Journal of Finance and Economics, 2021, Vol. 9, No. 3, 93-105
Available online at http://pubs.sciepub.com/jfe/9/3/1
Published by Science and Education Publishing
DOI:10.12691/jfe-9-3-1

Semi-Parametric Models for Longitudinal Data Analysis

Liu Yang[1,*], Xu-Feng Niu[2]

[1]Liberty Mutual Insurance, Boston, Massachusetts, USA
[2]Department of Statistics, Florida State University, Florida, USA
*Corresponding author: yl200906@gmail.com

Received April 23, 2021; Revised May 27, 2021; Accepted June 07, 2021

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 the repeated measurements from each subject, which cause time dependent correlations within subjects. Generalized Estimating Equations (GEE) can deal with correlated outcomes for longitudinal data through marginal effect. Our proposed model will be based on GEE, with a semi-parametric approach, to provide a flexible structure for regression models: coefficients for parametric covariates will be estimated and nuisance covariates will be fitted in kernel smoothers for the non-parametric part. The profile kernel estimator and the seemingly unrelated kernel estimator (SUR) will be used to obtain consistent and efficient semi-parametric estimators. We provide simulation results for estimating semi-parametric models with one or multiple non-parametric terms. Financial market data is a major component of data analysis; thus, we focus on the financial market in the application part. Credit card loan data will be used with the payment information for each customer across six months to investigate whether gender, income, age, or other factors will influence payment status significantly. Furthermore, we propose model comparisons to evaluate whether different models should be fitted for different subgroups of consumers, such as male and female.

Keywords: *longitudinal study, generalized estimating equations (GEE), semi-parametric model, profile-kernel estimator, the seemingly unrelated kernel estimator (SUR)*

Cite This Article: Liu Yang, and Xu-Feng Niu, "Semi-Parametric Models for Longitudinal Data Analysis." *Journal of Finance and Economics*, vol. 9, no. 3 (2021): 93-105. doi: 10.12691/jfe-9-3-1.

1. Introduction

For statistical scientific studies, experiment designs depend on the different types of system under study and the different goals for research. Longitudinal studies allow for the investigation of change over different time points and the effects of different factors on the change. One distinctive feature of longitudinal studies is repeated measurements at different time points within each subject (or cluster), which considers the time series correlation. For example, the financial market plays an important role in daily life. Financial institutions, such as commercial banks, investment banks, insurance companies, and brokerages are major players trading in financial markets. Most financial data analysis involves time series because time is valuable, and we want to track the temporal tendency of subjects. Therefore, once we obtain time changing measurements for each subject, as well as covariates, we can conduct longitudinal studies for financial data analysis.

A variety of longitudinal models have been applied in financial analysis. Petersen [1] pointed out that previous research focuses mainly on three major methods: the Fama-MacBeth procedure (Fama and MacBeth) [2] estimates, dummy variables in each cluster such as the fixed effect model and adjustments within cluster correlation such as Generalized Estimating Equations (GEE). Different methods should be applied depending on different interests. For a subject specified effect, the Generalized Linear Mixed Model (GLMM) will provide a nice estimator for individual subjects. When covariates are involved in general factor or policy, GEE can be applied to investigate the relationship between the response and covariates.

In order to capture the complex relationship in longitudinal data analysis, semi-parametric and non-parametric models have been developed for financial data analysis in longitudinal studies. Sam and Jiang [3] propose a non-parametric estimator for a short rate diffusion process with yields in longitudinal structure. In this paper we will introduce a class of semi-parametric regression models with GEE, which provide a flexible structure for longitudinal data analysis. Simulation studies will be conducted to compare the performance of our proposed models with other types of models. The semi-parametric regression models will be applied to credit card loan data and models for different subgroups will be examined.

Different estimation methods have been developed for non-parametric and semi-parametric regression models when observations of the response are independent. For non-parametric regression models, kernel estimation methods based on local likelihoods and splines based on penalized likelihoods can be used; for semi-parametric regression models, partial linear models, which specify the mean of the outcome variable as a parametric function with

respect to some covariates and non-parametric functions with respect to other covariates, can be used. 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 a different weight for neighborhood observations. Smoothing splines fit the non-parametric function through 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 it puts the penalty of smoothing splines on regression splines.

For longitudinal data analysis, non-parametric and semi-parametric regression should be able to deal with within-subject correlation for repeated measurements. Estimating equations 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 [4] proposed a kernel GEE estimator through local polynomial kernel estimating equations by the extension of a generalized linear model. Unlike the parametric GEE developed by Zeger and Liang [5], kernel GEE has limited conditions for a consistent estimator and cannot reach efficiency bound if accounting for within-subject association. Wang [6] provided the seemingly unrelated kernel (SUR) estimator which fulfills 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 the smoothing spline estimator has a close relationship with linear mixed models.

Whether semi-parametric regression can be applied in marginal models and linear mixed models will depend on the goal. If we focused on semi-parametric regression in marginal models, several estimation methods have been developed to deal with the within-subject correlations. Lin and Carroll [7] developed profile-kernel estimating equations which estimate the parametric part by a profile method and the non-parametric part by the kernel GEE with local polynomial kernels, which we mentioned above. The estimator from profile-kernel methods is consistent only when ignoring within-subject correlation and it is not semi-parametric efficient even without the within-subject correlation for non-parametric part. Wang, Carroll, and Lin [8] used the SUR kernel model for the non-parametric part and remained estimating the parametric part with the profile method, providing an estimator with consistency and semi-parametric efficiency. For semi-parametric linear mixed models, we can also use the profile SUR kernel methods to fit the model and the spline method as well.

The rest of this article will be organized as follows. In Section 2, we will display mathematical details for the semi-parametric model 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) [7] and the SUR kernel estimator (Wang, Carroll, and Lin) [8]. In Section 3, we will show a simulation study that follows the models in Section 2. 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 with separated training and testing datasets. Section 4 would be data application part. 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 we provided results for analysis when fitting model separately based on different level of factors. Section 5 shows conclusion and discussion, we will provide a summary and we will also discuss some challenges when we conduct semi-parametric models.

2. Models

In this section we propose semi-parametric models for GEE. We will provide local polynomial kernel GEE estimator and the seemingly unrelated kernel estimator, which are two main tools for model fitting. The difference of consistency and efficiency between those estimators will be displayed when accounting for association within subjects.

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

$$g\left(\mu_{ij}\right) = X_{ij}^T \beta + \sum_{d=1}^{q} \theta\left(Z_{ijd}\right)$$

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

2.1. Profile-Kernel Estimating Equations with Two Kernel Smoothers

We follow Lin and Carroll's [7] method, using a back-fitting algorithm to calculate the profile-kernel estimator, which had three steps in general: for a given β and other kernel smoother terms, we can estimate one of the non-parametric terms, using non-parametric estimating equations. After we estimate that non-parametric term, we can estimate the rest kernel smoother terms and after we have finished the estimator of all non-parametric terms, a traditional generalized estimating equation can be used to obtain β estimator.

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

$$g\left(\mu_{ij}\right)=X_{ij}^{T}\beta+\theta_1\left(Z_{ij1}\right)+\theta_2\left(Z_{ij2}\right)$$

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

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

$$\sum_{i=1}^{N}Z_{i2}(z)^{T}\,\Delta_i\left(X_i,z,Z_{i1}\right)K_{ih}^{\frac{1}{2}}(z)V_{1i}^{-1}\left(X_i,z,Z_{i1}\right)K_{ih}^{\frac{1}{2}}(z)$$
$$\times\left\{Y_i-\mu_i\left(X_i,z,Z_{i1}\right)\right\}=0$$

where $Z_{i2}(z)$ is an $n_i\times(r+1)$ matrix with the j^{th} row is $1,\left(Z_{ij2}-z\right),\ldots,(Z_{ij2}-z)^{r^{T}}$. $\boldsymbol{Y}_i$ and $\boldsymbol{\mu}_i$ are vectors: $\boldsymbol{Y}_{ij}^{T}=(X_{i1},\ldots,X_{in_i})$, $\boldsymbol{\mu}_{ij}^{T}=(\mu_{i1},\ldots,\mu_{in_i})$, $\mu_{ij}=E\left(\boldsymbol{Y}_{ij}\right)=\mu_{ij}(\beta)=g^{-1}(\boldsymbol{X}_{ij}^{T}\beta+\theta_1\left(Z_{ij1}\right)+\theta_2(z))$ and we use the identity link function. $K_{ih}(z)=diagK_h(Z_{ij2}-z)$ are kernel weighs of the target value for i^{th} subject. $\Delta_i(X_i,z,\boldsymbol{Z}_{i1})=diag\left\{\mu_{ij}^{(1)}\right\}$ and $\mu_{ij}^{(1)}(.)$ is the first derivative of $\mu(.)$.

$$V_{1i}\left(X_i,z,Z_{i1}\right)=S_i^{1/2}\left(X_i,z,Z_{i1}\right)R_{1i}S_i^{1/2}\left(X_i,z,Z_{i1}\right)$$

and $S_i(X_i,z,\boldsymbol{Z}_{i1})=diag\phi\omega_{ij}^{-1}V_i$, 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, the local average kernel GEE estimator has a closed form solution:

$$\hat{\theta}_{2K}(z)$$
$$=\frac{\sum_{i=1}^{N}1_i^{T}K_{ih}^{\frac{1}{2}}(z)V_{1i}^{-1}K_{ih}^{\frac{1}{2}}(z)\times\left\{Y_i-X_{ij}^{T}\beta-\theta_1\left(Z_{ij1}\right)\right\}}{\sum_{i=1}^{N}1_i^{T}K_{ih}^{\frac{1}{2}}(z)V_{1i}^{-1}K_{ih}^{\frac{1}{2}}(z)1_i^{T}}$$

Step 2: After we obtain $\hat{\theta}_2(Z_{ij1})$ and given β, we can proceed to calculate $\theta_1(z)$ by another estimating equation. Still, through the estimating equations, the local average kernel GEE estimator $\hat{\theta}_1(Z_{ij1})$ has a closed form solution.

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:

$$\sum_{i=1}^{N}\frac{\partial\mu X_i\beta+\hat{\theta}_1\left(Z_{1i};\beta\right)+\hat{\theta}_2\left(Z_{2i};\beta\right)}{\partial\beta}\times V_{3i}^{-1}\left(X_i,Z_{i1},Z_{i2}\right)$$
$$\times\left[Y_i-\mu\left\{X_i\beta+\hat{\theta}_1\left(Z_{1i};\beta\right)+\hat{\theta}_2\left(Z_{2i};\beta\right)\right\}\right]=0$$

where

$$\hat{\theta}_1\left(Z_{1i};\beta\right)=\hat{\theta}\left(Z_{i11};\beta\right),\ldots,\hat{\theta}\left(Z_{in_i1};\beta\right);$$
$$\hat{\theta}_2\left(Z_{2i};\beta\right)=\hat{\theta}\left(Z_{i12};\beta\right),\ldots,\hat{\theta}\left(Z_{in_i2};\beta\right),$$

$$V_{3i}\left(X_i,Z_{i1},Z_{i2}\right)$$
$$=S_i^{1/2}\left(X_i,Z_{i1},Z_{i2}\right)R_{2i}S_i^{1/2}\left(X_i,Z_{i1},Z_{i2}\right),$$
$$S_i\left(X_i,Z_{i1},Z_{i2}\right)$$
$$=diag\phi\omega_{ij}^{-1}V[\mu X_{ij}^{'}\beta+\hat{\theta}\left(Z_{i1};\beta\right)+\hat{\theta}\left(Z_{i2};\beta\right)]$$

and R_{3i} is a working correlation matrix.

We followed Fan and Li [9], providing that the estimating equation has a closed form solution for β:

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

where X is covariates matrix and Y is response variable. A_{1k} is the coefficient for the non-parametric regression estimator $\theta_1\left(Z_{ij1}\right)$, A_{2k} is the coefficient for the non-parametric regression estimator $\theta_2\left(Z_{ij2}\right)$ and $\tilde{V}=diag(V)$. If we write $\hat{\beta}_K=H_KY$, then $cov\left(\hat{\beta}_K\right)=H_K\tilde{\Sigma}H_K^{'}$, $\tilde{\Sigma}=diag(\tilde{\Sigma})$ 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.

2.2. Profile SUR Kernel Estimator with Two Kernel Smoothers

Following Wang, Carroll, and Lin's [8] method, we propose the SUR kernel estimator for semi-parametric model with two kernel smoother terms. Still, a back-fitting three-step iteration can be used for the estimation.

Step 1: Let $\tilde{\theta}_2(.)$ be the current estimator of $\theta_2(.)$. Given β and $\theta_1\left(Z_{ij1}\right)$, let

$$\hat{\alpha}=\hat{\alpha}\left(z,\beta,Z_{ij2}\right)$$
$$=\hat{\alpha}_0\left(z,\beta,Z_{ij2}\right),\hat{\alpha}_2\left(z,\beta,Z_{ij2}\right),\ldots,\hat{\alpha}_r\left(z,\beta,Z_{ij2}\right)^{T}$$

be the solution to the kernel equation

$$\sum_{i=1}^{N}\sum_{j=1}^{n_i}K_h\left(z-Z_{ij2}\right)\mu_{ij}^{(1)}\left(\beta,\hat{\alpha}\right)G_{ij}^{'}(z)V_i^{-1}$$
$$\times\left[Y_i-\mu^{*}z,X_i,Z_{i1},\beta,\hat{\alpha},\hat{\theta}_2\left(Z_{i2},\beta\right)\right]=0$$

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

$$\mu\left[\begin{array}{l}X_{ik}^{T}\beta+I(k=j)\left\{\hat{\alpha}_0+\dfrac{\hat{\alpha}_1\left(z-Z_{ij2}\right)+\ldots+\hat{\alpha}_d\left(z-Z_{ij2}\right)}{h}\right\}\\+I(k\neq j)\hat{\theta}\left(Z_{i2k};\beta\right)\end{array}\right]$$

and $\mu^{(1)}$ is the first derivative of the function $\mu(.)=g^{-1}(.)$ evaluated at

$$\left(X_{ij}^{'}\beta+\hat{\alpha}_0+\frac{\hat{\alpha}_1\left(z-Z_{ij2}\right)+\ldots+\hat{\alpha}_d\left(z-Z_{ij2}\right)}{h}\right).$$

The updated estimator of $\theta_2(z)$ is $\hat{\theta}_2\left(z,\beta,Z_{ij1}\right)=\widehat{\alpha_0}\left(z,\beta,Z_{ij1}\right)$ and $G_{ij}(z)$ is an $n_i\times(r+1)$ matrix of

zeros except the j^{th} column is $e_j \times \{(z - Z_{ij2})^r\}^T$, 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:

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

Where

$$K_{wh}(z) = \sum_{i=1}^{N}\sum_{i=1}^{n_i} K_h(Z_{ij2} - z)v_i^{jj-1}K_h(Z_{112} - z), \ldots, K_h(Z_{NnN_2} - z)^T$$

Here $K_{wh}(z)$ is an $N \times 1$ vector and N^* is the total number of observations, is

$$K_w = K_{wh}(Z_{112}), \ldots, K_{wh}(Z_{n_i n_i 2})^{'}$$

an $N^* \times N^*$ matrix, $\tilde{V} = diag(V_1, \ldots, V_N)$, $V_i^d = diag(v_i^{jj}) = diag(V^{-1})$, $\tilde{V}^d = diag(V_1^d, \ldots, V_1^n)$ and $Y = (Y_1^{'}, \ldots, Y_N^{'})^{'}$.

Step 2: Given β and $\hat{\theta}_2(Z_{ij2})$ we obtained from last step, the estimator of $\theta_1(z)$ can be calculated by another kernel equation with a closed form solution.

Step 3: After we obtain the estimators for two kernel smoothers, we can calculate β by solving the adjusted estimating equation. And still, we can update β by:

$$\theta_{1K}^{*} = A_{1K}^{*}\left(Y - X\beta^{*} - \theta_{2K}^{*}\right)$$

$$\theta_{2K}^{*} = A_{2K}^{*}\left(Y - X\beta^{*} - \theta_{1K}^{*}\right)$$

$$\beta_K^{*} = \left\{X'\left(I - A_{1K}^{*} - A_{2K}^{*}\right)\tilde{V}^{-1}\left(I - A_{1K}^{*} - A_{2K}^{*}\right)X\right\}^{-1} X'\left(I - A_{1K}^{*} - A_{2K}^{*}\right)\tilde{V}^{-1}\left(I - A_{1K}^{*} - A_{2K}^{*}\right)Y$$

Where A_{1K}^{*} is the coefficient for the non-parametric regression estimator $\theta_1(Z_{ij1})$, A_{2K}^{*} is the coefficient for the non-parametric regression estimator $\theta_2(Z_{ij2})$.

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

3. Simulation Results

In this section, simulations are conducted to compare 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 the local polynomial kernel GEE estimator and the SUR estimator will be used to display when and which unbiased estimator will achieve the least standard deviation under given conditions. For estimating the non-parametric part, the Gaussian density kernel will be used to construct kernel weights in the non-parametric smoother and the least square cross validation method (Silverman) will be used to select bandwidth parameter h which is critical for kernel regression models. 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. A training dataset and test dataset will be used to evaluate the performance of semi-parametric models and parametric models.

3.1. Semi-Parametric Model with One Kernel Smoother

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

$$Y_{ij} = X_{ij}^{'}\beta + \theta(Z_{ij}) + \epsilon_{ij}, i = 1, \ldots, n \text{ and } j = 1, \ldots, m. \quad (3.1)$$

where i denotes the i^{th} subject and j denotes the j^{th} time point. In the equation, $\theta(.)$ is a kernel smooth function, Z_{ij} denotes covariates in the non-parametric part, X_{ij} denotes covariates in the parametric part, and β is the coefficient vector. 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 $(Z_{ij}) = \exp(2/Z_{ij})$ in the second setup.
- X_{ij} and Z_{ij} are both scalars and time-varying covariates with $X_{ij} = b_{ij} + e_{1ij}$, $Z_{ij} = b_{ij} + e_{2ij}$. $b_{ij} \sim U[0,1]$, where e_{1ij} and e_{1ij} are independent to each other and follow uniform distribution $U[0,1]$.
- $\epsilon_{ij} = (\epsilon_{i1,\ldots,\epsilon_{in}})^{'}$ 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 (3.1), a 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, will be used in this simulation:

$$\boldsymbol{para1}: Y_{ij} = X_{ij}\beta + Z_{ij} + \epsilon_{ij},$$

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

$$\boldsymbol{para3}: Y_{ij} = X_{ij}\beta + Z_{ij} + Z^2{}_{ij} + Z^3{}_{ij} + \epsilon_{ij},$$

3.1.1. 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, as we discussed in Section 3.1.

Table 1 shows β estimates in the semi-parametric model (3.1) and parametric models in (para1–para3). The results for two setups show 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 ρ = 0.7 are lower than those for ρ = 0.3 in the parametric estimators and semi-parametric estimator for the first setup, while in the second setup, standard errors are higher when ρ = 0.7. For overall fitting mean square errors in the semi-parametric model (3.1) and parametric models (para1-para3). The mean square errors in the parametric estimators are larger than the mean square errors in semi-parametric estimators for both training dataset and test dataset. The gain is larger when we applied the second setup.

Table 1. β estimates and overall MSE with 4 time points, kernel estimator

setup1	**ρ =0.3**				**ρ =0.7**			
	mean	se	mse-train	mse-test	mean	se	mse-train	mse-test
semi-WI	0.998	0.019	1.17	1.235	0.994	0.014	1.173	1.229
semi-True	0.993	0.021	1.184	1.245	0.997	0.016	1.2	1.25
para1	1.005	0.068	1.488	1.504	1.003	0.065	1.499	1.513
para2	1.009	0.068	1.49	1.503	1.005	0.065	1.502	1.512
para3	0.992	0.069	1.457	1.507	0.997	0.066	1.467	1.519
setup2	**ρ =0.3**				**ρ =0.7**			
	mean	se	mse-train	mse-test	mean	se	mse-train	mse-test
semi-WI	1.001	0.013	1.31	1.391	1	0.014	1.304	1.389
semi-True	1.002	0.013	1.314	1.394	1.001	0.015	1.309	1.393
para1	1.05	0.087	3.635	4.175	1.045	0.096	3.609	4.205
para2	1.053	0.086	2.984	3.872	1.041	0.096	2.955	3.862
para3	1.07	0.088	2.227	2.835	1.068	0.088	2.203	2.808

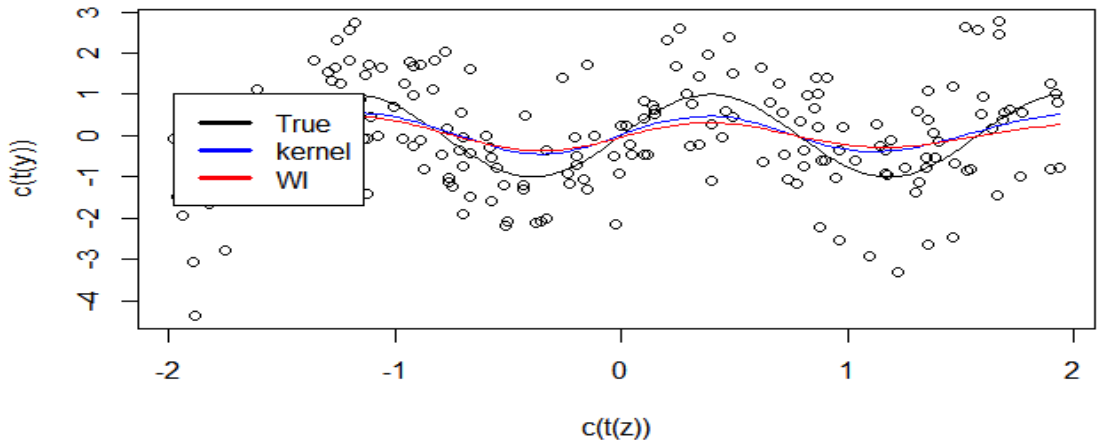

Figure 1. Fitted curves for the non-parametric part (kernel smoother)

Figure 1 above shows the non-parametric part fitting when $\theta(Z_{ij}) = \sin(4 \times Z_{ij})$ by the profile kernel estimator. The black line shows the true value, the blue line shows the fitting result using the independence working correlation matrix, while the red line shows fitting result using the true working correlation matrix. The three lines almost overlapped, which indicates that the results, using different working correlation matrices, deliver similar results in the non-parametric fitting part.

Table 2. β estimates and overall MSE with 10 time points, kernel estimator

setup1	**ρ =0.3**				**ρ =0.7**			
	mean	se	mse-train	mse-test	mean	se	mse-train	mse-test
semi-WI	0.998	0.006	1.105	1.133	0.997	0.005	1.08	1.116
semi-True	0.997	0.007	1.105	1.147	0.999	0.006	1.08	1.138
para1	1.004	0.041	1.488	1.506	1.001	0.046	1.467	1.488
para2	1.006	0.041	1.49	1.506	1.001	0.046	1.465	1.488
para3	0.994	0.041	1.469	1.498	0.998	0.045	1.443	1.489
setup2	**ρ =0.3**				**ρ =0.7**			
	mean	se	mse-train	mse-test	mean	se	mse-train	mse-test
semi-WI	1	0.005	1.291	1.34	1	0.004	1.278	1.352
semi-True	1	0.006	1.291	1.342	1	0.005	1.278	1.357
para1	1.004	0.051	3.652	4.269	1.003	0.05	3.663	4.339
para2	1.004	0.051	2.99	4.202	1.05	0.05	2.999	4.277
para3	1.008	0.05	2.233	3.299	1.004	0.05	2.228	3.377

Table 2 shows β estimates with Gaussian density kernel and ten time periods. The results are similar to results with four time periods, however the standard deviation for semi-parametric estimators are less than the situation when only four time points are involved, which indicates that the semi-parametric estimator gains more efficiency than parametric estimators when we have longer time periods. Moreover, we found that the standard errors for ρ = 0.7 are slightly higher than those for ρ = 0.3 in the parametric estimators. For the overall fitting mean square errors with the Gaussian density kernel and ten time periods, the results show that the mean square train and test errors for the semi-parametric approach are much lower than parametric estimates in all cases. Among parametric cases, the polynomial model performs best, but still much worse than semi-parametric fitting. Furthermore, mean square errors for semi-parametric estimators are less than the situation when only four time points are involved, which indicates that the semi-parametric estimator gains more accuracy than parametric estimators when we have longer time periods. Finally, we found that the mean square errors in training and testing datasets for ρ = 0.7 are slightly lower than those for ρ = 0.3 in the semi-parametric estimators. When we extend the time periods to ten, the coefficient estimators of the second setup have less bias and standard deviation when compared to the first setup. The semi-parametric estimator still gains more when we have longer time periods.

The result in the two tables shows that semi-parametric estimators with a stronger correlation, longer time period, and a more complicated pattern in the non-parametric part will benefit more when compared to parametric estimators with the same scenarios. According to the conclusion in Lin and Carroll [7] and results from our simulation, WI estimators perform better than an estimator compiled with a true correlation relationship, which conflicts with the properties of the GEE estimator. Another approach proposed by Wang [6] will be displayed in the next part, which delivers the estimator with the highest efficiency when fitting with true within subject association.

3.1.2. The SUR Estimator with One Kernel Smoother

The SUR estimator (Wang) [6] will be displayed in this part for the semi-parametric regression, running a simulation that follows the setups in the first part. Still, different setups will be applied in the simulation results.

Table 3. β estimates and overall MSE with 10 time points, SUR estimator

setup1	**ρ =0.3**				**ρ =0.7**			
	mean	se	mse-train	mse-test	mean	se	mse-train	mse-test
semi-WI	0.991	0.023	1.2	1.188	0.993	0.018	1.166	1.182
semi-True	0.999	0.01	1.022	1.014	0.999	0.01	0.992	1.006
para1	1	0.073	1.485	1.498	1.009	0.067	1.469	1.514
para2	0.998	0.073	1.483	1.5	1.008	0.066	1.466	1.514
para3	0.987	0.073	1.461	1.497	1.002	0.065	1.442	1.512
setup2	**ρ =0.3**				**ρ =0.7**			
	mean	se	mse-train	mse-test	mean	se	mse-train	mse-test
semi-WI	1.001	0.016	1.321	1.327	1.002	0.019	1.33	1.317
semi-True	1	0.014	1.252	1.257	1.001	0.014	1.256	1.24
para1	1.049	0.083	3.63	4.21	1.034	0.074	3.643	40214
para2	1.052	0.084	2.989	3.897	1.034	0.074	2.981	4.019
para3	1.067	0.085	2.21	2.856	1.034	0.073	2.211	3.068

Table 3 shows β estimates in the semi-parametric model (3.1) and parametric models in (para1–para3). The results show that the standard errors of the β estimates based on semi-parametric estimators are at least three 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. Similarly, we found that the standard errors for ρ = 0.7 are not higher than those for ρ = 0.3 in the parametric estimators and semi-parametric estimators. Among the parametric models, the polynomial model (para) has the smallest mean square error for the test dataset. Similar to the kernel estimation in the last part, in the second setup, the overall fitting accuracy in the semi-parametric model (3.1) gains more than the parametric models (para1–para3).

3.2. Semi-Parametric Model with Multiple Kernel Smoothers

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

$$Y_{ij} = X_{ij}^{'}\beta + \theta_1\left(Z_{ij1}\right) + \theta_2\left(Z_{ij2}\right) + \epsilon_{ij}, \quad i = 1,\ldots,n \text{ and } j = 1,\ldots,m \tag{3.2}$$

Still, i denotes the i^{th} subject and j denotes the j^{th} time point. In the equation, $\theta_1(.)$ and $\theta_2(.)$ are kernel smooth functions, Z_{ij1} and Z_{ij2} denote the covariates in the non-parametric part.

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

- Each run with 100 subjects, each subject with four time points and 200 replicates.
- The first setup is $\theta_1\left(Z_{ij1}\right) = \sin(4 \times Z_{ij1})$ in the first non-parametric term and $\theta_2\left(Z_{ij2}\right) = \sin(4 \times Z_{ij2})$ in the second non-parametric term; the second setup is $\theta_1\left(Z_{ij1}\right) = \exp(2/Z_{ij1})$ in the first non-parametric term and $\theta_2\left(Z_{ij2}\right) = \exp(2/Z_{ij2})$ in the second non-parametric term.
- X_{ij}, Z_{ij1} and Z_{ij2} are all scalars and time-varying covariates with $X_{ij} = b_{ij} + e_{1ij}$, $Z_{ij1} = b_{ij} + e_{2ij}$ and $Z_{ij2} = b_{ij} + e_{3ij}$. $b_{ij} \sim U[0,1]$, where e_{1ij}, e_{2ij}

and e_{3ij} are independent to each other and follow uniform distribution $U[-2,2]$.

We use the same setting for the working correlation matrix and parametric models as the last part: 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:

$$\boldsymbol{para4}: Y_{ij} = X_{ij}\beta + Z_{ij1} + Z_{ij2} + \epsilon_{ij},$$

$$\boldsymbol{para5}: Y_{ij} = X_{ij}\beta + \exp(Z_{ij1}) + \exp(Z_{ij2}) + \epsilon_{ij},$$

para 6:

$$Y_{ij} = X_{ij}\beta + Z_{ij1} + Z^2_{ij1} + Z^3_{ij1} + Z_{ij2} + Z^2_{ij2} + Z^3_{ij2} + \epsilon_{ij}$$

Similar to 3.1, in this simulation, we will focus on the estimation of β and the overall fitting of different estimators. Mean and standard deviation of the estimated β will be displayed. And we first show the results of the 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 the last paragraph.

Table 4. β estimates and overall MSE with 4 time points, two kernel smoothers

setup1	**ρ =0.3**				**ρ =0.7**			
	mean	se	mse-train	mse-test	mean	se	mse-train	mse-test
semi-WI	0.993	0.022	1.833	1.894	0.994	0.018	1.878	1.918
semi-True	0.989	0.023	1.868	2.25	0.994	0.022	1.904	2.175
Para4	1.015	0.09	1.981	2.205	1.01	0.079	1.987	2.043
Para5	1.008	0.091	1.975	2.207	1.005	0.08	1.978	2.043
Para6	0.991	0.092	1.929	2.013	0.989	0.08	1.917	2.031
setup2	**ρ =0.3**				**ρ =0.7**			
	mean	se	mse-train	mse-test	mean	se	mse-train	mse-test
semi-WI	1.012	0.029	2.51	2.52	1.011	0.027	2.72	2.77
semi-True	1.012	0.031	2.4	2.48	1.013	0.032	2.67	2.79
Para4	1.032	0.101	6.34	8.38	1.021	0.081	6.32	8.34
Para5	1.053	0.117	4.18	6.24	1.042	0.098	4.09	6.48
Para6	1.051	0.134	4.49	8.94	1.04	0.102	4.41	9.57

Table 4 above shows β estimates in the semi-parametric model (3.2) and parametric models in (para4–para6). The results in Table 4 based on the Gaussian Kernel density show that the standard errors of the β estimates based on semi-parametric estimators are at least three times less than the standard errors of the estimators from the three parametric models (para4–para6). The overall fitting mean square errors in the semi-parametric model (3.2) and parametric models (para4–para6) show that for the training dataset and test dataset, the mean square errors in the parametric estimators are higher than the mean square errors in semi-parametric estimators. Still, the result in this table shows that semi-parametric estimators with a stronger correlation and more complicated pattern in non-parametric part will benefit more compared to parametric estimators with the same scenarios. When compared to models with one kernel smoothers, the MSE for parametric models (para4-para6) increased at least four times, but the MSE of profile kernel GEE model (semi-WI and Semi-True) increased two times, indicating that profile kernel GEE estimator is more robust for MSE than parametric models.

4. Application

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

4.1. Description of the Dataset (Statistics and Data Analysis)

The dataset used in this application comes from UCI (University of California Irvine) Machine Learning Repository Website [10] with 30000 subjects and eight variables. A basic summary of statistics for those eight variables is as the follows:

1 Bill amount: 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.

2 Payment amount: Amount customer paid for current month, with minimum 0 and maximum 1684259.

3 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.

4 LIMIT BAL: Limit amount for each customer, with minimum 10000 and maximum 1000000.

5 SEX: With 1 denoting male and 2 denoting female.

6 EDUCATION: Education level for each customer: 1 denotes graduate school; 2 denotes university; 3 denotes high school; 4, 5, 6 denotes others.

7 MARRIAGE: Marital status: 1 denotes married; 2 denotes single; 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. Five variables are in the list left as predictors: 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 the response variable remaining amount and five predictors we discussed in the 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 matrices, and the results from semi-parametric models will be compared with the results from parametric models.

4.2. Results and Discussion: Overall Analysis

4.2.1. Using Remaining Amount as Response Variable

In this part, the remaining amount we defined in Section 4.1 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, marital status, and age. The following two parametric GEE models will be fitted:

Para 1 :

$$remaining\ amount = \beta_0 + \beta_1 limit\ balance + sex + education + marriage + \beta_2 age$$

Para 2 :

$$remaining\ amount = \beta_0 + \beta_1 limit\ balance + sex + education + marriage + \beta_2 age^2$$

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

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

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

Table 5. Parameter estimations for parametric and semi-parametric GEE in overall analysis

Para1	**Independence**		**Exchangeable**		**AR1**	
predictor	estimate	p-value	estimate	p-value	estimate	p-value
Intercept	-0.4809	< 2e-16	-0.481	< 2e-16	-0.484	< 2e-16
limit-balance	0.001977	< 2e-16	0.00198	< 2e-16	0.00199	< 2e-16
sex-female	-0.06279	3.25E-09	-0.0628	3.20E-09	-0.0662	7.20E-10
education-university	0.2064	< 2e-16	0.206	< 2e-16	0.213	< 2e-16
education-high school	0.1696	<2e-16	0.17	< 2e-16	0.176	< 2e-16
single	0.04873	0.0000582	0.0487	0.000058	0.0488	0.000066
age	0.0009698	0.142	0.00097	0.14	0.000937	0.16
Para2	**Independence**		**Exchangeable**		**AR1**	
predictor	estimate	p-value	estimate	p-value	estimate	p-value
Intercept	-0.467	< 2e-16	-0.467	< 2e-16	-0.471	< 2e-16
limit-balance	0.00198	< 2e-16	0.00198	< 2e-16	0.00199	< 2e-16
sex-female	-0.0625	3.40E-09	-0.0625	3.40E-09	-0.0659	7.40E-10
education-university	0.206	< 2e-16	0.206	< 2e-16	0.213	< 2e-16
education-high school	0.168	< 2e-16	0.168	< 2e-16	0.175	<2e-16
single	0.0496	0.000035	0.0496	0.000035	0.0496	0.00004
age^2	0.0000146	0.1	0.0000146	0.1	0.0000141	0.11
Semi	**Independence**		**Exchangeable**		**AR1**	
predictor	estimate	p-value	estimate	p-value	estimate	p-value
Intercept	-0.449	< 2e-16	-0.449	< 2e-16	-0.453	< 2e-16
limit-balance	0.00179	< 2e-16	0.00179	< 2e-16	0.0018	< 2e-16
sex-female	-0.0473	0.0000076	-0.0473	7.60E-06	-0.0506	0.0000022
education-university	0.216	< 2e-16	0.216	< 2e-16	0.222	< 2e-16
education-high school	0.166	< 2e-16	0.166	< 2e-16	0.172	<2e-16
single	0.0857	1.40E-15	0.0857	1.40E-15	0.086	2.10E-15

Table 5 shows the estimation results for the first parametric model (Para1) and the second parametric model (Para2) 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 marital status tend to have more remaining amount. The predictor limit balance has small coefficients, which denotes that limit balance has a positive correlation with remaining amount for both parametric model setups; age has a *p*-value larger than 0.05, showing that age is not statistically significant in both parametric models.

Different working correlation matrices such as independence, exchangeable, and AR1 are used in parametric models. The estimated parameters by those three working correlation matrices are quite similar across different settings in associations between time periods.

Since age is not significant with parametric patterns, such as linear and quadratic 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 and seeing if semi-parametric models are more advanced than pure parametric models.

The estimation results for the semi-parametric model with kernel smoother on the predictor age use different working correlation matrices. The result of the estimated coefficients is 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 marital status tend to have more remaining amount, and the coefficient for predictor single (0.086) is higher than the coefficient in parametric models (0.048 in Para3). The predictor limit balance has a coefficient of 0.002, which denotes that limit balance has a positive correlation with remaining amount.

4.2.2. Using Payment Status as Response Variable

In this part, the payment status, which is whether the client has a default we defined in Section 4.1 will be used as the response variable. 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, marital status, and age. We consider the parametric GEE model with linear form as the following:

Para 4 :

$$logit(p) = \beta_0 + \beta_1 limit\ balance + sex + education + marriage + \beta_2 age$$

where p is the probability of default.

Table 6. Parameter estimations for parametric GEE in overall analysis with payment status

	Independence		Exchangeable		AR1	
predictor	**estimate**	**p-value**	**estimate**	**p-value**	**estimate**	**p-value**
Intercept	1.42	<2e-16	1.42	<2e-16	1.41	< 2e-16
limit-balance	-3.32E-06	<2e-16	-3.32E-06	<2e-16	-3.20E-06	< 2e-16
sex-female	-0.275	<2e-16	-0.275	<2e-16	-0.251	< 2e-16
education-university	0.67	<2e-16	0.67	<2e-16	0.64	< 2e-16
education-high school	0.619	<2e-16	0.619	<2e-16	0.588	< 2e-16
single	0.14	4.00E-08	0.14	4.00E-08	0.134	1.20E-07
age	-0.0119	<2e-16	-0.0119	<2e-16	-0.011	3.10E-15

Table 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 marital status tend to have more probability to default. The predictor limit balance has a negative coefficient, which denotes that limit balance has a negative correlation with the probability of default, and age also has a negative correlation with the probability of default. Different working correlation matrices such as independence, exchangeable, and AR(1) matrix are used in Para1. The estimated parameters by those three working correlation matrices are quite similar.

4.3. Results and Discussion: Gender Analysis

In this section, we evaluate the difference between models for male customers and female customers. Following the overall analysis in Section 4.2, 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. 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.

4.3.1. Using Remaining Amount as Response Variable

Model Setups:

The remaining amount we defined in Section 4.1 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, marital status, and age for male and female customers. As an overall analysis in Section 4.2, the following three parametric GEE models will be fitted for male and female separately:

Para 1 for male :

$$remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_M + education \times I_M + marriage \times I_M + \beta_2 age \times I_M$$

Para 1 for female :

$$remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_F + education \times I_F + marriage \times I_F + \beta_2 age \times I_F$$

Para 2 for male :

$$remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_M + education \times I_M + marriage \times I_M + \beta_2 age^2 \times I_M$$

Para 2 for female :

$$remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_F + education \times I_F + marriage \times I_F + \beta_2 age^2 \times I_F$$

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

Semi for male :

$$remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_M + education \times I_M + marriage \times I_M + \theta(age) \times I_M$$

Semi for female :

$$remaining\ amount = \beta_0 + \beta_1 limit\ balance \times I_F + education \times I_F + marriage \times I_F + \theta(age) \times I_F$$

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

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

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

Table 7 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, the variables single and age are significant when using the independence working correlation matrix, while not significant when using exchangeable or AR(1) working correlation matrix; for female, age is not significant when using only 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. On the other hand, female customers with a high school degree have less remaining amount than those with a college school degree under all the three types of working correlation matrices.

Table 8 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 is significant only when using the independence working correlation matrix; for female, the quadratic term is significant when using the exchangeable working correlation matrix or AR(1) structure. Although the quadratic form of age is significant, it has a tiny impact on the response variable remaining amount because the number of coefficients 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 has more remaining amount. The marriage factor single is significant for male under the independence working correlation matrix, but it is not significant under any working correlation matrices for female customers.

Table 7. Parameter estimations for parametric GEE(para1) in gender analysis

	Independence: male		Independence: female		Exchangeable:male		Exchangeable:female		AR1:male		AR1:female	
predictor	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value
Intercept	-0.6126	< 2e-16	-0.39829	< 2e-16	-0.45529	1.00E-13	-0.1855	0.000011	-0.5682	< 2e-16	-0.33389	1.10E-14
limit-balance	0.00207	< 2e-16	0.001805	< 2e-16	0.002102	$<$ 2e-16	0.001874	< 2e-16	0.002128	< 2e-16	0.001821	< 2e-16
Education -university	0.14963	7.10E-11	0.20759	< 2e-16	0.148497	1.00E-10	0.208714	< 2e-16	0.163869	3.40E-12	0.212763	< 2e-16
education-high school	0.15242	6.30E-07	0.172856	1.70E-12	0.166683	5.20E-08	0.205231	< 2e-16	0.176602	1.40E-08	0.187171	5.10E-14
single	0.05519	0.0239	0.031259	0.087	0.01823	0.46	-0.01169	0.52	0.042972	0.085	0.020106	0.27606
age	0.00375	0.0041	-0.00158	0.113	-0.00015	0.91	-0.00748	3.60E-14	0.002241	0.087	-0.00347	0.00049

Table 8. Parameter estimations for parametric GEE(para2) in gender analysis

	Independence: male		Independence: female		Exchangeable: male		Exchangeable: female		AR1:male		AR1:female	
predictor	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value
Intercept	-0.54	< 2e-16	-0.428	< 2e-16	-0.467	2.00E-16	-0.335	2.00E-16	-0.527	< 2e-16	-0.403	< 2e-16
limit-balance	0.00208	< 2e-16	0.0018	< 2e-16	0.0021	2.00E-16	0.00185	2.00E-16	0.00213	< 2e-16	0.00181	< 2e-16
education-university	0.149	7.50E-11	0.208	< 2e-16	0.149	1.00E-10	0.209	2.00E-16	0.164	3.50E-12	0.213	< 2e-16
education-high school	0.152	0.0000007	0.173	2.10E-12	0.165	7.10E-08	0.203	2.20E-16	0.176	1.60E-08	0.186	8.80E-14
single	0.0529	0.0298	0.0324	0.071	0.0219	0.37	-0.00266	0.88	0.0424	0.088	0.0241	0.183
age^2	0.0000457	0.0092	-0.0000196	0.13	0.00000305	0.86	-0.0000864	1.30E-11	0.0000284	0.106	-0.0000405	0.0018

The results from two parametric models show that age may be not significant using some working correlation matrices, or has a tiny effect with parametric patterns, such as quadratic 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 9. Parameter estimations for semi-parametric GEE in gender analysis

	Independence: male		Independence: female		Exchangeable: male		Exchangeable: female		AR1: male		AR1:female	
predictor	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value
Intercept	-0.48386	< 2e-16	-0.485	< 2e-16	-0.48386	< 2e-16	-0.485	< 2e-16	-0.497014	< 2e-16	-0.486	< 2e-16
limit-balance	0.00218	< 2e-16	0.00182	< 2e-16	0.00218	< 2e-16	0.00182	< 2e-16	0.002215	< 2e-16	0.0018	< 2e-16
education-university	0.16564	< 2e-16	0.242	< 2e-16	0.16564	< 2e-16	0.242	< 2e-16	0.177118	< 2e-16	0.245	< 2e-16
education-high school	0.16031	1.50E-10	0.19	< 2e-16	0.16031	1.50E-10	0.19	< 2e-16	0.173217	7.40E-12	0.191	< 2e-16
single	0.04224	0.014	0.05	0.00031	0.04224	0.014	0.05	0.00031	0.041659	0.016	0.0513	0.00025

Table 10. Parameter estimations for parametric GEE(para4) with payment status

	Independence:male		Independence:female		Exchangeable:male		Exchangeable:female		AR1:male		AR1:female	
predictor	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value	estimate	p-value
Intercept	0.901231	1.80E-12	1.51457	<2e-16	0.901231	1.80E-12	1.51457	<2e-16	0.910599	4.50E-13	1.458421	< 2e-16
limit-balance	-0.003501	< 2e-16	-0.00337	<2e-16	-0.003501	< 2e-16	-0.00337	<2e-16	-0.003303	< 2e-16	-0.003312	< 2e-16
education-university	0.625329	< 2e-16	0.65678	<2e-16	0.625329	< 2e-16	0.65678	<2e-16	0.595055	< 2e-16	0.634428	< 2e-16
education-high school	0.53462	1.50E-14	0.64858	<2e-16	0.53462	1.50E-14	0.64858	<2e-16	0.50367	1.60E-13	0.615471	< 2e-16
single	0.238001	7.00E-06	0.0589	0.13	0.238001	7.00E-06	0.0589	0.13	0.2269	1.40E-05	0.073722	0.06
age	0.002483	0.38	-0.02031	<2e-16	0.002483	0.38	-0.02031	<2e-16	0.002179	0.43	-0.017833	1.20E-15

Table 9 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 the estimated coefficients is 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.

4.3.2. Using Payment Status as Response Variable

In this part, the payment status we defined in Section 4.1 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, marital 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.

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

Para 4 for male :

$$logit(p) = \beta_0 + \beta_1 limit\ balance \times I_M + education \times I_M + marriage \times I_M + \beta_2 age \times I_M$$

Para 4 for female :

$$logit(p) = \beta_0 + \beta_1 limit\ balance \times I_F + education \times I_F + marriage \times I_F + \beta_2 age \times I_F$$

where p is the probability of default.

Table 10 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 a single marital status tend to have more probability to default. The predictor limit balance has a negative coefficient for male and female, which denotes that limit balance has a 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, and AR(1) are used in Para4 for both male and female. The estimated coefficients by those three working correlation matrices are quite similar.

We did the same approach for the Education analysis, using two parametric models and one semi-parametric model to detect the impact for different education levels: customers with high school degree or university/graduate degree. When remaining balance is the response variable, the results from the parametric models show that age may be significant or has a tiny effect with parametric patterns. If we apply the semi-parametric models with kernel smoother on the predictor age, investigating whether there is a difference for customers with different education levels, the result of the estimated coefficients has 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 a positive coefficient, which denotes that limit balance has a positive correlation with remaining amount. Except for single, all other predictors are significant for all models

when using any types of three working correlation matrices. When payment status is the response variable, we only used the parametric model and found that for all models, female customers tend to have less probability to default because of negative coefficients. The predictor limit balance has a negative coefficient, which denotes that limit balance has a negative correlation with the probability of default. Age is not significant for customers with high school degrees but 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.

We also did an approach for Marriage analysis, using two parametric models and one semi-parametric model to detect the impact for different marital status: single customers or married customers. When remaining balance is the response variable, the results from two parametric models show that age may be not significant using some working correlation matrices or has tiny effect with parametric patterns, such as quadratic terms. Still, if we apply the semi-parametric models with kernel smoother on the predictor age, investigating whether there is a difference for customers with different marital status, the result for the estimated coefficients has some similarities to the estimation in parametric models. Based on the signs of the estimated coefficients, we found that for Limit balance, university and graduate are significant and positive correlated with remaining amount while variable female is significant and negative correlated with remaining amount for both single and married customers. When payment status is the response variable, we still only used the parametric model and found that the result is very similar to the Education analysis.

5. Conclusion and Discussion

In summary, the simulation result shows semi-parametric estimators are more robust with less standard error comparing to parametric estimators. For overall fitting, semi-parametric models have less mean square errors. Furthermore, semi-parametric estimators with stronger correlation, longer time period, and a more complicated pattern in the non-parametric part will benefit more when comparing to parametric estimators with the same scenarios.

In application part, we run the analysis based on credit card loan data and the result display that the parametric estimators will show clear patten when we treat some features as kernel smoother. We recommend applying parametric model first, figuring out the non-significant features and setting up them with kernel smoothers. By modeling with semi-parametric structure, we find the different behavior for customers with different gender, education level and marriage status.

In semi-parametric GEE study, estimating the working correlation matrix is critical when using data from the real world. One challenge comes from the application part, which is the dataset resource. Most financial datasets used in longitudinal studies reach the individuals level, which violates the privacy policies in most institutions in the United States. Our dataset, which comes from the UCI website, is based on credit information in Taiwan. In the future, we would like to use our model on other available credit loan datasets or other types of financial datasets in the United States.

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

The third challenge comes from the scheme of the semi-parametric approach. When assigning the non-parametric term in the semi-parametric approach, most applications in biological datasets use previous experience. In our approach, we used age as a 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 a non-parametric term.

The last challenge is a traditional issue for the GEE approach: estimating the working correlation matrix. In a semi-parametric GEE study, estimating the working correlation matrix is critical but more difficult than the parametric GEE approach. Fan, Huang, and Li [11] 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 to investigate whether this estimation method will provide more efficient semi-parametric estimators with fully specified working correlation matrix when we applied it in financial datasets.

Copyrights

References

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

[2] Fama, E. and MacBeth, J. (1973), *Risk, return, and equilibrium: Empirical tests*, The Journal of Political Economy, 81.3, 607–636.

[3] Sam, A.G. and Jiang, G. (2009), *Nonparametric estimation of the short rate diffusion from a panel of yields*, Journal of Financial and Quantitative Analysis (JFQA), Vol. 44, No. 5.

[4] 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.

[5] Zeger, S.L. and Liang, K.Y. (1986), *Longitudinal data analysis for discrete and continuous outcomes*, Biometrika, 43, 121-130.

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

[7] Lin, X. and Carroll, R.J. (2001), *Semiparametric regression for clustered data*, Biometrika, 88.4, 1179-1185.

[8] Wang, N., Carroll, R.J., and Lin, X. (2004), *Efficient semiparametric marginal estimation for longitudinal/clustered*

data, Journal of the American Statistical Association, 100, 147-157.

[9] 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.

[10] 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.

[11] Fan, J., Huang, T., and Li, R. (2007), *Analysis of longitudinal data with semiparametric estimation of covariance function*, Journal of the American Statistical Association, 102.478, 632-641.