarity, we found that the standard errors for $\rho = 0.7$ are lower than those for $\rho = 0.3$ in the parametric estimators and semi-parametric estimators. On the other hand, the standard errors in the semi-parametric estimators with Gaussian kernel density are smaller than those with Epanechnikov kernel density as we showed in table 4.1.

Table 4.4 shows overall fitting mean square errors in the semi-parametric model (4.2) and parametric models (4.3-4.6). The results in Table 4.4 based on the Gaussian Kernel density are very similar to the results in Table 4.2 using the Epanechnikov kernel density: for the training dataset and test dataset, the mean square errors in the parametric estimators and GAM estimator are larger than the mean square errors in semi-parametric estimators. Among the parametric models, exponential model (para2 in equation (4.4)) has the smallest mean square error for test dataset. In addition, the mean square errors in the semi-parametric estimators with Gaussian kernel density are smaller than those with Epanechnikov kernel density as we showed in table 4.2.

Figure 4.1 shows the non-parametric part fitting when $\theta_{ij} = sin(4 \times Z_{ij})$ by profile kernel estimator. Black line shows true value, blue line shows the fitting result using independence working correlation matrix while red line shows fitting result using true working correlation matrix.

Table 4.4 Overall MSE using the Gaussian kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta_{ij} = sin(4 \times Z_{ij})$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.170	1.184	1.488	1.490	1.457	1.475
$\rho = 0.7$	1.173	1.200	1.499	1.502	1.467	1.483
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.235	1.245	1.504	1.503	1.507	1.509
$\rho = 0.7$	1.229	1.250	1.513	1.512	1.519	1.527

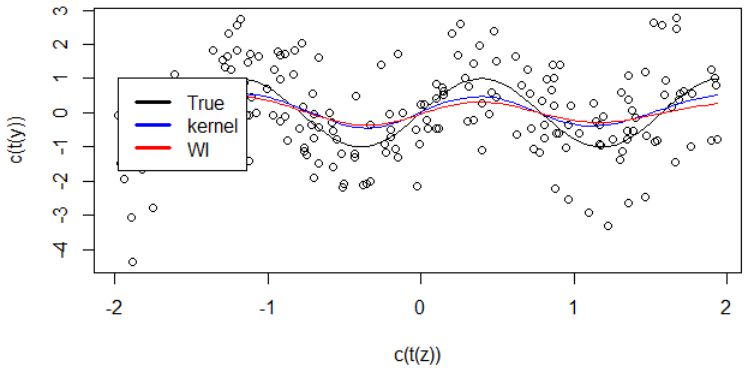

Figure 4.1: Plot for non-parametric part fitting $\theta_{ij} = sin(4 \times Z_{ij})$

Three lines almost overlapped with each other, which indicates the results using different working correlation matrices delivers similar results in non-parametric fitting part.

Table 4.5 shows β estimates when $\theta(Z_{ij}) = exp(sin(2/Z_{ij}^2))$ and Epanechnikov kernel density are used in the calculations. From the table, we can see that the estimated mean of semi-parametric estimators are more closer to true β than the estimated mean of parametric estimators; semi-WI estimates have the smallest standard error 0.015, which is slightly better than the standard error of the semi-True estimates. More over, the standard errors of the β estimates based on semi-WI and semi-True are at least 5 times less than the standard errors of the estimators from the three parametric models (para1-para3). In addition, we found that the standard errors for $\rho = 0.7$ are slightly lower than those for $\rho = 0.3$ in the semi-parametric estimators.

Table 4.6 shows overall fitting mean square errors in the semi-parametric model (4.2) and parametric models (4.3-4.6). For the training dataset and test dataset, the mean square errors in

Table 4.5 β estimates using the Epanechnikov kernel and 200 replications. Each of the 100 subjects with 4 time points and $\theta(Z_{ij}) = exp(sin(2/Z_{ij}^2))$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	1.001	0.015	1.001	0.011
semi-True	1.001	0.016	1.001	0.013
para1	1.047	0.091	1.026	0.078
para2	1.050	0.091	1.027	0.078
para3	1.067	0.088	1.031	0.073

Table 4.6 Overall MSE using the Epanechnikov kernel and 200 replications. Each of the 100 subjects with 4 time points and $\theta(Z_{ij}) = exp(sin(2/Z_{ij}^2))$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.344	1.346	3.592	2.948	2.168	1.546
$\rho = 0.7$	1.346	1.350	3.614	2.967	2.211	1.533
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.418	1.418	4.210	3.899	2.788	1.589
$\rho = 0.7$	1.438	1.439	4.255	4.066	3.130	1.635

the parametric estimators and GAM estimator are higher than the mean square errors in semi-parametric estimators. GAM model (equation (4.6)) has the smallest mean square error in non-semiparametric estimators and among the parametric models, polynomial estimator (para3 in equation (4.5)) has the smallest mean square error. More over, we found that the mean square errors for $\rho = 0.7$ are slightly higher than those for $\rho = 0.3$ in the semi-parametric estimators.

Table 4.7 shows β estimates in the semi-parametric model (4.2) and parametric models in (4.3-4.6). The results in Table 4.7 based on the Gaussian Kernel density are very similar to the results in Table 4.5 using the Epanechnikov kernel density: the standard errors of the β estimates based on semi-parametric estimators are at least 5 times less than the standard errors of the estimators from the three parametric models (para1-para3). More over, we found that the standard errors for $\rho = 0.7$ are slightly higher than those for $\rho = 0.3$ in the parametric estimators and semi-parametric estimators.

Table 4.8 shows overall fitting mean square errors in the semi-parametric model (4.2) and parametric models (4.3-4.6). The results in Table 4.8 based on the Gaussian Kernel density are very similar to the results in Table 4.6 using the Epanechnikov kernel density: for the training dataset and test dataset, the mean square errors in the parametric estimators and GAM estimator

Table 4.7 β estimates using the Gaussian kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta(Z_{ij}) = exp(sin(2/Z_{ij}^2))$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	1.001	0.013	1.000	0.014
semi-True	1.002	0.013	1.001	0.015
para1	1.050	0.087	1.045	0.096
para2	1.053	0.086	1.041	0.096
para3	1.070	0.088	1.068	0.088

Table 4.8 Overall MSE using the Gaussian kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta(Z_{ij}) = exp(sin(2/Z_{ij}^2))$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.310	1.314	3.635	2.984	2.227	1.541
$\rho = 0.7$	1.304	1.309	3.609	2.955	2.203	1.545
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.391	1.394	4.175	3.872	2.835	1.601
$\rho = 0.7$	1.389	1.393	4.205	3.862	2.808	1.602

are higher than the mean square errors in semi-parametric estimators. GAM model (equation (4.6)) has the smallest mean square error in non-semiparametric estimators and among the parametric models, polynomial estimator (para3 in equation (4.5)) has the smallest mean square error. In addition, the mean square errors in the semi-parametric estimators with Gaussian kernel density are smaller than those with Epanechnikov kernel density as we showed in table 4.6.

If time periods are extended to 10, using Gaussian density kernel we can obtain the results of the first setup in table 4.9 and 4.10 and the second setup in table 4.11 and 4.12.

Table 4.9 shows β estimates with Gaussian density kernel and 10 time periods. The results in table 4.9 are similar to results with 4 time periods, however the standard deviation for semi-parametric estimators are less than the situation when only 4 time points are involved, which indicates semi-parametric estimator gains more efficiency than parametric estimators when we have longer time periods. More over, we found that the standard errors for $\rho = 0.7$ are slightly higher than those for $\rho = 0.3$ in the parametric estimators.

Table 4.10 shows overall fitting mean square errors with Gaussian density kernel and 10 time periods. The results in Table 4.10 shows that the mean square train and test errors for semi-parametric approach are much lower than parametric estimates in all cases. Among parametric

Table 4.9 β estimates for Gaussian kernel with 10 time points when $\theta_{ij} = sin(4 \times Z_{ij})$

Entry	Lower		Upper	
	mean	se	mean	se
semi-WI	0.998	0.006	0.997	0.005
semi-True	0.997	0.007	0.999	0.006
para1	1.004	0.041	1.001	0.046
para2	1.006	0.041	1.001	0.046
para3	0.994	0.041	0.998	0.045

Table 4.10 Overall MSE for Gaussian kernel with 10 time points when $\theta_{ij} = sin(4 \times Z_{ij})$

Train	semi-WI	semi-True	para1	para2	para3	gam
Low	1.105	1.105	1.488	1.490	1.469	1.482
Upper	1.080	1.080	1.467	1.465	1.443	1.451
Test	semi-WI	semi-True	para1	para2	para3	gam
Low	1.133	1.147	1.506	1.506	1.498	1.509
Upper	1.116	1.138	1.488	1.488	1.489	1.503

cases, polynomial model performs best, but still much worse than semi-parametric fitting. GAM estimator has no advantage comparing to either parametric approach or semi-parametric approach. Furthermore, mean square errors for semi-parametric estimators are less than the situation when only 4 time points are involved, which indicates semi-parametric estimator gains more accuracy than parametric estimators when we have longer time periods. More over, we found that the mean square errors in training and testing datasets for $\rho = 0.7$ are slightly lower than those for $\rho = 0.3$ in the semi-parametric estimators.

Table 4.11 and table 4.12 show the results for semi-parametric model and parametric model with 10 time points and Gaussian kernel density in the second setup. The results are similar to the first setup, but coefficient estimators have less bias and standard deviation comparing to the first setup. Semi-parametric estimator still gains more when we have longer time periods. For

Table 4.11 β estimates for Gaussian kernel with 10 time points when $\theta(Z_{ij}) = exp(sin(2/Z_{ij}^2))$

Entry	Lower		Upper	
	mean	se	mean	se
semi-WI	1.000	0.005	1.000	0.004
semi-True	1.000	0.006	1.000	0.005
para1	1.004	0.051	1.003	0.050
para2	1.004	0.051	0.050	0.050
para3	1.008	0.050	1.004	0.050

Table 4.12 Overall MSE for Gaussian kernel with 10 time points when $\theta(Z_{ij}) = exp(sin(2/Z_{ij}^2))$

Train	semi-WI	semi-True	para1	para2	para3	gam
Low	1.291	1.291	3.652	2.990	2.233	1.549
Upper	1.278	1.278	3.663	2.999	2.228	1.541
Test	semi-WI	semi-True	para1	para2	para3	gam
Low	1.340	1.342	4.269	4.202	3.299	1.584
Upper	1.352	1.357	4.339	4.277	3.377	1.617

different kernel densities in semi-parametric approach, Gaussian kernel densities deliver estimators with less bias and more efficiency; semi-parametric estimators with stronger correlation, longer time period and more complicated pattern in non-parametric part will benefit more comparing to parametric estimators with the same scenarios. According to the conclusion in Lin and Carroll (2004) and results from our simulation, WI estimators performs better than estimator compiled with true correlation relationship, which conflicts the properties of GEE estimator. Another approach proposed by Wang (2005) will be displayed in the next part, which deliver the estimator with highest efficiency when fitting with true within subject association.

4.3 The SUR Estimator with One Kernel Smoother

The SUR estimator (Wang 2005) will be displayed in this part for semi-parametric regression, running simulation follow the setups in the first part. Still, different densities like Gaussian and Epanechnilov, different entries and different time periods will be applied in simulation results.

Table 4.13 shows β estimates in the semi-parametric model (4.2) and parametric models in (4.3-4.6) and local kernel average estimators with Epanechnikov kernel density and least square cross validation methods are used in the calculations. From the table, we can see that the semi-True estimates have the smallest standard error 0.011, which is better than the standard error of the semi-WI estimates. More over, the standard errors of the β estimates based on semi-WI and semi-True are at least 3 times less than the standard errors of the estimators from the three parametric models (para1-para3). In addition, we found that the standard errors for $\rho = 0.7$ are slightly lower than those for $\rho = 0.3$ in the semi-parametric estimators.

Table 4.14 shows overall fitting mean square errors in the semi-parametric model (4.2) and parametric models (4.3-4.6). For the training dataset, the mean square errors in the parametric estimators and GAM estimator are higher than the mean square errors in semi-parametric estima-

Table 4.13 β estimates using the Epanechnikov kernel and 200 replications. Each of the 100 subjects with 4 time points and $\theta_{ij} = sin(4 \times Z_{ij})$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	0.990	0.024	0.992	0.022
semi-True	0.998	0.012	0.997	0.011
para1	1.005	0.076	1.002	0.069
para2	1.002	0.076	1.001	0.070
para3	0.990	0.076	0.994	0.069

Table 4.14 Overall MSE using the Epanechnikov kernel and 200 replications. Each of the 100 subjects with 4 time points and $\theta_{ij} = sin(4 \times Z_{ij})$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.234	1.218	1.462	1.461	1.441	1.451
$\rho = 0.7$	1.247	1.234	1.479	1.476	1.450	1.460
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.270	1.224	1.518	1.518	1.511	1.526
$\rho = 0.7$	1.226	1.216	1.501	1.501	1.496	1.516

tors. For the test dataset, the mean square errors in parametric estimators and GAM estimator are larger than those for the semi-parametric estimators. Among the parametric models, polynomial model (para3 in equation (4.5)) has the smallest mean square error. For semi-parametric estimators, semi-True has smaller mean square errors than semi-WI. Mean square errors for training data are slightly higher than testing data for all models.

Table 4.15 shows β estimates in the semi-parametric model (4.2) and parametric models in (4.3-4.6). The results in Table 4.15 based on the Gaussian Kernel density are very similar to the results in Table 4.13 using the Epanechnikov kernel density: the standard errors of the β estimates based on semi-parametric estimators are at least 3 times less than the standard errors of the estimators from the three parametric models (para1-para3) and for semi-parametric estimators, semi-True has smaller standard errors than semi-WI. Similarity, we found that the standard errors for $\rho = 0.7$ are not higher than those for $\rho = 0.3$ in the parametric estimators and semi-parametric estimators. On the other hand, the standard errors in the semi-parametric estimators with Gaussian kernel density are smaller than those with Epanechnikov kernel density as we showed in table 4.13.

Table 4.16 shows overall fitting mean square errors in the semi-parametric model (4.2) and parametric models (4.3-4.6). The results in Table 4.16 based on the Gaussian Kernel density are very similar to the results in Table 4.14 using the Epanechnikov kernel density: for the training

Table 4.15 β estimates using the Gaussian kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta_{ij} = sin(4 \times Z_{ij})$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	0.991	0.023	0.993	0.018
semi-True	0.999	0.010	0.999	0.010
para1	1.000	0.073	1.009	0.067
para2	0.998	0.073	1.008	0.066
para3	0.987	0.073	1.002	0.065

Table 4.16 Overall MSE using the Gaussian kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta_{ij} = sin(4 \times Z_{ij})$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.200	1.022	1.485	1.483	1.461	1.472
$\rho = 0.7$	1.166	0.992	1.469	1.466	1.442	1.453
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.188	1.014	1.498	1.500	1.497	1.509
$\rho = 0.7$	1.182	1.006	1.514	1.514	1.512	1.514

dataset and test dataset, the mean square errors in the parametric estimators and GAM estimator are larger than the mean square errors in semi-parametric estimators. Among the parametric models, polynomial model (para3 in equation (4.5)) has the smallest mean square error for test dataset. In addition, the mean square errors in the semi-parametric estimators with Gaussian kernel density are smaller than those with Epanechnikov kernel density as we showed in table 4.14. Table 4.17 shows β estimates when $\theta(Z_{ij}) = exp(sin(2/Z_{ij}))$ and Epanechnikov kernel density are used in the calculations. From the table, we can see that the estimated mean of semi-parametric estimators are more closer to true β than the estimated mean of parametric estimators; semi-True estimates have the smallest standard error 0.015, which is slightly better than the standard error of the semi-WI estimates. More over, the standard errors of the β estimates based on semi-WI and semi-True are at least 4 times less than the standard errors of the estimators from the three parametric models (para1-para3). In addition, we found that the standard errors for $\rho = 0.7$ are slightly lower than those for $\rho = 0.3$ in the parametric estimators.

Table 4.18 shows overall fitting mean square errors in the semi-parametric model (4.2) and parametric models (4.3-4.6). For the training dataset and test dataset, the mean square errors in the parametric estimators and GAM estimator are higher than the mean square errors in semi-parametric estimators. GAM model (equation (4.6)) has the smallest mean square error

Table 4.17 β estimates using the Epanechnikov kernel and 200 replications. Each of the 100 subjects with 4 time points and $\theta(Z_{ij}) = exp(sin(2/Z_{ij}^2))$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	1.001	0.019	1.004	0.020
semi-True	0.999	0.015	1.002	0.014
para1	1.047	0.096	1.029	0.080
para2	1.050	0.097	1.030	0.081
para3	1.061	0.097	1.029	0.077

Table 4.18 Overall MSE using the Epanechnikov kernel and 200 replications. Each of the 100 subjects with 4 time points and $\theta(Z_{ij}) = exp(sin(2/Z_{ij}^2))$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.734	1.372	3.634	2.987	2.210	1.574
$\rho = 0.7$	1.563	1.343	3.613	2.961	2.217	1.539
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	3.021	1.354	4.202	3.869	2.826	1.601
$\rho = 0.7$	1.567	1.352	4.234	4.037	3.111	1.615

in non-semiparametric estimators and among the parametric models, polynomial estimator (para3 in equation (4.5)) has the smallest mean square error. For semi-parametric models, semi-True has smaller mean square errors than semi-WI in training and test datasets.

Table 4.19 shows β estimates in the semi-parametric model (4.2) and parametric models in (4.3-4.6). The results in Table 4.19 based on the Gaussian Kernel density are very similar to the results in Table 4.17 using the Epanechnikov kernel density: the standard errors of the β estimates based on semi-parametric estimators are at least 4 times less than the standard errors of the estimators from the three parametric models (para1-para3). More over, we found that the standard errors for $\rho = 0.7$ are slightly higher than those for $\rho = 0.3$ in the parametric estimators and semi-parametric estimators.

Table 4.20 shows overall fitting mean square errors in the semi-parametric model (4.2) and parametric models (4.3-4.6). The results in Table 4.20 based on the Gaussian Kernel density are very similar to the results in Table 4.18 using the Epanechnikov kernel density: for the training dataset and test dataset, the mean square errors in the parametric estimators and GAM estimator are higher than the mean square errors in semi-parametric estimators. GAM model (equation (4.6)) has the smallest mean square error in non-semiparametric estimators and among the parametric

Table 4.19 β estimates using the Gaussian kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta(Z_{ij}) = exp(sin(2/Z_{ij}^2))$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	1.001	0.016	1.002	0.019
semi-True	1.000	0.014	1.001	0.014
para1	1.049	0.083	1.034	0.074
para2	1.052	0.084	1.034	0.074
para3	1.067	0.085	1.034	0.073

Table 4.20 Overall MSE using the Gaussian kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta(Z_{ij}) = exp(sin(2/Z_{ij}^2))$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.321	1.252	3.630	2.989	2.210	1.554
$\rho = 0.7$	1.330	1.256	3.643	2.981	2.211	1.543
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.327	1.257	4.210	3.897	2.856	1.608
$\rho = 0.7$	1.317	1.240	4.214	4.019	3.068	1.594

models, polynomial estimator (para3 in equation (4.5)) has the smallest mean square error. In addition, the mean square errors in the semi-parametric estimators with Gaussian kernel density are smaller than those with Epanechnikov kernel density as we showed in table 4.18.

4.4 Semi-Parametric Model with Multiple Kernel Smoothers

In this section, we first show the results of local polynomial kernel GEE estimator with two kernel smoothers for semi-parametric regression and other estimators for parametric regression with various scenarios such as different kernel densities, correlation entries and time periods as we discussed in Section 4.1.2.

Table 4.21 shows β estimates in the semi-parametric model (4.7) and parametric models in (4.9-4.12) and local kernel average estimators with Epanechnikov kernel density and least square cross validation methods are used in the calculations. From the table, we can see that the semi-WI estimates have the smallest standard error 0.022, which is slightly better than the standard error of the semi-True estimates. More over, the standard errors of the β estimates based on semi-WI and semi-True are at least 4 times less than the standard errors of the estimators from the three parametric models (para1-para3). In addition, we found that the standard errors for $\rho = 0.7$ are

Table 4.21 β estimates using the Epanechnikov kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta_{1ij} = sin(4 \times Z_{1ij})$,$\theta_{2ij} = sin(4 \times Z_{2ij})$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	0.993	0.022	0.994	0.018
semi-True	0.989	0.023	0.994	0.022
para1	1.015	0.090	1.010	0.079
para2	1.008	0.091	1.005	0.080
para3	0.991	0.092	0.989	0.080

Table 4.22 Overall MSE using the Epanechnikov kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta_{1ij} = sin(4 \times Z_{1ij})$,$\theta_{2ij} = sin(4 \times Z_{2ij})$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.833	1.868	1.981	1.975	1.929	1.960
$\rho = 0.7$	1.878	1.904	1.987	1.978	1.917	1.953
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.894	2.250	2.025	2.027	2.013	2.030
$\rho = 0.7$	1.918	2.175	2.043	2.043	2.031	2.049

slightly lower than those for $\rho = 0.3$ in the semi-parametric estimators. Comparing to models with one kernel smoothers, the standard errors of β estimators for parametric models (para1-para3) increased slightly, but the standard errors of β estimators for profile kernel GEE model (semi-WI and Semi-True) did not increase, indicating that profile kernel GEE estimator is robust when fitting with two kernel smoothers.

Table 4.22 shows overall fitting mean square errors in the semi-parametric model (4.7) and parametric models (4.9-4.12). For the training dataset, the mean square errors in the parametric estimators and GAM estimator are higher than the mean square errors in semi-parametric estimators. For the test dataset, the mean square errors in parametric estimators and GAM estimator are larger than those for the semi-parametric estimators. Among the parametric models, polynomial model (para2 in equation (4.11)) has the smallest mean square error. Mean square errors for training data are slightly higher than testing data for all models.

Table 4.23 shows β estimates in the semi-parametric model (4.7) and parametric models in (4.9-4.12). The results in Table 4.3 based on the Gaussian Kernel density are very similar to the results in Table 4.1 using the Epanechnikov kernel density: the standard errors of the β estimates based on semi-parametric estimators are at least 3 times less than the standard errors of the estimators from the three parametric models (para1-para3). We found that the standard errors for $\rho = 0.7$

Table 4.23 β estimates using the Gaussian kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta_{1ij} = sin(4 \times Z_{1ij})$,$\theta_{2ij} = sin(4 \times Z_{2ij})$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	0.992	0.019	0.988	0.025
semi-True	0.989	0.020	0.989	0.026
para1	1.017	0.083	0.998	0.074
para2	1.008	0.086	0.994	0.075
para3	0.989	0.087	0.983	0.075

Table 4.24 Overall MSE using the Gaussian kernel and 200 replications.
Each of the 100 subjects with 4 time points and $\theta_{1ij} = sin(4 \times Z_{1ij})$,$\theta_{2ij} = sin(4 \times Z_{2ij})$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.835	1.815	1.990	1.985	1.930	1.964
$\rho = 0.7$	1.896	1.880	1.970	1.963	1.915	1.940
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	1.887	1.863	2.052	2.051	2.048	2.056
$\rho = 0.7$	1.933	1.963	2.057	2.061	2.061	2.074

are slightly higher than those for $\rho = 0.3$ in the semi-parametric estimators. On the other hand, the standard errors in the parametric estimators with Gaussian kernel density are slightly smaller than those with Epanechnikov kernel density as we showed in table 4.21.

Table 4.24 shows overall fitting mean square errors in the semi-parametric model (4.7) and parametric models (4.9-4.12). The results in Table 4.24 based on the Gaussian Kernel density are very similar to the results in Table 4.22 using the Epanechnikov kernel density: for the training dataset and test dataset, the mean square errors in the parametric estimators and GAM estimator are larger than the mean square errors in semi-parametric estimators. Among the parametric models, quadratic model (para2 in equation (4.10)) has the smallest mean square error for test dataset. In addition, the mean square errors in the parametric and semi-parametric estimators with Gaussian kernel density are similar to those with Epanechnikov kernel density as we showed in table 4.22.

Table 4.25 shows β estimates when $\theta(Z_{1ij}) = exp(sin(2/Z_{ij}))$ and $\theta(Z_{2ij}) = exp(sin(2/Z_{2ij}))$ and Epanechnikov kernel density are used in the calculations. From the table, we can see that the estimated mean of semi-parametric estimators are more closer to true β than the estimated mean of parametric estimators; semi-WI estimates have the smallest standard error 0.033, which is slightly better than the standard error of the semi-True estimates. More over, the standard errors

Table 4.25 β estimates using the Epanechnikov kernel and 200 replications. Each of the 100 subjects with 4 time points and $\theta(Z_{1ij}) = exp(sin(2/Z_{1ij}))$,$\theta(Z_{2ij}) = exp(sin(2/Z_{2ij}))$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	1.014	0.033	1.012	0.027
semi-True	1.015	0.034	1.015	0.030
para1	1.030	0.093	1.034	0.092
para2	1.057	0.115	1.061	0.107
para3	1.062	0.123	1.052	0.113

Table 4.26 Overall MSE using the Epanechnikov kernel and 200 replications. Each of the 100 subjects with 4 time points and $\theta(Z_{1ij}) = exp(sin(2/Z_{1ij}))$,$\theta(Z_{2ij}) = exp(sin(2/Z_{2ij}))$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	2.41	2.33	6.37	4.17	4.48	2.76
$\rho = 0.7$	2.37	2.68	6.48	4.14	4.46	2.75
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	2.39	2.38	8.28	6.18	8.86	2.83
$\rho = 0.7$	2.38	2.72	8.32	6.46	9.59	2.84

of the β estimates based on semi-WI and semi-True are at least 3 times less than the standard errors of the estimators from the three parametric models (para1-para3). In addition, we found that the standard errors for $\rho = 0.7$ are slightly lower than those for $\rho = 0.3$ in the semi-parametric estimators.

Table 4.26 shows overall fitting mean square errors in the semi-parametric model (4.7) and parametric models (4.9-4.12) for the second setup. For the training dataset and test dataset, the mean square errors in the parametric estimators and GAM estimator are higher than the mean square errors in semi-parametric estimators. GAM model (equation (4.12)) has the smallest mean square error in non-semiparametric estimators and among the parametric models, exponential estimator (para2 in equation (4.10)) has the smallest mean square error. More over, we found that the mean square errors for $\rho = 0.7$ are slightly higher than those for $\rho = 0.3$ in the semi-parametric estimators. Comparing to models with one kernel smoothers, the MSE for parametric models (para1-para3) increased at least 4 times, but the MSE of profile kernel GEE model (semi-WI and Semi-True) increased less than 2 times, indicating that profile kernel GEE estimator is more robust for MSE than parametric models.

Table 4.27 β estimates using the Gaussian kernel and 200 replications.
Each of the 100 subjects with 4 time points and
$\theta(Z_{1ij}) = exp(sin(2/Z_{1ij}))$,$\theta(Z_{2ij}) = exp(sin(2/Z_{2ij}))$

	$\rho = 0.3$		$\rho = 0.7$	
	mean	se	mean	se
semi-WI	1.012	0.029	1.011	0.027
semi-True	1.012	0.031	1.013	0.032
para1	1.032	0.101	1.021	0.081
para2	1.053	0.117	1.042	0.098
para3	1.051	0.134	1.040	0.102

Table 4.28 Overall MSE using the Gaussian kernel and 200 replications.
Each of the 100 subjects with 4 time points and
$\theta(Z_{1ij}) = exp(sin(2/Z_{1ij}))$,$\theta(Z_{2ij}) = exp(sin(2/Z_{2ij}))$

Train	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	2.51	2.4	6.34	4.18	4.49	2.76
$\rho = 0.7$	2.72	2.67	6.32	4.09	4.41	2.73
Test	semi-WI	semi-True	para1	para2	para3	gam
$\rho = 0.3$	2.52	2.48	8.38	6.24	8.94	2.84
$\rho = 0.7$	2.77	2.79	8.34	6.48	9.57	2.82

Table 4.27 shows β estimates in the semi-parametric model (4.7) and parametric models in (4.9-4.12). The results in Table 4.27 based on the Gaussian Kernel density are very similar to the results in Table 4.25 using the Epanechnikov kernel density: the standard errors of the β estimates based on semi-parametric estimators are at least 3 times less than the standard errors of the estimators from the three parametric models (para1-para3).

Table 4.28 shows overall fitting mean square errors in the semi-parametric model (4.7) and parametric models (4.9-4.12). The results in Table 4.28 based on the Gaussian Kernel density are very similar to the results in Table 4.26 using the Epanechnikov kernel density: for the training dataset and test dataset, the mean square errors in the parametric estimators and GAM estimator are higher than the mean square errors in semi-parametric estimators. GAM model (equation (4.12)) has the smallest mean square error in non-semiparametric estimators and among the parametric models, exponential estimator (para2 in equation (4.10)) has the smallest mean square error. Comparing to models with one kernel smoothers, the MSE for parametric models (para1-para3) increased at least 4 times, but the MSE of profile kernel GEE model (semi-WI and Semi-True) increased 2 times, indicating that profile kernel GEE estimator is more robust for MSE than parametric models.

CHAPTER 5

DATA DESCRIPTION AND APPLICATION

Credit card loan data are a major type of financial data owned by banks and other financial institutes and play an important role for longitudinal data analysis as we discussed in Chapter 1: for each subject, which is the customer, we have records of monthly payment history for multiple time points. The semi-parametric models and Generalized Estimating Equations method can be applied to this dataset and in this Chapter, we first give the detailed description of a credit card loan dataset. Our main purpose for this application is to investigate which factors will influence customer's payment status by using different approaches and explore the difference between parametric estimators and semi-parametric estimators.

5.1 Description of the Dataset

The dataset used in this application comes from UCI (University of California Irvine) Machine Learning Repository Website with 30000 subjects and 8 variables. A basic summary statistics for those eight variables are as the follows:

- **Bill amount ("BILL AMT")**: Amount should be paid by each customer for current month, with minimum -339603 and maximum 1664089. A negative number shows there are credits from last month.
- **Payment amount ("PAY AMT")**: Amount customer paid for current month, with minimum 0 and maximum 1684259.
- **"PAY"**: A categorical variable with values from -2 to 8 (11 categories), denoting how many delayed periods the customer had. A negative number shows that payment is made before due day.
- **"LIMIT BAL"**: Limit amount for each customer, with minimum 10000 and maximum 1000000.
- **"SEX"**: With 1 denoting male and 2 denoting female.

- **"EDUCATION"**": Education level for each customer: 1 denotes graduate school; 2 denotes university; 3 denotes high school, and $4, 5, 6$ denotes others.
- **"MARRIAGE"**: Marital status: 1 denotes married; 2 denotes single and 3 denotes others.

From the eight variables, two variables can be constructed to address our main concerns. The first response variable called "remaining amount", is the difference between bill amount and payment amount, showing whether the customer made full payment or not. The second response variable is the delayed pay periods denoted by "PAY": "PAY"= 1 denoting there is a delay, no matter how long for that delay and "PAY"= 0 denoting no delay, which means payment was made duly or before due day. There are 5 variables in the list left as predictors, including gender, education, marriage, age and limit balance.

Primary parametric GEE regression will be conducted as the first step for analyzing credit card loan data. For example, after fitting a linear GEE regression with response variable "remaining amount"and 5 predictors we have discussed on last paragraph, we get a result that four predictors are statistically significant with p-values less than 0.05 while the variable "age"is not statistically significant. In our semi-parametric models, the four significant predictors can be used in the parametric part while the variable "age"will be treated as a non-parametric covariate. Different semi-parametric models will be estimated with different working correlation matrix, and the results from semi-parametric models will be compared with the results from parametric models.

5.2 Results and Discussion: Overall Analysis

5.2.1 Using Remaining Amount as Response Variable

In this part, the remaining amount we defined on section 5.1 will be used as the response variable to explore the relationship between the amount of owed payments and other predictors: such as gender, education level, limit balance, marriage status, and age. The following three parametric GEE models will be fitted:

$$Para1: \ remaining\ amount = \beta_0 + \beta_1 limit\ balance + sex + education + marriage + \beta_2 age \quad (5.1)$$

$$Para2: \ remaining\ amount = \beta_0 + \beta_1 limit\ balance + sex + education + marriage + \beta_2 age + \beta_3 age^2 \quad (5.2)$$

$$Para3: \ remaining\ amount = \beta_0 + \beta_1 limit\ balance + sex + education + marriage + \beta_2 exp(age) \tag{5.3}$$

and we consider a semi-parametric model with non-parametric form on the predictor age:

$$Semi: remaining\ amount = \beta_0 + \beta_1 limit\ balance + sex + education + marriage + \theta(age) \tag{5.4}$$

where $\theta(.)$ is a kernel smoother.

Table 5.1 shows the estimation results for the first parametric model (Para1) using different working correlation matrices. Based on the signs of the estimated coefficients, we found that relative to male consumers, female consumers have less remaining amount. Relative to consumers with graduate degrees, customers with only college degrees or high school degrees have more remaining amount. Relative to married customers, customers with single marriage status tend to have more remaining amount. The predictor limit balance has coefficient of 0.002, which denotes that limit balance has positive correlation with remaining amount; age has p-value larger than 0.05, showing that age is not statistically significant in Para1.

Different working correlation matrices such as independence, exchangeable, AR1 and unstructured matrix are used in Para1. The estimated parameters by those four working correlation matrices are quite similar while the estimates using unstructured working correlation matrix has smallest standard errors among all other estimates.

Table 5.2 shows the estimation result for the second parametric model (Para2) using different working correlation matrices and a quadratic term for the predictor age. Based on the signs of the estimated coefficients, we found that the result is similar to the result in the first parametric GEE model: female customers tends to have less remaining amount comparing to male customers; customers with university degrees or high school degrees have more remaining amount comparing to customers with graduate degrees; single customers has more remaining amount than married customers. The predictor age has p-value larger than 0.05, showing that age is still not statistically significant this model.

Different working correlation matrices such as independence, exchangeable, AR1 and unstructured matrix are used in Para2. The estimated parameters share the same properties as Para1 when using these four working correlation matrices: the coefficients are quite similar while the estimates using unstructured working correlation matrix has smallest standard errors among the estimates using different working correlation matrices.

Table 5.3 shows the estimation result for the third parametric model (Para3) using different working correlation matrix and an exponential term on age. The estimated coefficients and standard errors are similar to the estimation results in Para1 and Para2. The exponential term age is significant in this model but the estimated coefficient for this exponential term is nearly zero $(2.61e-34)$, denoting that age has tiny effect on the response variable.

The results from three parametric models show that age is not significant or has tiny effect with parametric patterns, such as linear, quadratic or exponential terms. We consider semi-parametric models with kernel smoother on the predictor age, investigating the changes on estimated coefficients for other predictors fitted with linear patterns, seeing if semi-parametric models are more advanced than pure parametric models.

Table 5.4 shows the estimation results for the semi parametric model with kernel smoother on the predictor age using different working correlation matrices. The result of the estimated coefficients are similar to the estimation in parametric models. Based on the signs of the estimated coefficients, we found that relative to male consumers, female consumers have less remaining amount. Relative to consumers with graduate degrees, customers with only college degrees or high school degrees have more remaining amount. Relative to married customers, the semi-parametric model shows that customers with single marriage status tend to have more remaining amount, and the coefficient for predictor single (0.086) is higher than the coefficient in parametric models(0.042 in Para3). The predictor limit balance has coefficient of 0.002, which denotes that limit balance has positive correlation with remaining amount.

Table 5.5 shows the mean square error results for parametric models and semi-parametric model with kernel smoother on the predictor of age. The overall MSE for training dataset is lower than testing dataset. The MSE for semi-parametric model in training dataset(0.928) is slightly higher than the MSE for parametric models (0.926) while the MSE in testing dataset for semi-parametric model is lower than the MSE in testing dataset for parametric models. Among parametric models, the parametric model with exponential term (Para3) has highest testing error and different working correlation matrices has the same MSE.

5.2.2 Using Payment Status as Response Variable

In this part, the payment status, which is whether the client has a default we defined on section 5.1 will be used as the response variable.

Table 5.1 Parameter estimations for parametric GEE model (Para1)

	Independence			Exchangeable		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.4809	0.03072	< 2e-16	-0.481	0.0307	< 2e-16
$limit - balance$	0.001977	0.00006513	< 2e-16	0.00198	0.0000651	< 2e-16
$sex - female$	-0.06279	0.01061	3.25e-09	-0.0628	0.0106	3.2e-09
$education - university$	0.2064	0.01251	< 2e-16	0.206	0.0125	< 2e-16
$education - highschool$	0.1696	0.01587	<2e-16	0.17	0.0159	< 2e-16
single	0.04873	0.01212	0.0000582	0.0487	0.0121	0.000058
age	0.0009698	0.0006611	0.142	0.00097	0.000661	0.14
	AR1			Unstructured		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.484	0.031	< 2e-16	-0.488	0.0296	< 2e-16
$limit - balance$	0.00199	0.0000667	< 2e-16	0.00187	0.0000629	< 2e-16
$sex - female$	-0.0662	0.0107	7.2e-10	-0.0519	0.0102	0.0000004
$education - university$	0.213	0.0126	< 2e-16	0.186	0.012	< 2e-16
$education - highschool$	0.176	0.016	< 2e-16	0.151	0.0153	< 2e-16
single	0.0488	0.0122	0.000066	0.0448	0.0117	0.00013
age	0.000937	0.000664	0.16	0.000833	0.000637	0.19076

Table 5.2 Parameter estimations for parametric GEE model (Para2)

	Independence			Exchangeable		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.467	0.0222	< 2e-16	-0.467	0.0222	< 2e-16
$limit - balance$	0.00198	0.000065	< 2e-16	0.00198	0.000065	< 2e-16
$sex - female$	-0.0625	0.0106	3.4e-09	-0.0625	0.0106	3.40e-09
$education - university$	0.206	0.0125	< 2e-16	0.206	0.0125	< 2e-16
$education - highschool$	0.168	0.0159	< 2e-16	0.168	0.0159	< 2e-16
$single$	0.0496	0.012	0.000035	0.0496	0.012	0.000035
age^2	0.0000146	0.00000891	0.1	0.0000146	0.00000891	0.1
	AR1			Unstructured		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.471	0.0224	< 2e-16	-0.476	0.0214	< 2e-16
$limit - balance$	0.00199	0.0000665	< 2e-16	0.00187	0.0000627	< 2e-16
$sex - female$	-0.0659	0.0107	7.4e-10	-0.0516	0.0102	4.20e-07
$education - university$	0.213	0.0126	< 2e-16	0.186	0.012	< 2e-16
$education - highschool$	0.175	0.016	<2e-16	0.15	0.0153	< 2e-16
$single$	0.0496	0.0121	0.00004	0.0457	0.0116	0.000076
age^2	0.0000141	0.00000893	0.11	0.0000128	0.0000086	0.14

Table 5.3 Parameter estimations for parametric GEE model (Para3)

	Independence			Exchangeable		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.444	0.0171	$<$ 2e-16	-0.444	0.0171	$<$ 2e-16
$limit - balance$	0.00198	0.0000652	$<$ 2e-16	0.00198	0.0000652	$<$ 2e-16
$sex - female$	-0.0642	0.0106	1.2e-09	-0.0642	0.0106	1.2e-09
$education - university$	0.207	0.0125	$<$ 2e-16	0.207	0.0125	$<$ 2e-16
$education - highschool$	0.175	0.0155	$<$ 2e-16	0.175	0.0155	$<$ 2e-16
$single$	0.0415	0.0107	0.00011	0.0415	0.0107	0.00011
$exp(age)$	2.62e-34	2.41e-36	$<$2e-16	2.62e-34	2.41e-36	$<$ 2e-16
	AR1			Unstructured		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.449	0.0173	$<$ 2e-16	-0.457	0.0164	$<$ 2e-16
$limit - balance$	0.00199	0.0000667	$<$ 2e-16	0.00188	0.0000629	$<$ 2e-16
$sex - female$	-0.0675	0.0107	2.5e-10	-0.053	0.0102	0.00000019
$education - university$	0.213	0.0126	$<$ 2e-16	0.187	0.012	$<$ 2e-16
$education - highschool$	0.181	0.0156	$<$ 2e-16	0.156	0.0149	$<$ 2e-16
$single$	0.0418	0.0108	0.00011	0.0386	0.0103	0.00019
$exp(age)$	2.61e-34	2.45e-36	$<$ 2e-16	2.67e-34	2.46e-36	$<$ 2e-16

Table 5.4 Parameter estimations for semi-parametric GEE model

	Independence			Exchangeable		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.449	0.0171	< 2e-16	-0.449	0.0171	< 2e-16
$limit - balance$	0.00179	0.0000651	< 2e-16	0.00179	0.0000651	< 2e-16
$sex - female$	-0.0473	0.0106	0.0000076	-0.0473	0.0106	7.60e-06
$education - university$	0.216	0.0125	< 2e-16	0.216	0.0125	< 2e-16
$education - highschool$	0.166	0.0155	< 2e-16	0.166	0.0155	< 2e-16
$single$	0.0857	0.0107	1.4e-15	0.0857	0.0107	1.40e-15
	AR1			Unstructured		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.453	0.0173	< 2e-16	-0.461	0.0164	< 2e-16
$limit - balance$	0.0018	0.0000667	< 2e-16	0.00168	0.0000629	< 2e-16
$sex - female$	-0.0506	0.0107	0.0000022	-0.0361	0.0102	0.0004
$education - university$	0.222	0.0126	< 2e-16	0.196	0.012	< 2e-16
$education - highschool$	0.172	0.0156	<2e-16	0.147	0.0149	< 2e-16
$single$	0.086	0.0108	2.1e-15	0.0829	0.0103	1.10e-15

Table 5.5 Overall MSE for parametric models and semi-parametric model

ar1	para1	para2	para3	semi
training	0.926	0.926	0.926	0.928
testing	0.971	0.971	1.21	0.959
unstructured	para1	para2	para3	semi
training	0.926	0.926	0.926	0.928
testing	0.961	0.961	1.28	0.955

We would like to explore the relationship between whether the customer will default to pay the bills and other predictors: such as gender, education level, limit balance, marriage status, and age. We consider parametric GEE model with linear form as the following:

$$Para4: \ logit(p) = \beta_0 + \beta_1 limit\ balance + sex + education + marriage + \beta_2 age \tag{5.5}$$

where p is the probability of default.

Table 5.6 shows the estimation results for the parametric model (Para4) using different working correlation matrices. Based on the signs of the estimated coefficients, we found that relative to male consumers, female consumers have less probability to default. Relative to consumers with graduate degrees, customers with college degrees or high school degrees have more probability to default. Relative to married customers, customers with single marriage status tend to have more probability to default. The predictor limit balance has negative coefficient, which denotes that limit balance has negative correlation with the probability of default and age also has negative correlation with the probability of default.

Different working correlation matrices such as independence, exchangeable, AR(1) and unstructured matrix are used in Para1. The estimated parameters by those four working correlation matrices are quite similar while the estimates using unstructured working correlation matrix has smallest standard errors among all other estimates.

Table 5.7 shows the predictive accuracy tables in parametric model when the response variable is delay status. The testing accuracy (0.672) is slightly lower than training accuracy (0.686) and when the response variable is 0, the predictive error is higher than when the response variable is 1. Different working correlation matrices, such as AR(1) or unstructured will deliver similar result for predictive accuracy.

Table 5.6 Parameter estimations for parametric GEE model (Para4)

	Independence			Exchangeable		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	1.42	0.0644	¡2e-16	1.42	0.0644	<2e-16
$limit - balance$	-3.32e-06	9.12e-08	<2e-16	-3.32e-06	9.12e-08	<2e-16
$sex - female$	-0.275	0.0234	<2e-16	-0.275	0.0234	<2e-16
$education - university$	0.67	0.0252	<2e-16	0.67	0.0252	<2e-16
$education - highschool$	0.619	0.0356	<2e-16	0.619	0.0356	<2e-16
single	0.14	0.0255	4.00e-08	0.14	0.0255	4.00E-08
age	-0.0119	0.00141	<2e-16	-0.0119	0.00141	<2e-16
	AR1			Unstructured		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	1.41	0.0637	< 2e-16	1.45	0.0636	< 2e-16
$limit - balance$	-3.20e-06	9.01e-08	< 2e-16	-3.30e-06	8.99e-08	< 2e-16
$sex - female$	-0.251	0.0232	< 2e-16	-0.266	0.0231	< 2e-16
$education - university$	0.64	0.025	< 2e-16	0.657	0.0249	< 2e-16
$education - highschool$	0.588	0.0353	< 2e-16	0.604	0.0352	< 2e-16
single	0.134	0.0253	1.20e-07	0.134	0.0253	1.10e-07
age	-0.011	0.0014	3.10e-15	-0.0115	0.00139	< 2e-16

Table 5.7 Predictive Accuracy for Parametric GEE model (Para4)

training	ar1			accuracy	testing	ar1			accuracy
		pred					pred		
	true	0	1			*true*	0	1	
	0	7070	31801			0	3497	15113	
	1	6094	75035	0.686		1	2941	33427	0.672
	unstructured					unstructured			
		pred					pred		
	true	0	1			*true*	0	1	
	0	7070	31801			0	3486	15124	
	1	6106	75023	0.686		1	2952	33416	0.671

5.3 Results and Discussion: Gender Analysis

In this part, we evaluate the difference between models for male customers and female customers. Following the overall analysis in section 5.2, three parametric models and one semi-parametric model are fitted for analysis and we used two different outcomes: remaining amount and payment status as the response variable. The BIC values for the four models for male and female separately and together are calculated and used for model comparison. Estimated coefficients for all models are reported for the purpose of exploring the difference among the fitted models for different gender. We provide mean square error as the evaluation measurement for the comparison of parametric and semi-parametric models when using remaining amount as response variable and we use predictive accuracy as the evaluation measurement when using payment status as response variable.

5.3.1 Using Remaining Amount as Response Variable

Model Setups

The remaining amount we defined on section 5.1 will be used as the response variable to explore the relationship between the amount of owed payments and some predictors: such as education level, limit balance, marriage status, and age for male and female customers. As overall analysis in section 5.2, the following three parametric GEE models will be fitted for male and female separately:

$$Para1\ for\ male:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_M + education \times I_M + marriage \times I_M + \beta_2 age \times I_M \tag{5.6}$$

$$Para1\ for\ female:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_F + education \times I_F + marriage \times I_F + \beta_2 age \times I_F \tag{5.7}$$

$$Para2\ for\ male:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_M + education \times I_M + marriage \times I_M + \beta_2 age^2 \times I_M \tag{5.8}$$

$$Para2\ for\ female:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_F + education \times I_F + marriage \times I_F + \beta_2 age^2 \times I_F \tag{5.9}$$

$$Para3\ for\ male:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_M + education \times I_M + marriage \times I_M + \beta_2 exp(age) \times I_M \tag{5.10}$$

$$Para3\ for\ female:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_F + education \times I_F + marriage \times I_F + \beta_2 exp(age) \times I_F \tag{5.11}$$

and we consider a semi-parametric model with non-parametric form on the predictor age:

$$Semi\ for\ male: remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_M + education \times I_M + marriage \times I_M + \theta(age) \times I_M \tag{5.12}$$

$$Semi\ for\ female: remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_F + education \times I_F + marriage \times I_F + \theta(age) \times I_F \tag{5.13}$$

where $\theta(.)$ is a kernel smoother, I_M and I_F are indicator variables, defined as following:

$$I_M = \begin{cases} 0 & if \quad gender \quad is \quad female \\ 1 & if \quad gender \quad is \quad male \end{cases}$$

$$I_F = \begin{cases} 0 & if \quad gender \quad is \quad male \\ 1 & if \quad gender \quad is \quad female \end{cases}$$

Model selection criterion

There are various methods for goodness of fit in model selection. Jones (2010) proposed a modified Bayesian information criterion in Longitudinal and Clustered data. Traditional BIC assume the response variable is independent to each other but in this paper, they deal with mixed model by using effective sample size from Fisher's information instead of regular sample size. More specifically, the effective sample size changes depends on the number of subjects and the number of observations. They developed several formulas for different within subject association, such as compound (exchangeable) structure and AR(1) structure for balanced or unbalanced data in generalized mixed model.

Liu and Yang (2011) pointed out when using BIC (Schwarz, 1978), we assume the true model is parametrically and if we use Akaike's information criterion (AIC, Akaike, 1973), we assume true model is in non-parametric approach. They proposed PI (parametricness index), which is a new criterion for model selection. They pointed out that PI has nice properties in theory: it converges in probability to 1 for non-parametric models while converges to infinity to parametric models. They also showed that the model not picked by PI has a much smaller confidence intervals which raise the issue of uncertainty but the model selected by PI is convinced. They claimed that in application, PI provided solid evidence for choosing between AIC and BIC and it may also applied in other criterion such as FIC for a better estimator. They suggested using BIC for estimation when PI is high while using AIC when PI tends to 1. They also showed some limitations for PI, such as only for the model under Gaussian assumptions and different purposes for model selection should also be considered.

Table 5.8 BIC: Gender Analysis

model	para1	para2	para3	semi
overall	-11116.26	-11122.03	-11278.38	-10966.76
separate	-11374.06	-11375.28	-11440.83	-11564.94

BIC analysis

We use BIC (Bayesian Information Criterion) to explore what is the difference when we fit models for male and female separately comparing to we fit a model with male and female together. The BIC defined as following:

$$BIC = n \times ln(\hat{\sigma_e}^2) + (p+1) \times ln(n) \tag{5.14}$$

In (5.14), n denotes the number of observations and p denotes the number of predictors we used in the model and $\hat{\sigma_e}^2$ is the error variance, defined as following:

$$\hat{\sigma_e}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{5.15}$$

where y_i is the observed value for the remaining amount and $\hat{y_i}$ is the estimated remaining amount.

When fitting the model with male and female together, we used BIC calculated from models in section 5.2. While fitting models for male and female separately, we use BIC derived from models in (5.6)-(5.13), combining error variance from models with indicator variables I_M and I_F as following:

$$BIC_{separate} = n \times log(\hat{\sigma_{e_M}}^2 + \hat{\sigma_{e_F}}^2) + 2(p+1) \times ln(n) \tag{5.16}$$

where error variance, n and p are defined as the same as in BIC (5.10).

Table 5.8 shows the BIC values calculated in (5.16) when using different models for male and female separately and the BIC in (5.14) for models fitting with male and female together. A small BIC denotes a better model fitting. The BIC values for the four types of models fitted with male and female separately are smaller than the BIC values when the models are fitted for the overall model with male and female together. Furthermore, semi-parametric models with male and female separately is the best model, since it has the smallest BIC.

Table 5.9 shows the estimation results for the first parametric model (Para1) using different working correlation matrices for male and female separately. We found that unlike the overall

analysis, different working correlation matrices will identify different significant variables: for male, variable single and age are significant when using independence working correlation matrix and unstructured working correlation matrix while they are not significant when using exchangeable or AR(1) working correlation matrix; for female, age is not significant only when using the independence working correlation matrix. Relative to those with graduate degrees, male and female customers with only college degrees or high school degrees have more remaining amount. Furthermore, male customer with high school degree have slightly more remaining amount than those with college degrees under independence, exchangeable and AR(1) working correlation matrices, but not under the unstructured working correlation matrix. On the other hand, female customers with high school degree have less remaining amount comparing to those with college school degree under all the four types of working correlation matrices.

Table 5.10 shows the estimation results for the second parametric model (Para2) using different working correlation matrices for male and female separately and a quadratic form on age. We found that unlike the overall analysis, still, different working correlation matrices will identify different significant variables: for male, the quadratic form on age are significant only when using independence working correlation matrix and for female, the quadratic term is significant when using exchangeable working correlation matrix or AR(1) structure. Although the quadratic form of age is significant, it has tiny impact on the response variable remaining amount because the number of coefficient is nearly zero. For male and female customers, still, relative to those with graduate degrees, the one with only college degrees or high school degrees have more remaining amount. The marriage factor single is significant for male under independence working correlation matrix but it is not significant under any working correlation matrices for female customers.

Table 5.11 shows the estimation results for the third parametric model (Para3) using different working correlation matrix and an exponential term on age for male and female separately. We found that using different working correlation matrices will identify similar significant properties for the exponential term on age: for male and female, the variable age with exponential term are significant when using four types of working correlation matrices. For male and female customers, still, relative to those with graduate degrees, the one with only college degrees or high school degrees have more remaining amount. Although the exponential term of age is significant under most cases, it still has tiny impact on the response variable remaining amount just like using quadratic form

in the second parametric model (Para2), because the number of coefficient is nearly zero. The marriage factor single is not significant under any type of working correlation matrices for male but it is significant under all four types of working correlation matrices for female.

The results from three parametric models show that age may be not significant using some working correlation matrices or has tiny effect with parametric patterns, such as quadratic or exponential terms. We consider semi-parametric models with kernel smoother on the predictor age, investigating whether semi-parametric models are more advanced for male or female than pure parametric models.

Table 5.12 shows the estimation results for the semi parametric model (Semi) with kernel smoother on the predictor age using different working correlation matrices. The result of estimated coefficients are similar to the estimation in parametric models. Based on the signs of the estimated coefficients, we found that male and female have the same direction for all predictors. Relative to consumers with graduate degrees, male and female customers with only college degrees or high school degrees have more remaining amount. Relative to married customers, the semi-parametric model shows that customers with single marriage status for either male or female tend to have more remaining amount.

Table 5.13 and Table 5.14 shows the mean square error results for parametric models and semi-parametric model with kernel smoother on the predictor of age. The overall MSE for training dataset is lower than testing dataset for male and female. For male, the MSE for semi-parametric model in training dataset(0.919) is higher than the MSE for parametric models (0.872) while the MSE in testing dataset for semi-parametric model is slightly lower than the MSE in testing dataset for parametric models. For female, the parametric models have similar testing error to semi-parametric models. Models fit with female customers have less MSE in testing data set than models fit with male customers.

5.3.2 Using Payment Status as Response Variable

In this part, the payment status we defined on section 5.1 will be used as the response variable to evaluate the difference between male and female customers for whether the customers will default to pay the bills. Predictors such as education level, limit balance, marriage status, and age will be used in our models. Especially, we would like to investigate whether male customer and female customer should be fitted with different models.

Table 5.9 Parameter estimations for parametric GEE model (Para1) in Gender Analysis

		Independence:male			Independence:female	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.6126	0.06129	< 2e-16	-0.398286	0.042846	< 2e-16
limit − balance	0.00207	0.00012	< 2e-16	0.001805	0.000107	< 2e-16
education − university	0.14963	0.02296	7.1e-11	0.20759	0.019351	< 2e-16
education − highschool	0.15242	0.0306	0.00000063	0.172856	0.024488	1.7e-12
single	0.05519	0.02443	0.0239	0.031259	0.018293	0.087
age	0.00375	0.0013	0.0041	-0.001575	0.000995	0.113
		Exchangeable:male			Exchangeable:female	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.455291	0.061194	1e-13	-0.185496	0.042118	0.000011
limit − balance	0.002102	0.000121	< 2e-16	0.001874	0.000107	< 2e-16
education − university	0.148497	0.022977	1e-10	0.208714	0.019376	< 2e-16
education − highschool	0.166683	0.030612	0.000000052	0.205231	0.024631	< 2e-16
single	0.01823	0.02442	0.46	-0.011694	0.018213	0.52
age	-0.000151	0.001306	0.91	-0.007483	0.000988	3.6e-14
		AR1:male			AR1:female	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.568203	0.062117	< 2e-16	-0.333886	0.043207	1.1e-14
limit − balance	0.002128	0.000124	< 2e-16	0.001821	0.000113	< 2e-16
education − university	0.163869	0.023542	3.4e-12	0.212763	0.019598	< 2e-16
education − highschool	0.176602	0.031135	0.000000014	0.187171	0.024859	5.1e-14
single	0.042972	0.024975	0.085	0.020106	0.018459	0.27606
age	0.002241	0.001309	0.087	-0.003474	0.000997	0.00049
		Unstructured:male			Unstructured:female	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.604305	0.051038	< 2e-16	-0.378	0.037	< 2e-16
limit − balance	0.002025	0.000096	< 2e-16	0.00177	0.0000838	< 2e-16
education − university	0.152826	0.018729	3.3e-16	0.217	0.0157	< 2e-16
education − highschool	0.131736	0.024389	0.000000066	0.187	0.0196	< 2e-16
single	0.0619	0.01965	0.0016	0.0267	0.0149	0.07276
age	0.002768	0.001071	0.0097	-0.00302	0.00085	0.00044

Table 5.10 Parameter estimations for parametric GEE model (Para2) in Gender Analysis

	Independence:male			Independence:female		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.54	0.0431	< 2e-16	-0.428	0.0301	< 2e-16
$limit - balance$	0.00208	0.00012	< 2e-16	0.0018	0.000106	< 2e-16
$education - university$	0.149	0.023	7.5e-11	0.208	0.0194	< 2e-16
$education - highschool$	0.152	0.0307	0.0000007	0.173	0.0246	2.1e-12
single	0.0529	0.0243	0.0298	0.0324	0.0179	0.071
age^2	0.0000457	0.0000175	0.0092	-0.0000196	0.000013	0.13
	Exchangeable:male			Exchangeable:female		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.467	0.043	< 2e-16	-0.335	0.0296	< 2e-16
$limit - balance$	0.0021	0.00012	< 2e-16	0.00185	0.000106	< 2e-16
$education - university$	0.149	0.023	1e-10	0.209	0.0194	< 2e-16
$education - highschool$	0.165	0.0307	0.000000071	0.203	0.0247	2.2e-16
single	0.0219	0.0243	0.37	-0.00266	0.0179	0.88
age^2	0.00000305	0.0000175	0.86	-0.0000864	0.0000128	1.3e-11
	AR1:male			AR1:female		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.527	0.0439	< 2e-16	-0.403	0.0307	< 2e-16
$limit - balance$	0.00213	0.000124	< 2e-16	0.00181	0.000113	< 2e-16
$education - university$	0.164	0.0235	3.5e-12	0.213	0.0196	< 2e-16
$education - highschool$	0.176	0.0312	0.000000016	0.186	0.025	8.8e-14
single	0.0424	0.0249	0.088	0.0241	0.0181	0.183
age^2	0.0000284	0.0000175	0.106	-0.0000405	0.000013	0.0018
	Unstructured:male			Unstructured:female		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.53	0.0419	< 2e-16	-0.432	0.0345	< 2e-16
$limit - balance$	0.00201	0.000116	< 2e-16	0.00176	0.000104	< 2e-16
$education - university$	0.142	0.0222	1.5e-10	0.186	0.0187	< 2e-16
$education - highschool$	0.143	0.0296	0.0000012	0.159	0.0247	1.1e-10
single	0.0351	0.0234	0.134	0.0286	0.0184	0.12
age^2	0.0000292	0.000017	0.086	-0.0000219	0.0000177	0.22

Table 5.11 Parameter estimations for parametric GEE model (Para3) in Gender Analysis

	Independence:male			Independence:female		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.462	0.0291	< 2e-16	-0.455	0.0246	< 2e-16
limit − balance	0.00209	0.000121	< 2e-16	0.00179	0.000106	< 2e-16
education − university	0.15	0.023	7.2e-11	0.207	0.0193	< 2e-16
education − highschool	0.167	0.0301	0.000000029	0.164	0.0235	3.1e-12
single	0.0209	0.0204	0.31	0.0426	0.0166	0.01
exp(*age*)	1.68e-34	1.33e-36	< 2e-16	-8.05e-34	2.64e-35	< 2e-16
	Exchangeable:male			Exchangeable:female		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.462	0.0291	< 2e-16	-0.455	0.0246	< 2e-16
limit − balance	0.0021	0.000121	< 2e-16	0.00179	0.000106	< 2e-16
education − university	0.149	0.023	8.1e-11	0.207	0.0193	< 2e-16
education − highschool	0.167	0.0301	0.000000031	0.164	0.0235	3.1e-12
single	0.0205	0.0204	0.32	0.0426	0.0166	0.01
exp(*age*)	1.1e-34	8.43e-37	< 2e-16	-6.78e-34	1.37e-35	< 2e-16
	AR1:male			AR1:female		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.478	0.03	< 2e-16	-0.459	0.0254	< 2e-16
limit − balance	0.00214	0.000125	< 2e-16	0.00178	0.000113	< 2e-16
education − university	0.164	0.0235	3.1e-12	0.212	0.0196	< 2e-16
education − highschool	0.186	0.0307	1.5e-09	0.168	0.0239	1.9e-12
single	0.0227	0.0209	0.28	0.0453	0.0169	0.0073
exp(*age*)	1.39e-34	1.11e-36	< 2e-16	-6.22e-34	1.52e-35	< 2e-16
	Unstructured:male			Unstructured:female		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.48	0.0281	< 2e-16	-0.462	0.0239	< 2e-16
limit − balance	0.00201	0.000117	< 2e-16	0.00174	0.000105	< 2e-16
education − university	0.142	0.0221	1.3e-10	0.185	0.0187	< 2e-16
education − highschool	0.153	0.0291	0.00000014	0.149	0.0231	9.6e-11
single	0.0148	0.0197	0.45	0.0405	0.0161	0.012
exp(*age*)	1.34e-34	1e-36	< 2e-16	2.43-33	1.75e-35	< 2e-16

Table 5.12 Parameter estimations for parametric GEE model (Semi) in Gender Analysis

		Independence:male			Independence:female	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.48386	0.02491	< 2e-16	-0.485	0.0206	< 2e-16
$limit - balance$	0.00218	0.0001	< 2e-16	0.00182	0.0000858	< 2e-16
$education - university$	0.16564	0.01941	< 2e-16	0.242	0.0163	< 2e-16
$education - highschool$	0.16031	0.02502	1.5e-10	0.19	0.0195	< 2e-16
single	0.04224	0.01711	0.014	0.05	0.0139	0.00031
		Exchangeable:male			Exchangeable:female	
Intercept	-0.48386	0.02491	< 2e-16	-0.485	0.0206	< 2e-16
$limit - balance$	0.00218	0.0001	< 2e-16	0.00182	0.0000858	< 2e-16
$education - university$	0.16564	0.01941	< 2e-16	0.242	0.0163	< 2e-16
$education - highschool$	0.16031	0.02502	1.5e-10	0.19	0.0195	< 2e-16
single	0.04224	0.01711	0.014	0.05	0.0139	0.00031
		AR1:male			AR1:female	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.497014	0.025293	< 2e-16	-0.486	0.0209	< 2e-16
$limit - balance$	0.002215	0.000102	< 2e-16	0.0018	0.0000886	< 2e-16
$education - university$	0.177118	0.019685	< 2e-16	0.245	0.0164	< 2e-16
$education - highschool$	0.173217	0.025292	7.4e-12	0.191	0.0196	< 2e-16
single	0.041659	0.017318	0.016	0.0513	0.014	0.00025
		Unstructured:male			Unstructured:female	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.472	0.024	< 2e-16	-0.475	0.0199	< 2e-16
$limit - balance$	0.00204	0.0000968	< 2e-16	0.00174	0.0000834	< 2e-16
$education - university$	0.151	0.0188	6.7e-16	0.216	0.0157	< 2e-16
$education - highschool$	0.141	0.0242	5.3e-09	0.17	0.0189	< 2e-16
single	0.0367	0.0166	0.027	0.0484	0.0134	0.00029

Table 5.13 Overall MSE for parametric models and semi-parametric model:male

ar1	para1	para2	para3	semi
training	0.872	0.871	0.870	0.919
testing	1.019	1.020	1.020	1.018
unstructured	para1	para2	para3	semi
training	0.921	0.872	0.871	0.919
testing	1.022	1.024	1.025	1.021

Table 5.14 Overall MSE for parametric models and semi-parametric model:female

ar1	para1	para2	para3	semi
training	0.851	0.850	0.851	0.847
testing	0.916	0.916	0.917	0.945
unstructured	para1	para2	para3	semi
training	0.850	0.853	0.853	0.848
testing	0.946	0.916	0.917	0.946

We first consider three parametric GEE model with linear form of age for male and female as the following:

$$Para4\ for\ male:\ logit(p) = \beta_0 + \beta_1 limit\ balance \times I_M + education \times I_M + marriage \times I_M + \beta_2 age \times I_M \tag{5.17}$$

$$Para4\ for\ female:\ logit(p) = \beta_0 + \beta_1 limit\ balance \times I_F + education \times I_F + marriage \times I_F + \beta_2 age \times I_F \tag{5.18}$$

where p is the probability of default.

Table 5.15 shows the estimation results for the parametric model (Para4) using different working correlation matrices. Relative to male and female consumers with graduate degrees, customers with college degrees or high school degrees have more probability to default. For male customers, relative to married customers, customers with single marriage status tend to have more probability to default. The predictor limit balance has negative coefficient for male and female, which denotes that limit balance has negative correlation with the probability of default. For male, age is not significant under any types of working correlation matrices but for female, age is significant with negative coefficients with the probability of default for all four types of working correlation matrices. Different working correlation matrices such as independence, exchangeable, AR(1) and unstructured matrix are used in Para4 for both male and female. The estimated coefficients by those four working correlation matrices are quite similar.

Table 5.16 and 5.17 shows the predictive accuracy tables in parametric model when the response variable is delay status for male and female. For male, the testing accuracy (0.702) is lower than training accuracy (0.706) and for female, testing accuracy (0.65) is slightly lower than training accuracy (0.672). The model with female customers has lower accuracy in training and testing dataset comparing to male customers.

5.4 Results and Discussion: Education Analysis

In this part, we evaluate the difference between models for customers with different education levels. Following the overall analysis in section 5.2, three parametric models and one semi-parametric model are fitted for analysis and we used two different outcomes: remaining amount and payment status as the response variable. The BIC values for the four models for fitting different education levels separately and together are calculated and used for model comparison. Estimated coefficients for all models are reported for the purpose of exploring the difference among the fitted models for different education levels. We provide mean square error as the evaluation measurement for the comparison of parametric and semi-parametric models when using remaining amount as response variable and we use predictive accuracy as the evaluation measurement when using payment status as response variable.

5.4.1 Using Remaining Amount as Response Variable

Model Setups

The remaining amount we defined on section 5.1 will be used as the response variable to explore the relationship between the amount of owed payments and some predictors: such as gender, limit balance, marriage status, and age for customers with different education levels. As overall analysis in section 5.2, the following three parametric GEE models will be fitted for customers with different education levels separately:

$$Para1\ for\ highschool:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_H + gender \times I_H + marriage \times I_H + \beta_2 age \times I_H \tag{5.19}$$

$$Para1\ for\ advancedegree:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_A + gender \times I_A + marriage \times I_A + \beta_2 age \times I_A \tag{5.20}$$

$$Para2\ for\ highschool:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_H + gender \times I_H + marriage \times I_H + \beta_2 age^2 \times I_H \tag{5.21}$$

$$Para2\ for\ advancedegree:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_A + gender \times I_A + marriage \times I_A + \beta_2 age^2 \times I_A \tag{5.22}$$

$$Para3\ for\ highschool:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_H + gender \times I_H + marriage \times I_H + \beta_2 exp(age) \times I_H \tag{5.23}$$

$$Para3\ for\ advancedegree:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_A + gender \times I_A + marriage \times I_A + \beta_2 exp(age) \times I_A \tag{5.24}$$

and we consider a semi-parametric model with non-parametric form on the predictor age:

$$Semi\ for\ highschool: remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_H + gender \times I_H + marriage \times I_H + \theta(age) \times I_H \tag{5.25}$$

$$Semi\ for\ advancedegree: remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_A + gender \times I_A + marriage \times I_A + \theta(age) \times I_A \tag{5.26}$$

where $\theta(.)$ is a kernel smoother, I_H and I_A are indicator variables, defined as following:

$$I_H = \begin{cases} 0 & if\ \ degree\ \ is\ \ unversity/graduate \\ 1 & if\ \ degree\ \ is\ \ highschool \end{cases}$$

$$I_A = \begin{cases} 0 & if\ \ degree\ \ is\ \ highschool \\ 1 & if\ \ degree\ \ is\ \ unversity/graduate \end{cases}$$

BIC Analysis

We use BIC (Bayesian Information Criterion) to explore what is the difference when we fit models for customers with different education levels separately comparing to we fit a model with all customers together. The BIC defined as the same as in 5.14. When fitting the all customers with different education levels together, we used BIC calculated from models in section 5.2. While fitting models for different education levels separately, we use BIC derived from models in (5.19)-(5.26), combining error variance from models with indicator variables I_H and I_A as following:

$$BIC_{separate} = n \times log(\hat{\sigma_{e_H}}^2 + \hat{\sigma_{e_A}}^2) + 2(p+1) \times ln(n) \tag{5.27}$$

where error variance, n and p are defined as the same as in BIC (5.10).

Table 5.15 Parameter estimations for parametric GEE model (Para4) in Gender Analysis

		Independence:male			Independence:female	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	0.901231	0.127831	1.80e-12	1.51457	0.09293	<2e-16
limit − balance	-0.003501	0.000174	< 2e-16	-0.00337	0.00015	<2e-16
education − university	0.625329	0.050368	< 2e-16	0.65678	0.03909	<2e-16
education − highschool	0.53462	0.069524	1.50e-14	0.64858	0.05766	<2e-16
single	0.238001	0.052976	7.00e-06	0.0589	0.03933	0.13
age	0.002483	0.002808	0.38	-0.02031	0.00224	<2e-16
		Exchangeable:male			Exchangeable:female	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	0.901231	0.127831	1.80e-12	1.51457	0.09293	<2e-16
limit − balance	-0.003501	0.000174	< 2e-16	-0.00337	0.00015	<2e-16
education − university	0.625329	0.050368	< 2e-16	0.65678	0.03909	<2e-16
education − highschool	0.53462	0.069524	1.50e-14	0.64858	0.05766	<2e-16
single	0.238001	0.052976	7.00e-06	0.0589	0.03933	0.13
age	0.002483	0.002808	0.38	-0.02031	0.00224	<2e-16
		ar1:male			ar1:female	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	0.910599	0.1258	4.50e-13	1.458421	0.092645	< 2e-16
limit − balance	-0.003303	0.000172	< 2e-16	-0.003312	0.000149	< 2e-16
education − university	0.595055	0.049767	< 2e-16	0.634428	0.038782	< 2e-16
education − highschool	0.50367	0.068254	1.60e-13	0.615471	0.057415	< 2e-16
single	0.2269	0.052231	1.40e-05	0.073722	0.039148	0.06
age	0.002179	0.002767	0.43	-0.017833	0.002228	1.20e-15
		unstructured:male			unstructured:female	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	0.95624	0.126489	4.00e-14	1.514816	0.091951	<2e-16
limit − balance	-0.003466	0.000172	< 2e-16	-0.003343	0.000148	<2e-16
education − university	0.617262	0.049916	< 2e-16	0.640811	0.038587	<2e-16
education − highschool	0.52332	0.068711	2.60e-14	0.626088	0.057129	<2e-16
single	0.230574	0.052501	1.10e-05	0.065367	0.038924	0.093
age	0.002166	0.002782	0.44	-0.018843	0.002211	<2e-16

Table 5.16 Predictive Accuracy for Parametric GEE model (Para4):male

training	ar1			accuracy	testing	ar1			accuracy
		pred					pred		
	true	0	1			*true*	0	1	
	0	2039	11643			0	1150	5864	
	1	1987	30633	0.706		1	1028	15106	0.702
	unstructured					unstructured			
		pred					pred		
	true	0	1			*true*	0	1	
	0	2133	11549			0	1197	5817	
	1	2043	30577	0.706		1	1053	15081	0.703

Table 5.17 Predictive Accuracy for Parametric GEE model (Para4):female

training	ar1			accuracy	testing	ar1			accuracy
		pred					pred		
	true	0	1			*true*	0	1	
	0	4682	19191			0	2759	10513	
	1	3868	42609	0.672		1	2161	20105	0.65
	unstructured					unstructured			
		pred					pred		
	true	0	1			*true*	0	1	
	0	4644	19229			0	2711	10201	
	1	3798	42679	0.673		1	2125	10141	0.65

Table 5.18 shows the BIC values calculated in (5.27) when using different models for different education levels separately and the BIC in (5.14) for models fitting with all customers with different education levels together. A small BIC denotes a better model fitting. The BIC values for the four types of models fitted with different education levels separately are larger than the BIC values when the models are fitted for the overall model with all education levels together. Furthermore, parametric models fitting with exponential term on age with all education levels together is the best model, since it has the smallest BIC.

Table 5.18 BIC: Education Analysis

model	para1	para2	para3	semi
total	-11116.26	-11122.03	-11278.38	-10966.76
separate	-10046.14	-10201.68	-10051.09	-9931.403

Table 5.19 shows the estimated results for the first parametric model (Para1) using different working correlation matrices for customer with different education levels separately. We found that for customers with high school degree, variable single and age are not significant using any types of working correlation matrices. For customers have university and graduate degrees, variable single is not significant when using any types of working correlation matrices but age is significant only when using exchangeable working correlation matrix. For all models, limit balance is positive correlated with response variable remaining amount. Variable female is negative correlated with remaining amount which indicates that relative to male customers, females have less remaining amount because of a negative estimated coefficients. The estimated coefficients for female in customers with advanced degrees is as half as the coefficients for high school customers.

Table 5.20 shows the estimation results for the second parametric model (Para2) using different working correlation matrices and a quadratic term on age for customer with different education levels separately. We found that unlike the previous model (Para1), different working correlation matrices will identify the same significant results: variable single and age are not significant using any types of working correlation matrices. For all models, still, limit balance is positive correlated with response variable remaining amount while variable female is negative correlated with remaining amount. The estimated coefficients for female in customers with university or graduate degrees is as half as the coefficients for high school customers.

Table 5.21 shows the estimation result for the third parametric model (Para3) using different working correlation matrix and an exponential term on age for customers with different education levels separately. We found that using different working correlation matrices will identify similar significant variables for the exponential term on age: for customers with high school degrees, the variable age with exponential term are not significant when using any types of working correlation matrices;for customers with university or graduate degrees, the variable age with exponential term are significant when using all four types of working correlation matrices. For all models, limit balance is positive correlated with response variable remaining amount while female is negative correlated with remaining amount. Although the exponential term of age is significant under models fitting for customers with university or graduate degrees, it has tiny impact on the response variable remaining amount, because the number of coefficient is nearly zero. Variable single is not significant under any types of working correlation matrices for all models

The results from three parametric models show that age may be significant or has tiny effect with parametric patterns, such as quadratic or exponential terms. We consider semi-parametric models with kernel smoother on the predictor age, investigating whether there is a difference for customers with different education levels, seeing whether semi-parametric models are better than pure parametric models for education analysis.

Table 5.22 shows the estimation results for the semi parametric model with kernel smoother on the predictor age using different working correlation matrices. The result of the estimated coefficients have some similarities to the estimation in parametric models. Based on the signs of the estimated coefficients, we found that relative to male consumers, female consumers have less remaining amount. The predictor limit balance has positive coefficient, which denotes that limit balance has positive correlation with remaining amount. Except single, all other predictors are significant for all models when using any types of four working correlation matrices.

Table 5.23 and Table 5.24 shows the mean square error results for parametric models and semi-parametric model with kernel smoother on the predictor of age. The overall MSE for training dataset is lower than testing dataset for customers with all education levels. For high school customers, the MSE for semi-parametric model in training dataset(0.912) is higher than the MSE for parametric models (0.823) while the MSE in testing dataset for semi-parametric model is slightly lower than the MSE in testing dataset for parametric models. For customers with university or

graduate degrees, the parametric models have similar testing error to semi-parametric models (1.03). MSE for models fitting with customers have university or graduate degrees are smaller than the models for high school customers.

5.4.2 Using Payment Status as Response Variable

In this part, the payment status we defined on section 5.1 will be used as the response variable to evaluate the difference between customers with different education levels for whether the customers will default to pay the bills. Predictors such as gender, limit balance, marriage status, and age will be used in our models. Especially, we would like to investigate whether customer with high school degrees and customer with university or graduate degrees should be fitted with different models. We first consider three parametric GEE model with linear form of age for high school customers and university or graduate customers as the following:

$$Para4\ for\ highschool:\ logit(p) = \beta_0 + \beta_1 limit\ balance \times I_H + sex \times I_H + marriage \times I_H + \beta_2 age \times I_H \tag{5.28}$$

$$Para4\ for\ advancedegree:\ logit(p) = \beta_0 + \beta_1 limit\ balance \times I_A + sex \times I_A + marriage \times I_A + \beta_2 age \times I_A \tag{5.29}$$

where p is the probability of default.

Table 5.25 shows the estimation results for the parametric model (Para4) using different working correlation matrices. For all models, female customers tend to have less probability to default because of negative coefficients. The predictor limit balance has negative coefficient, which denotes that limit balance has negative correlation with the probability of default. Age is not significant for customers with high school degrees but it is significant for customers with university or graduate degrees. The variable single is not significant for all models under any types of working correlation matrices.

Different working correlation matrices such as independence, exchangeable, AR1 and unstructured matrix are used in Para4 for all customers. The estimated parameters using those four working correlation matrices are quite similar.

Table 5.26 and 5.27 shows the predictive accuracy tables in parametric model when the response variable is delay status for customers with different education levels. For customers with high school degree, the testing accuracy (0.722) is lower than training accuracy (0.737) and for customers with advanced degree, testing accuracy (0.647) is slightly lower than training accuracy (0.665).

Table 5.19 Parameter estimations for parametric GEE model (Para1) in Education Analysis

	Independence:highschool			Independence:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.20659	0.080575	0.0103	-0.277	0.0478	0.000000007
$limit - balance$	0.002668	0.00025	<2e-16	0.00163	0.0000806	<2e-16
single	-0.013749	0.032611	0.6733	0.0145	0.0161	0.3655
$female$	-0.091633	0.031598	0.0037	-0.0395	0.0142	0.0053
age	0.000497	0.00136	0.7149	0.0011	0.000971	0.2587
	Exchangeable:highschool			Exchangeable:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.0995	0.08088	0.219	-0.12	0.04777	0.0116
$limit - balance$	0.00268	0.00025	<2e-16	0.00167	0.0000808	<2e-16
single	-0.03146	0.03262	0.335	-0.0162	0.0161	0.3143
$female$	-0.09387	0.03162	0.003	-0.0494	0.0142	0.0005
age	-0.00191	0.00136	0.162	-0.00266	0.00097	0.0061
	AR1:highschool			Exchangeable:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.159	0.08211	0.0528	-0.226	0.0486	0.0000033
$limit - balance$	0.002729	0.000258	<2e-16	0.00165	0.0000848	<2e-16
single	-0.017719	0.033436	0.5962	0.00596	0.0163	0.7142
$female$	-0.102418	0.032023	0.0014	-0.0437	0.0145	0.0026
age	-0.000294	0.001391	0.8326	-0.00012	0.000973	0.9015
	Unstructured:highschool			Exchangeable:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.263177	0.066796	0.000081	-0.241	0.038	2.2e-10
$limit - balance$	0.002414	0.000196	< 2e-16	0.00163	0.0000636	< 2e-16
single	-0.022155	0.025928	0.393	0.0136	0.013	0.3
$female$	-0.067418	0.025409	0.008	-0.048	0.0113	0.000021
age	0.000833	0.001145	0.467	-0.000411	0.000779	0.6

Table 5.20 Parameter estimations for parametric GEE model (Para2) in Education Analysis

	Independence:highschool			Independence:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.289	0.0471	8e-10	-0.3	0.0268	<2e-16
$limit - balance$	0.00267	0.00025	<2e-16	0.00163	0.0000802	< 2e-16
$single$	-0.0139	0.0322	0.6666	0.0154	0.0159	0.3324
$female$	-0.0916	0.0316	0.0037	-0.0391	0.0141	0.0056
age^2	0.00000624	0.0000171	0.7144	0.0000164	0.0000134	0.2221
	Exchangeable:highschool			Exchangeable:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.234	0.0471	0.00000069	-0.231	0.0267	< 2e-16
$limit - balance$	0.00267	0.000251	<2e-16	0.00166	0.0000804	<2e-16
$single$	-0.0294	0.0322	0.361	-0.0101	0.0159	0.52744
$female$	-0.0937	0.0316	0.003	-0.0476	0.0141	0.00076
age^2	-0.0000211	0.0000171	0.216	-0.000026	0.0000134	0.05164
	AR1:highschool			Exchangeable:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.268	0.0485	0.000000032	-0.279	0.0273	<2e-16
$limit - balance$	0.00273	0.000258	< 2e-16	0.00165	0.0000845	< 2e-16
$single$	-0.0172	0.033	0.602	0.00849	0.0161	0.598
$female$	-0.102	0.032	0.0014	-0.0429	0.0144	0.003
age^2	-0.00000291	0.0000175	0.8679	0.00000256	0.0000134	0.848
	Unstructured:high/school			Exchangeable:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.312	0.0472	4e-11	-0.307	0.0259	<2e-16
$limit - balance$	0.00263	0.00025	<2e-16	0.00158	0.0000778	< 2e-16
$single$	-0.0131	0.0314	0.6758	0.00395	0.0153	0.797
$female$	-0.0795	0.0307	0.0096	-0.0308	0.0136	0.023
age^2	-0.000000102	0.0000169	0.9952	0.00000171	0.000013	0.896

Table 5.21 Parameter estimations for parametric GEE model (Para3) in Education Analysis

	Independence:highschool			Independence:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.277	0.0292	<2e-16	-0.274	0.0159	<2e-16
$limit - balance$	0.00267	0.000251	<2e-16	0.00164	0.0000801	<2e-16
$single$	-0.017	0.03	0.571	0.00619	0.0138	0.6544
$female$	-0.0918	0.0317	0.0038	-0.0416	0.014	0.0029
$exp(age)$	4.56e-33	3.94e-33	0.2473	1.72e-34	1.1e-36	<2e-16
	Exchangeable:highschool			Exchangeable:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.277	0.0292	<2e-16	-0.274	0.0159	<2e-16
$limit - balance$	0.00267	0.000251	<2e-16	0.00164	0.0000801	<2e-16
$single$	-0.0171	0.03	0.5698	0.00598	0.0138	0.6651
$female$	-0.0919	0.0317	0.0037	-0.0418	0.014	0.0027
$exp(age)$	3.98e-33	2.84e-33	0.16	1.14e-34	8.84e-37	<2e-16
	AR1:highschool			Exchangeable:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.275	0.03	<2e-16	-0.275	0.0165	<2e-16
$limit - balance$	0.00273	0.000259	<2e-16	0.00164	0.0000844	<2e-16
$single$	-0.0152	0.0306	0.6196	0.00745	0.0141	0.5965
$female$	-0.102	0.0321	0.0015	-0.0427	0.0142	0.0027
$exp(age)$	3.97e-33	3.39e-33	0.2408	1.42e-34	9.59e-37	<2e-16
	Unstructured:highschool			Exchangeable:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.313	0.0291	<2e-16	-0.304	0.0153	<2e-16
$limit - balance$	0.00263	0.000251	<2e-16	0.00158	0.0000777	<2e-16
$single$	-0.0126	0.0292	0.67	0.00344	0.0133	0.796
$female$	-0.0792	0.0308	0.01	-0.0305	0.0135	0.024
$exp(age)$	4.73e-33	3.3e-33	0.15	1.36e-34	7.55e-37	<2e-16

Table 5.22 Parameter estimations for semi-parametric GEE model (Semi) in Education Analysis

		Independence:highschool			Independence:university/graduate	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.257924	0.024687	<2e-16	-0.28	0.0132	<2e-16
$limit - balance$	0.002331	0.000201	<2e-16	0.00151	0.0000656	<2e-16
single	0.006517	0.025091	0.7951	0.0647	0.0116	0.000000025
$female$	-0.071688	0.02624	0.0063	-0.0375	0.0116	0.0012
		Exchangeable:highschool			Exchangeable:university/graduate	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.257924	0.024687	<2e-16	-0.28	0.0132	<2e-16
$limit - balance$	0.002331	0.000201	<2e-16	0.00151	0.0000656	<2e-16
single	0.006517	0.025091	7.95e-01	0.0647	0.0116	0.000000025
$female$	-0.071688	0.02624	0.0063	-0.0375	0.0116	0.0012
		AR1:highschool			Exchangeable:university/graduate	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.256116	0.025051	<2e-16	-0.28	0.0136	<2e-16
$limit - balance$	0.002369	0.000204	<2e-16	0.00151	0.0000675	<2e-16
single	0.003693	0.025414	0.8845	0.0648	0.0117	0.000000032
$female$	-0.080116	0.026428	0.0024	-0.0397	0.0118	0.00075
		Unstructured:highschool			Exchangeable:university/graduate	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.293958	0.024131	<2e-16	-0.313	0.0128	<2e-16
$limit - balance$	0.002246	0.000195	<2e-16	0.00143	0.0000633	<2e-16
single	0.008067	0.024251	0.739	0.0621	0.0112	0.000000027
$female$	-0.060963	0.025294	1.60E-02	-0.0275	0.0112	0.014

Table 5.23 Overall MSE for parametric models and semi-parametric model:highschool

ar1	para1	para2	para3	semi
training	0.823	0.823	0.823	0.912
testing	1.105	1.104	1.106	1.101
unstructured	para1	para2	para3	semi
training	0.916	0.823	0.823	0.912
testing	1.101	1.103	1.105	1.103

Table 5.24 Overall MSE for parametric models and semi-parametric model:university/graduate

ar1	para1	para2	para3	semi
training	0.903	0.903	0.901	0.950
testing	1.043	1.042	1.043	1.044
unstructured	para1	para2	para3	semi
training	0.949	0.903	0.901	0.951
testing	1.045	1.046	1.045	1.048

Table 5.25 Parameter estimations for parametric GEE model (Para4) in Education Analysis

	Independence:highschool			Independence:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	2.119766	0.209298	<2e-16	2.238801	0.102844	<2e-16
$limit - balance$	-0.004669	0.000327	<2e-16	-0.003824	0.000117	<2e-16
$female$	-0.241688	0.074098	0.0011	-0.223287	0.031281	9.50e-13
single	0.024956	0.076318	0.7437	0.00605	0.033688	0.86
age	-0.002289	0.003489	0.5118	-0.014294	0.001944	1.90e-13
	Exchangeable:highschool			Exchangeable:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	2.243006	0.209331	<2e-16	2.394642	0.103012	<2e-16
$limit - balance$	-0.004661	0.000327	<2e-16	-0.003786	0.000117	<2e-16
$female$	-0.244425	0.074128	0.00098	-0.23278	0.031333	1.10e-13
single	0.005483	0.076293	0.9427	-0.024279	0.033735	0.47
age	-0.00505	0.003479	0.14657	-0.018028	0.001943	<2e-16
	ar1:highschool			ar1:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	2.01782	0.2052	<2e-16	2.232723	0.101835	< 2e-16
$limit - balance$	-0.0045	0.00032	<2e-16	-0.003671	0.000116	< 2e-16
$female$	-0.20183	0.07295	0.0057	-0.205443	0.031025	3.50e-11
single	0.04584	0.07602	0.5466	-0.004401	0.033448	0.9
age	-0.00192	0.00346	0.5792	-0.014837	0.001926	1.30e-14
	unstructured:highschool			unstructured:university/graduate		
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	2.183044	0.206479	<2e-16	2.308166	0.101469	<2e-16
$limit - balance$	-0.00465	0.000323	<2e-16	-0.003745	0.000115	<2e-16
$female$	-0.230581	0.073474	0.0017	-0.21921	0.030957	1.40e-12
single	0.032289	0.076019	0.671	-0.007885	0.03334	0.81
age	-0.003462	0.003457	0.3166	-0.015355	0.001915	1.10e-15

The model for customers with advanced degrees have lower accuracy in training and testing dataset comparing to high school customers.

5.5 Results and Discussion: Marriage Status Analysis

In this part, we evaluate the difference between models for customers with different marriage status. Following the overall analysis in section 5.2, three parametric models and one semi-parametric model are fitted for analysis and we used two different outcomes: remaining amount and payment status as the response variable. The BIC values for the four models for fitting single customers and married customers separately and together are calculated and used for model comparison. Estimated coefficients for all models are reported for the purpose of exploring the difference among the fitted models for single and married customers. We provide mean square error as the evaluation measurement for the comparison of parametric and semi-parametric models when using remaining amount as response variable and we use predictive accuracy as the evaluation measurement when using payment status as response variable.

5.5.1 Using Remaining Amount as Response Variable

Model Setups

The remaining amount we defined on section 5.1 will be used as the response variable to explore the relationship between the amount of owed payments and some predictors: such as gender, limit balance, education level, and age for customers with different marriage status. As overall analysis in section 5.2, the following three parametric GEE models will be fitted for single and married customers separately:

$$Para1\ for\ single:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_S + education \times I_S + gender \times I_S + \beta_2 age \times I_S \tag{5.30}$$

$$Para1\ for\ married:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_{MA} + education \times I_{MA} + gender \times I_{MA} + \beta_2 age \times I_{MA} \tag{5.31}$$

$$Para2\ for\ single:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_S + education \times I_S + gender \times I_S + \beta_2 age^2 \times I_S \tag{5.32}$$

Table 5.26 Predictive Accuracy for Parametric GEE model (Para4):highschool

training	ar1			accuracy	testing	ar1			accuracy
		pred					pred		
	true	0	1			*true*	0	1	
	0	480	4539			0	372	2359	
	1	474	13587	0.737		1	296	6513	0.722
	unstructured					unstructured			
		pred					pred		
	true	0	1			*true*	0	1	
	0	475	4544			0	372	2359	
	1	480	13581	0.737		1	300	6509	0.721

Table 5.27 Predictive Accuracy for Parametric GEE model (Para4):advanced degree

training	ar1			accuracy	testing	ar1			accuracy
		pred					pred		
	true	0	1			*true*	0	1	
	0	4799	27820			0	2759	10513	
	1	4884	60069	0.665		1	2161	20105	0.65
	unstructured					unstructured			
		pred					pred		
	true	0	1			*true*	0	1	
	0	4776	27843			0	2630	14482	
	1	4845	60108	0.665		1	2741	28933	0.647

$$Para2\ for\ married:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_{MA} + education \times I_{MA} + gender \times I_{MA} + \beta_2 age^2 \times I_{MA} \tag{5.33}$$

$$Para3\ for\ single:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_S + education \times I_S + gender \times I_S + \beta_2 exp(age) \times I_S \tag{5.34}$$

$$Para3\ for\ married:\ remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_{MA} + education \times I_{MA} + gender \times I_{MA} + \beta_2 exp(age) \times I_{MA} \tag{5.35}$$

and we consider a semi-parametric model with non-parametric form on the predictor age:

$$Semi\ for\ single: remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_S + education \times I_S + gender \times I_S + \theta(age) \times I_S \tag{5.36}$$

$$Semi\ for\ married: remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_{MA} + education \times I_{MA} + gender \times I_{MA} + \theta(age) \times I_{MA} \tag{5.37}$$

where $\theta(.)$ is a kernel smoother, I_S and I_{MA} are indicator variables, defined as following:

$$I_S = \begin{cases} 0 & if \quad marriagestatus \quad is \quad married \\ 1 & if \quad marriagestatus \quad is \quad single \end{cases}$$

$$I_{MA} = \begin{cases} 0 & if \quad marriagestatus \quad is \quad single \\ 1 & if \quad marriagestatus \quad is \quad married \end{cases}$$

BIC analysis

We use BIC (Bayesian Information Criterion) to explore what is the difference when we fit models for customers with different marriage status separately comparing to we fit a model with all customers together. The BIC defined as the same as in 5.14.

When fitting the all customers with different education levels together, we used BIC calculated from models in section 5.2. While fitting models for different education levels separately, we use BIC derived from models in (5.30)-(5.37), combining error variance from models with indicator variables I_S and I_{MA} as following:

$$BIC_{separate} = n \times log(\hat{\sigma_{e_S}}^2 + \hat{\sigma_{e_{MA}}}^2) + 2(p+1) \times ln(n) \tag{5.38}$$

Table 5.28 BIC: Marriage Analysis

model	para1	para2	para3	semi
overall	-11116.26	-11122.03	-11278.38	-10966.76
separate	-11523.47	-11525.86	-11636.53	-11557.32

where error variance, n and p are defined as the same as in BIC (5.10).

Table 5.28 shows the BIC values calculated in (5.38) when using different models for single and married customers separately and the BIC in (5.14) for models fitting with all customers with different marriage status together. A small BIC denotes a better model fitting. The BIC values for the four types of models fitted with single and married customers separately are smaller than the BIC values when the models are fitted for the overall model with all the customers together. Furthermore, parametric models fitting with exponential term on age with single and married customers separately is the best model, since it has the smallest BIC.

Table 5.29 shows the estimation results for the first parametric model (Para1) using different working correlation matrices for single and married customers separately. We found that unlike the overall analysis, different working correlation matrices will identify different significant variables: for single customers, variable age is not significant when using independence, AR(1) or unstructured working correlation matrix while age with exchangeable working correlation matrix is significant; for married customers, age is significant only when using independence working correlation matrix. Limit balance has positive coefficients while female has negative coefficients, which indicates that female have less remaining amount relative to male. Education factor university and graduate are significant, which says that relative to customer with high school degree, customers with advanced degree have more remaining amount.

Table 5.30 shows the estimation results for the second parametric model (Para2) using different working correlation matrices for single and married customers separately and a quadratic form on age. We found that unlike the overall analysis, still, different working correlation matrices will identify different significant variables: for single customers, the quadratic form on age is significant only when using independence working correlation matrix and for married customers, the quadratic term is significant when using independence working correlation matrix. Although the quadratic form of age is significant under independence working correlation matrix, it has tiny impact on the response variable remaining amount because the number of coefficient is nearly zero. Limit

balance, university and graduate are significant and has positive coefficients with response variable while female has negative coefficient with remaining amount.

Table 5.31 shows the estimation results for the third parametric model (Para3) using different working correlation matrix and an exponential term on age for single and married customers separately. For single customers, all predictors including the exponential term of age are significant when using any four types of working correlation matrices while for married customers, the exponential term on age is significant only when using exchangeable working correlation matrix. Like the previous parametric models, limit balance, university and high school are significant with positive coefficients while female is negative correlated with remaining amount. Although the exponential term of age is significant, it still has tiny impact on remaining amount just like model (Para2), because the number of coefficient is nearly zero.

The results from three parametric models show that age may be not significant using some working correlation matrices or has tiny effect with parametric patterns, such as quadratic or exponential terms. We consider semi-parametric models with kernel smoother on the predictor age, investigating whether semi-parametric models are more advanced for single or married customers than pure parametric models.

Table 5.32 shows the estimation results for the semi parametric model with kernel smoother on the predictor age using different working correlation matrices. The result of the estimated coefficients are similar to the estimation in parametric models. Limit balance, university and graduate are significantly positive correlated with remaining amount while variable female are significant and negative correlated with remaining amount.

Table 5.33 and Table 5.34 shows the mean square error results for parametric models and semi-parametric model with kernel smoother on the predictor of age. The overall MSE for training dataset is lower than testing dataset for single and married. For single customers, the MSE for semi-parametric model in training dataset(0.943) is higher than the MSE for parametric models (0.913) while the MSE in testing dataset for semi-parametric model is slightly lower than the MSE in testing dataset for parametric models. For married customers, semi parametric models has more MSE in testing dataset than parametric models. For both single and married customers, different working correlation matrices deliver different MSE: the MSE calculated in AR(1) structure are lower than the MSE calculated under unstructured working correlation matrix.

Table 5.29 Parameter estimations for parametric GEE model (Para1) in Marriage Analysis

		Independence:single			Independence:married	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.2748613	0.0615815	0.00000807	-0.470722	0.04618	< 2e-16
limit − balance	0.0017218	0.0001104	< 2e-16	0.002088	0.000111	< 2e-16
education − university	0.1869339	0.024237	1.23e-14	0.177638	0.018478	< 2e-16
education − highschool	0.1836246	0.027587	2.81e-11	0.119591	0.027003	0.0000095
female	-0.0834457	0.0185955	0.00000721	-0.036156	0.017119	0.0347
age	-0.0009853	0.0010706	0.357	0.003148	0.001137	0.0056
		Exchangeable:single			Exchangeable:married	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.154932	0.06156	0.01184	-0.338316	0.046017	2e-13
limit − balance	0.001737	0.000111	< 2e-16	0.002144	0.000111	< 2e-16
education − university	0.183832	0.024242	3.4e-14	0.180813	0.018493	< 2e-16
education − highschool	0.192902	0.027599	2.8e-12	0.144131	0.027113	0.00000011
female	-0.091678	0.018619	0.00000085	-0.043472	0.017143	0.011
age	-0.003695	0.00107	0.00055	-0.001082	0.001135	0.341
		AR1:single			AR1:married	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.238995	0.063099	0.00015	-0.427495	0.046911	< 2e-16
limit − balance	0.001731	0.000117	< 2e-16	0.002134	0.000114	< 2e-16
education − university	0.191857	0.024536	5.3e-15	0.189278	0.018887	< 2e-16
education − highschool	0.19482	0.027851	2.6e-12	0.139806	0.027704	0.00000045
female	-0.089305	0.018962	0.0000025	-0.03991	0.017549	0.023
age	-0.001814	0.001077	0.09191	0.001534	0.001142	0.179
		Unstructured:single			Unstructured:married	
predictor	estimate	se	p-value	estimate	se	p-value
Intercept	-0.360936	0.053862	2.1e-11	-0.411	0.0375	< 2e-16
limit − balance	0.001772	0.000092	< 2e-16	0.00203	0.0000893	< 2e-16
education − university	0.202286	0.020219	< 2e-16	0.179	0.0152	
education − highschool	0.19701	0.02297	< 2e-16	0.105	0.0214	0.00000089
female	-0.06975	0.015573	0.0000075	-0.0483	0.0139	0.00051
age	-0.000583	0.000931	0.53	0.0012	0.000917	0.19032

Table 5.30 Parameter estimations for parametric GEE model (Para2) in Marriage Analysis

	Independence:single			Independence:married		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.388	0.0381	< 2e-16	-0.454	0.0242	< 2e-16
$limit - balance$	0.00172	0.00011	< 2e-16	0.0021	0.00011	< 2e-16
$education - university$	0.188	0.0242	9.9e-15	0.177	0.0185	< 2e-16
$education - highschool$	0.182	0.0277	4.6e-11	0.118	0.0271	0.000013
$female$	-0.0819	0.0186	0.00001	-0.0368	0.0171	0.031
age^2	-0.00000591	0.0000135	0.66	0.0000435	0.000016	0.0065
	Exchangeable:single			Exchangeable:married		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.335	0.038	< 2e-16	-0.408	0.024	<2e-16
$limit - balance$	0.00173	0.00011	< 2e-16	0.00214	0.00011	< 2e-16
$education - university$	0.185	0.0242	2.3e-14	0.18	0.0185	< 2e-16
$education - highschool$	0.191	0.0277	4.8e-12	0.141	0.0272	0.00000023
$female$	-0.0894	0.0186	0.0000015	-0.0423	0.0171	0.013
age^2	-0.0000357	0.0000135	0.0081	-0.00000665	0.0000159	0.675
	AR1:single			AR1:married		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.376	0.0392	< 2e-16	-0.444	0.0247	< 2e-16
$limit - balance$	0.00173	0.000117	< 2e-16	0.00214	0.000114	< 2e-16
$education - university$	0.193	0.0245	4.1e-15	0.189	0.0189	< 2e-16
$education - highschool$	0.193	0.0279	4.5e-12	0.138	0.0278	0.00000068
$female$	-0.0875	0.0189	0.0000038	-0.04	0.0175	0.022
age^2	-0.000015	0.0000136	0.27	0.0000234	0.000016	0.142
	Unstructured:single			Unstructured:married		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.346	0.0387	< 2e-16	-0.447	0.0231	< 2e-16
$limit - balance$	0.00161	0.000108	< 2e-16	0.00196	0.000105	< 2e-16
$education - university$	0.17	0.0248	6.5e-12	0.159	0.0177	< 2e-16
$education - highschool$	0.171	0.028	9.5e-10	0.114	0.0261	0.000011
$female$	-0.0765	0.0189	0.000054	-0.0216	0.0164	0.19
age^2	-0.0000372	0.0000145	0.01	0.0000153	0.0000151	0.31

Table 5.31 Parameter estimations for parametric GEE model (Para3) in Marriage Analysis

	Independence:single			Independence:married		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.4	0.0289	< 2e-16	-0.414	0.0202	< 2e-16
$limit - balance$	0.00171	0.000111	< 2e-16	0.00213	0.000109	< 2e-16
$education - university$	0.189	0.0242	5.2e-15	0.18	0.0184	< 2e-16
$education - highschool$	0.181	0.0272	2.8e-11	0.138	0.026	0.00000012
$female$	-0.0793	0.0184	0.000016	-0.0416	0.0171	0.015
$exp(age)$	1.7e-34	1.33e-36	< 2e-16	3.13e-32	1.18e-31	0.79
	Exchangeable:single			Exchangeable:married		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.4	0.0289	< 2e-16	-0.414	0.0202	<2e-16
$limit - balance$	0.00171	0.000111	< 2e-16	0.00213	0.000109	< 2e-16
$education - university$	0.189	0.0242	6e-15	0.18	0.0184	< 2e-16
$education - highschool$	0.181	0.0272	3.1e-11	0.138	0.026	0.00000011
$female$	-0.0797	0.0184	0.000015	-0.0415	0.017	0.015
$exp(age)$	1.13e-34	9.96e-37	< 2e-16	-7.34e-32	2.99e-32	0.014
	AR1:single			AR1:married		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.404	0.03	< 2e-16	-0.422	0.0206	< 2e-16
$limit - balance$	0.00172	0.000117	< 2e-16	0.00215	0.000113	< 2e-16
$education - university$	0.195	0.0245	1.8e-15	0.19	0.0188	
$education - highschool$	0.189	0.0275	5.6e-12	0.149	0.0268	0.000000029
$female$	-0.0828	0.0187	0.0000095	-0.0425	0.0175	0.015
$exp(age)$	1.41e-34	1.14e-36	< 2e-16	-1.6e-32	7.81e-32	0.838
	Unstructured:single			Unstructured:married		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.412	0.0287	< 2e-16	-0.432	0.0193	< 2e-16
$limit - balance$	0.00157	0.000109	< 2e-16	0.00197	0.000105	< 2e-16
$education - university$	0.174	0.0252	5.1e-12	0.16	0.0176	< 2e-16
$education - highschool$	0.159	0.028	0.000000013	0.121	0.025	0.0000013
$female$	-0.065	0.0193	0.00075	-0.0233	0.0164	0.16
$exp(age)$	1.21e-34	2.27e-36	< 2e-16	-3.04e-32	6.17e-32	0.62

Table 5.32 Parameter estimations for parametric GEE model (Semi) in Marriage Analysis

	Independence:single			Independence:married		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.438	0.0252	< 2e-16	-0.377	0.0172	< 2e-16
$limit - balance$	0.00176	0.0000927	< 2e-16	0.00183	0.0000921	< 2e-16
$education - university$	0.227	0.0207	< 2e-16	0.21	0.0157	< 2e-16
$education - highschool$	0.212	0.0234	< 2e-16	0.11	0.0212	0.00000023
$female$	-0.0688	0.0158	0.000013	-0.0296	0.0143	0.039
	Exchangeable:single			Exchangeable:married		
$Intercept$	-0.438	0.0252	< 2e-16	-0.377	0.0172	< 2e-16
$limit - balance$	0.00176	0.0000927	< 2e-16	0.00183	0.0000921	< 2e-16
$education - university$	0.227	0.0207	< 2e-16	0.21	0.0157	< 2e-16
$education - highschool$	0.212	0.0234	< 2e-16	0.11	0.0212	0.00000023
$female$	-0.0688	0.0158	0.000013	-0.0296	0.0143	0.039
	AR1:single			AR1:married		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.438	0.0257	< 2e-16	-0.384	0.0173	< 2e-16
$limit - balance$	0.00176	0.0000961	< 2e-16	0.00184	0.0000931	< 2e-16
$education - university$	0.229	0.0209	< 2e-16	0.22	0.0159	< 2e-16
$education - highschool$	0.216	0.0235	< 2e-16	0.115	0.0216	0.000000094
$female$	-0.0741	0.0159	0.0000033	-0.0306	0.0145	0.035
	Unstructured:single			Unstructured:married		
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	-0.454	0.0246	< 2e-16	-0.398	0.0166	< 2e-16
$limit - balance$	0.00172	0.0000919	< 2e-16	0.0017	0.0000884	< 2e-16
$education - university$	0.206	0.0201	< 2e-16	0.189	0.0152	< 2e-16
$education - highschool$	0.194	0.0226	< 2e-16	0.0941	0.0206	0.0000049
$female$	-0.0631	0.0153	0.000038	-0.0157	0.0138	0.26

Table 5.33 Overall MSE for parametric models and semi-parametric model:single

ar1	para1	para2	para3	semi
training	0.913	0.913	0.910	0.943
testing	1.030	1.031	1.030	1.029
unstructured	para1	para2	para3	semi
training	0.946	0.913	0.909	0.943
testing	1.031	1.035	1.034	1.030

Table 5.34 Overall MSE for parametric models and semi-parametric model:married

ar1	para1	para2	para3	semi
training	0.868	0.867	0.868	0.929
testing	1.046	1.045	1.046	1.049
unstructured	para1	para2	para3	semi
training	0.927	0.868	0.869	0.928
testing	1.049	1.052	1.053	1.055

5.5.2 Using Payment Status as Response Variable

In this part, the payment status we defined on section 5.1 will be used as the response variable to evaluate the difference between customers with different marriage status for whether the customers will default to pay the bills. Predictors such as gender, limit balance, education levels, and age will be used in our models. Especially, we would like to investigate whether single customers and married customers should be fitted with different models. We first consider three parametric GEE model with linear form of age for single customers and married customers separately as the following:

$$Para4\ for\ single:\ logit(p) = \beta_0 + \beta_1 limit\ balance \times I_S + sex \times I_S + education \times I_S + \beta_2 age \times I_S \tag{5.39}$$

$$Para4\ for\ married:\ logit(p) = \beta_0 + \beta_1 limit\ balance \times I_{MA} + sex \times I_{MA} + education \times I_{MA} + \beta_2 age \times I_{MA} \tag{5.40}$$

where p is the probability of default.

Table 5.35 shows the estimation results for the parametric model (Para4) using different working correlation matrices. All predictors are significant for either single or married customers with any four types of working correlation matrices. The predictor limit balance has negative coefficient for single and married customers, which denotes that limit balance has negative correlation with the probability of default and age is significant with positive correlation with the probability of default.

Table 5.35 Parameter estimations for parametric GEE model (Para4) in Marriage Analysis

		Independence:male			Independence:female	
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	1.917085	0.134934	< 2e-16	1.654273	0.110309	<2e-16
$limit - balance$	-0.003699	0.000162	< 2e-16	-0.003183	0.000156	<2e-16
$female$	-0.285626	0.041753	7.90e-12	-0.237037	0.038922	1.10e-09
$education - university$	0.648884	0.046156	< 2e-16	0.646036	0.040649	<2e-16
$education - highschool$	0.660288	0.0594	< 2e-16	0.485903	0.064968	7.50e-14
age	-0.0144	0.00233	6.40e-10	-0.007861	0.002551	0.0021
		Exchangeable:male			Exchangeable:female	
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	2.073819	0.135533	< 2e-16	1.76582	0.109761	<2e-16
$limit - balance$	-0.003682	0.000162	< 2e-16	-0.003137	0.000156	<2e-16
$female$	-0.296253	0.041853	1.50e-12	-0.24276	0.038943	4.60e-10
$education - university$	0.645383	0.046221	< 2e-16	0.648521	0.04068	<2e-16
$education - highschool$	0.673108	0.059601	< 2e-16	0.507377	0.065249	7.40e-15
age	-0.017937	0.00234	1.80e-14	-0.011435	0.002529	0.0000061
		ar1:male			ar1:female	
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	1.895092	0.132428	< 2e-16	1.647558	0.109387	< 2e-16
$limit - balance$	-0.003504	0.000158	< 2e-16	-0.003095	0.000155	< 2e-16
$female$	-0.265154	0.041204	1.20e-10	-0.211492	0.038595	4.30e-08
$education - university$	0.621909	0.045643	< 2e-16	0.615588	0.040331	< 2e-16
$education - highschool$	0.629508	0.05887	< 2e-16	0.477611	0.064646	1.50e-13
age	-0.01436	0.002296	4.00e-10	-0.008341	0.002535	0.001
		unstructured:male			unstructured:female	
predictor	estimate	se	p-value	estimate	se	p-value
$Intercept$	1.977575	0.132615	< 2e-16	1.729513	0.109007	<2e-16
$limit - balance$	-0.003625	0.000159	< 2e-16	-0.00315	0.000154	<2e-16
$female$	-0.278457	0.041229	1.40e-11	-0.230634	0.038563	2.20e-09
$education - university$	0.632864	0.045486	< 2e-16	0.632813	0.040302	<2e-16
$education - highschool$	0.639161	0.058773	< 2e-16	0.494598	0.064616	1.90e-14
age	-0.015085	0.002292	4.70e-11	-0.009272	0.002521	0.00024

Table 5.36 and 5.37 shows the predictive accuracy tables in parametric model when the response variable is delay status for single customers and married customers. For single customers, the testing accuracy is the same as training accuracy. The model with single customers has lower accuracy in training and testing dataset comparing to married customers.

Table 5.36 Predictive Accuracy for Parametric GEE model (Para4):single

training	ar1			accuracy	testing	ar1			accuracy
		pred					pred		
	true	0	1			*true*	0	1	
	0	5050	14852			0	2196	6418	
	1	3822	33576	0.674		1	1565	13071	0.657
	unstructured					unstructured			
		pred					pred		
	true	0	1			*true*	0	1	
	0	5054	14848			0	2215	6399	
	1	3834	33564	0.674		1	1578	13058	0.657

Table 5.37 Predictive Accuracy for Parametric GEE model (Para4):married

training	ar1			accuracy	testing	ar1			accuracy
		pred					pred		
	true	0	1			*true*	0	1	
	0	2156	16948			0	1124	8737	
	1	2267	41581	0.695		1	1209	20406	0.684
	unstructured					unstructured			
		pred					pred		
	true	0	1			*true*	0	1	
	0	2152	16952			0	1120	8741	
	1	2259	41589	0.695		1	1198	20417	0.684

CHAPTER 6

FUTURE STUDY

One challenge comes from the application part is the dataset resource. Most financial dataset used in longitudinal study reaches individuals level, which violates the privacy policies in most institutes in the United States. Our dataset comes from the UCI website is based on the credit information in Taiwan. In future, we would like to use our model on other available credit loan dataset or other types of financial datasets in the United States.

When using semi-parametric models, another challenge rises from the application part. Based on the evaluation metrics, such as BIC, Mean Square Error and predictive accuracy, we observed that the advantage of semi-parametric model with kernel smoother is not huge. We would like to use semi-parametric models in other financial dataset with longitudinal perspective, investigating whether our semi-parametric models with kernel smoother will be much better than any other types of parametric models for other financial data.

The third challenge comes from the scheme of semi-parametric approach. When assigning the non-parametric term in semi-parametric approach, most applications in biological dataset use previous experience. In our approach, we used age as non-parametric term because it is not significant under several parametric GEE approaches. We would like to try other continuous variables and create a robust approach for identifying which variable should be used as non-parametric term.

The last challenge is a traditional issue for GEE approach: estimating working correlation matrix. In semi-parametric GEE study, estimating working correlation matrix is critical but more difficult comparing to parametric GEE approach. Fan, Huang and Li (2007) proposed a scheme of estimation procedure, using profile weighted least squares approach to estimate working correlation matrix. We would like to try this approach in the future, investigating whether this estimation method will provide more efficient semi-parametric estimators with fully specified working correlation matrix when we applied it in financial datasets.

REFERENCES

[1] Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B.N.; Csáki, F., *2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, pp. 267–281.

[2] Akaike, H. (1974), "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, 19 (6): 716–723.

[3] Al Kadiri, M., Carroll, R. J., and Wand, M. P. (2010), "Marginal longitudinal semiparametric regression via penalized splines", *Statistics and probability letters*, 80(15), 1242-1252.

[4] Agresti A. (1997), "A model for repeated measurements of a multivariate binary response", Journal of the American Statistical Association, 92:315–321.

[5] Agresti,A. (2013), *Categorical Data Analysis*, WILEY-INTERSCIENCE.

[6] Balan, R. M. and Schiopu-Kratina, I. (2005), "Asymptotic results with generalized estimating equations for longitudinal data", *Annals of Statistics*, 32 522–541.

[7] Boneva, Lena, Oliver Linton, and Michael Vogt. (2015), "A semiparametric model for heterogeneous panel data with fixed effects", *Journal of Econometrics*, 188.2: 327-345.

[8] Chen, Jia, et al. (2015), "Semiparametric GEE analysis in partially linear single-index models for longitudinal data", *The Annals of Statistics*, 43.4: 1682-1715.

[9] Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford, U.K.: Oxford University Press.

[10] Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

[11] Fan, J. and Li, R. (2004), "New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis", *Journal of the American Statistical Association*, 99, 710–723.

[12] Fan, J. and Huang, T. (2005), "Profile likelihood inferences on semiparametric varying-coefficient partially linear models", *Bernoulli*, 11, 1031–1057.

[13] Fan, Jianqing, Tao Huang, and Runze Li.(2007), "Analysis of longitudinal data with semiparametric estimation of covariance function", *Journal of the American Statistical Association*, 102.478: 632-641.

[14] Fan J. and Wu Y. (2008), "Semiparametric estimation of covariance matrixes for longitudinal data", *Journal of the American Statistical Association*, 103, 1520–1533.

[15] Fieuws, S. and Verbeke, G. (2006), "Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles", *Biometrics*, 62(2), pp.424-431.

[16] Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008), *Longitudinal Data Analysis*, London: Chapman and Hall/CRC.

[17] Griffith, R., Redding, S. and Van Reenen, J. (1999), "Mapping the two faces of RD: productivity growth in a panel of OECD manufacturing industries", *Institute for Fiscal Studies*, mimeo.

[18] Hall, P. and J.S. Racine and Q. Li (2004), "Cross-validation and the estimation of conditional probability densities", *Journal of the American Statistical Association*, 99, 1015-1026.

[19] Hansen, L. (1982), "Large sample properties of generalized method of moments estimators", *Econometrica*, 50, 1029–1054.

[20] Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

[21] Hastie, T., Tibshirani, R., and Friedman, J. (2008), *The Elements of Statistical Learning*, New York: Springer.

[22] Henderson, Daniel J., Raymond J. Carroll, and Qi Li. (2008), "Nonparametric estimation and testing of fixed effects panel data models", *Journal of Econometrics*, 144.1: 257-275.

[23] Hu, Zonghui, Naisyin Wang, and Raymond J. Carroll. (2004), "Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data", *Biometrika*, 91.2: 251-262.

[24] Jones, R. H. (2011), "Bayesian information criterion for longitudinal and clustered data", *Statistics in medicine*, 30(25), 3050-3056.

[25] Li, Q. and J. Lin and J.S. Racine (2013), "Optimal bandwidth selection for nonparametric conditional distribution and quantile functions", *Journal of Business and Economic Statistics*, 31, 57-65.

[26] Li, Q. and J.S. Racine (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

[27] Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models", *Biometrika*, 73, 13–22.

[28] Liang K and Zeger S. (1989), "A class of logistic regression models for multivariate binary time series", *Journal of the American Statistical Association*, 84: 447–451.

[29] Lin, X., and Carroll, R. J. (2000), "Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error", *Journal of the American Statistical Association*, 95, 520–534.

[30] Lin, X. and Raymond J. Carroll.(2001), "Semiparametric regression for clustered data", *Biometrika*, 88.4: 1179-1185.

[31] Lin, X. and Carroll, R. J. (2006), "Semiparametric estimation in general repeated measures problems", *J. R. Statist. Soc. B*, 68, 69–88.

[32] Liu, W., and Yang, Y. (2011), "Parametric or nonparametric? A parametricness index for model selection", *The Annals of Statistics*, 2074-2102.

[33] McCullagh, P. (1983), "Quasi-likelihood functions", *Ann. Statist*, 11, 59-67.

[34] O'Brien L and Fitzmaurice G. (2004), "Analysis of longitudinal multiple-source binary data using generalized estimating equations", *Applied Statistics*, 53: 177–193.

[35] Petersen, Mitchell A. (2009), "Estimating standard errors in finance panel data sets: Comparing approaches", *The Review of Financial Studies*, 22.1 : 435-480.

[36] Qu, A., Lindsay, B.G. and Li, B. (2000), "Improving generalized estimating equations using quadratic inference functions", *Biometrika*, 87, 823–836.

[37] Rao Chaganty, N. and Joe, H. (2004), "Efficiency of generalized estimating equations for binary responses", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4), 851-860.

[38] Richards, David L., and Ronald D. Gelleny. (2006), "Banking Crises, Collective Protest and Rebellion", *Canadian Journal of Political Science / Revue Canadienne De Science Politique*, vol. 39, no. 4, pp. 777–801.

[39] Roy J and Lin X. (2000), "Latent variable models for longitudinal data with multiple continuous outcomes", *Biometrics*.

[40] Ruppert, D., Sheather, S. J. and Wand, M. P. (1995), "An effective bandwidth selector for local least squares regression", *Journal of the American Statistical Association*, 90, 1257–1270.

[41] Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-likelihood Estimation in Semiparametric Models", *Journal of the American Statistical Association*, 89, 501–511.

[42] Schwarz, Gideon E. (1978), "Estimating the dimension of a model", *Annals of Statistics*, 6 (2): 461–464.

[43] Ullah, Aman, and David EA Giles, eds. (2016), *Handbook of empirical economics and finance*, CRC Press.

[44] Verbeke G, Fieuws S, Molenberghs G, Davidian M. (2014), "The analysis of multivariate longitudinal data: a review", *Stat Methods Med Res*, 23(1):42–59.

[45] Wang, N. (2003), "Marginal nonparametric kernel regression accounting for within-subject correlation", *Biometrika*, 90, 43–52.

[46] Wang, N., Carroll, R.J., and Lin, X. (2004), "Efficient semiparametric marginal estimation for longitudinal/clustered data", *Journal of the Americal Statistical Association*, 100, 147–157.

[47] Wedderburn, R. W. M. (1974), "Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method", *Biometrika*, 61, 439-447.

[48] Yao, W. and Li, R. (2013), "New local estimation procedure for a non-parametric regression function for longitudinal data", *J. R. Stat. Soc. Ser. B. Stat. Methodol*, 75 123–138.

[49] Yeh, I. C., and Lien, C. H. (2009), "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients", *Expert Systems with Applications*, 36(2), 2473-2480.

[50] Yeh, I. C. and Lien, C. H. (2009), *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml], Irvine, CA: University of California, School of Information and Computer Science.

[51] Yeh, I. C., and Lien, C. H. (2009), "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients", *Expert Systems with Applications*, 36(2), 2473-2480.

[52] Zeger, S. L. and Liang, K. Y. (1986), "Longitudinal Data Analysis for Discrete and Continuous Outcomes", *Biometrika*, 43, 121-130.

[53] Zhang, Y. and Heagerty, P. J. (2005), "Partly Conditional Survival Models for Longitudinal Data", *Biometrics*, 61, 379-391.

BIOGRAPHICAL SKETCH

Liu Yang was born in 1986, China. She graduated in 2009 from Fudan University, Shanghai, China and got her bachelor degree in international finance. In 2011, she got her master degree in applied economics in Western Michigan University, Kalamazoo, Michigan. In 2013, she got her master degree of statistics in George Washington University, Washington DC. In 2013, she started her Ph.D. study in Florida State University in department of statistics.