Performance of Generalized Estimating Equations with Sparse Binary Outcomes and Multicollinearity Using Firth and SCAD Penalizations

Ellen Zhu[1,*], Qi Wang[1,*], Charlene Tang[1,*], Monica Ramsey[1,*] and Awan Afiaz[2,*]

[*]Group 1. Authors are listed in reverse alphabetical order

[1]Department of Statistics, University of Washington

[2]Department of Biostatistics, University of Washington

Abstract

Sparsity in binary outcomes and collinear regressors are common problems in longitudinal data analysis. The standard GEE fails to work when both such problems are present simultaneously, and several modified versions of the GEE have been proposed to address these shortcomings, however, no existing method accounts for the case where both sparsity and multicollinearity occur. In our work, we propose a double-penalized version of GEE, named P2GEE, combining both the Firth-type and SCAD penalties to jointly address sparsity and multicollinearity in binary longitudinal data. Through extensive simulation studies, we evaluate the performance of P2GEE in reducing bias in the estimates and standard errors of the regression coefficients. Our results show that the proposed P2GEE outperforms the standard GEE approach when the binary outcome is sparse and multicollinearity in regressors is present. To demonstrate the use of P2GEE within a real-data setting, we have applied our proposed method to the Framingham Heart Study. We conclude from our study that the P2GEE works very well when the data suffers from the two aforementioned problems, and note that a future direction for this research would be to improve on the standard sandwich variance estimator that works better in conjunction with P2GEE.

1 Introduction

Analyzing longitudinal binary data is common in biomedical research studies. In such settings, a binary outcome variable of interest is observed at multiple time points on all the subjects in the study along with a vector of covariates that is hypothesized to be associated with that outcome. For example, in a clinical trial, multiple subjects may be observed repeatedly during the study on some disease status (1 = diseased, 0 = non-diseased) as well as other important covariates regularly. These repeated measurements belonging to the same subject are typically correlated, which necessitates sophisticated statistical methods that take the within-subject correlation into account. Molenberghs et al. (2005) discussed three families of models to analyze such data which include marginal, conditional, and random effects models. Liang and Zeger (1986) proposed the Generalized Estimating Equations (GEE) as an extension to the generalized linear models (GLMs) to model the marginal mean of longitudinal outcome variables. GEE circumvents the need to specify the joint distribution of repeated measurements and utilizes the quasi-likelihood principle. This framework facilitates only specifying a working correlation structure that reflects the assumed correlation pattern within the data. Furthermore, GEE allows model robust estimation of the covariance matrix using a sandwich estimator.

GEE models the marginal mean of a longitudinal discrete outcome and provides regression coefficients that have population-average interpretation in comparison to the subject-specific interpretation provided by the Generalized Linear Mixed Models (GLMM). The main advantage of GEE is that if the sample is large enough and if the subjects/clusters are independent between themselves, the sandwich variance estimator produces consistent and unbiased standard errors (SEs) for the estimated regression coefficients, even when the covariance structure is incorrectly specified Liang and Zeger (1986). However, the limitations of GEE become apparent in small sample settings where the method's ability to generate consistent estimates of the regression coefficients is compromised and the corresponding SEs are inefficient (Paul and Zhang, 2014). This issue is further exacerbated by the presence of separation or sparsity within the binary longitudinal outcome, which occurs when specific outcomes (e.g., disease presence) exhibit highly infrequent events within subjects (Mondol and Rahman, 2019).

1.1 Separation problem in binary outcomes

Albert and Anderson (1984) first noted the problem of separation in binary outcomes when studying the existence of maximum likelihood estimates (MLEs) in logistic regression models. They observed that when a linear combination of the explanatory variables perfectly predicts the observed outcomes, then some of the regression coefficients do not exist and their estimates diverge during the iterative fitting process. Heinze (2006) defined separation as the case when a binary or continuous predictor or a linear combination of predictors perfectly separates (or predicts) events from the non-events. For logistic regression, although the log-likelihood may converge to some finite value, it cannot be maximized by a finite parameter value leading to infinite or zero maximum likelihood estimates. Based on Mondol and Rahman (2019) and Greenland et al. (2016), there exist three types of separation: complete, quasi-complete, and near separation (also called sparsity). Complete separation, which is also known as perfect prediction, is an extreme case of separation. As an illustrative example, consider a study where we have binary outcome variable y (0 = failure, 1 = success) and two predictors, x_1 and x_2, as presented in Table 1.

Table 1: Example of complete separation

x_1	1	2	3	3	5	6	10	11
x_2	3	2	-1	-1	2	4	1	0
y	0	0	0	0	1	1	1	1

Notice that in Table 1, all observations with $y = 0$ have values of $x_1 \leqslant 3$ and all observations with $y = 1$ have values of $x_1 > 3$. In other words, y separates x_1 perfectly. Alternatively, we can see that x_1 predicts y perfectly, since $x_1 \leqslant 3$ corresponds to $y = 0$ and $x_1 > 3$ corresponds to $y = 1$. If we use a cut point of $\leqslant 3$ to transform x_1 into a binary predictor, we would obtain y exactly; that is, we would find a perfect predictor x_1 for the outcome variable y. With respect to predicted probabilities, we can calculate $\Pr(y = 1|x_1 \leqslant 3) = 0$ and $\Pr(y = 1|x_1 > 3) = 1$ without the use of a model. In the case of 2×2 contingency tables, where we have a binary exposure and a binary outcome variable, complete separation would look like Case I of Table 2.

Table 2: Examples of separation types as shown by Mondol and Rahman (2019)

I. Complete separation			II. Quasi-complete separation			III. Near separation (sparsity)		
	Y			Y			Y	
	0	1		0	1		0	1
Non-exposed	20	0	Non-exposed	12	8	Non-exposed	12	8
Exposed	0	20	Exposed	0	20	Exposed	2	18

Quasi-complete or near-separation occurs when the outcome variable separates a predictor variable or a linear combination of predictor variables almost completely. For a 2×2 contingency table, quasi-complete separation means that one of the four cells has a zero value (see case II of table 2). Lastly, Mondol and Rahman (2019) defined the ’near-to-quasi-complete separation’ in such cases where there is at least one key covariate pattern (or, cell) with very few observations. This phenomenon was also termed as ’sparsity’ by Greenland et al. (2016) (see case III of table 2). Following Mondol and Rahman (2019), we defined near-separation as the case where there are non-zero cells with cell count being less than 15% of the total observations.

Problems of separation usually occur in small samples or in cases where there is an extreme split on the dependent variable based on one or more of the covariates. Moreover, it can also occur in case of rare outcomes (Allison, 2008). For example, consider the case of a logistic regression where the outcome variable is whether a person has some disease. Assume that the overall prevalence is less than 1 in 1000 and also that the explanatory variables include a set of five indicator variables representing five different categories. Even with a sample size of 10,000 people, it is reasonable to expect that no one would have the disease for at least one of the indicator variables. For correlated binary data, the intra-cluster correlation exerts influence over the responses of a cluster over time, and thus can also cause separation in the data.

1.2 Multicollinearity in longitudinal data

Multicollinearity refers to the linear relationship among two or more predictor variables, which also means a lack of orthogonality between them. As mentioned in a review by Chan et al. (2022), multicollinearity presents two primary challenges: first, the instability of estimates caused by the variables' interdependence, coupled with large standard errors of the regression coefficients, heavily undermines the reliability and precision of these estimates. Second, when two or more variables exhibit linear relationships, accurately determining the individual impact of a variable becomes difficult. Consequently, the model's ability to generalize weakens, leading to over-fitting, and thus the model performs inadequately on new, unseen data.

The presence of multicollinearity in binary logistic regression models, as well as similar regression models, can lead to increased standard errors for the model's coefficients, while the predictors themselves might remain largely unaffected. This results in decreased stability of the estimated parameters, which compromises their reliability ((Midi et al., 2013) King (2008)EF et al. (2017)). Hosmer Jr et al. (2013) expanded on this by demonstrating that multicollinearity may influence not only the standard errors, but also the estimated parameters. These findings suggest that the impact of multicollinearity might vary across different scenarios, where in one case it may inflate error terms for coefficients, and in another, it may alter parameter estimates.

In the context of longitudinal data analysis, researchers collect repeated measurements of certain variables over a defined period for various subjects or individuals. As mentioned in Rahmani et al. (2018), such types of data are commonly encountered in medical research where the responses are subject to various time-dependent and time-constant effects such as pre- and post-treatment types, gender effect, baseline measures, and many biomarkers. It is quite natural that these variables may exhibit some form of dependence between them. In practice, the primary objective of longitudinal studies is to assess the impact and significance of different factors. In this realm, Fitzmaurice and Laird (1993) and Sutradhar et al. (2014), have introduced several estimation methods grounded in likelihood and pseudo-likelihood techniques aimed at quantifying regression effects. However, the reliability of these estimators comes into question, especially when faced with multicollinearity among the predictors, as discussed by Eliot et al. (2011), Hossain et al. (2018), and Saleh et al. (2014).

1.3 Sparse outcomes and multicollinearity in longitudinal data

Sparsity and multicollinearity can indeed occur together in datasets, particularly in high-dimensional settings where the number of variables (p) is large compared to the number of observations (n), often referred to as the $p >> n$ scenario. This situation becomes particularly pronounced in the analysis of rare events, where the outcome variable exhibits sparsity, and is common in many fields of research such as genetic studies. For example, when researching a rare genetic disorder or phenotype, the dataset might consist of a large number of potential explanatory genetic features (such as gene expressions or mutations) against a backdrop of very few cases (observations) that exhibit the rare condition. Here, sparsity in the outcome variable means that the event of interest (such as the presence of a rare genetic disorder) is uncommon, resulting in a disproportionate number of zeros (no occurrence) compared to ones (occurrence). Simultaneously, multicollinearity can emerge within the explanatory variables due to biological or genetic interrelations, where certain genes or mutations are highly correlated with others because they participate in the same biological processes or pathways. This duality—sparse outcomes and multicollinearity among explanatory variables—complicates statistical modeling and analysis. The challenge lies in accurately identifying the few relevant predictors that are associated with the outcome of interest while navigating the intricate correlations among these predictors that can obscure their individual effects.

1.4 Literature review

1.4.1 Development of methods to address sparsity

Multiple studies have been conducted to investigate and improve the performance of GEE under small-sample scenarios. Paul and Zhang (2014) proposed a bias-corrective method (GEEBc) based on Cox and Snell (1968) that adds (or subtracts) a correction to the estimators and a bias-preventive method (GEEBr) based on Firth (1993) that introduces a bias term into the score function to alleviate small-sample bias and increase efficiency of the GEE. Lunardon and Scharfstein (2017) showed that the strategy proposed by Paul and Zhang (2014) is dependent on the correct specification of the working correlation matrix (WCM) and suggested a revised formulation, called the bias-corrected GEE (BCGEE), which remains valid under misspecification the working covariance matrix. However, none of these methods considered the problem of separation or sparsity in their methods. Mondol and Rahman (2019) proposed a penalized GEE (PGEE) method for longitudinal binary data to reduce the bias and account for the separation in binary outcomes by adding a Firth-type [1993] penalty to the GEE score equation that showed promising performance in the presence of separation, especially near-separation. It yielded very high convergence rates in the presence of complete or quasi-separation providing superior estimates, compared to the classical GEE which either failed to converge for complete/quasi-complete separation or provided a very large estimate in the case of near separation. It is important to note that the regression coefficients and their SE estimates were still moderately biased for the complete/quasi-complete separation scenarios using PGEE. However, in the presence of near-separation or sparsity, PGEE was shown to be superior in obtaining unbiased and consistent estimates of the regression coefficients. Motivated by the equivalence of FL and maximum likelihood estimation with iteratively augmented data, a recent study by Geroldinger et al. (2022) proposed two new extensions of Firth logistic (FL) to GEE, called the 'fully iterated' and 'single-step' augmented GEE (augGEE). They found that PGEE either matched the augGEE or often slightly outperformed the augGEE approaches, but had a higher burden of implementation. Gosho et al. (2023) conducted a comparative study between GEE, BCGEE, and PGEE in small-sample sparse longitudinal binary data settings and found that PGEE outperformed BCEE across all settings in correcting the bias caused by sparsity.

1.4.2 Development of methods to address multicollinearity

Historically, multicollinearity has been extensively examined in the realm of linear models, and various penalized regression techniques have been developed to mitigate its effects. Adopting a penalty function becomes beneficial when dealing with correlations among the covariates, as it helps to sidestep computational challenges and enhances the prediction accuracy and variable selection effectiveness of the model. These include the ridge estimator, which imposes a squared magnitude penalty; the Stein estimator, known for shrinking estimates; the bridge estimator, which allows for continuous shrinkage; and the Least Adaptive Shrinkage and Selection Operator (LASSO), which promotes sparsity (in the regressors) by penalizing the absolute size of the coefficients ((Tibshirani, 1996)). The Lasso, in particular, has proven to be versatile, finding utility in diverse statistical models such as linear regression, logistic regression, Cox proportional hazards models, and even within the architecture of neural networks.

However, a notable drawback of LASSO is the significant bias that can result from its shrinkage effect. To mitigate this, the Smoothly Clipped Absolute Deviation penalty (SCAD), (Fan and Li, 2001b) modified the LASSO approach by applying a variable penalization rate based on coefficient size. This means while smaller coefficients are penalized similarly to LASSO, larger coefficients are virtually unaffected by the penalty, thereby avoiding the introduction of unnecessary bias.

Building on the foundational understanding of addressing multicollinearity with various penalization techniques, several studies have extended these methods to apply penalties within Generalized Estimating Equations (GEE) for managing multicollinearity in longitudinal data. For instance, Fu (2003) introduced the bridge penalty to tackle collinearity issues in longitudinal studies through this model framework. Similarly, Dziak (2006) incorporated penalties like LASSO and SCAD into GEE for longitudinal data, offering a penalized approach to enhance model robustness.

In our work, we opt for the SCAD (Smoothly Clipped Absolute Deviation) penalty as the preferred penalty term to address multicollinearity in longitudinal data. As mentioned in Dziak (2006), the SCAD penalty has demonstrated superior capability in identifying models that are both parsimonious and sufficiently comprehensive. Unlike LASSO, which tends to either oversimplify or not adequately penalize complexity, SCAD navigates the balance between simplicity and adequacy more effectively. Moreover, SCAD's performance remains robust in high-dimensional settings where the number of variables significantly exceeds the number of observations. In such scenarios, exhaustive search methods become impractical due to the exponential growth in the number of possible model subsets. SCAD, with its reliance on a tuning parameter and an efficient optimization algorithm, offers a quicker and more stable alternative, making it particularly advantageous. Dziak (2006) also found that SCAD demonstrated its flexibility and empirical success in various simulations and asymptotic studies. Although no single model selection method outperformed all others across every scenario, SCAD consistently ranked highly, especially when fine-tuned with criteria like light BIC. This empirical evidence supports SCAD's effectiveness in diverse conditions and its ability to produce models that accurately reflect underlying data structures.

1.5 Study objectives and structure

This paper revisits penalization within the context of Generalized Estimating Equations (GEE) and delves into the efficacy of the Firth and SCAD penalty functions in addressing both sparsity and multicollinearity jointly within the context of binary longitudinal data. Our goal is to discern the most effective approach in preserving the integrity and interpretability of longitudinal data analyses when the data are plagued by sparse outcomes and collinear regressors. To that end, we attempt to develop a bias-reduced sparsity- and multicollinearity-proof penalized GEE method, which we call P2GEE, and provide empirical evidence in favor of this method.

The first chapter provides an introduction to the pertinent notions, definitions, overview, and background of the research study undertaken. It also contains an extensive literature review of the extant developments made in this area to identify the literature gap and finally discusses the general and specific study objectives. The second chapter discusses the methodology of the study in detail including discussions on the estimators their strengths. The third chapter chapter provides a detailed account of the extensive simulation study undertaken for this research by specifying the simulation design, simulation scenarios, and an in-depth look into the the subsequent results. The final chapter provides the discussion of the implications of the results obtained in the simulation study, the strengths and limits and finally concludes to provide recommendations to practitioners in the field.

2 Methodology

2.1 Standard GEE model for longitudinal data

In a longitudinal dataset of m independent subjects, we assume there are n_i observations over time for the ith subject. Let Y_{ij} and $\mathbf{X}_{ij}$ denote the outcome and covariates observed for the ith subject at time j, respectively, where $\mathbf{X}_{ij}$ is a $p \times 1$ vector. Liang and Zeger (1986) assumed that $\mu_{ij} = \mathbb{E}(Y_{ij}|\mathbf{X}_{ij}) = g^{-1}(\mathbf{X}_{ij}^\top \boldsymbol{\beta})$ and $\mathrm{Var}(Y_{ij}|\mathbf{X}_{ij}) = \phi v(\mu_{ij})$, where $g(\cdot)$ is a link function, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, ϕ is a scale parameter, and $v(\cdot)$ is a variance function. We assume a working correlation matrix of $\mathbf{R}_i(\boldsymbol{\alpha})$, which is parameterized by the parameter $\boldsymbol{\alpha}$. Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^\top$, $\mathbf{X}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{in_i})^\top$, $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{in_i})^\top$, $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$, $\mathbf{A}_i = \mathrm{diag}(\phi v(\mu_{ij}))$, and $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$. Then the regression parameter $\boldsymbol{\beta}$ can be estimated by solving the estimating equation

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{m} \mathbf{D}_i^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0, \tag{1}$$

where we replace ϕ and $\boldsymbol{\alpha}$ with their $\sqrt{m}$-consistent estimators. Moreover, (Liang and Zeger, 1986) showed that $\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically normal with mean 0 and covariance Σ, where

$$\Sigma = \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{D}_i^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)(\mathbf{Y}_i - \boldsymbol{\mu}_i)^\top \mathbf{V}_i^{-1} \mathbf{D}_i \right) \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1}.$$

This implies that when m goes to infinity,

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1} \left(\sum_{i=1}^{m} \mathbf{D}_i^\top \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)(\mathbf{Y}_i - \boldsymbol{\mu}_i)^\top \mathbf{V}_i^{-1} \mathbf{D}_i \right) \left(\sum_{i=1}^{m} \mathbf{D}_i^\top \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1}. \tag{2}$$

The consistent sandwich variance estimator for $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ can be obtained by plugging in consistent estimates for the parameters in (2).

In their seminal work, Liang and Zeger (1986) proved that the consistency of $\hat{\boldsymbol{\beta}}$ and $\hat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})$ only relied on the correct specification of the marginal mean μ_{ij} and did not depend on the working correlation matrix. However, these estimators are biased when the outcome is sparse, and the variance estimation is problematic when there are sparse outcomes or multicollinearity, or when both sparsity and multicollinearity are present. We will discuss two modified versions of GEE in Sections 2.2 and 2.3, which address the biases induced by sparsity and multicollinearity, respectively.

2.2 Penalized GEE for sparse longitudinal data

Mondol and Rahman (2019) proposed a penalized GEE method to overcome the bias in the parameter and variance estimation of the standard GEE in the presence of separation. Specifically, Mondol and Rahman (2019) added a Firth-type penalty to the standard estimating equation (1). Then, the estimates for regression coefficients $\hat{\boldsymbol{\beta}}_{FP}$ are obtained by solving this modified estimating equation, i.e.,

$$U_k^{\mathrm{F}} = U_k + \frac{1}{2}\mathrm{trace}\left(\mathbf{I}^{-1}\frac{\partial \mathbf{I}}{\partial \beta_k}\right) = 0,$$

where U_k is the kth element of $U(\boldsymbol{\beta})$ and $\mathbf{I}$ is the Fisher information matrix. Let X_{kij} be the kth element of $\mathbf{X}_{ij}$. The derivative of the Fisher information matrix is given by

$$\frac{\partial \mathbf{I}}{\partial \beta_k} = \sum_{i=1}^{m}\left(\mathbf{D}_i^{\top}\mathbf{Q}_{ik}\mathbf{V}_i^{-1}\mathbf{D}_i + \mathbf{D}_i^{\top}\mathbf{V}_i^{-1}\mathbf{Q}_{ik}\mathbf{D}_i\right),$$

where $\mathbf{Q}_{ik} = \mathrm{diag}(X_{ki1}(1/2-\mu_{i1}, \ldots, X_{kin_i}(1/2-\mu_{in_i}).$

We can derive the sandwich covariance estimator for $\hat{\boldsymbol{\beta}}_F$ by substituting $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}_F$ in (2), the sandwich estimator of standard GEE.

2.3 Penalized GEE for longitudinal data with multicollinearity

Wang et al. (2012) considered adding the non-convex SCAD penalty to the estimating equations of GEE. Specifically, we obtain the estimates of regression coefficients by solving

$$U_k^{\mathrm{SCAD}} = U_k - q_\lambda(|\beta_k|)\mathrm{sign}(\beta_k) = 0,$$

where $\mathrm{sign}(\tau) = I(\tau > 0) - I(\tau \leqslant 0)$ and $q_\lambda(\tau) = \lambda\left\{I(\tau \leqslant \lambda) + \frac{(a\lambda-\tau)_+}{(a-1)\lambda}I(\tau > \lambda)\right\}$. Wang et al. (2012) also derived a sandwich estimator for the covariance of $\hat{\boldsymbol{\beta}}_{SCAD}$.

The SCAD penalty can also potentially address the multicollinearity problem by constraining the coefficient norm and forcing some coefficient values to 0 (Wang et al., 2012; Blommaert et al., 2014).

2.4 Double-penalized GEE for longitudinal data with sparsity and multicollinearity (P2GEE)

When we have both sparse binary outcomes and multicollinearity in the regressors in the context of longitudinal data, we propose applying both Firth-type and SCAD penalties to the GEE. With this modification, the estimating equation is given by

$$U_k^{P2GEE} = U_k + \frac{1}{2}\text{trace}\left(\mathbf{I}^{-1}\frac{\partial \mathbf{I}}{\partial \beta_k}\right) - q_\lambda(|\beta_k|)\text{sign}(\beta_k) = 0.$$

We estimate the covariance of the estimator for regression coefficients $\hat{\boldsymbol{\beta}}_{F,SCAD}$ by replacing $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}_{F,SCAD}$ in the standard sandwich estimator (2).

3 Simulation study

To empirically investigate the performance of the proposed P2GEE method for binary longitudinal data in which sparsity (in outcome) and multicollinearity (in the regressors) are present, we conduct an extensive simulation study which we discuss in the following subsections.

3.1 Data generation

For our simulation study, we consider correlated binary outcome variable Y_{ij}, the associated vector of covariates X_{ij}, marginal mean μ_{ij}, and correlation matrix $\mathbf{R}_i(\boldsymbol{\alpha})$ as defined in Section 2.1. The observations $\{X_j, Y_j\}_{j=1}^n$ are grouped into m clusters of size n, and, following the framework of fLiang and Zeger (1986), the marginal mean is given by:

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij},$$

where X_{1ij}, our binary exposure of interest, follows a Bernoulli distribution with success probability η. X_{2ij} denotes the time point for each observation for a subject, where $X_{2ij} = 1, 2, \ldots, n$, and n is the number of measurement occasions (i.e. cluster size). X_{3ij} follows a normal distribution with mean μ_1 and standard deviation σ_1. $X_{4ij} = \tau X_3 + \epsilon$, where τ is a constant and ϵ follows a standard normal distribution with mean 0 and standard deviation 1.

To investigate the performance of the Firth and SCAD penalty functions in addressing sparsity and multicollinearity for binary longitudinal data, we conduct a series of simulations under different scenarios. Throughout this section and in subsequent sections, we will refer to the following four scenarios as follows: 1) Scenario 1: no sparsity in the binary outcome and no multicollinearity between the predictors; 2) Scenario 2: sparsity in the binary outcome but no multicollinearity between predictors; 3) Scenario 3: no sparsity in the binary outcome but multicollinearity between predictors; and 4) Scenario 4: both sparsity in the binary outcome and multicollinearity between predictors. For each of these scenarios, we fix the following simulation parameters: (i) $\eta = 0.3$; (ii) $\beta_1 = 1.8$, which controls the effect of the binary covariate X_{1ij}; (iii) $\beta_2 = -0.3$, which is the coefficient for time point X_{2ij}; (iv) $\beta_3 = 1.2$, the coefficient for the continuous covariate X_{3ij}; (v) $\beta_4 = 0.5$; and for $\mathbf{R}_i(\boldsymbol{\alpha})$, we assign an exchangeable working correlation structure. For our SCAD hyperparameters, we set $\alpha = 3.7$ and $\lambda = 2$, based on Fan and Li (2001a).

For each scenario, we vary each of the following parameters and consider all the combinations of values: cluster size $n \in \{3, 6\}$; number of subjects $m \in \{20, 40, 60\}$; and intra-cluster correlation coefficient (ICC) $\rho \in \{0.01, 0.1, 0.3\}$; and we define the working correlation matrix, $\mathbf{R}_i(\boldsymbol{\alpha})$ as both exchangeable and first-order autoregressive (AR(1)). By considering two different working correlation structures in our simulations, we can examine the performance of P2GEE when the WCM is correctly specified as exchangeable, and when it is misspecified as AR(1).

Finally, we vary the following scenario-based parameters: for sparse data, we set our intercept $\beta_0 = -7$; for non-sparse data, $\beta_0 = -3$; and for multicollinear data, we set $\tau = 2$; and for non-multicollinear data, $\tau = 0$. Note that in the case of no multicollinearity, this results in $X_4 \sim N(0, 1)$, which removes the linear relationship between X_3 and X_4. Following the framework outlined above, we simulate 1000 data sets for each of the four scenarios, which results in 36 simulations conducted for each scenario.

3.2 Model evaluation

In this section, we illustrate how to evaluate the performance of P2GEE. Since P2GEE is supposed to reduce both the bias in the estimated coefficients as well as the Standard Errors (SEs), we calculate the empirical bias as the average difference between the true coefficient and the estimated coefficients. Next, we compute the ratio between the average SE estimates and the empirical standard deviation to assess the performance of the sandwich estimator. Lastly, we derive the empirical coverage probability of 95% confidence intervals.

3.3 Software and packages

To implement P2GEE, we create a custom function based on the `geefirthr` R package, proposed in Mondol and Rahman (2019) and is publicly accessible on GitHub (repository: “mhmondol/geefirthr”). Specifically, we modify the `geefirth()` function to incorporate the SCAD penalty. All simulations and subsequent analyses are performed using R software (version 4.3.2).

3.4 Simulation results

In this section, we provide the results for each of the simulation scenarios using the P2GEE method, followed by simulation results for Scenario 4 in which we use the standard GEE method.

3.4.1 Results for Scenario 1: no sparsity in binary outcome and no multicollinearity in the regressors

The results for the simulated data, where both sparsity and multicollinearity are absent, are depicted in Figures 1, 2, 3, and 4. Specifically, Figure 1 illustrates outcomes for a smaller cluster size ($n = 3$) with an accurately specified working correlation structure (exchangeable). As the number of subjects m increases, the bias for all $\hat{\beta}$ values tends towards zero. Generally, configurations with higher ρ values exhibit lower bias. Nevertheless, $\hat{\beta}_1$ is observed to be more biased than the others, with its bias exacerbated at higher ρ levels, which is what we expect as the level of separation increases with high ICC. The ratio of the average sandwich estimate of the standard error to the empirical standard deviation is approximately 0.87, displaying slight instability marked by a minor decrease for $\hat{\beta}_1$ and $\hat{\beta}_2$. The coverage rate of the 95% confidence interval is commendable for $\hat{\beta}_0$, while the coverage for $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{\beta}_4$ hovers around 0.9. However, $\hat{\beta}_1$ demonstrates sub-optimal performance with a coverage level near 0.8. This is likely due to the high bias observed for $\hat{\beta}_1$. Increasing the cluster size, as shown in Figure 14, can ameliorate these issues for the most part. Furthermore, a higher number of subjects leads to better estimates and coverage across all settings.

In Figure 3, the outcomes for a small cluster size ($n = 3$) with an incorrectly specified working correlation structure (AR(1)) are presented. It's important to note that, in GEE, unlike mixed models, the correlation structure does not influence the marginal parameter estimates, resulting in outcomes that closely align with those from the correctly specified structure. However, this misalignment impacts the standard error estimates, causing them to be slightly more unstable than those seen in Figure 1, but the impact of misspecification is not significant. As with the previous situation, increasing the cluster size, as depicted in Figure 16, enhances stability and accuracy. Again, a higher number of subjects leads to better estimates and coverage across all settings.

3.4.2 Results for Scenario 2: sparse binary outcome but no multicollinearity in the regressors

The simulation results of the setting with sparsity in binary outcome but no multicollinearity are shown in Figures 5, 6, 7, and 8.

Figures 5 and 6 present the scenarios where the working correlation matrix is correctly specified. The bias of the estimates for regression coefficients converges to 0 as the number of subjects increases. Although a larger number of subjects decreases the bias of the sandwich variance estimator and improves the confidence interval, the

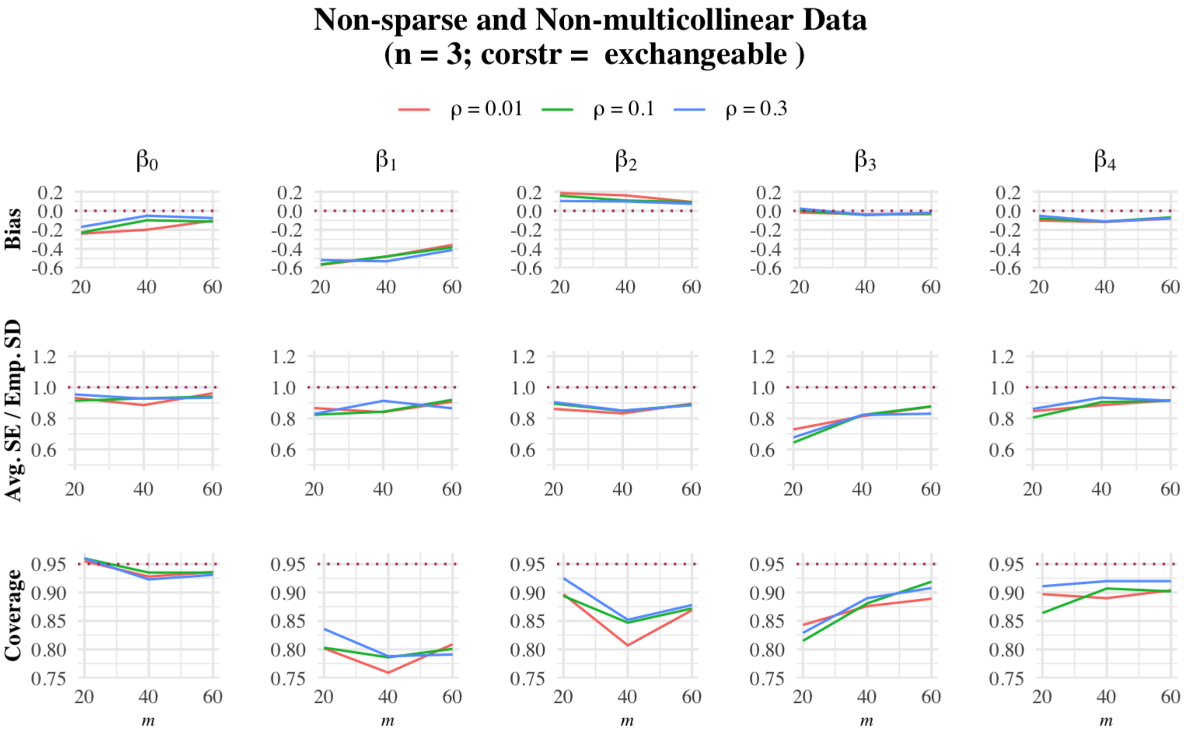

Figure 1: Data without sparsity and multicollinearity ($n = 3$ and exchangeable working correlation structure)

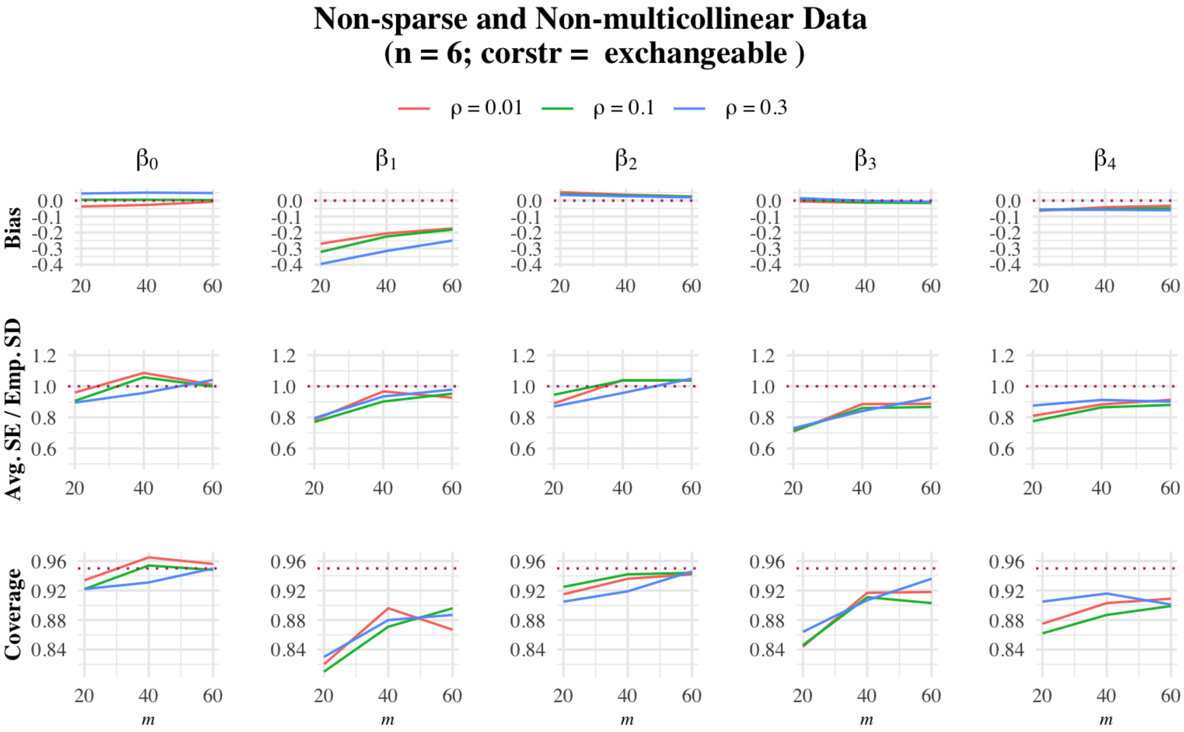

Figure 2: Data without sparsity and multicollinearity ($n = 6$ and exchangeable working correlation structure)

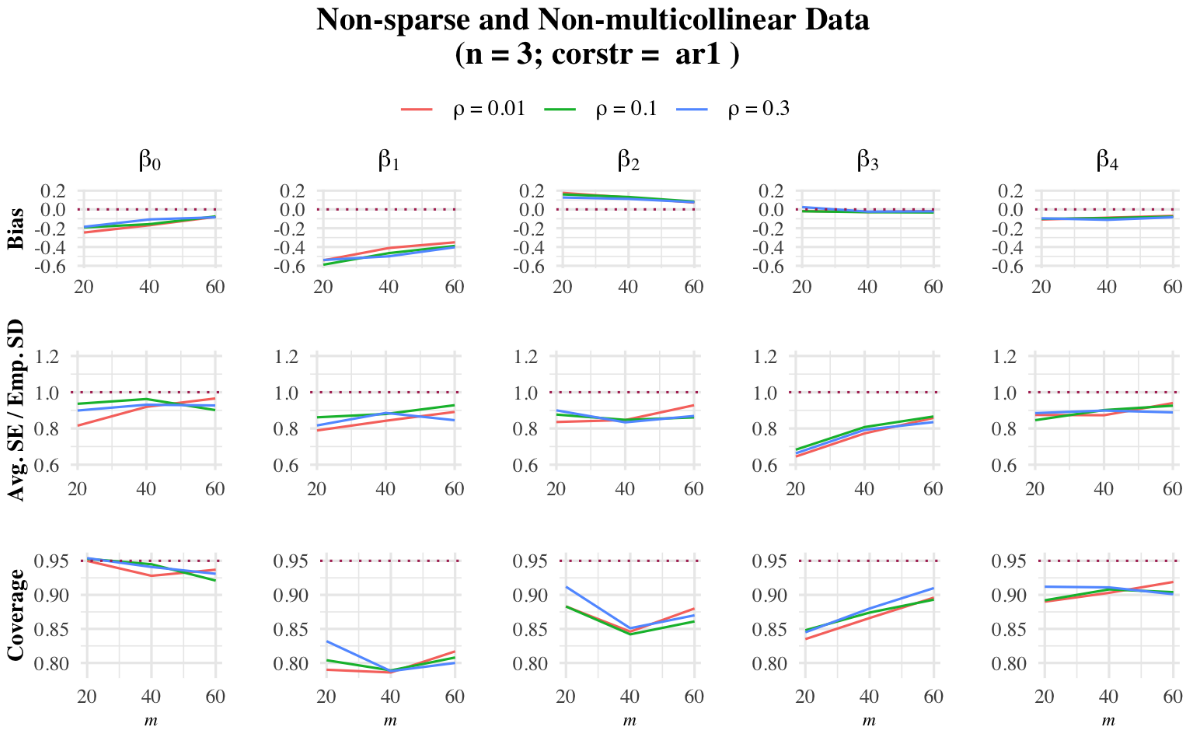

Figure 3: Data without sparsity and multicollinearity ($n = 3$ and $AR(1)$ working correlation structure)

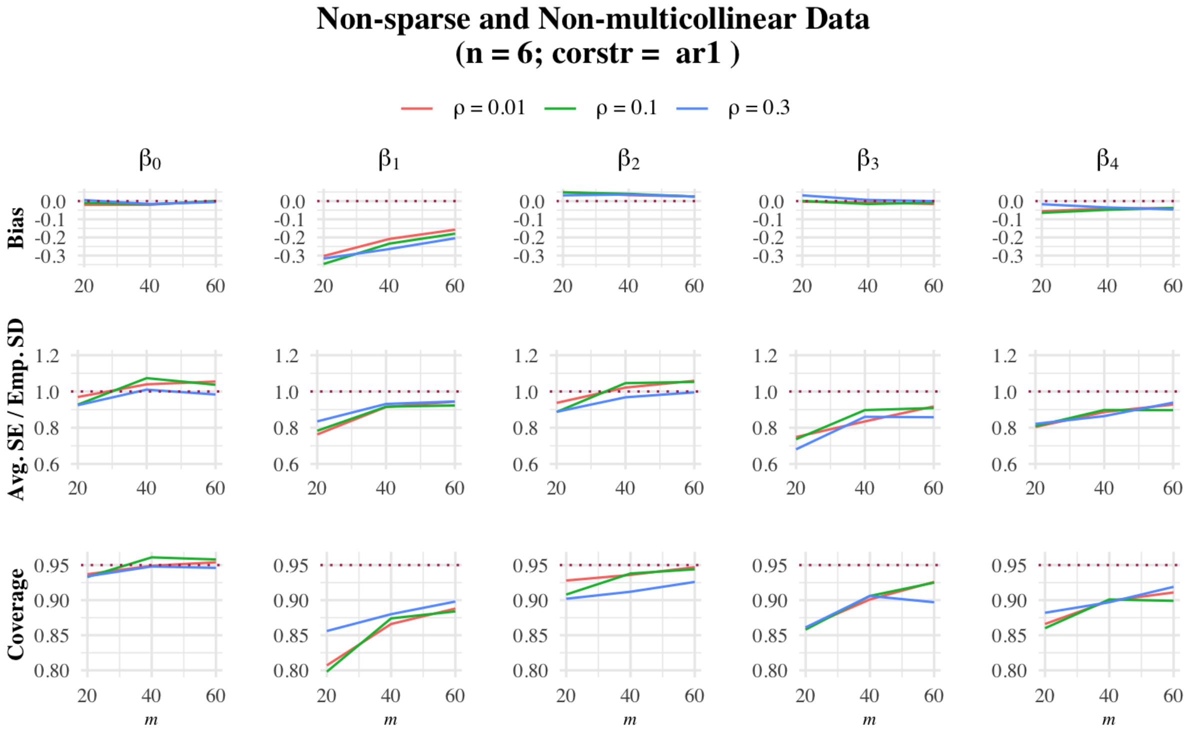

Figure 4: Data without sparsity and multicollinearity ($n = 6$ and $AR(1)$ working correlation structure)

sandwich estimator underestimates the true variance and the confidence interval has an under-coverage problem. This indicates that the estimator for regression coefficients is consistent but the sandwich estimator does not perform well, especially with small sample sizes. Additionally, we notice that a larger cluster size can lead to reduced bias and better performance of the sandwich estimator.

Figures 7 and 8 show the simulation results with the misspecified working correlation matrix. We can observe similar patterns as in the setting with the correctly specified working correlation matrix. On the other hand, the misspecification of the working correlation results in a slightly worse estimation of variance and performance of the confidence interval.

3.4.3 Results for Scenario 3: no sparsity in binary outcome but multicollinearity present in the regressors

The simulation results for the scenario in which our data are non-sparse and multicollinear are shown in Figures 9 and 10 for the models with a correctly specified exchangeable working correlation structure, and Figures 11 and 12 for the models with a misspecified AR(1) working correlation structure, in which we present the bias, the ratio between average sandwich estimation of standard error and the empirical standard deviation, as well as the coverage rate of a 95% confidence interval for the estimations.

For cluster size $n = 3$ and correct exchangeable working correlation structure, as depicted in Figure 9, we note that as the number of subjects m increases, the bias for β_2, β_3, and β_4 tends towards zero. For β_0, as m increases the bias tends towards zero for $\rho = 0.01$ and $\rho = 0.1$, but slightly increases for $\rho = 0.3$ at at $m = 40$ subjects. For a larger cluster size $n = 6$, and exchangeable working correlation structure, as depicted in Figure 10, we note an improvement in model performance, where the biases for all β values are closer to zero as m increases, when compared to the biases when $n = 3$.

For cluster size $n = 6$, the ratio between the average sandwich estimation of standard error and the empirical standard deviation tends toward one for all β values as m increases, and this trend is consistent across all values of ρ; however, when $n = 3$, the model does not perform as well. Specifically, the ratio between the average sandwich estimation of standard error and the empirical standard deviation tends toward one for β_0, β_1, and β_2 as sample size increases, but for β_3 and β_4 this ratio deviates away from one when $\rho = 0.01$ and $m = 60$. Moreover, when $n = 3$, the coverage rates tend to increase for all β coefficients as m and ρ increase, except for β_2, where we notice a general decrease in coverage as m increases. Additionally, when $\rho = 0.01$, the coverage rates for β_3 and β_4 plateau at $m = 40$. This discrepancy appears to be resolved when we increase the cluster size to $n = 6$, where the model performs better and the coverage rate for β_2 tends to increase as sample size increases.

When the working correlation structure is incorrectly specified as AR(1), the biases for all β coefficients follow

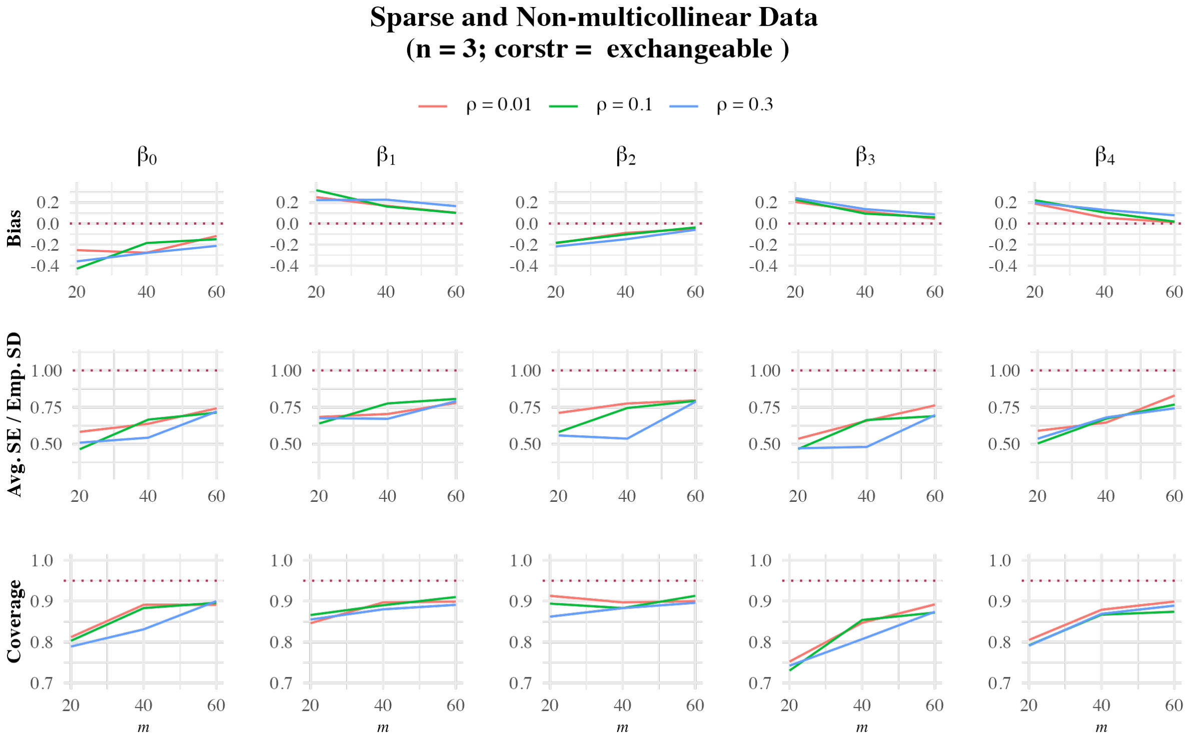

Figure 5: Data with sparsity but no multicollinearity ($n = 3$ and exchangeable working correlation structure)

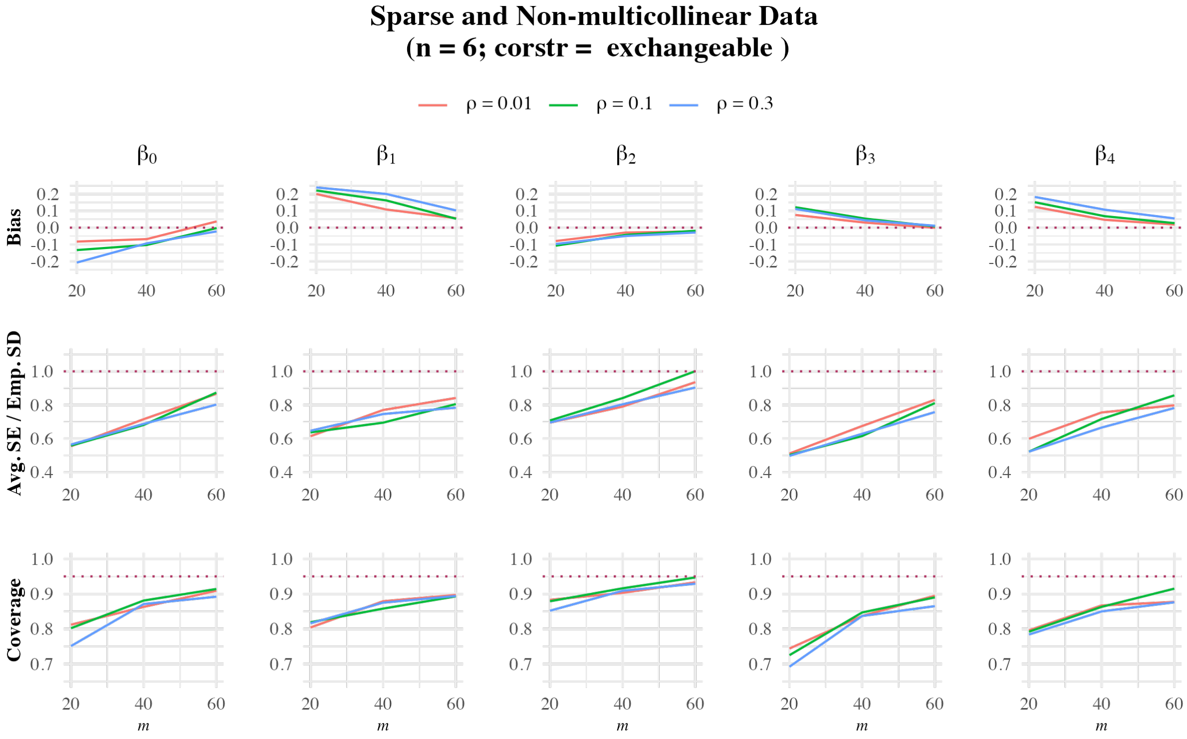

Figure 6: Data with sparsity but no multicollinearity ($n = 6$ and exchangeable working correlation structure)

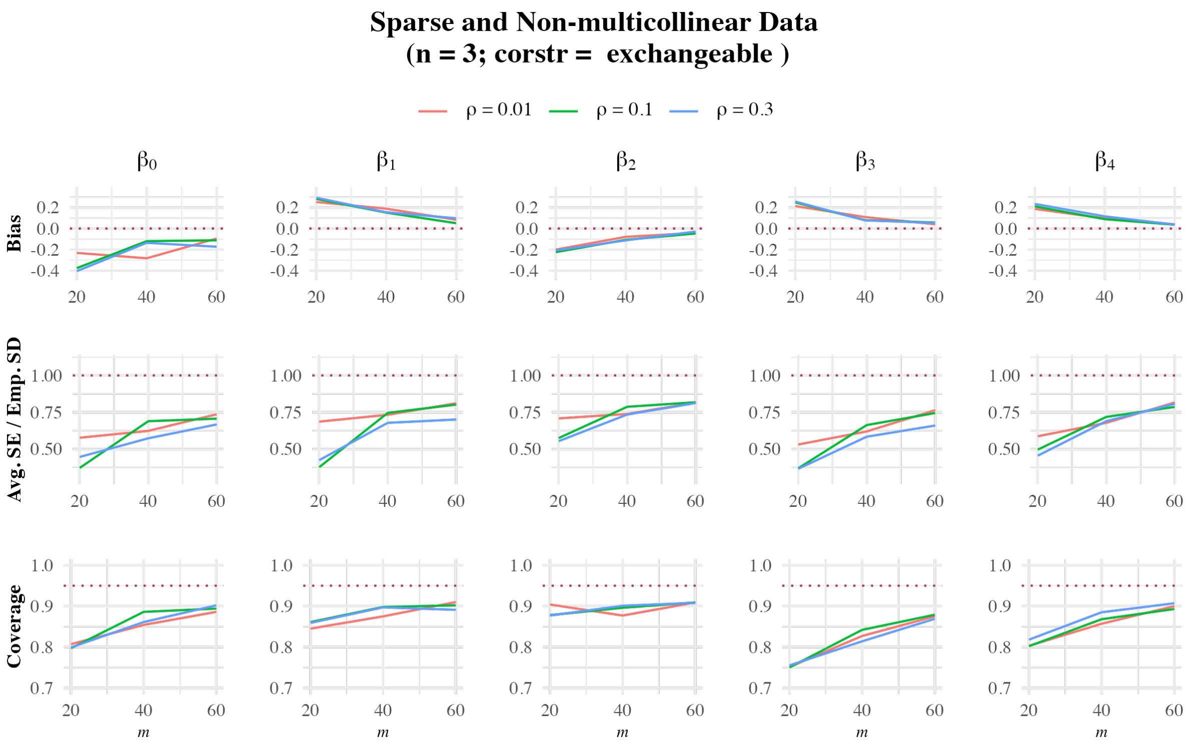

Figure 7: Data with sparsity but no multicollinearity ($n = 3$ and $AR(1)$ working correlation structure)

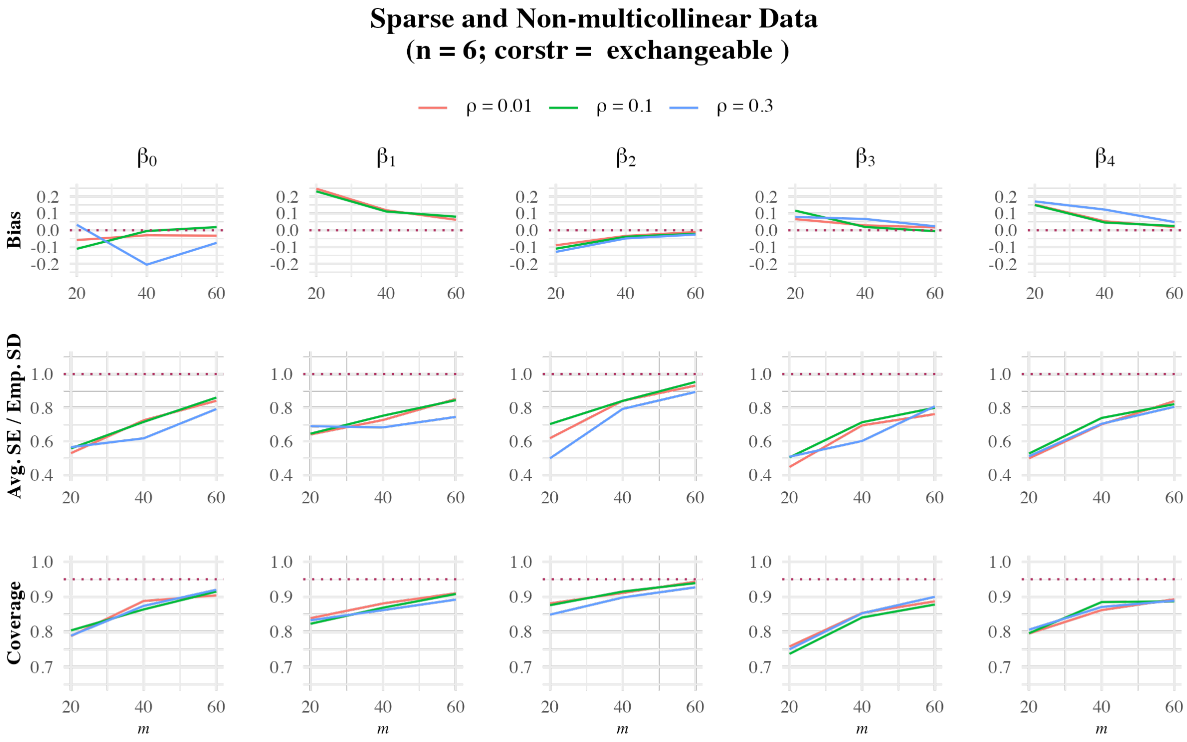

Figure 8: Data with sparsity but no multicollinearity ($n = 6$ and $AR(1)$ working correlation structure)

a similar trend as when the working correlation matrix is correctly specified, with model performance improving as sample size m increases, for both $n = 3$ and $n = 6$. Similarly, the ratio between the average sandwich estimation of standard error and the empirical standard deviation tends toward one as the sample size increases, and the model performs slightly better for a larger cluster size. Similar to what we observe with an exchangeable working correlation structure, when $n = 3$ the coverage rates generally follow an increasing trend for all coefficients except β_2, which consistently decreases as m increases. This performance is subsequently corrected when $n = 6$, as the rate of coverage is improved for β_2 and follows an increasing trend as m increases.

Based on the above, the P2GEE yields similar results for both the exchangeable and AR(1) working correlation structures, indicating that our proposed method is not affected by a misspecified working correlation structure. The performance is inconsistent for smaller values of n, but appears to improve as both cluster size and sample size increase.

3.4.4 Results for Scenario 4: sparse binary outcome and multicollinearity present in the regressors

The results for the simulated data with both sparse binary outcome and multicollinearity in the context are shown in Figures 13, 14, 15, and 16.

Figure 13 shows the results when n, the cluster size, is equal to 3, with exchangeable as the assumed working correlation structure. The bias approaches 0 as the number of subjects m increases when $\rho = 0.1$ and $\rho = 0.3$ for all β estimates except for β_4, the coefficient for one of the covariates that are correlated. The ratio between the average sandwich estimation of standard error and the empirical standard deviation is quite far away from 1, but still increases as m increases. This suggests that when the number of subjects is small, the naive sandwich estimator does not provide a good estimate of the standard deviation. Due to the dependency of confidence intervals and the sandwich estimation of standard deviation, the coverage rate of 95% confidence interval of β estimates is lower than the desired 95%. The cases when $\rho = 0.01$ is the most unstable when estimating β_1 and β_4 related performance measures.

Figure 14 shows the results when the cluster size 6, with exchangeable correlation structure. In general, the models are performing better compared to the case when setting cluster size as 3. The bias is smaller for the estimated β_1, β_2 and β_3. The bias for the estimates of β_0 and β_4 is at about the same level in both cases and has a general trend of approaching 0. The naive sandwich estimator does a slightly better job estimating the true standard deviation. The ratio between the average sandwich estimation of standard error and the empirical standard deviation and the coverage rate are both slightly higher.

Figures 15 and 16 show the cases when the models are fitted when the working correlation matrix is misspecified as AR(1). The bias has shown a similar trend as in the model that correctly specified the exchangeable correlation

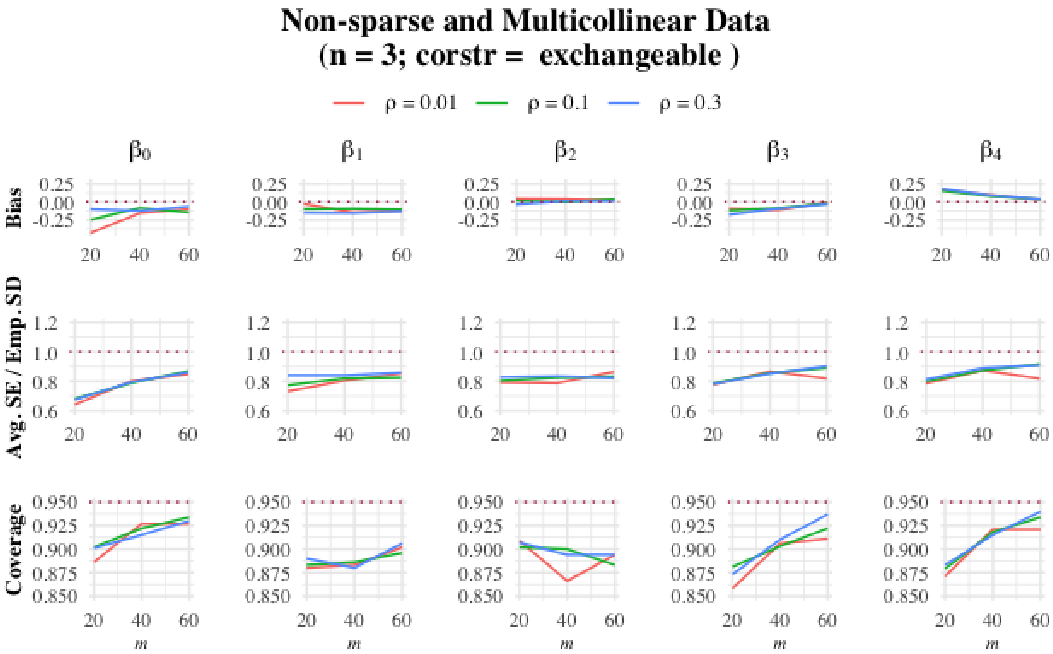

Figure 9: Data with non-sparsity and multicollinearity (n = 3 and exchangeable working correlation structure)

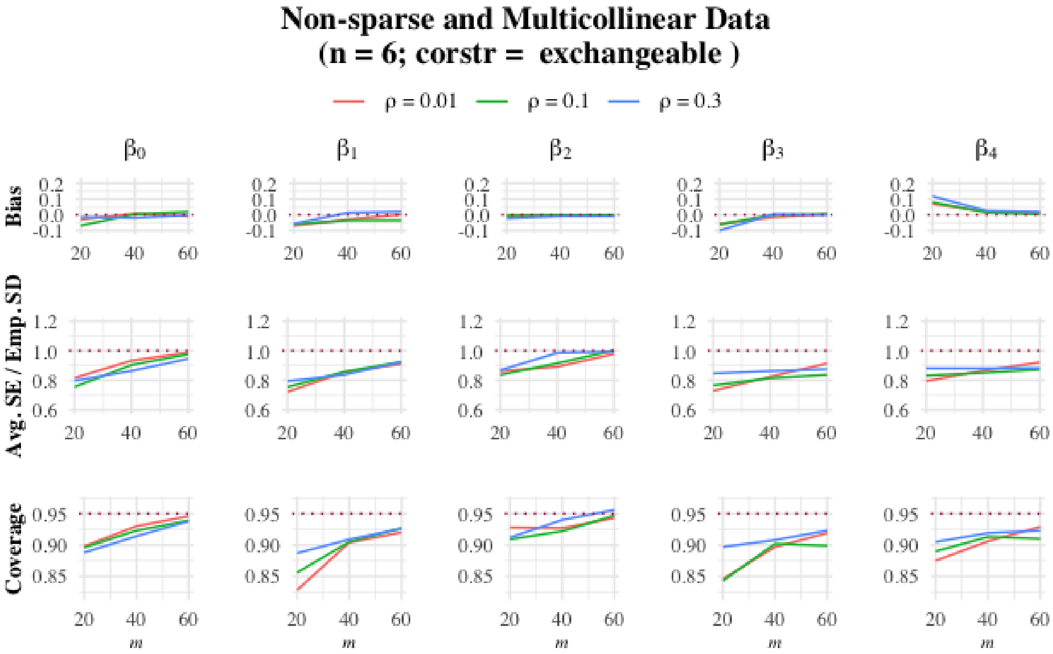

Figure 10: Data with non-sparsity and multicollinearity (n = 3 and exchangeable working correlation structure)

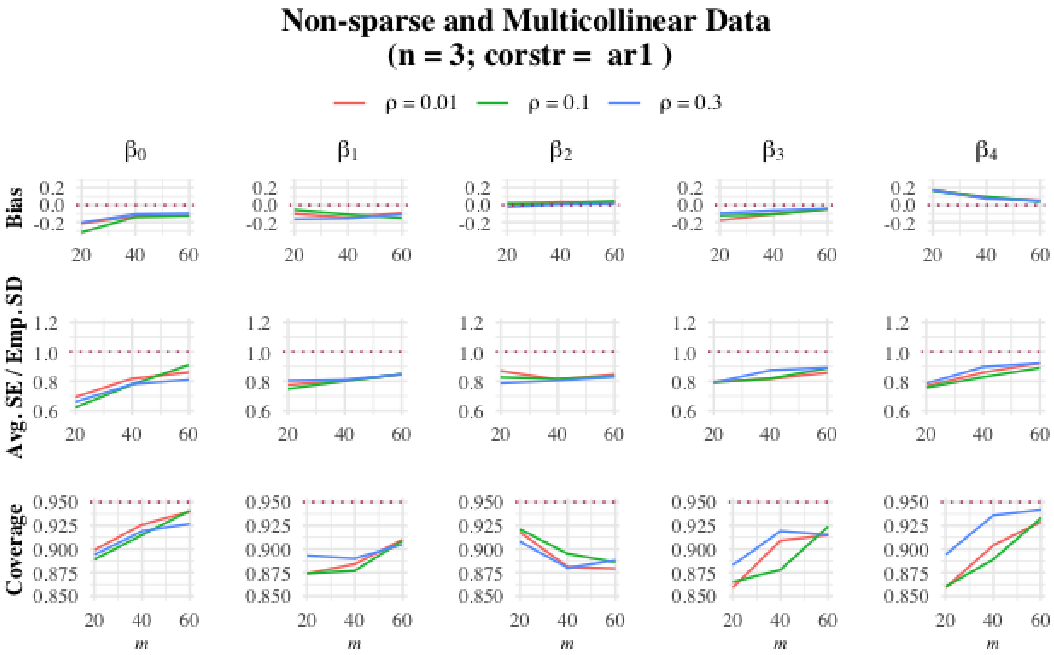

Figure 11: Data with non-sparsity and multicollinearity (n = 3 and AR(1) working correlation structure)

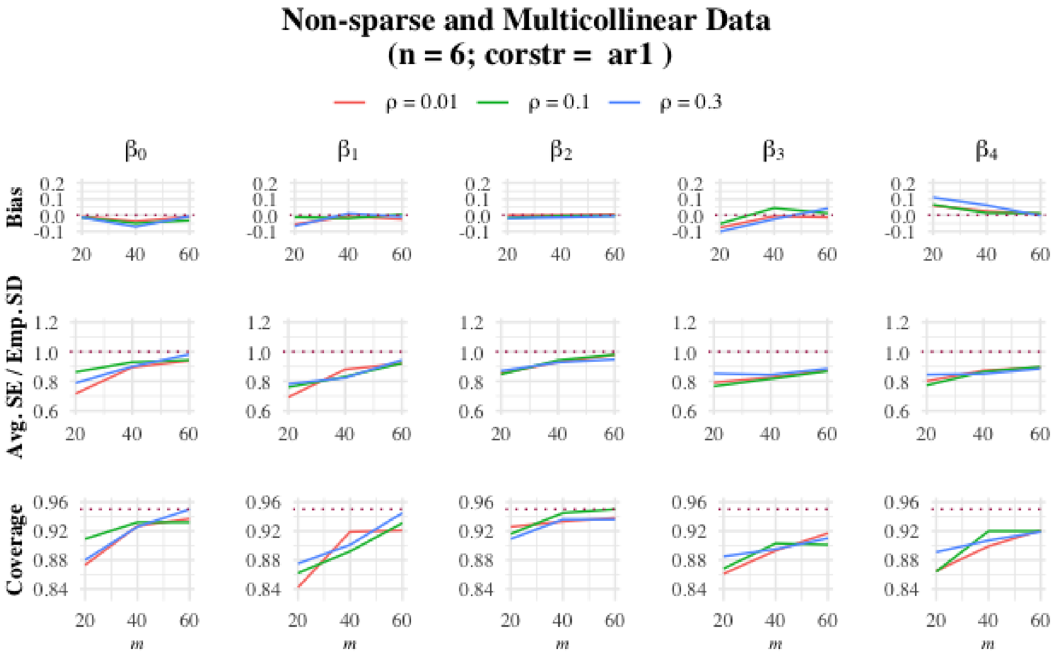

Figure 12: Data with non-sparsity and multicollinearity (n = 6 and AR(1) working correlation structure)

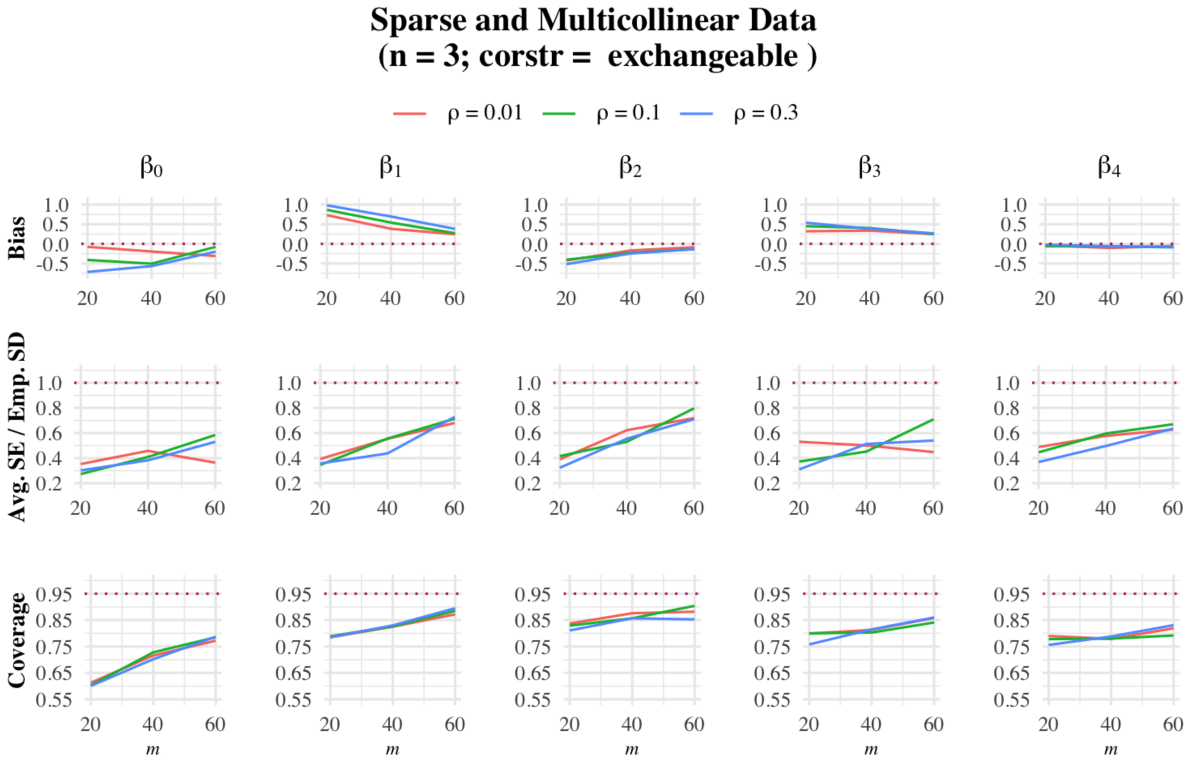

Figure 13: Data with both sparsity and multicollinearity ($n = 3$ and exchangeable working correlation structure)

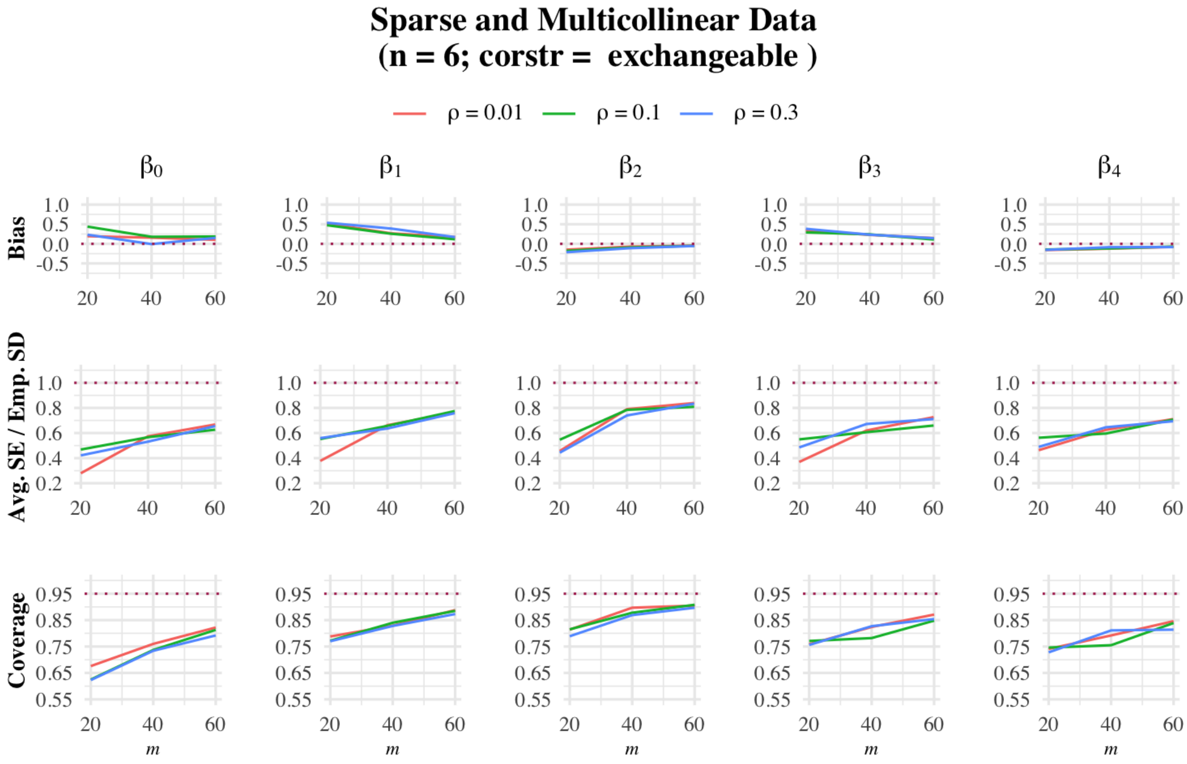

Figure 14: Data with both sparsity and multicollinearity ($n = 3$ and exchangeable working correlation structure)

structure. The coverage rate of 95% confidence interval is slightly affected by the misspecified working correlation matrix, which is most likely to occur when m is small.

3.4.5 Comparison with standard GEE

The results for the standard GEE models in the setting with both sparse binary outcomes and multicollinearity in the covariates are shown in Figures 17 and 18. Standard GEE models have a large and non-converging bias of the estimators for the regression coefficients and the sandwich estimator underestimates the true standard errors for small sample sizes. Moreover, the coverage rate of 95% confidence intervals is 0 for all the β's, which is likely due to the incredibly biased estimates of regression coefficients. These suggest that standard GEE models do not perform well for data with sparse binary outcomes and multicollinearity in the covariates.

4 Application to cardiovascular disease data

4.1 Overview of data

This section presents a case study using data from The Framingham Heart Study, a pioneering long-term investigation into the causes of cardiovascular disease in Framingham, Massachusetts. Beginning in 1948 with 5,209 participants, the study was groundbreaking for its prospective approach and identification of cardiovascular risk factors. Participants underwent biennial exams, including assessments of risk factors, health behaviors, and disease markers, while outcomes like Angina Pectoris, Myocardial Infarction, Heart Failure, and Stroke were rigorously monitored. This dataset comprises a subset of the study, featuring data on 4,434 participants from three exam periods between 1956 and 1968, tracking them over 24 years for cardiovascular events and mortality.

The subsample for the current study consisted of 100 participants, each with 3 observations, and included various variables as presented in Table 3. These were examined across four distinct scenarios characterized by differing levels of sparsity and multicollinearity. The characteristics of the variables and the proportion of outcomes (death) for each scenario are detailed in Tables 4, 5, 6, and 7. Note that sparsity happens when we only look at CURSMOKE and DEATH where the proportion of the case of CURSMOKE=1 and DEATH=1 is less than 15% as defined in Table 2.

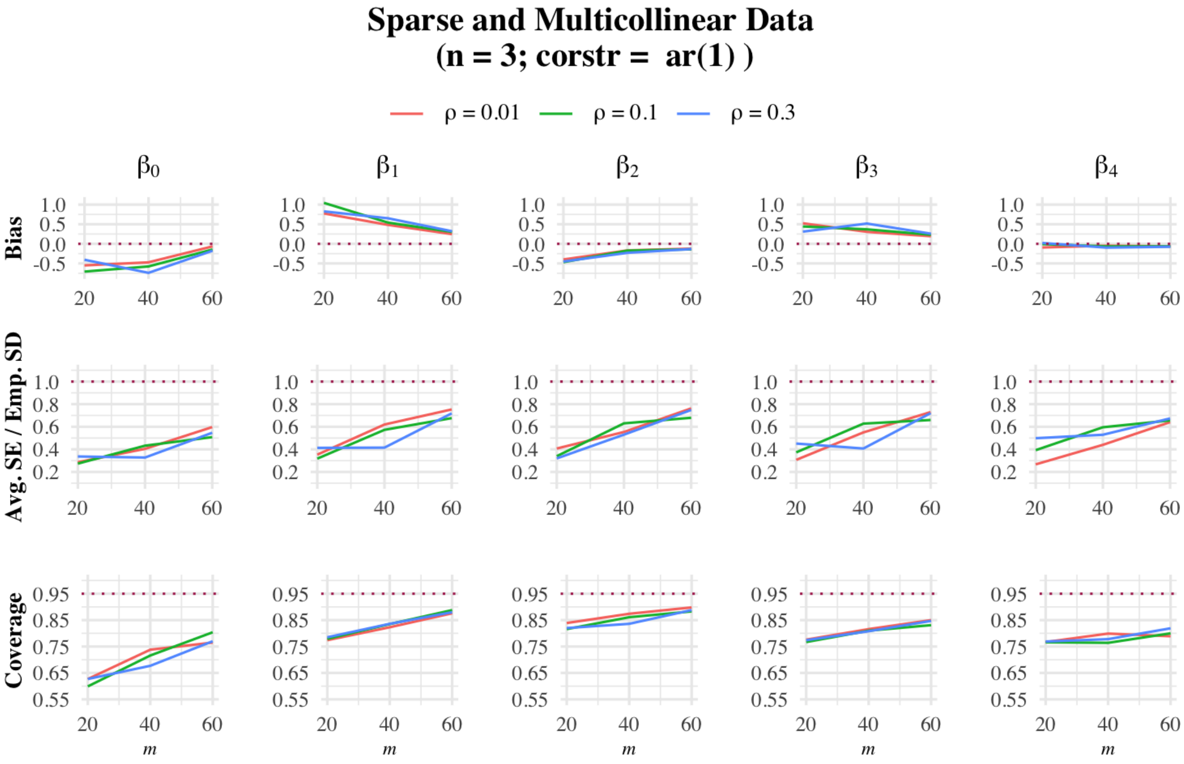

Figure 15: Data with both sparsity and multicollinearity ($n = 3$ and AR(1) working correlation structure)

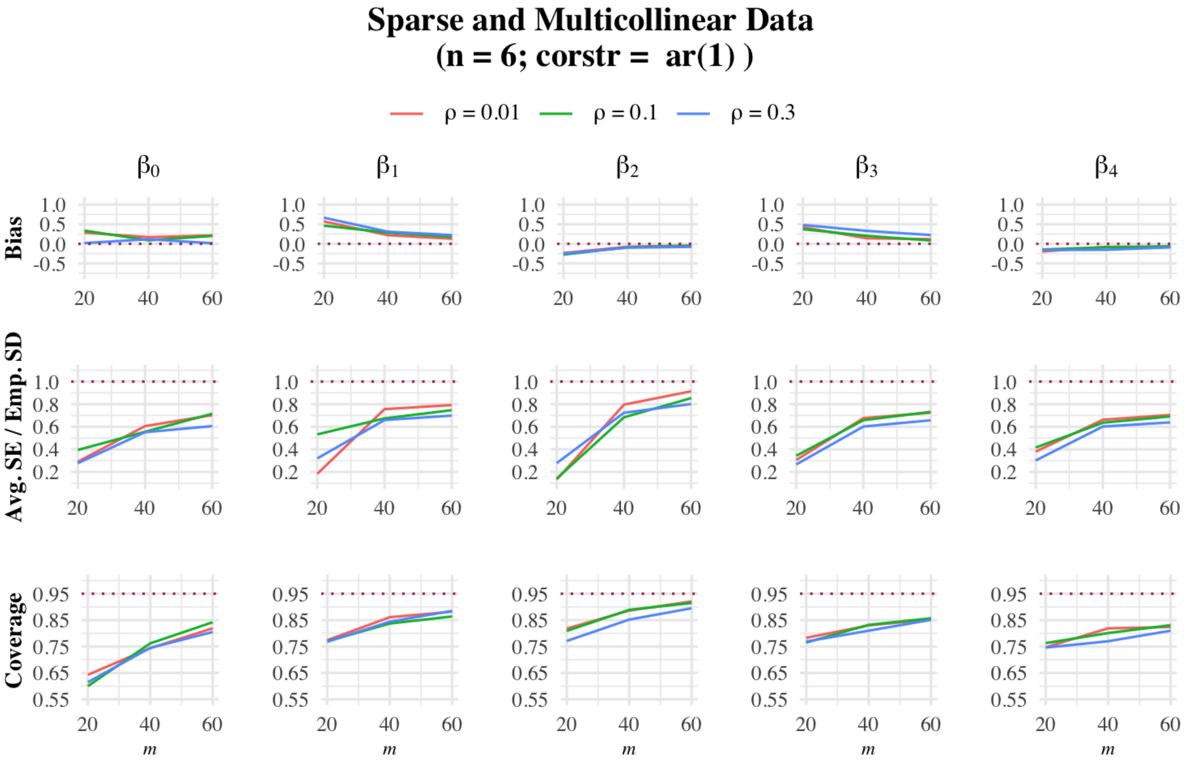

Figure 16: Data with both sparsity and multicollinearity ($n = 3$ and AR(1) working correlation structure)

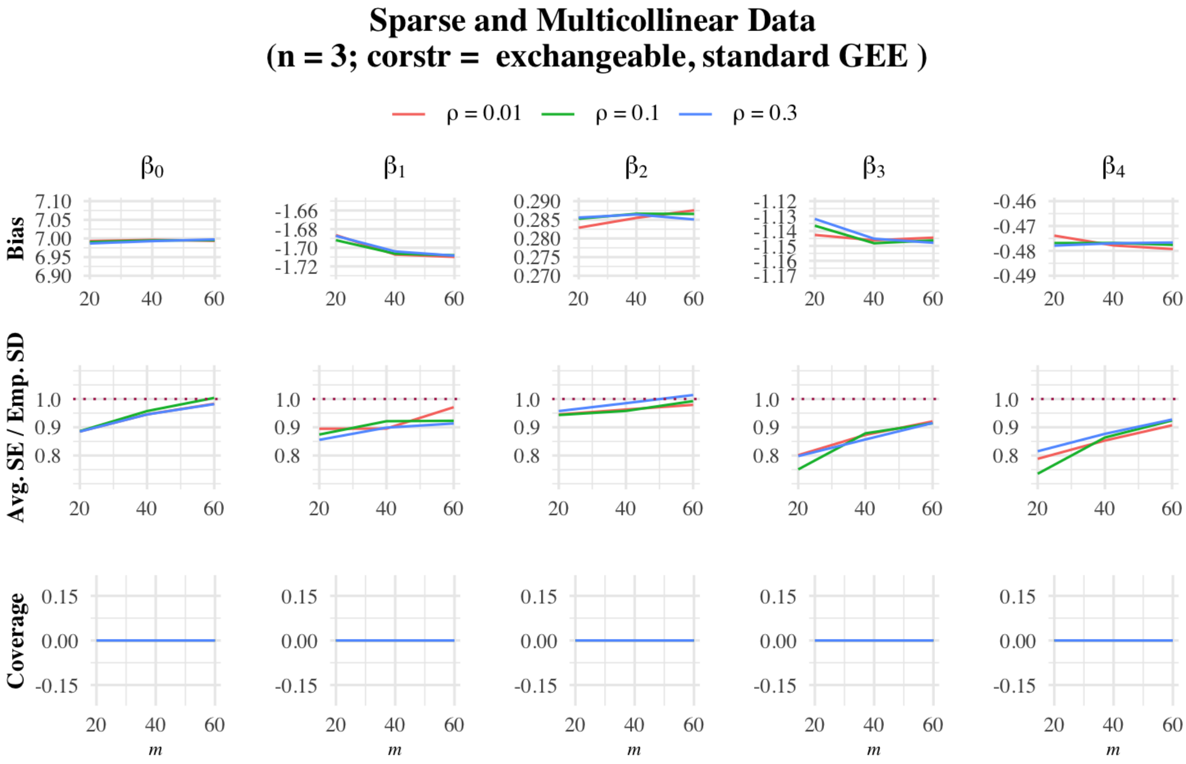

Figure 17: Data with both sparsity and multicollinearity (n = 3 and exchangeable working correlation structure with standard GEE)

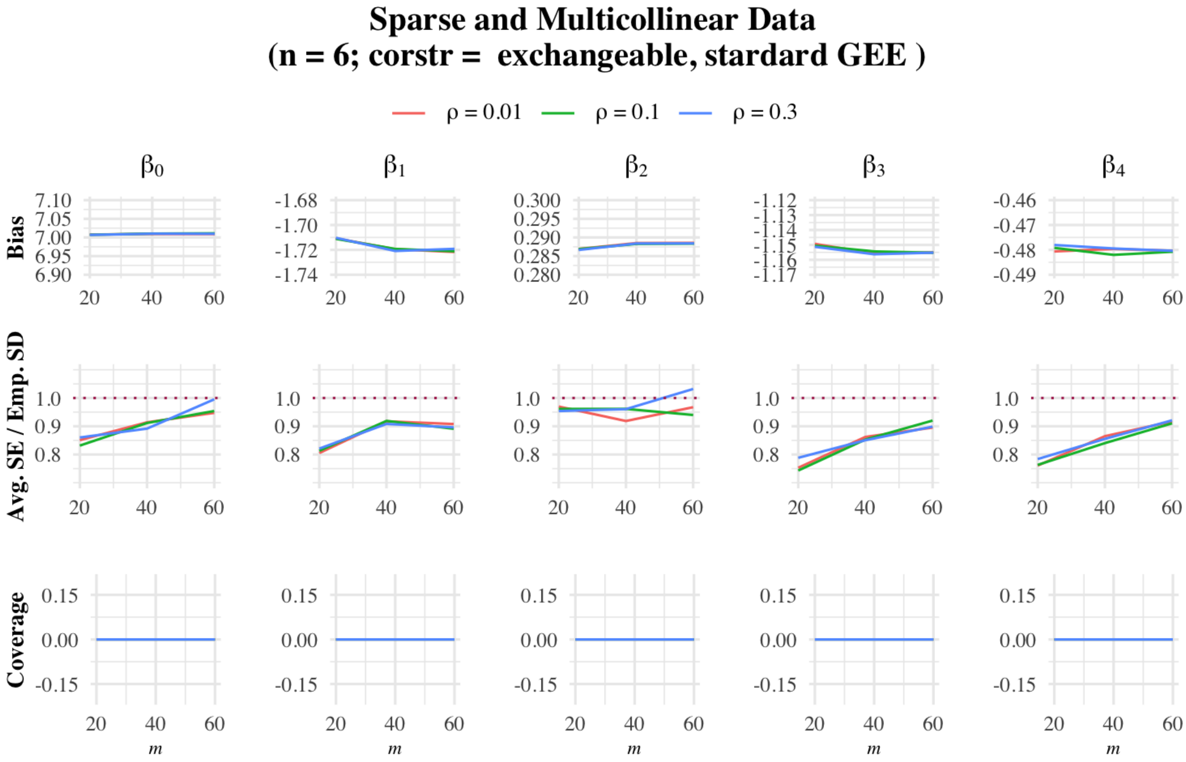

Figure 18: Data with both sparsity and multicollinearity (n = 6 and exchangeable working correlation structure with standard GEE)

Variable	
RANDID	Unique identification number for each participant.
SEX	Participant sex. 1 = Male, 2 = Female.
AGE	Age at exam (years).
BMI	Body Mass Index, weight in kilograms/height meters squared.
CURSMOKE	Current cigarette smoking at exam. 0 = Not current smoker, 1 = Current smoker.
SYSBP	Systolic Blood Pressure (mean of last two of three measurements) (mmHg).
SYSBP(modified)	SYSBP $\times 7 + \epsilon$, where $\epsilon \sim Norm(0, 1)$, used to create multicollinearity.
DEATH	Death from any cause. 0 = Did not occur during followup, 1 = Did occur during followup.

Table 3: Selected Variables

Variable	**Mean**	**SD**	**% of Indicator=1**
SEX	-	-	50
AGE	54.46	9.16	-
BMI	26.19	3.79	-
CURSMOKE	-	-	42.67
SYSBP	136	21.9	-
DEATH	-	-	33

Table 4: Descriptive statistics for covariates under scenario 1 – Data with non-sparsity and non-multicollinearity

4.2 Model Comparison

After setting up the four scenarios, we want to compare the coefficient estimates and standard errors obtained from standard GEE and our proposed P2GEE. The results are presented in Table 8, 9, 10, and 11. The results from P2GEE align well with the existing literature. For example, in three of the four scenarios, the estimates by P2GEE suggest that a higher level of systolic blood pressure (SYSBP) is correlated with a higher mortality rate. The review by Hajar (2016) supports this finding.

For the standard GEE models, the estimated coefficients and standard errors are mostly close to 0, which is due

Variable	**Mean**	**SD**	**% of Indicator=1**
SEX	-	-	50
AGE	54.98	10.36	-
BMI	25.98	3.56	-
CURSMOKE	-	-	37.33
SYSBP	132	20	-
DEATH	-	-	28

Table 5: Descriptive statistics for covariates under scenario 2 – Data with sparsity and non-multicollinearity

Variable	Mean	SD	% of Indicator=1
SEX	-	-	50
AGE	54.46	9.16	-
BMI	26.19	3.79	-
CURSMOKE	-	-	42.67
SYSBP	136	21.9	-
SYSBP(modified)	951.94	153.14	-
DEATH	-	-	33

Table 6: Descriptive statistics for covariates under scenario 3 – Data with non-sparsity and multicollinearity

Variable	Mean	SD	% of Indicator=1
SEX	-	-	50
AGE	54.98	10.36	-
BMI	25.98	3.56	-
CURSMOKE	-	-	37.33
SYSBP	132	20	-
SYSBP(modified)	926.66	140.28	-
DEATH	-	-	28

Table 7: Descriptive statistics for covariates under scenario 4 – Data with sparsity and multicollinearity

to that the numerical computation does not converge. This implies that the standard GEE does not perform well in the presence of either sparsity in binary outcomes or multicollinearity in covariates and confirms our conclusions in Section 3.4.5.

Variable	GEE(SE)	P2GEE(SE)
INTERCEPT	0.540 (0.000)	1.402 (0.676)
SEX	-0.140 (0.000)	-0.916 (0.422)
AGE	0.000 (0.000)	0.001 (0.001)
BMI	0.000 (0.000)	-0.020 (0.004)
CURSMOKE	0.000 (0.000)	0.019 (0.017)
SYSBP	0.000 (0.000)	0.000 (0.000)

Table 8: Coefficient estimates and standard error under scenario 1 – Data with non-sparsity and non-multicollinearity

5 Discussion and conclusions

The problem of near separation (sparsity) and multicollinearity is widespread in many biostatistical research problems. Sparsity in binary outcomes occurs from small sample size, rare outcomes, high ICC, or a combination of

Variable	GEE(SE)	P2GEE(SE)
INTERCEPT	0.280 (0.00)	0.445 (0.707)
SEX	0.000 (0.000)	-0.338 (0.419)
AGE	0.000 (0.000)	0.001 (0.001)
BMI	0.000 (0.000)	-0.020 (0.011)
CURSMOKE	-0.000 (0.000)	-0.013 (0.031)
SYSBP	-0.000 (0.000)	0.002 (0.001)

Table 9: Coefficient estimates and standard error under scenario 2 – Data with sparsity and non-multicollinearity

Variable	GEE(SE)	P2GEE(SE)
INTERCEPT	0.54 (0.152)	0.096 (0.680)
SEX	-0.140 (0.009)	-0.432 (0.433)
AGE	0.000 (0.000)	-0.001 (0.000)
BMI	-0.000 (0.000)	-0.004 (0.001)
CURSMOKE	0.000 (0.000)	-0.009 (0.006)
SYSBP	0.000 (0.000)	0.017 (0.007)
SYSBP(modified)	-0.000 (0.000)	-0.003 (0.001)

Table 10: Coefficient estimates and standard error under scenario 3 – Data with non-sparsity and multicollinearity

Variable	GEE(SE)	P2GEE(SE)
INTERCEPT	0.280 (0.000)	-0.349 (0.761)
SEX	-0.000 (0.000)	-0.164 (0.416)
AGE	0.000 (0.000)	0.005 (0.002)
BMI	0.000 (0.000)	-0.007 (0.015)
CURSMOKE	-0.000 (0.000)	-0.028 (0.062)
SYSBP	0.000 (0.000)	0.143 (0.040)
SYSBP(modified)	-0.000 (0.000)	-0.021 (0.006)

Table 11: Coefficient estimates and standard error under scenario 4 – Data with sparsity and multicollinearity

these issues while multicollinearity arises from building models with biomarkers and covariates that are governed by interrelated biochemical processes, similar behavioral practices or sociodemographic conditions and so on. These issues render many real-life problems difficult to study and therefore require statistical methods that can tackle such issues. This research investigated the consequences of sparsity and multicollinearity in GEE and addressed the problems by introducing a double penalized GEE (P2GEE). The P2GEE is derived by adding a Firth-type penalty term and treating GEE as if it were equivalent to a likelihood score equation, and a SCAD penalty term to address multicollinearity concerns. We have shown that P2GEE achieves convergence and provides finite estimates of the regression coefficients in the presence of separation, which was often not possible when using the standard GEE method. Furthermore, we have also demonstrated that the P2GEE can also attenuate the issues caused by multicollinearity.

Our simulation study showed that, when compared to the original GEE method, our proposed P2GEE offers a considerable improvement in model performance and addresses the problem of convergence, which poses a hurdle for the original GEE. While the original GEE failed to converge and resulted in biased estimates and inefficient SEs, our proposed P2GEE provided more stability in the presence of sparsity and multicollinearity. In each of the four simulation scenarios comprising of different sparsity-multicollinearity combinations, P2GEE exhibited the standard theme that a larger sample size (i.e. increase in both m and n yielded better estimates. This empirically indicates the fact that P2GEE retains the asymptotic properties of the standard GEE while improving upon its shortcomings. However, the standard sandwich variance estimator appeared biased in estimating the SEs of P2GEE estimates of the regression coefficients and hence provided poor empirical coverage, even for near-zero bias. Our findings on the sandwich variance estimator suggest that further study is required to correctly estimate the SE and coverage probabilities associated with the P2GEE estimates. There exist multiple modified sandwich variance estimators in the literature to attenuate the bias in small sample sandwich variance estimation. Finding a sandwich variance estimator that could optimally accompany P2GEE and provide efficient variance estimates is therefore a possible future direction of research.

In summary, the findings of our current paper suggest that the P2GEE might be an alternative solution to the standard GEE, as our proposed method showed substantial improvement over GEE by achieving convergence and reducing bias while simultaneously addressing multicollinearity. However, the sandwich variance estimation can still be improved for better inference. Thus we recommend the use of P2GEE for larger samples, especially if used in conjunction with the usual sandwich variance estimator.

References

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Allison, P. D. (2008). Convergence failures in logistic regression. In *SAS Global Forum*, volume 360(1), page 11.

Blommaert, A., Hens, N., and Beutels, P. (2014). Data mining for longitudinal data under multicollinearity and time dependence using penalized generalized estimating equations. *Computational Statistics & Data Analysis*, 71:667–680.

Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., and Chen, Y.-L. (2022). Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics*, 10(8):1283.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):248–265.

Dziak, J. J. (2006). *Penalized Quadratic Inference Functions for Variable Selection in Longitudinal Research.* PhD thesis, The Pennsylvania State University.

EF, S., NJ, P., SL, M., KA, A., and EM., M. (2017). *Collinearity and Causal Diagrams: A Lesson on the Importance of Model Specification.* Epidemiology.

Eliot, M., Ferguson, J., Reilly, M. P., and Foulkes, A. S. (2011). Ridge regression for longitudinal biomarker data. *Int. J. Biostat.*, 7(1):Art. 37, 11.

Fan, J. and Li, R. (2001a). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Li, R. (2001b). Variable selection via nonconcave penalizedlikelihood and its oracle properties. *Journal of the American Statistical Association.*

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80(1):141–151.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics.*

Geroldinger, A., Blagus, R., Ogden, H., and Heinze, G. (2022). An investigation of penalization and data augmentation to improve convergence of generalized estimating equations for clustered binary outcomes. *BMC Medical Research Methodology*, 22(1):168.

Gosho, M., Ishii, R., Noma, H., and Maruo, K. (2023). A comparison of bias-adjusted generalized estimating equations for sparse binary data in small-sample longitudinal studies. *Statistics in Medicine*.

Greenland, S., Mansournia, M. A., and Altman, D. G. (2016). Sparse data bias: a problem hiding in plain sight. *bmj*, 352.

Hajar, R. (2016). Framingham contribution to cardiovascular disease. *Heart Views*, 17(2):78–81.

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in medicine*, 25(24):4216–4226.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Hossain, S., Thomson, T., and Ahmed, E. (2018). Shrinkage estimation in linear mixed models for longitudinal data. *Metrika*, 81(5):569–586.

King, J. E. (2008). *Best Practices in Quantitative Methods.* SAGE Publications, Inc.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lunardon, N. and Scharfstein, D. (2017). Comment on 'small sample gee estimation of regression parameters for longitudinal data'. *Statistics in medicine*, 36(22):3596–3600.

Midi, H., Rana, S., and Imon, A. (2013). On a robust estimator in heteroscedastic regression model in the presence of outliers. In *Proceedings of the world congress on engineering*, volume 1.

Molenberghs, G., Verbeke, G., et al. (2005). Models for discrete longitudinal data.

Mondol, M. H. and Rahman, M. S. (2019). Bias-reduced and separation-proof gee with small or sparse longitudinal binary data. *Statistics in Medicine*, 38(14):2544–2560.

Paul, S. and Zhang, X. (2014). Small sample gee estimation of regression parameters for longitudinal data. *Statistics in medicine*, 33(22):3869–3881.

Rahmani, M., Arashi, M., Mamode Khan, N., and Sunecher, Y. (2018). Improved mixed model for longitudinal data analysis using shrinkage method. *Mathematical Sciences*, 12(4).

Saleh, A. K. M. E., Arashi, M., and Tabatabaey, S. M. M. (2014). *Statistical inference for models with multivariate t-distributed errors.* Hoboken, NJ: John Wiley & Sons.

Sutradhar, B. C., Jowaheer, V., and Rao, R. P. (2014). Remarks on asymptotic efficient estimation for regression effects in stationary and nonstationary models for panel count data. *Brazilian Journal of Probability and Statistics*, 28(2):241 – 254.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360.