mathematics

MDPI

Article

Multicollinearity and Linear Predictor Link Function Problems in Regression Modelling of Longitudinal Data

Mozhgan Taavoni [1], Mohammad Arashi [1,*] and Samuel Manda [2]

[1] Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad 9177948974, Iran
[2] Department of Statistics, Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria 0028, South Africa
* Correspondence: arashi@um.ac.ir

Abstract: In the longitudinal data analysis we integrate flexible linear predictor link function and high-correlated predictor variables. Our approach uses B-splines for non-parametric part in the linear predictor component. A generalized estimation equation is used to estimate the parameters of the proposed model. We assess the performance of our proposed model using simulations and an application to an analysis of acquired immunodeficiency syndrome data set.

Keywords: generalized estimating equations; longitudinal data; multicollinearity; partially generalized linear models; ridge regression

MSC: 62J07; 62H12; 32H30

check for updates

Citation: Taavoni, M.; Arashi, M.; Manda, S. Multicollinearity and Linear Predictor Link Function Problems in Regression Modelling of Longitudinal Data. *Mathematics* **2023**, *11*, 530. https://doi.org/10.3390/math11030530

Academic Editor: Leonid V. Bogachev

Received: 1 October 2022
Revised: 13 December 2022
Accepted: 26 December 2022
Published: 18 January 2023

1. Introduction

Generalized linear models (GLMs) [1], which have a link function connecting the predictors linearly, are now part of regression models toolbox. Assuming a linear predictor link function could be very restrictive as the true relationship could be non-linear. Generalized partial linear models (GPLMs) accommodates both parametric and nonparametric connections. For example, in modelling longitudinal data, the GPLMs have proved useful as the model account for possible dependencies in the data [2–6].

One of the basic assumptions in the regression analysis is that all the explanatory variables are linearly independent. However two or more of the explanatory variables could be correlated with one another, resulting into a multicollinearity problem. Thus, it becomes harder to distinguish between effects of the independent variables on the outcome variable. It also results in the inflation of the variance of the regression parameter estimates. Ridge regression is widely used in regression model analyses with a large number of highly correlated independent variables [7–17]. While ridge regression techniques have widely been used in modelling cross-section data, there have been very few studies and applications in longitdinal data [18–22].

In this paper, we consider two typical problems name: of multicollinearity among predictor variables and linear predictor link function in the analysis of longitudinal non-normal data. For the former, we employ ridge regression and for the later we adopt the use of B-splines for nonparametric component of the linear predictor in an integrated approach. We concentrate on the estimation of population averaged model parameters. For this, we have the marginal mean model specification and account for the possible dependencies in the longitudinal data in a nonparametric manner through a convenient working within-subject covariance and employ the generalised estimation equations (GEE) for estimating the parameters.

In Section 2, we specify the underlying longitudinal model with the nonparametric part using splines. The estimation and asymptotic properties of the model parameters are also

presented in Section 2. Simulation studies and an application to acquired immunodeficiency syndrome data set are in Section 3. We conclude the paper in Section 4.

2. Model and Estimation Procedure

2.1. GPLMs for Longitudinal Data

Suppose we have n subjects and subject i has n_i observations denoted by y_{ij} ($i = 1, \ldots, n, j = 1, \ldots, n_i$) for a total of $N = \sum_{i=1}^{n} n_i$ observations. Also, X_{ij} be a vector of time-varying covariate. Thus, the total observed data set for the analysis is $\{(\boldsymbol{X}_{ij}, y_{ij}, t_{ij}), i = 1, \ldots, n, j = 1, \ldots, n_i\}$. Further let $\mathrm{E}(y_{ij}) = \mu_{ij}$ and $\mathrm{Var}(y_{ij}) = \phi v(\mu_{ij})$, where ϕ is a scale parameter and $v(.)$ is a known variance function. We model the longitudinal data with a GPLM, and specify a marginal model on the first two moments of y_{ij}. Especially, the marginal mean μ_{ij} is modeled as

$$\eta_{ij} = g(\mu_{ij}) = \boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta} + f(t_{ij}); \qquad \mu_{ij} = g^{-1}(\eta_{ij}), \tag{1}$$

where g(.) is a link function for the GLM, $\boldsymbol{\beta}$ is the regression coefficient vector with dimension p, and $f(.)$ is an unknown smooth function. We also assume independency between observations from different subjects. Finally, we assume t_{ij} are all scaled into the interval $[0, 1]$.

Similar to [3,23,24], we approximate the unspecified smooth function by the following polynomial spline

$$f(t_{ij}) = \alpha_0 + \alpha_1 t_{ij} + \ldots + \alpha_d t_{ij}^d + \sum_{l=1}^{L_n} \alpha_{(d+1)+l}(t_{ij} - t_i^{(l)})_+^d = \boldsymbol{B}^{\top}(t_{ij})\boldsymbol{\alpha},$$

where d is the degree of the polynomial component, L_n is the number of interior knots (rate of L_n will be specified in Remark 1), $t_i^{(l)}$ are knots of the ith subject, $\boldsymbol{B}(t_{ij}) = \left(1, t_{ij}, \ldots, t_{ij}^d, (t_{ij} - t_i^{(1)})_+^d, \ldots, (t_{ij} - t_i^{(L_n)})_+^d\right)$ is a $h_n \times 1$ vector of basis functions, h_n is the number of basis functions used to approximate $f(t_{ij})$, $h_n = d + 1 + L_n$, $(a)_+ = \max(0, a)$, and $\boldsymbol{\alpha}_n = (\alpha_0, \ldots, \alpha_d, \alpha_{d+1}, \ldots, \alpha_{d+1+L_n})^{\top}$ is the spline coefficients vector of dimension h_n. The nonparametric part in the linear prediction part is set as the basis functions with pseudo-design variables. In this way, the regression model problem in (1) could be linearised

$$\eta_{ij}(\boldsymbol{\theta}) = g(\mu_{ij}(\boldsymbol{\theta})) = \boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta} + \boldsymbol{B}(t_{ij})^{\top}\boldsymbol{\alpha} = \boldsymbol{D}_{ij}^{\top}\boldsymbol{\theta}. \tag{2}$$

$\boldsymbol{D}_{ij} = \left(\boldsymbol{X}_{ij}^{\top}, \boldsymbol{B}(t_{ij})^{\top}\right)^{\top}$ is a $(p + h_n) \times 1$ design matrix, which combines both the fixed and spline effects for the jth outcome of the ith subject. The combined regression coeffieicnts $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\top}, \boldsymbol{\alpha}^{\top})^{\top}$ has dimension $(p + h_n) \times 1$. Let $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^{\top}$, $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{in_i})^{\top}$, where $\mu_{ij} = g^{-1}(\boldsymbol{D}_{ij}^{\top}\boldsymbol{\theta})$, and $\boldsymbol{D}_i = (\boldsymbol{X}_i^{\top}, \boldsymbol{B}(\boldsymbol{t}_i)^{\top})^{\top}$. By the linear form of the GPLM in (1), using the spline approach, any computational algorithm developed for the GLM can be used for the GPLM.

Remark 1. *In spline smoothing, it is important to select the knots efficiently. Concentrating on the estimation of $\boldsymbol{\beta}$, Ref. [3] noticed that knot selection is more important for estimating $f(\cdot)$ rather than $\boldsymbol{\beta}$. Because in most of the studies, the focus is on $\boldsymbol{\beta}$ and providing sufficient statistical inference, and one only needs some basic information about $f(.)$. Therefore they particularly used the sample quantiles of $\{t_{ij}, i = 1, \ldots, n; k = 1, \ldots, n_i\}$ as knots. For instance, with three internal knots, we take three quartiles of the observed t_{ij}. Considering splines of order 4, they applied cubic splines with the integer part of $M^{1/5}$, the number of internal knots, where M is the number of distinct values in t_{ij}. Another study, Ref. [25], proposed that the number of distinct knots should increase with sample size to achieve asymptotic consistency. One must note that having too many knots increases the variance of estimators. Thus, the number of knots must be appropriately selected. When n goes to ∞,*

the number of knots should increase at $n^{1/(2m+1)}$. Thus, here, we use $L_n \approx n^{1/(2m+1)}$ with integer value m as the number of internal knots. We fix $m = 2$, and choose $L_n \approx n^{1/5}$, for asymptotic consistency. However, it is mainly based on practical experience and a desire for simplicity and is not an optimal choice. We considered a similar procedure in simulation and real data.

2.2. Ridge Generalized Estimating Equation (RGEE)

In most applications of GPLMs, the primary research interest is to make statistical inferences on the regression coefficient $\boldsymbol{\theta}$, along with and understanding of basic features of $f(t)$. For the ridge GEE procedure for $\boldsymbol{\beta}$, we briefly review the GEE method. For the estimating equation of $\boldsymbol{\theta}$, we have

$$n^{-1}\sum_{i=1}^{n}\frac{\partial \mu_{ij}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\top}}\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)=\mathbf{0}, \tag{3}$$

where $\boldsymbol{V}_i$ is a covariance matrix of $\boldsymbol{Y}_i$. For most applications, the actual intracluster covariance is regularly unknown. We take the working correlation matrix as $\boldsymbol{R}(\boldsymbol{\tau}) : \boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}(\boldsymbol{\theta})\boldsymbol{R}(\boldsymbol{\tau})\boldsymbol{A}_i^{1/2}(\boldsymbol{\theta})$, with the finite-dimensional parameter $\boldsymbol{\tau}$. Some commonly used working correlation structures include independence, autocorrelation (AR)-1, equally correlated (also called compound symmetry), or unstructured correlation. For a given working correlation structure, $\boldsymbol{\tau}$ can be estimated using the residual-based moment method. Here, similar to [26], the marginal density of Y_{ij} follows a canonical exponential family. Consequently, $\mu_{ij}(\boldsymbol{\theta}) = a(\theta_{ij})$ and $\sigma_{ij}^2(\boldsymbol{\theta}) = \phi a^{'}(\theta_{ij})$, where $\theta_{ij} = \boldsymbol{D}_{ij}^{\top}\boldsymbol{\theta}$, for a differentiable function $a()$ and a scaling constant ϕ. Assume $\widehat{\boldsymbol{R}}$ is the estimated working correlation matrix. Then, (3) simplifies to

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}_i^{\top}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})\widehat{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\theta})\big(\boldsymbol{Y}_i-\boldsymbol{\mu}_i(\boldsymbol{\theta})\big)=0. \tag{4}$$

We formally define the GEE estimator as the solution $\widehat{\boldsymbol{\beta}}$ of the above-estimating equations. For ease of exposition, we assume $\phi = 1$ in the rest of the article.

To account for multicollinearity in longitudinal data, we use the ridge GEE in the GPLM in (1) for parameter estimation. We do this by adding a shrinkage term $\lambda\boldsymbol{\beta}^{\top}\boldsymbol{\beta}$ to the objective function for handling correlated predictors. The ridge GEE has form

$$\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\theta}) - \lambda\boldsymbol{\beta}, \tag{5}$$

where

$$\boldsymbol{S}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}_i^{\top}\boldsymbol{A}_i^{1/2}(\boldsymbol{\theta})\widehat{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\theta})\big(\boldsymbol{Y}_i-\boldsymbol{\mu}_i(\boldsymbol{\theta})\big), \tag{6}$$

are the estimating functions defining the GEE. Here, λ is the tuning parameter that determines the shrinkage amount. The RGEE estimator $\widehat{\boldsymbol{\beta}}_R$ is the solution to $\boldsymbol{U}(\boldsymbol{\theta}) = \mathbf{0}$. We use the Newton–Raphson algorithm along with (6) to get the following iterative algorithm

$$\widehat{\boldsymbol{\theta}}^{k+1} = \widehat{\boldsymbol{\theta}}^{k} + \Big[\boldsymbol{H}(\widehat{\boldsymbol{\theta}}^{k}) + n\lambda\boldsymbol{E}(\widehat{\boldsymbol{\theta}}^{k})\Big]^{-1} \times \Big[\boldsymbol{S}(\widehat{\boldsymbol{\theta}}^{k}) - n\lambda\boldsymbol{E}(\widehat{\boldsymbol{\theta}}^{(k)}\widehat{\boldsymbol{\theta}}^{k})\Big].$$

Here, $\boldsymbol{H}(\widehat{\boldsymbol{\theta}}^{k}) = n^{-1}\sum_{i=1}^{n}\boldsymbol{D}_i^{\top}\boldsymbol{A}_i^{1/2}(\boldsymbol{\theta})\widehat{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\theta})\boldsymbol{D}_i$, $\boldsymbol{E}(\widehat{\boldsymbol{\theta}}^{k}) = \text{diag}\left\{\mathbf{1}_p, \mathbf{0}_{N_k}\right\}$. Further, $\mathbf{1}_p$ and $\mathbf{0}_{N_k}$ represent a vector of 1 with dimension p, and a zero vector of dimension N_k, respectively. The suggested estimation approach can be implemented step by step, and the detailed computation procedure can be summarized in Algorithm 1, describing the combination of ridge regression into the Newton–Raphson iterative algorithm of GEE.

With prespecified λ and initial value $\boldsymbol{\beta}$, the above algorithm is repeated to update $\widehat{\boldsymbol{\beta}}^{k+1}$ untill convergence.

Algorithm 1 Monte Carlo Newton–Raphson (MCNR) algorithm

Step 1. Approximate each predictor trajectory $f(t_{ij})$, by regression splines technique where described in Section 2.1. The smoothed predictor trajectories are then denoted as $f(t_{ij}) = \pi(t_{ij})^\top \boldsymbol{\alpha}$.

Step 2. Set $k = 0$. Choose initial values for parameter space $\boldsymbol{\theta}^0 = (\boldsymbol{\beta}^{0\top}, \boldsymbol{\alpha}^{0\top})^\top$ and correlation parameter $\boldsymbol{\tau}^0$ where determine covariance matrix $\boldsymbol{V}_i^0 = \boldsymbol{A}_i^{1/2}(\boldsymbol{\theta}^0)\boldsymbol{R}(\boldsymbol{\tau}^0)\boldsymbol{A}_i^{1/2}(\boldsymbol{\theta}^0)$. Use these values to find the RGEE estimates of $\boldsymbol{\theta}$. Specially,

(a) Compute $\boldsymbol{\theta}^{(k+1)}$ from the expression

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \Big[\boldsymbol{H}(\boldsymbol{\theta}^k) + n\lambda \mathrm{E}(\boldsymbol{\theta}^k)\Big]^{-1} \times \Big[\boldsymbol{S}(\boldsymbol{\theta}^k) - n\lambda \mathrm{E}(\boldsymbol{\theta}^{(k)}\boldsymbol{\theta}^k)\Big],$$

where

$$\begin{aligned} \boldsymbol{H}_n(\boldsymbol{\theta}^k) &= n^{-1}\sum_{i=1}^{n} \boldsymbol{D}_i^\top \boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}^k)\boldsymbol{R}^{-1}(\boldsymbol{\tau}^k)\boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}^k)\boldsymbol{D}_i, \\ \boldsymbol{S}_n(\boldsymbol{\theta}^k) &= \sum_{i=1}^{n} \boldsymbol{D}_i^\top \boldsymbol{A}_i^{\frac{1}{2}}(\boldsymbol{\theta}^k)\widehat{\boldsymbol{R}}^{-1}(\boldsymbol{\tau}^k)\boldsymbol{A}_i^{-\frac{1}{2}}(\boldsymbol{\theta}^k))(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta}^k)). \end{aligned}$$

(b) Compute $\boldsymbol{V}_i^{k+1}$ by $\boldsymbol{A}_i^{1/2}(\boldsymbol{\theta}^{k+1})\boldsymbol{R}(\boldsymbol{\tau}^{k+1})\boldsymbol{A}_i^{1/2}(\boldsymbol{\theta}^{k+1})$ where $\boldsymbol{\tau}^{k+1}$ For a given working correlation structure, $\boldsymbol{\tau}$ can be estimated using the residual-based moment method according to the prespecified working correlation structure. For more details refer to

(c) Set $k = k + 1$.

Step 3. Go to step 2 until convergence is achieved. Choose $\boldsymbol{\theta}^{k+1}$ and $\boldsymbol{V}_i^{k+1}$ to be the RGEE estimates of $\boldsymbol{\theta}$ and $\boldsymbol{V}_i$.

It is critical to choose a suitable value of tuning parameter λ to achieve satisfactory performance of the selection procedure. Many authors have introduced many methods for choosing an optimal tuning parameter within a given set of candidates. Traditional model selection criteria, such as AIC and BIC, have several limitations. The generalized cross validation (GCV) suggested by [27], later [28] proposed a Bayesian information criterion (BIC). How to choose λ for high-dimensional data discussed by [29]. They proposed a modified BIC. Further [30] extended the BIC information criterion. For the selection of the tuning parameter λ, here, we apply the suggested GCV of [27], given by

$$\mathrm{GCV}_\lambda = \frac{\frac{1}{n}\mathrm{RSS}(\lambda_n)}{\Big(1 - \frac{1}{n}d(\lambda_n)\Big)^2},$$

where

$$\mathrm{RSS}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \big(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\widehat{\boldsymbol{\theta}})\big)^\top \widehat{\boldsymbol{V}}_i^{-1}\big(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\widehat{\boldsymbol{\theta}})\big),$$

is the residual sum of squares, and effective number is equal to

$$d(\lambda) = \mathrm{tr}\Big[\Big\{\boldsymbol{H}(\widehat{\boldsymbol{\theta}}) + n\mathrm{E}(\widehat{\boldsymbol{\theta}})\Big\}^{-1} \times \boldsymbol{H}_n(\widehat{\boldsymbol{\theta}})\Big].$$

The optimal parameter denoted by λ_{opt} is the minimizer of the GCV_λ. In practical implementation, one can use PGEE package of R software, where function CVfit computes cross-validated tuning parameter value for longitudinal data. In numerical studies of the current paper, we used R codes similarly to compute λ_{opt}.

2.3. Asymptotics

We now discuss the asymptotic properties of the estimators $\boldsymbol{\beta}$ and $f(.)$ for the ridge GEE. Assuming (A.1)–(A.4) in Appendix A, the following theorems state the large sample property for $f(.)$ and $\boldsymbol{\beta}$, respectively. For the proofs, refer to Appendix A.

Theorem 1. *For $k_n = n^{1/(2m+1)}$ we have*

$$N^{-1}\sum_{i,j}\left(\widehat{f}(t_{ij}) - f(t_{ij})\right)^2 = O_p(n^{-\nu}),$$

where $\sum_{i,j} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}$ and $\nu = 2m/(2m+1)$.

Theorem 2. *Under the named regularity Conditions (A.1)–(A.4), $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} + b(\boldsymbol{\beta})\right)$ has the asymptotic p-variate normal distribution with the zero-vector mean and covariate matrix $\boldsymbol{V}$, where*

$$\begin{aligned} b(\boldsymbol{\beta}) &= -\lambda \boldsymbol{H}^{-1}(\boldsymbol{\beta})\boldsymbol{\beta}, \\ \boldsymbol{V} &= \boldsymbol{R}^{'}(\lambda)\boldsymbol{H}^{-1}(\boldsymbol{\beta})\boldsymbol{R}(\lambda), \\ \boldsymbol{H}(\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^{\top}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})\widehat{\boldsymbol{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta})\boldsymbol{X}_i, \\ \boldsymbol{R}(\lambda) &= \left(\mathbf{1}_p + \lambda \boldsymbol{H}^{-1}(\boldsymbol{\beta})\right)^{-1}. \end{aligned}$$

3. Numerical Analyses

3.1. Simulations

Here, we assess the performance of the GEE compared to its counterpart, the ridge GEE, for multiple correlated predictor variables. We generate the explanatory variables using

$$x_{ik} = (1-\gamma^2)^{\frac{1}{2}}\omega_{ik} + \gamma^2\omega_{ip}, \quad i = 1,\ldots,n = 100, \qquad k = 1,\ldots,p-1,\ p = 5, \tag{7}$$

where ω_{ik} are assumed to be independent and generated from a normal distribution with zero mean and unit variance. The parameter γ reflects the correlation such that any two explanatory variables correlate equally to γ^2. We generality, we consider $\gamma \in \{0.70, 0.80, 0.90, 0.99\}$. The nonparametric part of (1) has form $f(t_{ij}) = 2sin(2\pi t_{ij})$. In the entire process, $n_i = 4$ for each subject i and use the following GPLM for simulation

$$y_{ij} = x_{i,1,j}\beta_1 + x_{i,2,j}\beta_2 + x_{i,3,j}\beta_3 + x_{i,4,j}\beta_4 + x_{i,5,j}\beta_5 + f(t_{ij}) + \varepsilon_{ij}. \tag{8}$$

We set the true $\boldsymbol{\beta}$ as $\boldsymbol{\beta}^T = (0.5, 1, 1.5, 2, 0.1, 0.2)$. We generate t_{ij} from the uniform distribution over $(0,1)$. The ε_{ij} is generated from a normal distribution with zero mean, a common marginal variance $\sigma^2 = 1$. Moreover, the correlation structure is AR(1), i.e., $\text{corr}(\varepsilon_{is}, \varepsilon_{it}) = \rho^{|t-s|}$ for $s \neq t$, $\rho = 0.9$. Each simulated data set is fitted separately by the GEE approach of [26] and our ridge GEE using Algorithm 1. Then, 200 replications are run for each combination of ρ and γ.

To assess the behavior of both estimators encountering misspecified correlation structures, we conduct a comparison between $\widehat{\beta}_{GEE}$ and $\widehat{\beta}_{RGEE}$. We test the exchangeable working correlation structure (GEE-I) or (RGEE-I) when the true correlation structure is AR(1), (GEE-C), or (RGEE-C). For each of the estimators, we measure the accuracy in estimation using the mean squared error (MSE) given by $\text{MSE} = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\top}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. We recall the tuning parameter λ was obtained using the GCV, where λ_{opt} was the minimizer of the GCV_{λ}. Alternatively, one can use the ridge trace to find λ_{opt}. Figure 1 illustrates the ridge trace for the first generated random sample. As can be seen, the minimizer occurs at $K = 0.05$, which is the same value as the minimizer of GCV. The simulation results for MSE are presented in the latest column of Table 1. Table 1 reports the empirical biases and standard deviations (SDs) of the estimated β from the GEE and RGEE methods. We can

take the following observations: Our proposed RGEE has a superior performance in the MSE criterion. For misspecified correlation structures, the RGEE outperforms. However, it has more bias and offers a smaller SD in most cases compared to the GEE. By increasing the correlation among predictors, the increase in RGEE MSE is lesser than the GEE for all considered criteria. The conclusion is evident at the extreme level of correlation $\gamma = 0.99$.

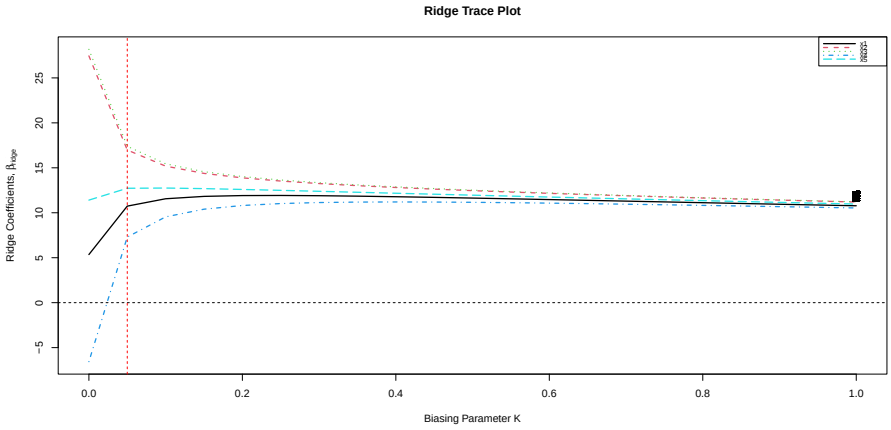

Figure 1. The ridge trace plot, for the first simulated data.

Table 1. Estimated regression coefficients for the important variables; bias (SD) based on 200 replications.

Methods	**Parameters**	$\gamma = 0.70$	$\gamma = 0.80$	$\gamma = 0.90$	$\gamma = 0.99$
RGEE-C	β_1	0.028(0.032)	0.033(0.038)	0.046(0.052)	0.141(0.142)
	β_2	0.024(0.027)	0.029(0.032)	0.039(0.044)	0.122(0.126)
	β_3	0.056(0.027)	0.067(0.032)	0.092(0.044)	0.284(0.127)
	β_4	0.055(0.029)	0.066(0.035)	0.090(0.048)	0.279(0.137)
	β_5	0.080(0.035)	0.097(0.045)	0.137(0.068)	0.446(0.209)
	MSE	0.243	0.291	0.404	1.271
RGEE-I	β_1	0.041(0.033)	0.049(0.039)	0.068(0.053)	0.209(0.150)
	β_2	0.057(0.029)	0.068(0.034)	0.093(0.047)	0.289(0.137)
	β_3	0.055(0.029)	0.065(0.034)	0.090(0.047)	0.278(0.137)
	β_4	0.061(0.031)	0.073(0.037)	0.101(0.051)	0.311(0.146)
	β_5	0.070(0.037)	0.076(0.048)	0.092(0.073)	0.249(0.236)
	MSE	0.284	0.331	0.444	1.335
GEE-C	β_1	0.027(0.055)	0.033(0.066)	0.045(0.090)	0.129(0.279)
	β_2	0.024(0.044)	0.028(0.052)	0.039(0.072)	0.113(0.223)
	β_3	0.058(0.051)	0.069(0.061)	0.096(0.084)	0.308(0.259)
	β_4	0.055(0.050)	0.065(0.060)	0.089(0.082)	0.244(0.254)
	β_5	0.082(0.053)	0.100(0.068)	0.143(0.103)	0.449(0.391)
	MSE	0.246	0.295	0.411	1.243
GEE-I	β_1	0.040(0.057)	0.048(0.068)	0.065(0.094)	0.184(0.289)
	β_2	0.055(0.048)	0.065(0.057)	0.088(0.078)	0.237(0.242)
	β_3	0.057(0.052)	0.069(0.062)	0.096(0.085)	0.320(0.262)
	β_4	0.061(0.056)	0.072(0.067)	0.099(0.092)	0.271(0.284)
	β_5	0.072(0.055)	0.079(0.071)	0.099(0.110)	0.262(0.423)
	MSE	0.286	0.333	0.447	1.275

3.2. AIDS Data Analysis

For illustration, in this section, the proposed model is used to analyze the CD4 cell data. From the number of 369 patients, 2376 CD4 measurements are recorded. The population's average time course of CD4 decay is regressed on the following covariates: packs per day for an indication of smoking; binary variable recreational drug use; SEXP as an indication of the number of sexual partners; and depression symptoms as measured by the CESD

scale (larger values indicate increased depressive symptoms). Similar to most literature, we take the square root of CD4 numbers. For the reason of the latter transformation, the reader is referred to [31,32]. For the correlation structure, we follow the approach of [31] and fit the compound symmetry covariance, where $\rho = 0.509$. We then used our proposed RGEE compared with the GEE for parameter estimation. We computed the standard errors (SDs) calculated using the bootstrap method. Table 2 provides the parameter estimates. From the result of this table, our proposed RGEE gives effectively smaller SD values compared to the GEE.

Table 2. Regression coefficient estimates (SD) in the analysis of the CD4 data.

Coefficients	Methods		Coefficients	Methods	
	RGEE	GEE		RGEE	GEE
AGE	3.987 (0.006)	4.298 (0.009)	AGE*CESD	−0.268 (0.008)	−0.262 (0.001)
SMOKE	32.780 (0.053)	32.916 (0.062)	SMOKE*DRUG	−16.204 (0.046)	−16.221 (0.055)
DRUG	17.949 (0.066)	18.254 (0.075)	SMOKE*SEXP	4.051 (0.002)	4.057 (0.005)
SEXP	2.801 (0.009)	2.797 (0.013)	SMOKE*CESD	−0.268 (0.003)	−0.251 (0.002)
CESD	−3.077 (0.002)	−3.077 (0.005)	DRUG*SEXP	−1.205 (0.005)	−1.292 (0.013)
AGE*SMOKE	0.039 (0.002)	−0.007 (0.003)	DRUG*CESD	0.274 (0.003)	0.273 (0.005)
AGE*DRUG	−1.006 (0.003)	−1.017 (0.009)	SEXP*CESD	0.033 (0.008)	0.026 (0.001)
AGE*SEXP	−0.565 (0.003)	−0.596 (0.001)			

4. Concluding Remarks and Discussion

We considered a generalized partially linear model (GPLM) and ridge regression to tackle the problems of multicollinearity and non-linearity in the relationship between the mean response and covariates in the longitudinal data analysis. The generalized estimation Equation (GEE) methods were used to estimate parameters in our proposed model. Using simulation studies, our methods resulted in smaller biases for estimating parameters than could have been obtained in a standard GEE. The performance of our proposed method decreased with increased dependencies between the model predictors. We also applied our model to a specific data set on AIDS data analysis.

Author Contributions: Conceptualization, M.T., M.A. and S.M.; Funding acquisition, M.A. and S.M.; Methodology, M.T. and M.A.; Software, M.T.; Supervision, M.A. and S.M.; Visualization, M.T., M.A. and S.M.; Formal analysis, M.T. and M.A.; Writing—original draft preparation, M.T.; Writing—review and editing, M.T., M.A. and S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was based upon research supported in part by the National Research Foundation (NRF) of South Africa, SARChI Research Chair UID: 71199, the South African DST-NRF-MRC SARChI Research Chair in Biostatistics (Grant No. 114613) and STATOMET at the Department of Statistics at the University of Pretoria, South Africa. The opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

Data Availability Statement: The data is publicly available.

Acknowledgments: We sincerely thank two anonymous reviewers for their constructive comments that significantly improved the presentation and led to putting many details in the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

To study the asymptotic properties of RGEE estimators, the following regularity conditions are required.

(A.1) Number of observations over time (n_i) is a bounded sequence of positive integers, and the distinct values of t_{ij} form a quasi-uniform sequence that grows dense on $[0, 1]$, and the kth derivative of $f(t_{ij})$ is bounded for some $k \geq 2$;

(A.2) The covariates $\boldsymbol{X}_{ij}$, $1 \leq i \leq n$, $1 \leq j \leq m$ are uniformly bounded;

(A.3) The unknown parameter $\boldsymbol{\beta}$ belongs to a compact subset $\mathcal{B} \subseteq \mathcal{R}^p$, the true parameter value lies in the interior of $\mathcal{B}$;

(A.4) There exist two positive constants, b_1 and b_2, such that

$$b_1 \leq \lambda_{\min}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^{\top}\boldsymbol{X}_i\right) \leq \lambda_{\max}\left(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^{\top}\boldsymbol{X}_i\right) \leq b_2,$$

where $\lambda_{\min}$(resp. $\lambda_{\max}$) denotes the minimum (resp. maximum) eigenvalue of a matrix.

To verify Theorem 1, we need the following lemma.

Lemma A1. *Under Condition (A.1), there exists a constant C depending only on k_n such that*

$$\sup_{t\in[0,1]} |f(t) - \pi(t)\boldsymbol{\alpha}| \leq Ck_n^{-m}.$$

The proof of this lemma follows readily from Theorem 12.7 of Schumaker [33].

Proof of Theorem 1. By Lemma A1, we approximate $f(t_{ij})$ by $\pi(t)\boldsymbol{\alpha}$, then by choosing $k_n \approx n^{1/(2m+1)}$ we have

$$\begin{aligned}\left(\widehat{f}(t) - f(t)\right)^2 &= |\widehat{f}(t) - f(t)||\widehat{f}(t) - f(t)| \\ &\leq \sup_{t\in[0,1]}|\widehat{f}(t) - f(t)| \sup_{t\in[0,1]}|\widehat{f}(t) - f(t)| \\ &\leq Ck_n^{-m}Ck_n^{-m} = C^2n^{-2m/(2m+1)} = O_p(n^{-2m/(2m+1)}),\end{aligned}$$

which proves Theorem 1. □

Proof of Theorem 2. To proof Theorem 2 define the linear operator $\boldsymbol{R}(\lambda) = \left(\mathbf{1}_p + \lambda\boldsymbol{H}^{-1}(\boldsymbol{\beta})\right)^{-1}$. It is straightforward to calculate that the ridge estimator $\widehat{\boldsymbol{\beta}}_{RGEE}$ can be expressed as $\boldsymbol{R}(\lambda)\widehat{\boldsymbol{\beta}}_{GEE}$ where $\widehat{\boldsymbol{\beta}}_{GEE}$ is ordinary GEE estimator. The expectation of the ridge estimator can be expressed as

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}_{RGEE}) = \mathrm{E}(\boldsymbol{R}(\lambda)\widehat{\boldsymbol{\beta}}_{GEE}) = \boldsymbol{\beta} - \lambda\boldsymbol{H}^{-1}(\boldsymbol{\beta})\boldsymbol{\beta}.$$

Clearly, $\mathrm{E}(\widehat{\boldsymbol{\beta}}_{RGEE} - \boldsymbol{\beta}) = -\lambda\boldsymbol{H}^{-1}(\boldsymbol{\beta})\boldsymbol{\beta} \neq 0$ for any $\lambda > 0$. Hence, the ridge estimator is biased with $-\lambda\boldsymbol{H}^{-1}(\boldsymbol{\beta})\boldsymbol{\beta} = b(\boldsymbol{\beta})$. The variance of the RGEE estimator is straightforwardly obtained when exploiting its linearly relation with the GEE estimator. Then,

$$\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}_{RGEE}\right) = \mathrm{Var}\left(R(\lambda)\widehat{\boldsymbol{\beta}}_{GEE}\right) = R^{'}(\lambda)\,\mathrm{Var}\left(\widehat{\beta}_{GEE}\right)R(\lambda) = R^{'}(\lambda)\boldsymbol{H}^{-1}(\boldsymbol{\beta})R(\lambda),$$

where $R^{'}(\lambda)\boldsymbol{H}^{-1}(\boldsymbol{\beta})R(\lambda) = \boldsymbol{V}$. Combining the expectation and variance terms, the proof is complete. □

References

1. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman and Hall: London, UK, 1989.
2. He, X.; Zhu, Z.Y.; Fung, W.K. Estimation in a Semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **2002**, *89*, 579–590. [CrossRef]
3. He, X.M.; Fung, W.K.; Zhu, Z.Y. Robust estimation in a generalized partially linear model for cluster data. *J. Am. Stat. Assoc.* **2005**, *100*, 1176–1184. [CrossRef]
4. Qin, G.; Bai, Y.; Zhu, Z. Robust empirical likelihood inference for generalized partial linear models with longitudinal data. *J. Multivar. Anal.* **2012**, *105*, 32–44. [CrossRef]

5. Chen, B.; Zhou, X.H. Generalized partially linear models for incomplete longitudinal data in the presence of population-Level information. *Biometrics* **2013**, *69*, 386–395. [CrossRef]
6. Zhang, J.; Xue, L. Empirical likelihood inference for generalized partially linear models with longitudinal data. *Open J. Stat.* **2020**, *10*, 188–202. [CrossRef]
7. Hoerl, A.; Kennard, R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
8. Hoerl, A.; Kennard, R. Ridge regression: Application to nonorthogonal problems. *Technometrics* **1970**, *12*, 69–82. [CrossRef]
9. Theobald, C.M. Generalization of mean squer error applied to ridge regresion. *J. R. Stat. Soc.* **1974**, *36*, 103–106.
10. Tikhonov, A. On the stability of inverse problems. *Proc. USSR Acad. Sci.* **1943**, *39*, 267–288.
11. Saleh, A.K.M.; Kibria, B.M.G. Performances of some new preliminary test ridge regression estimators and their properties. *Commun. Stat.—Theory Methods* **1993**, *22*, 2747–2764. [CrossRef]
12. Kibria, B.M.G.; Saleh, A.K.M.E. Effect of W,LR and LM tests on the performance of preliminary test ridge regression estimators. *J. Jpn. Stat. Soc.* **2003**, *33*, 119–136. [CrossRef]
13. Kibria, B.M.G.; Saleh, A.K.M.E. Preliminary test ridge regression estimators with student's /errors and conflicting test-statistics. *Metrika* **2004**, *59*, 105–124. [CrossRef]
14. Arashi, M.; Tabatabaey, S.M.M.; Iranmanesh, A. Improved estimation in stochastic linear models under elliptical symmetry. *J. Appl. Probab. Stat.* **2010**, *5*, 145–160.
15. Bashtian, H.M.; Arashi, M.; Tabatabaey, S.M.M. Using improved estimation strategies to combat multicollinearity. *J. Stat. Comput. Simul.* **2011**, *81*, 1773–1797. [CrossRef]
16. Bashtian, H.M.; Arashi, M.; Tabatabaey, S.M.M. Ridge estimation under the stochastic restriction. *Commun. Stat.—Theory Methods* **2011**, *40*, 3711–3747. [CrossRef]
17. Arashi, M.; Tabatabaey, S.M.M.; Soleimani, H. Simple regression in view of elliptical models. *Linear Algebra Its Appl.* **2012**, *437*, 1675–1691. [CrossRef]
18. Zhang, B.; Horvath, S. Ridge regression based hybrid genetic algorithms for multi-locus quantitative trait mapping. *Bioinform. Res. Appl.* **2006**, *1*, 261–272. [CrossRef]
19. Malo, N.; Libiger, O.; Schork, N. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am. J. Hum. Genet.* **2008**, *82*, 375–385. [CrossRef]
20. Eliot, M.; Ferguson, J.; Reilly, M.P.; Foulkes, A.S. Ridge regression for longitudinal biomarker data. *Int. J. Biostat.* **2011**, *7*, 37. [CrossRef]
21. Rahmani, M.; Arashi, M.; Mamode Khan, N.; Sunecher, Y. Improved mixed model for longitudinal data analysis using shrinkage method. *Math. Sci.* **2018**, *12*, 305–312. [CrossRef]
22. Taavoni, M.; Arashi, M. Semiparametric ridge regression for longitudinal data. In Proceedings of the 14th Iranian Statistics Conference, Shahrood University of Technology, Shahrood, Iran, 25–27 August 2018.
23. Qin, G.Y.; Zhu, Z.Y. Robustified maximum likelihood estimation in generalized partial linear mixed model for longitudinal data. *Biometrics* **2009**, *65*, 52–59. [CrossRef] [PubMed]
24. Taavoni, M.; Arashi, M. High-dimensional generalized semiparametric model for longitudinal data. *Statistics* **2021**, *55*, 831–850. [CrossRef]
25. Qin, G.Y.; Zhu, Z.Y. Robust estimation in generalized semiparametric mixed models for longitudinal data. *J. Multivar. Anal.* **2007**, *98*, 1658–1683. [CrossRef]
26. Liang, K.Y.; Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–22. [CrossRef]
27. Fan, J.Q.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
28. Wang, H.S.; Li, R.Z.; Tcai, C.L. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **2007**, *94*, 553–568. [CrossRef] [PubMed]
29. Wang, H.S.; Li, B.; Leng, C.L. Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 671–683. [CrossRef]
30. Li, G.R.; Peng, H.; Zhu, L.X. Nonconcave penalized M-estimation with a diverging number of parameters. *Stat. Sin.* **2011**, *21*, 391–419.
31. Zeger, S.L.; Diggle, P.J. Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **1994**, *50*, 689–699. [CrossRef]
32. Wang, N.; Carroll, R.; Lin, X.H. Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Am. Stat. Assoc.* **2005**, *100*, 147–157. [CrossRef]
33. Schumaker, L.L. *Spline Functions*; Wiley: New York, NY, USA, 1981.