Nonparametric Regression for Correlated Data

NOOR AKMA IBRAHIM
Universiti Putra Malaysia
Institute for Mathematical Research
UPM Serdang, Selangor Darul Ehsan 43400
MALAYSIA
nakma@putra.upm.edu.my

SULIADI
Bandung Islamic University
Department of Statistics
Jl. Tamansari No. 1 Bandung 40116
INDONESIA
suliadi@gmail.com

Abstract: This paper considers nonparametric regression to analyze correlated data. The correlated data could be longitudinal or clustered data. Some developments of nonparametric regression have been achieved for longitudinal or clustered categorical data. For data with exponential family distribution, nonparametric regression for correlated data has been proposed using GEE-Local Polynomial Kernel (LPK). It was showed that in order to obtain an efficient estimator, one must ignore within subject correlation. This means within subject observations should be assumed independent, hence the working correlation matrix must be an identity matrix. Thus to obtained efficient estimates we should ignore correlation that exist in longitudinal data, even if correlation is the interest of study. In this paper we propose GEE-Smoothing spline to analyze correlated data and study the properties of the estimator such as the bias, consistency and efficiency. We use natural cubic spline and combine with GEE in estimation. We want to study numerically, whether the properties of GEE-Smoothing spline are better than of GEE-Local Polynomial Kernel. Several conditions have been considered. i.e. several sample sizes and several correlation structures. Using simulation we show that GEE-Smoothing Spline is better than GEE-local polynomial. The bias of pointwise estimator is decreasing with increasing sample size. The pointwise estimator is also consistent even with incorrect correlation structure, and the most efficient estimate is obtained if the true correlation structure is used. We also give example using real data, and compared the result of the proposed method with parametric method and GEE-Smoothing Spline under independent assumption.

Key–Words: Nonparametric regression, Longitudinal binary data, Generalized estimating equation, Natural cubic spline, Properties of estimator.

1 Introduction

Nowadays many studies are conducted in the wide area that consist of many districts. In these studies, subjects are drawn from many districts. These studies are very common in economics, epidemiology or clinical trials. In the studies related to area, area usually can be split into some clusters where within cluster subjects are homogeneous but between cluster, subjects are heterogeneous. These studies are usually in economic or epidemiological research. For example, Gonzaléz et al. [4] studied the behavior obesity of childhood in developing countries. In this case, the subjects in a district are correlated whilst subjects from different districts are not. Other research of this type was conducted by Raimundo & Venturino [12] that studied the drug resistent impact on tubercolis transition. In this study dependency among subjects within an area must be considered in the model.

Another type of study that is common in epidemiology, biology, and clinical trial is a study that is related to time. In this study, subjects are followed over time or several occasions to collect response variables. This study is commonly known as longitudinal study. Example for longitudinal study is given by Adina et al. [2]. Adina et al. [2] studied the prognosis factor in metastatic breast cancer, carried out on 120 patients admitted at the Cluj-Napoca Institue of Oncology between January 2000 and December 2005. Subjects were followed on several observations. Data from the same subject are more similar than from different subject, meaning within subject observations are dependent whilst between subject observations are independent.

The characteristic of these data is that they are no longer independent. In the clustered data, there are correlations among subjects in a cluster whilst subjects from different cluster are independent. In longitudinal sudy, within subject observations are correlated whilst between subject observations are independent. Another characteristic is that the variances usually are not homogeneous.

Methods in the class of generalized linear model (GLM) are no longer valid for these data, since GLM

assumes that observations are independent. Some developments have been proposed to analyze such data, that can be classified into three types of model, marginal model, subject specific effect, and transition model (Davis [3]). In the class of marginal model, Liang and Zeger [8] and Zeger and Liang [13] extended quasi-likelihood estimation of Weddernburn [14] by introducing "working correlation" to accommodate within subject correlation, which is called generalized estimating equation (GEE). GEE yields consistent estimates of the regression coefficients and their variances even though there is misspecification of the working correlation structure, provided the mean function is correctly specified.

GEE is part of the class of parametric estimation, in which the model can be stated in a linear function and the function is known. Very often the effect of the covariate cannot be specified in the specific function. Nonparametric regression can accommodate this problem by relaxing relationship between covariate and response. In nonparametric regression, we assume that the effect of the covariate follows an unknown function without specific term, that is just a smooth function. To date there are several methods in nonparametric regression, for example: local polynomial kernel regression, penalized splines regression, and smoothing splines. Green and Silverman [5] gave a simple algorithm for nonparametric regression using cubic spline by penalized least square estimation. They also gave nonparametric and semiparametric methods for independent observations for class of generalized linear models.

Some developments of nonparametric and semiparametric regression for longitudinal or clustered data have been achieved. Lin and Carroll [9] considered nonparametric regression using longitudinal data GEE-Local Polynomial Kernel (LPK). They showed that for kernel regression, in order to obtained an efficient estimator, one must ignore within subject correlation. This means within subject observations should be assumed independent, hence the working correlation matrix must be an identity matrix. This result was definitely different from GEE of Liang & Zeger's, in which the GEE estimator was consistent even there are misspecification of the true correlation as working correlation. Lin and Carroll [10] also studied the behavior of local polynomial kernel which was applied to semiparametric-GEE for longitudinal data. The result was the same as in nonparametric GEE-LPK in Lin and Carroll [9]. Welsh et al. [15] studied the locality of the kernel method for nonparametric regression and compared it to P-splined regression and smoothing splines. The result was that the kernel is local even when the correlation is taken into account. The result was different for smoothing splines, in which if there is no within subject correlation then smoothing splines is local, and if within subject correlation increases, than smoothing splines become more nonlocal. This implies that for smoothing splines, within subject correlation must be taken into account in the working correlation.

This paper considers nonparametric regression to analyze longitudinal data. In this paper we propose GEE-Smoothing spline to analyze longitudinal data and study the properties of the estimator such as the bias, consistency and efficiency. We use natural cubic spline and combine this with GEE of Liang & Zeger's in estimation. We want to study numerically, whether the properties of GEE-Smoothing spline are better than of GEE-Local Polynomial Kernel proposed by Lin & Carrol [9]. Simulation study was carried out to investigate these properties.

The outline of this paper is follows. We give a short review of GEE in section 2.1. Section 2.2 considers brief review of smoothing splines. The algorithm of the proposed method is considered in section 3.1. Section 3.2 considers smoothing parameter selection. Properties of GEE-smoothing spline estimator using simulation are given in section 4. In section 5 we illustrate the application to real data and compare to the parametric GEE and GEE-Smoothing Spline. The conclusion and discussion are given in Section 6.

2 Generalized Estimating Equation and Smoothing Splines

2.1 Generalized estimating equation

Suppose there are K subjects, and the *i-th* subject is observed n_i times for the responses and covariates. Let $y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$ be the $n_i \times 1$ vector of response variable and $X_i = (x_{i1}, \ldots, x_{in_i})^T$ be $n_i \times p$ matrix of covariate for the *i-th* subject, and $x_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$. It is assumed that the marginal density of y_{ij} follows exponential family with probability density function

$$f(y_{ij}) = \exp\left(\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi)\right)$$

The first two moments of y_{ij} are

$$\mathrm{E}(y_{ij}) = b'(\theta_{ij}) = \mu_{ij}$$

and

$$\mathrm{Var}(y_{ij}) = b''(\theta_{ij})a(\phi),$$

where θ_{ij} is canonical parameter. It is assumed that between subject, observations are independent. The

relationship between μ and covariates through the link function is

$$g(\mu_{ij}) = \eta_{ij} = x_{ij}^T \beta \quad (1)$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ be $p \times 1$ vector of regression coefficient.

Generalized estimating equation to solve β was given by Liang and Zeger [8] as follows:

$$\sum_{i=1}^{K} D_i^T V_i^{-1} S_i = 0 \quad (2)$$

where

$$\begin{aligned} D_i &= \frac{\partial(b'(\theta_i))}{\partial\beta} = \frac{\partial\mu_i}{\partial\beta} \\ &= \frac{\partial\mu_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta} \\ &= A_i \Delta_i X_i, \end{aligned}$$

$$\Delta_i = \frac{\partial\theta_i}{\partial\eta_i},$$

and

$$V_i = A_i^{1/2} R(\alpha) A_i^{1/2}.$$

A_i is an $n_i \times n_i$ diagonal matrix with diagonal elements var(y_{ij}). $R(\alpha)$ is also called a "working correlation", an $n_i \times n_i$ symmetric matrix which fulfills the requirement of being a correlation matrix, and $S_i = y_i - \mu_i$. The estimating equation (2) is similar to the quasi-likelihood estimating equation, except the form of V_i. Thus it can be seen as an estimating equation of β by letting Φ as the "quasi-likelihood" score function of the $y_1, y_2, \ldots, y_K$. Solution of β can be obtained by minimizing Φ subject to β . Thus the estimating equation is

$$\frac{\partial\Phi}{\partial\beta} = \sum_{i=1}^{n} D_i^T V_i^{-1} S_i = 0$$

Liang and Zeger [8] gave the iterative procedure using modified Fisher scoring for β and moment estimation method of α and ϕ . Given the current estimates of $\hat{\alpha}$ and $\hat{\phi}$ then the iterative procedure for β is

$$\begin{aligned} \hat{\beta}_{s+1} =& \hat{\beta}_s + \left[\sum_{i=1}^{n} D_i^T(\hat{\beta}_s)\tilde{V}_i^{-1} D_i(\hat{\beta}_s)\right]^{-1} \\ &\times \left[\sum_{i=1}^{n} D_i^T(\hat{\beta}_s)\tilde{V}_i^{-1} S_i(\hat{\beta}_s)\right] \end{aligned} \quad (3)$$

where $\tilde{V}_i(\beta) = \tilde{V}_i\{\beta, \alpha(\beta, \hat{\phi}(\beta))\}$. The close form of moment estimator for α and ϕ for some correlation structures can be seen in Liang & Zeger [8].

2.2 Smoothing spline

Green and Silverman [5] gave simple approach in estimating smooth function f in interval $[a, b]$ using natural cubic splines. Suppose given n real number $t_1, t_2, \ldots, t_n$ on the interval $[a, b]$ and satisfying $a < t_1 < \cdots < t_n < b$. A function f on $[a, b]$ is cubic spline if two conditions are satisfied. First, f is cubic polynomial on each interval $(a, t_1), (t_1, t_2), \ldots, (t_n, b)$; second, the polynomial pieces fit together at the points t_i in such a way that f itself and its first and second derivative are continuous at each t_i, thus the function is continuous on the whole of $[a, b]$. It is said to be natural cubic spline (NCS), if its second and third derivative are zero at a and b. Suppose $f_i = f(t_i)$ and $\gamma_i = f''(t_i)$ for $i = 1, 2, \ldots, n$. By definition of NCS, the second derivative of f at t_1 and t_n are zero, so $\gamma_1 = \gamma_n = 0$. Let $\boldsymbol{f} = (f_1, f_2, \ldots, f_n)^T$ and $\gamma = (\gamma_2, \ldots, \gamma_{n-1})^T$. Vector γ is numbered in non standard way, starting at $i = 2$. The vector f and vector γ completely specify the curve f. These two vectors are related and specified by two matrices Q and R defined below.

Let $h_i = t_{i+1} - t_i$, for $i = 1, 2, \ldots, n-1$. Let Q be the $n \times (n-2)$ matrix with elements q_{ij}, $i = 1, \ldots, n$, and $j = 2, \ldots, n-1$, given by

$$q_{j-1,j} = h_{j-1}^{-1},$$

$$q_{jj} = -h_{j-1}^{-1} - h_j^{-1},$$

and

$$q_{j+1,j} = h_j^{-1}.$$

The R matrix is defined by the $(n-2) \times (n-2)$ symmetric matrix with elements r_{ij}, for i and j running from 2 to $(n-1)$, given by

$$r_{ii} = (h_{i-1} + h_i)/3, \text{ for } i = 2, 3, .., n-1$$

$$r_{i,i+1} = r_{i+1,i} = h_i/6, \text{ for } i = 2, 3, .., n-1$$

Matrix R and Q are numbered in non standard way. The matrix R is strictly diagonal dominant, in which $|r_{ii}| > \sum_{i \neq j} |r_{ij}|$. Thus R is strictly positive-definite, hence R^{-1} exists. Defined a matrix G by

$$G = QR^{-1}Q^T \quad (4)$$

The important result is the theorem below (Greean & Silverman [5]):

Theorem 1 *The vector $\boldsymbol{f}$ and γ specify a natural cubic spline f, if and only if the condition*

$$Q^T \boldsymbol{f} = R\gamma$$

is satisfied. If condition above is satisfied then the roughness penalty will satisfy

$$\int_a^b [f''(t)]^2 dt = \gamma^T R \gamma = \boldsymbol{f}^T G \boldsymbol{f} \quad (5)$$

The proof of this theorem can be seen in Green and Silverman [5].

Green and Silverman [5] proposed smoothing spline for several conditions, e.g nonparametric and semiparametric regressions for independent continuous data, nonparametric and semiparametric generalized linear models for independent data, and quasi-likelihood for independent data. They also considered method for correlated continuous data. For quasi-likelihood approach, the important result is the solution of the function f for nonparametric regression and parameter β in semiparametric regression, obtained by maximizing "penalized quasi-likelihood":

$$\Pi = \Phi - \frac{1}{2}\lambda \int [f''(t)]^2 dt \tag{6}$$

Thus the solution of f is obtained by maximizing (6).

3 Generalized Estimating Equation-Smoothing Spline

3.1 Estimation of GEE-smoothing spline

Suppose there are K subjects and the measurement of the *i-th* subject taken n_i times. Let $y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$ be a vector of responses of the *i-th* subject, corresponding to the vector of covariate $t_i = (t_{i1}, t_{i2}, \ldots, t_{in_i})^T$ and $y_{ij} \in \{0, 1\}$ be Bernoulli distributed and comes from exponential family distribution with canonical parameter θ_{ij}. Thus $E(y_{ij}) = b'(\theta_{ij}) = \mu_{ij}$ and $Var(y_{ij}) = b''(\theta_{ij})a(\phi) = \mu_{ij}(1 - \mu_{ij})$.

Consider the population average model, where the systematic component of the exponential family is nonparametric, rather than parametric, that is

$$g(\mu_{ij}) = \eta_{ij} = f(t_{ij}),$$

$$i = 1, 2, ..., K;\ j = 1, 2, ..., n_i$$

We replace the systematic component in (1) with unknown smooth function, i.e. natural cubic splines, rather than linear (known) function. In this paper we use the canonical link function $\theta_{ij} = \eta_{ij}$. Suppose X_i an $n_i \times q$ incidence matrix of all t_{ij}'s that can be constructed as follows. Let all t_{ij}'s have q different values that can be ordered to be $t_{(1)} < t_{(2)} < \cdots < t_{(q)}$ with relation to x_{ijk} is $x_{ijk} = 1$, if $t_{ij} = t_{(k)}$ and $x_{ijk} = 0$, if $t_{ij} \neq t_{(k)}$ for $k = 1, 2, \ldots, q$.

Let $x_{ij} = (x_{ij1}, x_{ij2} \ldots, x_{ijq})^T$ and vector of the functions f at different points denoted by $\boldsymbol{f} = [f(t_{(1)}), f(t_{(2)}), \ldots, f(t_{(q)})]^T$. Then the function f at point t_{ij} can be expressed as $f(t_{ij}) = x_{ij}^T \boldsymbol{f}$. Set

$$\begin{aligned} X_i &= (x_{i1}, x_{i2}, \ldots, x_{in_i})^T, \\ y_i &= (y_{i1}, y_{i2}, \ldots, y_{in_i})^T, \\ \eta_i &= (\eta_{i1}, \eta_{i2}, \ldots, \eta_{in_i})^T, \\ \mu_i &= (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})^T \end{aligned}$$

Since function f can be any arbitrary smooth function, then to maximize "quasi-likelihood" score function Φ (see Sub-Section 2.1), one might take y_{ij} as the estimates of $f(t_{ij})$ and the Φ will be maximum. But the function obtained, $\hat{f}$, is just an interpolation of the y_{ij}'s and the function is too rough or wiggly. One might want a smooth function by adding roughness penalty to the objective function. This is called penalized "quasi-likelihood" function defined by

$$\Pi = \Phi - \frac{1}{2}\lambda \int_a^b [f''(t)]^2 dt \tag{7}$$

From (2), (3), and (5), the estimating equation that maximizing penalized "quasi-likelihood" function (7) is defined as

$$\begin{aligned} \frac{\partial \Pi}{\partial f} &= \sum_{i=1}^{K} D_i^T V_i^{-1} S_i - \frac{\partial}{\partial f}\left[\frac{1}{2}\lambda \int [f''(t)]^2 dt\right] \\ &= \sum_{i=1}^{K} D_i^T V_i^{-1} S_i - \lambda G \boldsymbol{f} = 0 \end{aligned}$$

where

$$\begin{aligned} D_i &= \frac{\partial (b'(\theta_i))}{\partial \beta} = \frac{\partial \mu_i}{\partial \beta} \\ &= \frac{\partial \mu_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta} \\ &= A_i \Delta_i X_i, \end{aligned}$$

and $S_i = y_i - \mu_i$(see subsection 2.1).

Given the current estimates of $\hat{\alpha}$ and assuming canonical link function is used, following Liang and Zeger [8] as in (3), then the iterative procedure using modified Fisher scoring for $\boldsymbol{f}$, is

$$\begin{aligned} \hat{\boldsymbol{f}}_{s+1} =& \hat{\boldsymbol{f}}_s + \left[\sum_{i=1}^{K} D_i^T \tilde{V}_i^{-1} D_i + \lambda G\right]^{-1} \\ & \times \left[\sum_{i=1}^{K} D_i^T \tilde{V}_i^{-1} S_i - \lambda G \hat{\boldsymbol{f}}_s\right] \end{aligned} \tag{8}$$

where D_i, $\tilde{V}_i$, and S_i are evaluated using $\hat{\boldsymbol{f}}_s$. The association parameter α can be estimated using method of moment [8].

We may use sandwich variance estimator for the estimate suggested by Liang & Zeger [8]. This estimator is robust due to the misspecification of the correlation structure. The sandwich variance estimator of $\hat{\boldsymbol{f}}$ is defined by

$$\text{Var}_S(\hat{\boldsymbol{f}}) = \Sigma_0^{-1}\Sigma_1\Sigma_0^{-1}, \tag{9}$$

where

$$\Sigma_0^{-1} = \left[\sum_{i=1}^{K} D_i^T \tilde{V}_i^{-1} D_i + \lambda G\right]^{-1} \quad \text{and}$$

$$\Sigma_1 = \sum_{i=1}^{K} D_i^T \tilde{V}_i^{-1} S_i S_i^T D_i \tag{10}$$

A special case using canonical link funtion, the $\partial\theta_i/\partial\eta_i = I_{n_i}$. Thus the form of (10) becomes

$$\Sigma_0^{-1} = \left[\sum_{i=1}^{K} X_i^T A_i \tilde{V}_i^{-1} A_i X_i + \lambda G\right]^{-1} \quad \text{and}$$

$$\Sigma_1 = \sum_{i=1}^{K} X_i^T A_i \tilde{V}_i^{-1} S_i S_i^T \tilde{V}_i^{-1} A_i X_i$$

Another possibility of $\text{Var}(\hat{\boldsymbol{f}})$ is model based covariance obtained from (8), this is also called naive estimator. The naive estimator is defined by the inverse hessian matrix, i.e

$$\text{Var}_N(\hat{\boldsymbol{f}}) = \Sigma_0^{-1}. \tag{11}$$

3.2 Smoothing parameter selection

Smoothing parameter (λ) is an important part in GEE-Smoothing Spline. The parameter measures the "trade off" or exchange between goodness of fit and the roughness or the smoothness of the curve. Hence, the performance of the estimator depends on this parameter. In selecting smoothing parameter, we use a method proposed by Wu & Zhang ([16], p326) which is called *leave-one-subject-out cross validated deviance* (SCVD). Smoothing parameter λ is chosen that minimizes SCVD score, where

$$SCVD(\lambda) = \sum_{i=1}^{K}\sum_{j=1}^{n_i} d(y_{ij}, \hat{\mu}_{ij}^{(-i)})$$

where d is "deviance" and $\mu_{ij}^{(-i)} = g^{-1}(X_i\hat{\boldsymbol{f}}^{(-i)})_{ij}$ is the estimate value for the *i-th* subject and the *j-th* time observation using $\hat{\boldsymbol{f}}^{(-i)}$. The $\hat{\boldsymbol{f}}^{(-i)}$ is f obtained without the *i-th* observation. Since GEE is based on quasi-likelihood thus the deviance is also based on quasi-likelihood (see: Hardin & Hilbe [6], Ch. 4; McCullagh & Nelder [11], Ch. 9).

Direct computation of $\hat{\boldsymbol{f}}^{(-i)}$ is time consuming. Wu & Zhang [16] suggested using approximate of $\hat{\boldsymbol{f}}^{(-i)}$ computed as follows. Suppose from the final iteration of (8), we have D_i, $\tilde{V}_i^{-1}$, S_i and $\hat{f}_s$. Then the $\hat{\boldsymbol{f}}^{(-i)}$ is approximated by

$$\hat{\boldsymbol{f}}^{(-i)} = \hat{\boldsymbol{f}}_s + \left[\sum_{i\neq r}^{K} D_r^T \tilde{V}_r^{-1} D_r + \lambda G\right]^{-1} \times \left[\sum_{i\neq r}^{K} D_r^T \tilde{V}_r^{-1} S_r - \lambda G\hat{\boldsymbol{f}}_s\right]$$

We still need to compute $\hat{\boldsymbol{f}}^{(-i)}$ for $i = 1, 2, \ldots, K$, but we do not need to iterate (8) from the beginning.

4 Simulation Study

The objective of this simulation is to study the properties of GEE-smoothing spline, such as biasness, consistency, and efficiency, considering different sample sizes with correct and incorrect correlation structure in estimation. In this simulation we only consider binary data using logit link function.

4.1 Model and structure of data

We generated correlated binary data using R language version 2.7.1 (see: Leisch et al [7]). Three correlation structures were considered: (i) autoregressive with $corr(y_{ij}, y_{i(j+1)}) = 0.7$, for $j = 1, 2, \ldots, n_i$; (ii) exchangeable with $corr(y_{ij}, y_{ij'}) = 0.35$, for $j', j = 1, 2, \ldots, n_i$ and $j' \neq j$; and (iii) independency with $corr(y_{ij}, y_{ij'}) = 0$, for $j', j = 1, 2, \ldots, n_i$ and $j' \neq j$. Each subject is considered to be measured ten times, $t = 7.5, 25.5, 43.5, \ldots, 169.5$. The function is $f(t) = \sin(\pi t/90)$. Response variable, y_{ij}, related to covariate, t, through canonical link function is as follows,

$$E(y_{ij}) = \mu_{ij} \quad \text{and} \quad \text{logit}\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = f(t_{ij})$$

We considerd three sample sizes $n = 15$, $n = 30$, and $n = 50$. For each correlation structure, we estimated function f using the three correlation structure: autoregressive, exchangeable, and independency. Thus for each one, there are nine combinations of sample sizes and correlation structure. Each combination was run 250 times.

The purpose of this simulation is to study the properties of the estimator, such as biasness, consistency, and efficiency.

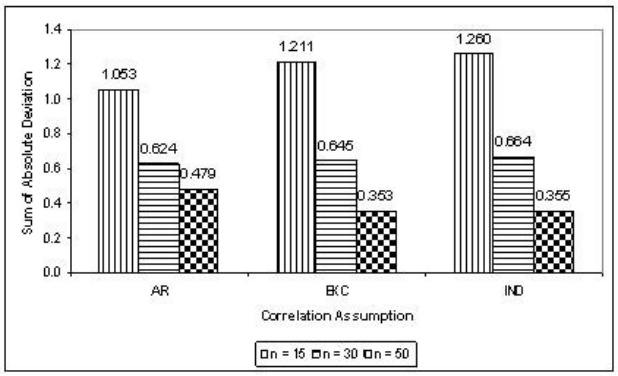
4.2 Simulation results

In order to assess the biasness of the estimator we use pointwise sum of absolute deviation (SAD). SAD is defined as follows. Suppose the estimate of f at point t for the *r-th* replication is $\hat{f}_t^{(r)}$ and $\hat{f}_t^*$ is the average of $\hat{f}_t^{(r)}$ of 250 replications, thus $\hat{f}_t^* = \sum_{r=1}^{250} \hat{f}_t^{(r)}/250$, and the true f at point t is f_t. SAD is defined as $SAD = \sum_{j=1}^{10} |\hat{f}_{t_j}^* - f_{t_j}|/10$. Thus SAD shows the size of bias of the estimates. Figure 1 (a), (b), and (c) show the SAD for true correlation structure of autoregressive, exchangeable, and independency respectively.

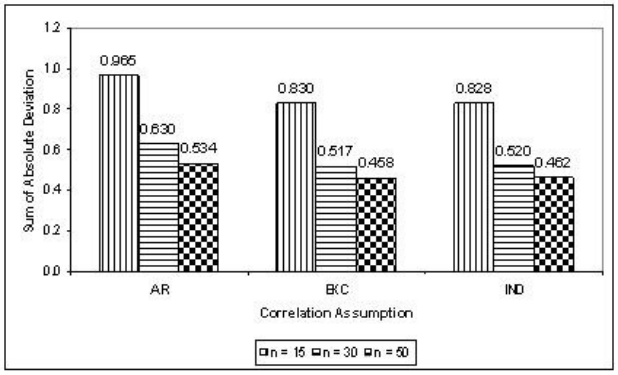
From Figure 1 we can see the biasness of the estimators. Refering to the correlation structure, there is no pattern for the size of bias whether we use correct or incorrect correlation structure. The degree of biasness is related to the sample size. Whether using correct or incorrect correlation structure, the bias will decrease when sample size increases. This pattern is the same for data that have high correlation (autoregressive, $\alpha = 0.7$), moderate correlation (Exchangeable, $\alpha = 0.35$), and independent.

We used standard deviation of 250 replication at each point estimates to study the consistency and efficiency. The estimator is consistent if standard deviation tends to zero when sample size is infinity, i.e. standard deviation decreases while sample size increases. This standard deviation can also be used to study the efficiency, that is small standard deviation indicates the efficiency of the estimator. Figure 2 and Table 1 show the standard deviation of 250 pointwise function estimates.

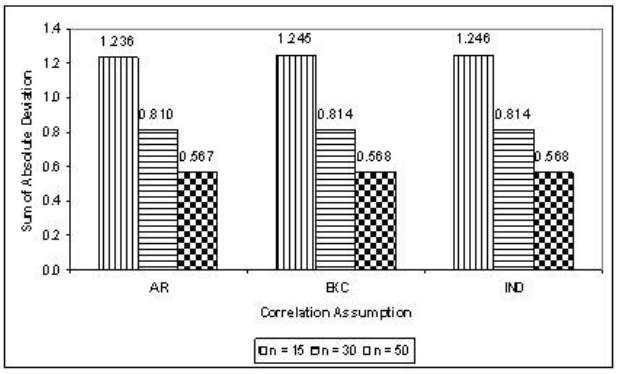
From Figure 2 and Table 1 we can see the consistency of the estimator. The pattern of standard deviation for all true correlation structures is the same. It decreases when sample size increases. The same pattern is also observed for all correlation structures, using correct or incorect correlation structure. This means that the estimators are consistent and the consistency still holds even if we use incorrect correlation structure. The rate of the decreasing of standard deviation from $n = 15$ to $n = 30$, and from $n = 30$ to $n = 50$, are the same for all true correlation structures. This indicates the convergency rate is (almost) the same for all conditions of true correlation structures. From the standard deviation we can also study the efficiency of the estimator. From the result of the efficiency study we may conclude whether we need to take into account the correlation into the model or just ignore the dependency. The method that has smaller variance or standard deviaton of estimator is more efficient than others.



(a) True correlation is AR-1



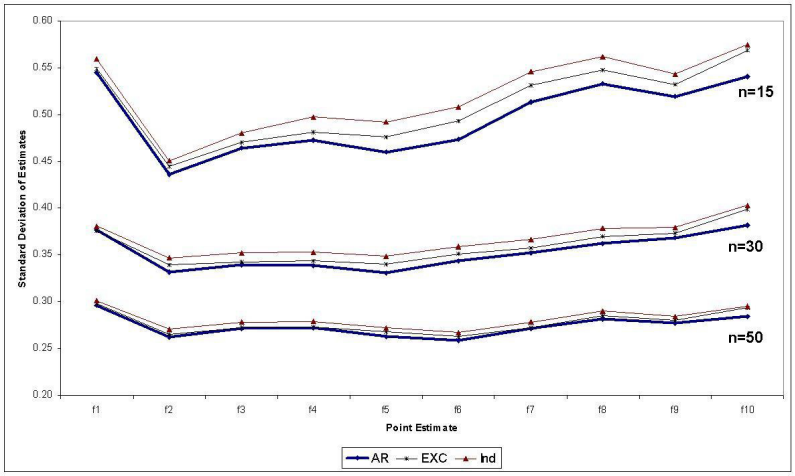
(b) True correlation is Exchangeable



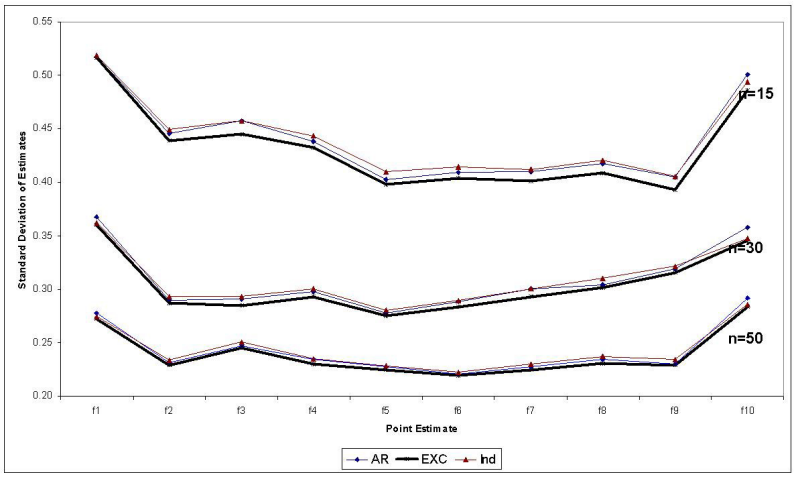
(c) True correlation is Independent

Figure 1: Sum of Absolute Deviation of the Three of True Correlation Structures

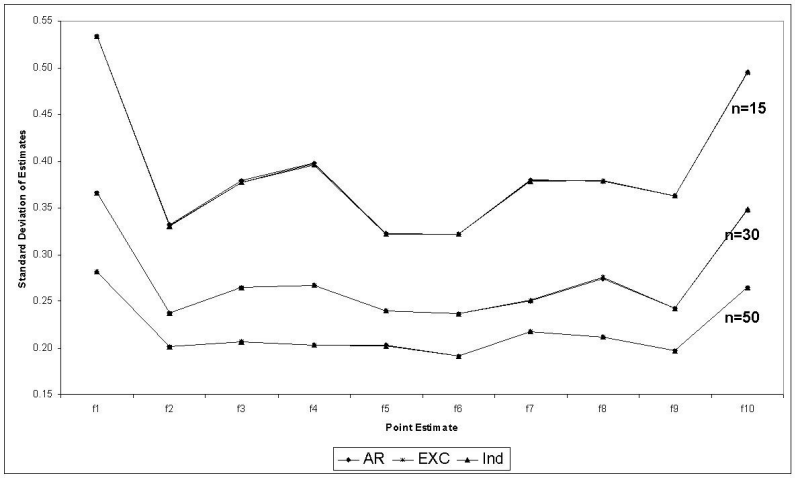
Figure 2 and Table 1 show that if data are correlated (true correlation is autoregressive or exchangeable), for specific sample size, the biggest standard deviation is obtained if one assumes that the data are independent. Whilst using true correlation structure, the standard deviation is the smallest. This means that taking into account the dependency into the model is better than assuming data are independent, even we use incorrect correlation structure. The most efficient estimate is obtained if we use true correlation structure. The difference between standard deviations of



(a) True correlation is AR-1



(b) True correlation is Exchangeable



(c) True correlation is Independent

Figure 2: Standard Deviation of 250 Replications of Pointwise Function Estimates

correlation structure (AR1, EXC, and IND) tends to get closer when we increase the sample size, hence we can make a conjecture that the efficiency of correct or incorrect correlation structure is almost similar if sample size is large. If true correlation structure is independent, the standard deviation of AR1, EXC,

Sample K	Assuming Corr.	Estimate Point $f(t_1)$	$f(t_2)$	$f(t_3)$	$f(t_4)$	$f(t_5)$	$f(t_6)$	$f(t_7)$	$f(t_8)$	$f(t_9)$	$f(t_{10})$
		Correlation Structure AR-1									
15	AR1	0.545	0.436	0.464	0.472	0.460	0.474	0.513	0.533	0.519	0.541
	EXC	0.549	0.445	0.470	0.481	0.476	0.493	0.531	0.547	0.532	0.568
	IND	0.559	0.451	0.480	0.497	0.493	0.508	0.545	0.562	0.543	0.575
30	AR1	0.377	0.332	0.339	0.338	0.331	0.343	0.352	0.362	0.368	0.381
	EXC	0.376	0.339	0.342	0.344	0.340	0.351	0.357	0.370	0.373	0.399
	IND	0.381	0.347	0.352	0.353	0.349	0.359	0.366	0.378	0.380	0.403
50	AR1	0.296	0.262	0.271	0.273	0.263	0.259	0.271	0.282	0.278	0.284
	EXC	0.298	0.265	0.272	0.273	0.268	0.262	0.273	0.285	0.280	0.293
	IND	0.301	0.270	0.278	0.279	0.273	0.267	0.278	0.290	0.284	0.295
		Correlation Structure Exchangeable									
15	AR1	0.517	0.445	0.458	0.439	0.403	0.409	0.410	0.417	0.405	0.501
	EXC	0.516	0.439	0.445	0.432	0.398	0.404	0.401	0.409	0.393	0.486
	IND	0.519	0.449	0.458	0.444	0.410	0.415	0.412	0.420	0.405	0.494
30	AR1	0.368	0.290	0.291	0.297	0.278	0.289	0.301	0.304	0.319	0.358
	EXC	0.360	0.287	0.285	0.293	0.275	0.283	0.293	0.301	0.315	0.345
	IND	0.363	0.293	0.293	0.301	0.281	0.289	0.300	0.310	0.322	0.348
50	AR1	0.277	0.231	0.247	0.234	0.227	0.221	0.227	0.234	0.230	0.292
	EXC	0.272	0.229	0.245	0.229	0.224	0.219	0.225	0.230	0.229	0.283
	IND	0.275	0.233	0.250	0.235	0.228	0.223	0.229	0.237	0.234	0.286
		Correlation Structure Independent									
15	AR1	0.534	0.332	0.379	0.398	0.323	0.322	0.380	0.378	0.363	0.495
	EXC	0.534	0.331	0.377	0.397	0.322	0.322	0.379	0.379	0.363	0.495
	IND	0.534	0.330	0.377	0.396	0.322	0.322	0.378	0.379	0.363	0.495
30	AR1	0.367	0.237	0.265	0.267	0.240	0.236	0.250	0.274	0.242	0.348
	EXC	0.366	0.237	0.265	0.267	0.239	0.236	0.251	0.275	0.242	0.348
	IND	0.366	0.237	0.265	0.267	0.239	0.236	0.251	0.275	0.242	0.348
50	AR1	0.282	0.201	0.206	0.203	0.203	0.191	0.217	0.211	0.197	0.265
	EXC	0.282	0.201	0.206	0.203	0.202	0.191	0.217	0.212	0.197	0.265
	IND	0.282	0.201	0.206	0.203	0.202	0.191	0.217	0.212	0.197	0.265

Table 1: Standard Deviation of the Estimate Points of Function for Nonparametric Components

and IND are almost similar, for all sample sizes. Thus in this case, the efficiency of using incorrect correlation structures is almost similar to the efficiency of using correct correlation structure.

5 Application to Real Data

As an application of the proposed method, we used data of A5055 Long-Term Viral Dynamic Data. The data were generated from AIDS Clinical Trials Group, ACTG 5055 study, which was sponsored by NIAID/NIH. More details of this clinical study can be found in Acosta et al. [1]. Among the total 44 patients accrued in this study, 42 subjects were included in the analysis; of the remaining two subjects, one was excluded from the analysis because the some covariates were not obtained and the other was excluded because the phenotype assay could not be completed on this subject. Detection limit of the viral load (HIV RNA copies) assay is 50 copies per ml blood. If it is below detectable, it is imputed as 25 in the data set. Data were recoded as

$$y_{ij} = \begin{cases} 1, & \text{if RNA} < 50 \text{ copies per ml blood} \\ 0, & \text{otherwise} \end{cases}$$

The covariate is day after treatment. For dependent model, we assumed that the structure of correlation is exchangeable. This means that the within subject correlations for different lag-time are the same.

As comparison we did three scenarios: (i) parametric approach; (ii) nonparametric (smoothing

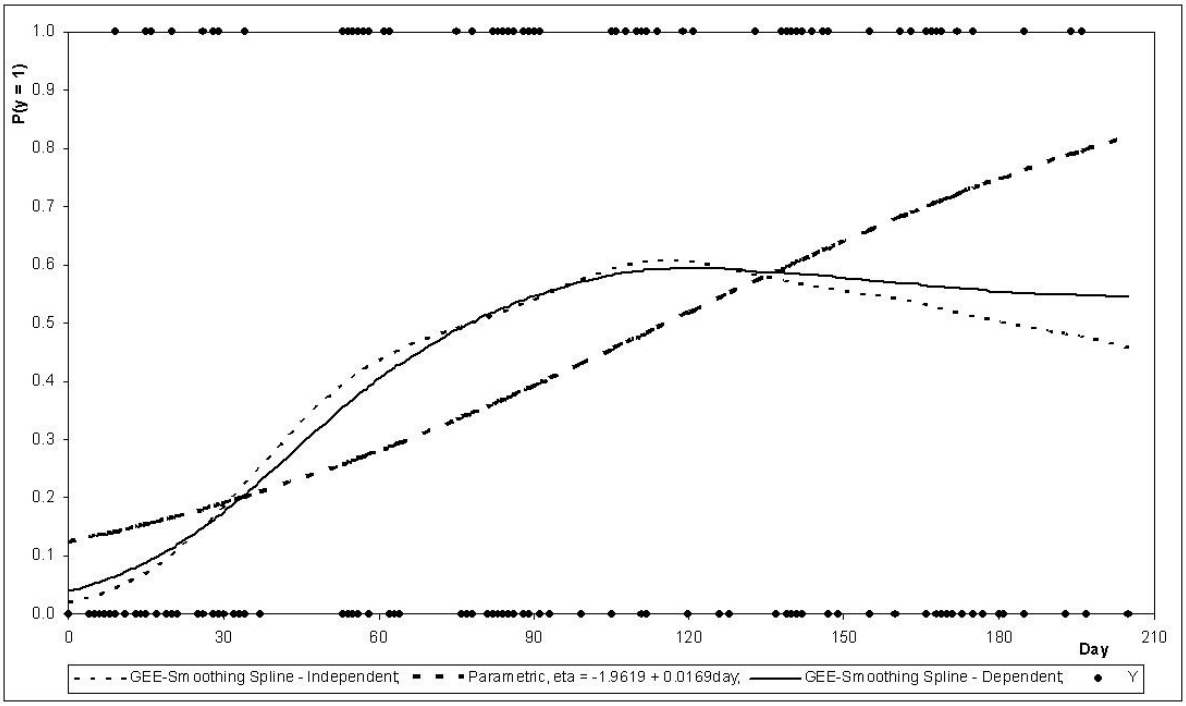


Figure 3: Comparison of the Result: the Parametric Approach, Independent, and Dependent GEE-Smoothing Spline

spline) approach with assumption within subject observations are independent; and (iii) nonparametric approach with assumption within subject observations are dependent, using GEE-Smoothing Spline. We used PROC GENMOD for parametric approach, for smoothing spline we used PROC GAM in SAS 9.0, and SAS IML for GEE-Smoothing Spline.

Results of Parametric and Nonparametric model are completely different. GEE-parametric model is $\eta_{ij} = -1.9619 + 0.0169$day, and correlation coefficient is $r = 0.2467$. The P-value of the intercept and covariate are less than .0001 respectively. The curve of this model shows that P(y=1) will increase slowly with respect to the increasing of day.

Result of nonparametric approach is definitely different with parametric ones (see Figure 3). Nonparametric models showed that the relationship of $P(y = 1)$ and day is almost quadratic. Results of independent and dependent assumption are almost the same. From beginning of the day after treatment until day ≈ 130, those two curves are similar. The difference between those two assumptions started from this point, in which the curve of the dependent assumption will decrease faster than independent assumption. Another result is $P(y = 1)$ for dependent assumption is higher than independent one. For dependent assumption, the estimate of within subject correlation is 0.2208.

6 Conclusion and Discussion

From section 4, it can be concluded that GEE-smooting spline has better properties than GEE-local polynomial kernel proposed by Lin & Carroll [9]. The pointwise estimates of GEE-smoothing are consistent, even we use incorrect correlation structure. The convergency rates of consistency for independent data (no correlation), moderate correlation, and high correkation are the same. If data are correlated, ignoring this correlation in the model, will give the most inefficient estimate. Taking into account the dependency into the model is better than ignoring it, even using incorrect correlation structure. If data are independent, the efficiency of using correct or incorrect correlation structures is almost similar. Hence, since in true situation the correlation is unknown, then it is better to assumme the data are correlated rather than to assume data are independent . We have shown by simulation that the estimator of GEE-smoothing spline has good properties. As an extension for future research, it is imperative these properties should be shown analitically.

Acknowledgements: This research is supported by Science Fund grant Vote No. 5450434 from Ministry of Science, Technology and Innovation Malaysia.

References:

[1] Acosta, E.P, H. Wu, A. Walawander, J. Eron, C. Pettinelli, S. Yu, D. Neath, E. Ferguson, A. J. Saah, D. R. Kuritzkes, and J. G. Gerber, for the Adult ACTG 5055 Protocol Team, Comparison of two indinavir/ritonavir regimens in treatment-experienced HIV-infected individuals. *Journal of Acquired Immune Deficiency Syndromes*, **37**, 2004, pp1358-1366

[2] Adina, Man Milena, C. Bondor, I. Neagoe, M. Pop, A. Trofor, D. Alexandrescu, and O. Arghir, Prognosis Factors in the Evaluation of Metastatic Breast Cancer, *WSEAS TRANSACTIONS on BIOLOGY and MEDICINE*, **5**, 2008, pp281-292.

[3] Davis, Charles S, *Statistical Methods for the Analysis of Repeated Measurements*, Springer-Verlag, New York, USA, 2002.

[4] Gonzaléz, Gilberto, Lucas Jódar, Rafael Villanuera, and Fransisco Santonja, Random Modeling of Population Dynamics with Uncertainty, *WSEAS TRANSACTIONS on BIOLOGY and MEDICINE*, **5**, 2008, pp34-45.

[5] Green, P. J. and B. W. Silverman, *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach*, Chapman & Hall/CRC, New York, USA, 1994.

[6] Hardin, James W. and Joseph M Hilbe, *Generalized Estimating Equations*, Chapman & Hall/CRC, Washington DC, USA, 2003.

[7] Leisch, Friedrich, Andreas Weingessel and Kurt Hornik, On the Generation of Correlated Artificial Binary Data. *Working Paper. No 13*. SFB. Adaptive Information System and Modeling in Economics and Management Science, Vienna University of Economics and Business Administration, Austria, 1998.

[8] Liang, K. Y. and S. L. Zeger, Longitudinal Data Analysis using Generalized Linear Models, *Biometrika*, **73**, 1986, pp13-22.

[9] Lin, Xihong and Raymond J. Carroll, Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error, *Journal of the American Statistical Association*, **95**, 2000, pp520-534.

[10] Lin, Xihong and Raymond J. Carroll, Semiparametric Regression for Clustered Data Using Generalized Estimating Equations, *Journal of the American Statistical Association*, **96**, 2001, pp1045-1056.

[11] McCullagh, P. and J. A. Nelder, *Generalized Linear Models* 2nd Edition Chapman and Hall, London, UK, 1989.

[12] Raimundo,Silvia Martorano and Ezio Venturino, Drug Resistants Impact on Tuberculosis Transmission, *WSEAS TRANSACTIONS on BIOLOGY and BIOMEDICINE*, **5**, 2008, pp85-95.

[13] Zeger, Scott L. and Kung-Yee Liang, Longitudinal Data Analysis for Discrete and Continuous Outcomes, *Biometrics*, **42**, 1986, pp121-130.

[14] Weddernburn, R. W. M, Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method, *Biometrika*, **61**, 1974, pp439-447.

[15] Welsh, Alan H., Xihong Lin, and Raymond J. Carroll, Marginal Longitudinal Nonparametric Regression: Locality and Effeciency of Spline and Kernel Methods. *Journal of the American Statistical Association*, **97**, 2002, 482-493.

[16] Wu, Hulin and Jin-Ting Zhang, *Nonparametric Regression Methods for Longitudinal Data Analysis*, John Wiley & Sons, New Jersey, USA, 2006.