RESEARCH ARTICLE

Variable selection in multivariate multiple regression

Asokan Mulayath Variyath☯*, **Anita Brobbey**☯

Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL, Canada

☯ These authors contributed equally to this work.
* variyath@mun.ca

Abstract

Introduction

In many practical situations, we are interested in the effect of covariates on correlated multiple responses. In this paper, we focus on estimation and variable selection in multi-response multiple regression models. Correlation among the response variables must be modeled for valid inference.

Method

We used an extension of the generalized estimating equation (GEE) methodology to simultaneously analyze binary, count, and continuous outcomes with nonlinear functions. Variable selection plays an important role in modeling correlated responses because of the large number of model parameters that must be estimated. We propose a penalized-likelihood approach based on the extended GEEs for simultaneous parameter estimation and variable selection.

Results and conclusions

We conducted a series of Monte Carlo simulations to investigate the performance of our method, considering different sample sizes and numbers of response variables. The results showed that our method works well compared to treating the responses as uncorrelated. We recommend using an unstructured correlation model with the Bayesian information criterion (BIC) to select the tuning parameters. We demonstrated our method using data from a concrete slump test.

Citation: Variyath AM, Brobbey A (2020) Variable selection in multivariate multiple regression. PLoS ONE 15(7): e0236067. https://doi.org/10.1371/journal.pone.0236067

Editor: Feng Chen, Tongii University, CHINA

Received: April 12, 2020

Accepted: June 26, 2020

Published: July 17, 2020

Data Availability Statement: The data is available at the reference paper indicated in the paper Yeh, I-C, 2006, 2007, 2008.

Funding: NO funding.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Multivariate multiple regression analysis is often used to assess covariate effects when one or multiple response variables are collected in observational or experimental studies. Many multivariate regression techniques are designed for univariate responses. A common way to deal

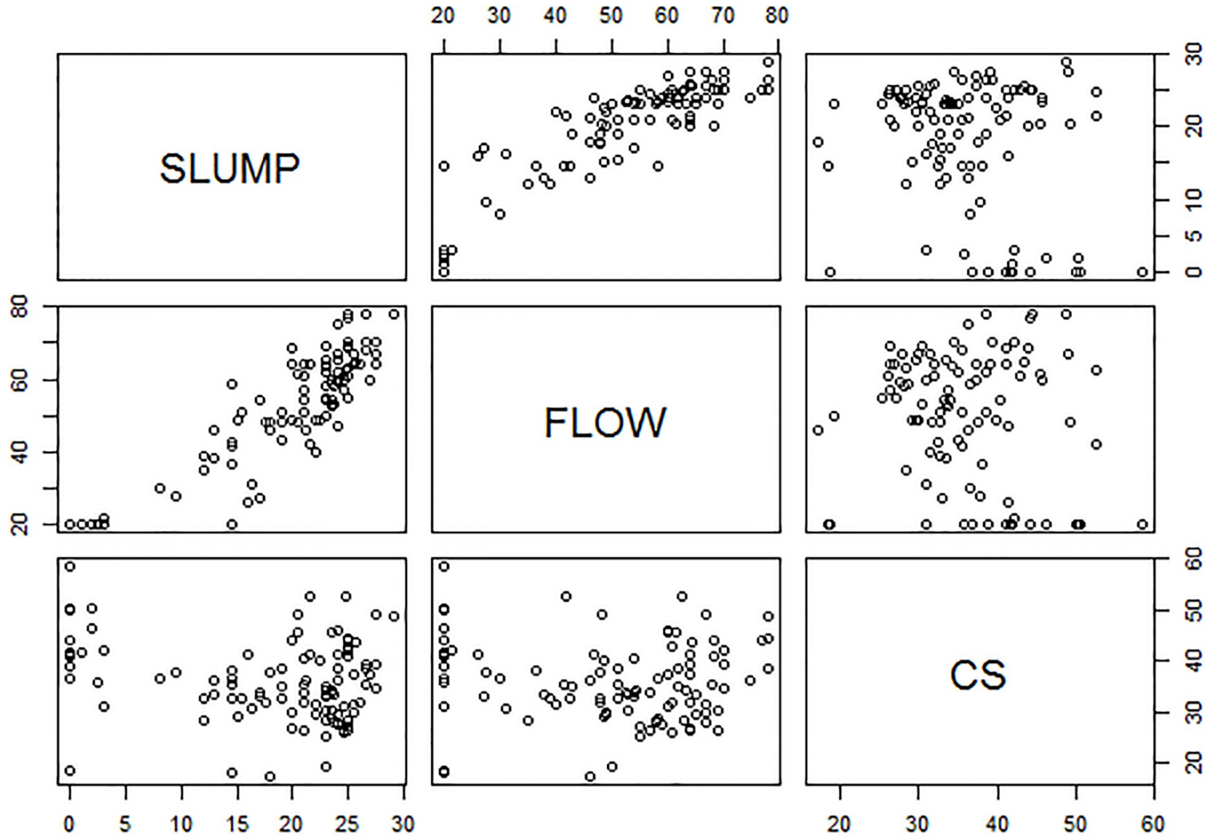

Fig 1. Scatter plot indicating the relationship between slump (Y_1), flow (Y_2), and compressive strength (Y_3).

https://doi.org/10.1371/journal.pone.0236067.g001

with multiple response variables is to apply the univariate technique separately to each variable, ignoring the joint correlation among the responses.

Consider the concrete slump test study reported in [1], [2], and [3]. The data set consists of three continuous output variables (slump, flow, and 28-day compressive strength (CS)). We wish to model these responses as a function of seven concrete ingredients (covariates): cement (X_1), fly ash (X_2), blast furnace slag (X_3), water (X_4), super plasticizer (X_5), coarse aggregate (X_6), and fine aggregate (X_7). The responses are correlated, and separate regression analysis will not take into account the importance of covariance on the response variables. Fig 1 shows the correlation among the output variables; in particular, slump and flow are highly correlated.

The joint model for all responses results in 27 parameters that must be estimated. Some of the covariates have no influence on the response variable(s), and excluding them results in a simpler model with better interpretive and predictive value.

The multi-response regression problem has been studied by various researchers in the generalized linear model (GLM) framework. For instance, the curd and whey method [4] uses the correlation among the response variables to improve the predictive accuracy. Multivariate modeling methods have been extensively used in transportation and accident analysis, especially for binary outcomes [5–7]. Some researchers have explored multivariate modeling with consideration of correlation in a Bayesian framework. For example, see [8] for the modeling of multivariate spatio-temporal Tobit regression and [9] for an approach based on spatial analysis.

The analysis of multivariate outcomes is more difficult when there are multiple types of outcomes. These occur frequently in the investigation of, e.g., dose–response experiments in toxicology [10, 11], birth defects in teratology [12], and pain in public health research [12, 13]. The methodologies used for mixed outcomes include factorization-based approaches on extensions of the general location model [14, 15]. However, these approaches depend on parametric distributional assumptions. Approaches based on latent variables include [16] and [17]. Modified generalized estimating equations (GEEs) [18] have been used to model longitudinal data; these approaches are of great interest because of their simplicity.

The GEE approach of [18] provides flexible modeling of multivariate observations based on a quasi-likelihood (QL) approach. In QL modeling, one assumes the existence of the first two moments of the responses of interest. It extends the GEE methodology to simultaneously analyze binary, count, and continuous outcomes with nonlinear models that incorporate the intra-subject correlation. The method uses a working correlation matrix. The incorporation of the intra-subject correlation makes this approach attractive. However, when we apply a joint model for all responses, many regression parameters must be estimated, and some have little or no influence on the responses. Large models can be difficult to interpret, so variable selection for multi-response modeling is of great interest.

We first systematically study the GEE approach in a cross-sectional set-up with multiple responses [11, 19]. Simultaneous parameter estimation and variable selection [20] has been used in many areas, including longitudinal data analysis [21]. We have extended this method to multivariate multiple regression using a penalized GEE methodology. We use the Bayesian information criterion (BIC) and generalized cross validation (GCV) to find the tuning parameters. Our simulation studies show that our methodology performs well.

The remainder of the paper is organized as follows. In the next section, we review the GEE for multiple responses and introduce our penalized GEE and the computational procedures. We discuss the distributional properties of the estimates and presents the simulation studies, subsequently and provides concluding remarks in the last section.

Materials and methods

GEE for multiple outcomes

We now discuss the GEE model based on the marginal distributions of the response for the analysis of longitudinal data. In a cross-sectional study with multiple responses, [12] used the GEE approach to estimate the parameters. Let the observations (y_i^m, x_i^m) denote the response and covariate respectively for the mth response ($m = 1, 2, \ldots, M_i$) measured on subject $i = 1, \ldots, n$. The QL approach requires us to specify the first two moments of the data (y_i^m). We define

$$E(y_i^m) = \mu_i^{(m)} = f(x_i^m, \boldsymbol{\beta}^{(m)})$$

$$var(y_i^{(m)}) = s^{(m)} h^{(m)}(\mu_i^{(m)}) = \sigma_i^{2(m)}$$

where $h^{(m)}(\cdot)$ is a known function, $s^{(m)}$ is a scaling parameter, $f^{(m)}(\cdot)$ is a nonlinear function of the coefficients, and $\boldsymbol{\beta}^{(m)}$ is a $p^{(m)} \times 1$ vector of model coefficients for the mth response variable. Let $\boldsymbol{y_i} = (y_i^{(1)}, \ldots, y_i^{(M_i)}), \boldsymbol{\mu_i} = (\mu_i^{(1)}, \ldots, \mu_i^{(M_i)})$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)^T}, \boldsymbol{\beta}^{(2)^T}, \ldots, \boldsymbol{\beta}^{(M)^T})^T$ be the $p \times 1$ vector of model parameters for all M outcomes, where $p = (p^{(1)} + p^{(2)} + \cdots + p^{(M)})$. In the QL framework with multiple outcomes, the regression coefficients $\boldsymbol{\beta}$ can be estimated by solving the GEEs

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_{\boldsymbol{i}}^{T} \boldsymbol{V}_{\boldsymbol{i}}^{-1} \boldsymbol{\varepsilon}_{\boldsymbol{i}} = \boldsymbol{0}. \tag{1}$$

For each subject i, let $\mathbf{D}_i$ be an $M_i \times p$ full-rank derivative diagonal block matrix such that $\mathbf{D}_i = \mathrm{diag}(\frac{\partial \mu_i^{(1)}}{\partial \boldsymbol{\beta}^{(1)T}}, \frac{\partial \mu_i^{(2)}}{\partial \boldsymbol{\beta}^{(2)T}}, \cdots, \frac{\partial \mu_i^{(M_i)}}{\partial \boldsymbol{\beta}^{(M_i)T}})$, $\boldsymbol{\varepsilon}_{\boldsymbol{i}} = (\boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{\mu}_{\boldsymbol{i}})$ be an $M_i \times 1$ vector of residuals, and $\boldsymbol{V}_{\boldsymbol{i}} = \boldsymbol{A}_{\boldsymbol{i}}^{\mathbf{1/2}} \boldsymbol{R}_{\boldsymbol{i}}(\boldsymbol{\alpha}) \boldsymbol{A}_{\boldsymbol{i}}^{\mathbf{1/2}}$ be the $M_i \times M_i$ working covariance matrix of $\boldsymbol{y}_{\boldsymbol{i}}$. Here, $\boldsymbol{A}_{\boldsymbol{i}} = \mathrm{diag}(\sigma_i^{2(1)}, \sigma_i^{2(2)}, \ldots, \sigma_i^{2(M_i)})$ is an $M_i \times M_i$ diagonal matrix of $\mathrm{var}(y_i^{(m)})$ and $\boldsymbol{R}_{\boldsymbol{i}}(\boldsymbol{\alpha})$ is an $M_i \times M_i$ working correlation matrix parameterized with the parameter vector $\boldsymbol{\alpha}$. The GEE estimator $\hat{\boldsymbol{\beta}}$ is asymptotically consistent as n goes to infinity.

Penalized GEE

To perform parameter estimation and variable selection simultaneously in the presence of mixed discrete and continuous outcomes, we propose a penalized version of the extended GEEs [12, 19]. Penalized likelihood methods such as LASSO [22] and SCAD [20] have been successful both theoretically and in practice. All the variables are considered at the same time, which may lead to a better global submodel. The penalized GEE has the feature that the consistency of the model holds even if the working correlation is misspecified. However, to improve the statistical efficiency of the coefficient, we recommend a covariance matrix based on the estimate of the unstructured working correlation. The regression coefficients $\boldsymbol{\beta}$ can be estimated by solving the penalized GEEs defined by

$$\sum_{i=1}^{n} \boldsymbol{D}_{\boldsymbol{i}}^{T} \boldsymbol{V}_{\boldsymbol{i}}^{-1} \boldsymbol{\varepsilon}_{\boldsymbol{i}} - nP'_{\lambda}(\boldsymbol{\beta}) sign(\beta) = \boldsymbol{0} \tag{2}$$

where $P'_{\lambda}(\boldsymbol{\beta}) = \partial P_{\lambda}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ is the vector derivative of the penalty function $P_\lambda(\beta)$ with λ being the vector of tuning parameters.

Although different penalty functions can be adopted, we consider only LASSO and SCAD. The former has the sparsity property, and the latter simultaneously achieves the three desirable properties of an ideal penalty: sparsity, unbiasedness, and continuity [20]. The LASSO penalty defined as $P_\lambda(\beta) = ||\beta||$ as per [22], where as per [20], the derivative of the SCAD penalty is defined as

$$P'_{\lambda,a}(\beta) = \lambda \left\{ I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\} \; for\ some\ a > 2\ and\ \beta > 0.$$

where a and λ are tuning parameters.

Computational algorithm. To compute $\hat{\boldsymbol{\beta}}$, we use the local quadratic approximation (LQA) algorithm [20]. With the aid of the LQA, the optimization of (2) can be carried out using a modified Newton–Raphson (MNR) algorithm. The estimate of $\hat{\boldsymbol{\beta}}$ at the $(r + 1)$th iteration is

$$\hat{\boldsymbol{\beta}}_{r+1} = \hat{\boldsymbol{\beta}}_{r} - \left\{ \frac{\partial S(\hat{\boldsymbol{\beta}}_r)}{\partial \boldsymbol{\beta}} - n\Sigma_{\lambda}(\hat{\boldsymbol{\beta}}_r) \right\}^{-1} \{ S(\hat{\boldsymbol{\beta}}_r) - n\boldsymbol{U}_{\lambda}(\hat{\boldsymbol{\beta}}_r) \} \tag{3}$$

where

$$\Sigma_\lambda(\beta_r) = \mathrm{diag}(P'_\lambda(|\beta_{1r}|)/|\beta_{1r}|, \ldots, P'_\lambda(|\beta_{pr}|)/|\beta_{pr}|),$$

$$\frac{\partial S(\hat{\boldsymbol{\beta}}_r)}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i;\,; \quad U_\lambda(\beta_r) = \Sigma_\lambda(\beta_r)\beta_r.$$

Given a tuning parameter λ, we repeat the above algorithm to update $\hat{\boldsymbol{\beta}}_r$ until we achieve convergence.

Correlation structure. Many researchers (e.g., [23–25]) have shown that an incorrectly specified correlation structure reduces the estimation efficiency. Thus, we suggest using an unstructured correlation structure $\boldsymbol{R_u(\alpha)}$ to estimate each variance and covariance uniquely. This structure can be estimated using a residual-based moment method. Let $\widehat{\boldsymbol{V}(\alpha)} = \hat{A}^{1/2} diag(\hat{R}_u, \ldots, \hat{R}_u)\hat{A}^{1/2}$ be the unstructured covariance matrix estimate. We have

$$\hat{R}_u = \frac{1}{n}\sum_{i=1}^{n} \hat{A}_i^{-1/2} \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^T \hat{A}_i^{-1/2}.$$

Tuning parameter selection

We set $a = 3.7$ for SCAD penalty as per [20]. Thus, we tune **λ** for both LASSO and SCAD. We define the GCV [26] criterion via

$$GCV(\lambda) = \frac{1}{n}\frac{D}{(1 - n^{-1}df(\lambda))^2}$$

and the BIC (see [27] and [6]) via

$$BIC(\lambda) = log\left(\frac{D}{n}\right) + \left(\frac{log(n)}{n}\right) df(\lambda)$$

where D is the deviance of the model and $df(\lambda) = tr\{X(X^T X + n\Sigma_\lambda)^{-1} X^T\}$. We choose the tuning parameter λ that minimizes GCV(λ) and BIC(λ).

Properties of estimates

Let $\beta = (\boldsymbol{\beta}_{\mathcal{A}}, \boldsymbol{\beta}_{\mathcal{N}})$ be the true vector of the regression coefficients. Under some necessary regularity conditions [28, 29] for sufficiently large n, the parameter estimates of the penalized GEE with the LASSO ($\lambda = O_p(n^{-1/2})$) and SCAD ($\lambda = o_p(1)$) penalties are consistent and asymptotically normal, i.e.,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{\Omega_0^{-1}\Omega_1\Omega_0^{-1}})$$

where $\boldsymbol{\Omega_0^{-1}\Omega_1\Omega_0^{-1}}$ is the sandwich variance estimator, with $\boldsymbol{\Omega_0} = \sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i - n\boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\beta}})$ and $\boldsymbol{\Omega_1} = \sum_{i=1}^{n} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \mathrm{cov}(\mathrm{y_i}) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i$.

The cov($\mathrm{y_i}$) can be replaced by $\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^T$.

Results

Performance analysis

We conducted a series of simulation studies to investigate the performance of our variable selection approach on continuous, binary, and count response outcomes using the LASSO and SCAD penalty functions. The simulations were conducted using the R software. For faster optimization of the tuning parameter λ, we used the warm-starting principle, where the initial value of $\boldsymbol{\beta}$ is replaced by $\hat{\boldsymbol{\beta}}_{(\lambda+\delta\lambda)}$ for the MNR algorithm. We select the model with minimum BIC(λ) or GCV(λ). We assess the model performance using the model error (ME) [20] as well as the standard error and the correct and incorrect deletions. The ME is due to the lack of fit of an underlying model and is denoted by $ME(\hat{\beta})$. Its size reflects how well the model fits the data:

$$ME(\hat{\beta}) = E_x\{\mu(\boldsymbol{X\beta}) - \mu(\boldsymbol{X}\hat{\boldsymbol{\beta}})\}^2$$

where $\mu(\boldsymbol{X}|\boldsymbol{\beta}) = E(\boldsymbol{y}|\mathbf{X})$. The ME has been expressed as the median relative model error (MRME). The relative model error is defined via

$$RME = \frac{ME}{ME_{full}},$$

where ME_{full} is the ME calculated by fitting the data with the full model. The correct deletions are the average number of true zero coefficients correctly estimated as zero, and the incorrect deletions are the average number of true nonzero coefficients erroneously set to zero. In the tables, the estimated values for correct and incorrect deletions are reported in the columns "Correct" and "Incorrect". For comparison purposes, we estimated the covariance matrix of the response variables based on both the unstructured working correlation (UWC) and the independent working correlation (IWC). We simulated 1000 data sets consisting of $n = 50$ and $n = 100$ observations from the response model

$$g(E(Y)) = X_{ij}^T\beta$$

with $i = 1, 2, \ldots n$ subjects and $j = 1, 2, \ldots, m$ responses. For binary outcomes we use a logit link; for count outcomes we use a log link; and for continuous (normal) outcomes we use the identity link function. We generated the covariates X_{ij} from the multivariate normal distribution with marginal mean 0, marginal variance 1, and AR(1) correlation with $\rho_x = 0.5$. For the simulations, we considered the following three cases of continuous, binary, and count response outcomes with different $\boldsymbol{\beta}$ values and correlation ρ_y between the responses and with $\sigma_y^2 = 1$.

Case 1: Three correlated cormal responses. We consider correlated normal responses ($m = 3$) with AR(1) true correlation. We set $\rho_y = 0.7$ and consider two covariates ($k = 2$) with $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)}) = ((3, 1.5), (0, 0), (2, 0))$. The simulation results are summarized in Table 1 for IWC and Table 2 for UWC. The tables show that the nonzero estimates of both SCAD and LASSO are close to the true values, i.e., $\beta_1^{(1)} = 3$, $\beta_2^{(1)} = 1.5$, and $\beta_1^{(3)} = 2$. However, the standard errors of the estimates in Table 2 are lower, which can be attributed to the correlation between the responses. For both $n = 50$ and $n = 100$, the mean ME and its standard error are smaller for SCAD than LASSO. The average number of zero coefficients increases as n increases in Table 2, especially for SCAD. This indicates that SCAD performs better than LASSO.

Case 2: Two correlated normal responses and one independent binary response. We consider three outcomes ($m = 3$): two continuous and one binary. The continuous outcomes were generated from a normal distribution and were correlated with AR(1) true correlation. We set $\rho_y = 0.7$ and consider the binary outcome from an independent binary observation and two covariates ($k = 2$) with $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)}) = ((3, 1.5), (0, 0), (2, 0))$. The simulation results are summarized in Tables 3 and 4. The tables show that the nonzero estimates for IWC are similar to those for UWC. However, because of the large correlation (0.7) between the continuous responses, the standard errors of $\beta_1^{(1)} = 3$, $\beta_2^{(1)} = 1.5$ are smaller for UWC. Again, the

Table 1. Simulations results for correlated normal responses (Case 1) with IWC.

Selection	Penalty	MRME	Correct	Incorrect
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	0.064	1.297	0.000
	LASSO	0.092	0.982	0.001
$\hat{\lambda}_{BIC}$	SCAD	0.053	1.532	0.000
	LASSO	0.113	1.180	0.002
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	0.030	1.298	0.000
	LASSO	0.038	0.871	0.001
$\hat{\lambda}_{BIC}$	SCAD	0.025	1.538	0.000
	LASSO	0.043	1.066	0.000
Selection	Penalty	$\hat{\beta}_1^{(1)}$	$\hat{\beta}_2^{(1)}$	$\hat{\beta}_1^{(3)}$
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	2.998(0.171)	1.496(0.168)	1.993(0.154)
	LASSO	2.898(0.203)	1.388(0.219)	1.831(0.229)
$\hat{\lambda}_{BIC}$	SCAD	2.998(0.171)	1.496(0.168)	1.992(0.147)
	LASSO	2.866(0.236)	1.356(0.244)	1.789(0.266)
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	2.998(0.115)	1.506(0.116)	1.996(0.105)
	LASSO	2.931(0.170)	1.438(0.154)	1.891(0.152)
$\hat{\lambda}_{BIC}$	SCAD	2.998(0.115)	1.506(0.115)	1.998(0.100)
	LASSO	2.898(0.216)	1.403(0.192)	1.857(0.190)

https://doi.org/10.1371/journal.pone.0236067.t001

Table 2. Simulations results for correlated normal responses (Case 1) with UWC.

Selection	Penalty	MRME	Correct	Incorrect
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	0.045	1.457	0.000
	LASSO	0.079	1.214	0.001
$\hat{\lambda}_{BIC}$	SCAD	0.035	1.661	0.000
	LASSO	0.079	1.261	0.011
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	0.022	1.513	0.000
	LASSO	0.040	1.265	0.000
$\hat{\lambda}_{BIC}$	SCAD	0.017	1.696	0.000
	LASSO	0.040	1.318	0.000
Selection	Penalty	$\hat{\beta}_1^{(1)}$	$\hat{\beta}_2^{(1)}$	$\hat{\beta}_1^{(3)}$
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	2.999(0.155)	1.496(0.145)	1.992(0.137)
	LASSO	2.884(0.200)	1.427(0.156)	1.842(0.185)
$\hat{\lambda}_{BIC}$	SCAD	3.000(0.145)	1.496(0.131)	1.993(0.122)
	LASSO	2.861(0.212)	1.421(0.164)	1.823(0.236)
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	2.998(0.102)	1.505(0.098)	1.996(0.091)
	LASSO	2.921(0.122)	1.457(0.100)	1.892(0.125)
$\hat{\lambda}_{BIC}$	SCAD	2.999(0.092)	1.504(0.090)	1.996(0.083)
	LASSO	2.917(0.122)	1.454(0.100)	1.887(0.124)

https://doi.org/10.1371/journal.pone.0236067.t002

Table 3. Simulations results for correlated normal and independent binary responses (Case 2) with IWC.

Selection	Penalty	MRME	Correct	Incorrect
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	0.059	1.755	0.007
	LASSO	0.129	1.663	0.024
$\hat{\lambda}_{BIC}$	SCAD	0.054	2.143	0.030
	LASSO	0.154	1.787	0.051
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	0.027	1.816	0.001
	LASSO	0.072	1.799	0.023
$\hat{\lambda}_{BIC}$	SCAD	0.023	2.122	0.003
	LASSO	0.095	2.002	0.043
Selection	Penalty	$\hat{\beta}_1^{(1)}$	$\hat{\beta}_2^{(1)}$	$\hat{\beta}_1^{(3)}$
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	2.995(0.171)	1.494(0.165)	2.192(0.799)
	LASSO	2.888(0.188)	1.381(0.201)	0.772(0.423)
$\hat{\lambda}_{BIC}$	SCAD	2.996(0.171)	1.494(0.165)	2.069(0.919)
	LASSO	2.864(0.204)	1.355(0.218)	0.687(0.419)
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	2.997(0.115)	1.506(1.113)	2.078(0.487)
	LASSO	2.906(0.145)	1.413(0.144)	0.903(0.435)
$\hat{\lambda}_{BIC}$	SCAD	2.997(0.115)	1.506(0.113)	2.060(0.470)
	LASSO	2.876(0.159)	1.381(0.167)	0.731(0.383)

https://doi.org/10.1371/journal.pone.0236067.t003

Table 4. Simulations results for correlated normal and independent binary responses (Case 2) with UWC.

Selection	Penalty	MRME	Correct	Incorrect
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	0.056	1.829	0.005
	LASSO	0.094	1.762	0.006
$\hat{\lambda}_{BIC}$	SCAD	0.037	2.209	0.037
	LASSO	0.097	1.824	0.008
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	0.025	1.825	0.001
	LASSO	0.057	1.880	0.002
$\hat{\lambda}_{BIC}$	SCAD	0.015	2.336	0.001
	LASSO	0.063	2.091	0.002
Selection	Penalty	$\hat{\beta}_1^{(1)}$	$\hat{\beta}_2^{(1)}$	$\hat{\beta}_1^{(3)}$
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	2.995(0.156)	1.492(0.148)	2.192(0.815)
	LASSO	2.918(0.148)	1.429(0.141)	0.782(0.391)
$\hat{\lambda}_{BIC}$	SCAD	2.998(0.142)	1.488(0.133)	2.076(0.936)
	LASSO	2.912(0.150)	1.424(0.140)	0.739(0.364)
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	2.999(0.108)	1.501(1.002)	2.079(0.480)
	LASSO	2.938(0.102)	1.453(0.094)	0.882(0.388)
$\hat{\lambda}_{BIC}$	SCAD	3.002(0.096)	1.498(0.102)	2.066(0.469)
	LASSO	2.927(0.097)	1.445(0.091)	0.767(0.299)

https://doi.org/10.1371/journal.pone.0236067.t004

Table 5. Simulations results for correlated normal and binary responses (Case 3) with IWC.

Selection	Penalty	MRME	Correct	Incorrect
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	0.071	1.916	0.209
	LASSO	0.092	1.343	0.173
$\hat{\lambda}_{BIC}$	SCAD	0.070	2.446	0.301
	LASSO	0.119	1.509	0.258
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	0.034	1.775	0.066
	LASSO	0.050	1.449	0.084
$\hat{\lambda}_{BIC}$	SCAD	0.047	2.430	0.151
	LASSO	0.056	1.622	0.152
Selection	Penalty	$\hat{\beta}_1^{(1)}$	$\hat{\beta}_2^{(1)}$	$\hat{\beta}_1^{(3)}$
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	2.997(0.167)	1.499(0.171)	0.543(0.520)
	LASSO	2.899(0.202)	1.395(0.214)	0.241(0.224)
$\hat{\lambda}_{BIC}$	SCAD	2.997(0.167)	1.499(0.170)	0.246(0.461)
	LASSO	2.886(0.219)	1.361(0.238)	0.212(0.222)
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	2.998(0.114)	1.503(0.116)	0.633(0.201)
	LASSO	2.918(0.149)	1.421(0.157)	0.287(0.194)
$\hat{\lambda}_{BIC}$	SCAD	2.998(0.113)	1.503(0.115)	0.309(0.432)
	LASSO	2.892(0.166)	1.393(0.188)	0.253(0.185)

https://doi.org/10.1371/journal.pone.0236067.t005

average number of zero coefficients is higher for UWC than for IWC. As the SCAD sample size increases, the mean ME and its standard error decrease for both GCV and BIC. The LASSO estimates for $\beta_3^{(1)}$ are not close to the true value, but the SCAD estimates of the nonzero coefficients are all close to the true values. Thus, SCAD performs better than LASSO.

Case 3: Two correlated normal responses and one binary response. We consider three outcomes ($m = 3$): two continuous and one binary. They are generated using an unstructured correlation structure with the parameters $\rho_{12} = 0.3$, $\rho_{13} = 0.4$, and $\rho_{23} = 0.6$, and we consider two covariates ($k = 2$) with $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)}) = ((3, 1.5), (0, 0), (2/3, 0))$. We set the $\boldsymbol{\beta}$ values for the binary outcome smaller than before to avoid numerical instability. The correlated normal and binary outcomes were generated in R using the **BinNor** package [30] for generating multiple binary and normal variables simultaneously given marginal characteristics and association structure; it is based on the methodology of [31]. The simulation results are summarized in Tables 5 and 6. The tables show that if the sample size is increased, the mean ME and its standard error are reduced. Again, the standard errors of the nonzero parameter estimates are lower for UWC than IWC. The average numbers of zero coefficients using SCAD with BIC for all sample sizes are close to the target value of three, and for SCAD with GCV the nonzero estimated coefficients are close to the true values for $n = 50$ and $n = 100$.

Overall, Tables 1 to 6 show that the nonzero estimates are unbiased regardless of the correlation structure. However, the unstructured correlation resulted in lower standard errors compared to the estimates based on an independent working correlation. The average number of zero coefficients is higher in the unstructured case. We notice a decrease in the mean ME when the sample size increases from 50 to 100 for both LASSO and SCAD. SCAD has a smaller mean ME than LASSO in all cases. We conclude that SCAD with BIC performs well.

Table 6. Simulations results for correlated normal and binary responses (Case 3) with UWC.

Selection	Penalty	MRME	Correct	Incorrect
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	0.065	1.975	0.167
	LASSO	0.098	1.538	0.117
$\hat{\lambda}_{BIC}$	SCAD	0.059	2.493	0.242
	LASSO	0.106	1.601	0.241
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	0.031	1.980	0.041
	LASSO	0.059	1.578	0.057
$\hat{\lambda}_{BIC}$	SCAD	0.037	2.537	0.094
	LASSO	0.063	1.700	0.079
Selection	Penalty	$\hat{\beta}_1^{(1)}$	$\hat{\beta}_2^{(1)}$	$\hat{\beta}_1^{(3)}$
$n = 50$				
$\hat{\lambda}_{GCV}$	SCAD	2.998(0.153)	1.496(0.153)	0.574(0.498)
	LASSO	2.883(0.178)	1.417(0.173)	0.209(0.237)
$\hat{\lambda}_{BIC}$	SCAD	2.993(0.147)	1.495(0.145)	0.287(0.464)
	LASSO	2.872(0.180)	1.407(0.181)	0.190(0.219)
$n = 100$				
$\hat{\lambda}_{GCV}$	SCAD	2.998(0.105)	1.500(0.106)	0.643(0.337)
	LASSO	2.907(0.121)	1.442(0.113)	0.256(0.211)
$\hat{\lambda}_{BIC}$	SCAD	2.990(0.100)	1.499(0.097)	0.357(0.433)
	LASSO	2.894(0.126)	1.421(0.122)	0.216(0.184)

https://doi.org/10.1371/journal.pone.0236067.t006

Case study

We now revisit the concrete slump test data set discussed in Section 1. From Fig 1, we see that slump (Y_1) and flow (Y_2) are highly correlated. We therefore used penalized GEE to perform the variable selection and parameter estimation. The resulting estimates are given in Tables 7 to 9. The second and third columns of the tables give the performance using penalized GEE with IWC for SCAD and LASSO. The fourth and fifth columns give the performance using penalized GEE with UWC. For the model selection procedures, both unweighted BIC and GCV were used to estimate the regression coefficients; their performance was similar. Therefore, we present only the results based on the unweighted BIC. Table 7 shows that SCAD with IWC selected 5 of the 7 covariates for slump (Y_1), whereas LASSO with IWC selected 4 covariates. The difference is that LASSO omitted fine aggregate (X_7). SCAD and LASSO with UWC obtained the same estimates for all the variables: they retained fine aggregate (X_7) but forced fly ash (X_2) and coarse aggregate (X_6) to zero. Table 8 shows that both SCAD and LASSO with IWC selected fly ash (X_2), water (X_4), and coarse aggregate (X_6) for flow (Y_2), but SCAD and LASSO with UWC selected only fly ash (X_2) and water (X_4). The standard errors of the estimates is lower with UWC. Table 9 shows that LASSO with IWC selected all the covariates except coarse aggregate (X_6) for CS (Y_3), whereas the other methods dropped coarse aggregate (X_6) and superplasticizer (X_5).

Concrete slump test data with artificial binary response. For illustration purposes, we create an artificial binary response variable to indicate whether or not a specimen can sustain a heavy load before distortion. For this analysis, we consider that concrete with a compressive strength below 35 is of poor quality. We therefore convert this continuous response to a binary based on the quality. Let $Y_3 = 1$ if the compressive strength is above 35, and $Y_3 = 0$

Table 7. Estimates of regression coefficients for slump (Y_1), with standard error in parentheses.

Variable	IWC		UWC	
	SCAD	LASSO	SCAD	LASSO
X_1	–	–	–	–
	–	–	–	–
X_2	-0.0297	-0.0375	–	–
	(0.0021)	(0.0013)	–	–
X_3	-0.0061	-0.0098	-0.0023	-0.0023
	(0.0001)	(0.0010)	(0.0003)	(0.0003)
X_4	0.0866	0.1222	0.0278	0.0278
	(0.0003)	(0.0025)	(0.0015)	(0.0015)
X_5	–	–	–	–
	–	–	–	–
X_6	-0.0011	-0.0017	–	–
	(0.0000)	(0.000)	–	–
X_7	0.0070	–	0.0163	0.0163
	(0.0000)	–	(0.0000)	(0.0000)

https://doi.org/10.1371/journal.pone.0236067.t007

otherwise. The goal is to apply variable selection to model the correlated continuous and binary outcomes. The resulting estimates are given in Tables 10 to 12 (the columns of these tables are the same as those for Tables 7 to 9).

Table 10 shows that SCAD with IWC selected 5 of the 7 covariates for slump (Y_1), whereas LASSO with IWC selected 4 covariates. The difference is that LASSO omitted fine aggregate (X_7). These results are similar to the independent results in Table 7, which confirms the use of IWC. SCAD with UWC forced fly ash (X_2) to zero whereas SCAD with IWC did not. LASSO with UWC selected the same variables as LASSO with IWC. Table 11 shows that all the methods selected fly ash (X_2), water (X_4), and aggregate (X_6) for flow (Y_2). Table 12 shows that all the methods except LASSO with IWC selected 5 covariates for the binary CS (Y_3). The estimates obtained with UWC have lower standard errors.

Table 8. Estimates of regression coefficients for flow (Y_2), with standard error in parentheses.

Variable	IWC		UWC	
	SCAD	LASSO	SCAD	LASSO
X_1	–	–	–	–
	–	–	–	–
X_2	-0.0529	-0.0715	-0.0169	-0.0169
	(0.0024)	(0.2544)	(0.0022)	(0.0022)
X_3	–	–	–	–
	–	–	–	–
X_4	0.2868	0.3341	0.2507	0.2507
	(0.0004)	(0.0077)	(0.0000)	(0.0000)
X_5	–	–	–	–
	–	–	–	–
X_6	-0.0033	-0.0121	–	–
	(0.0000)	(0.0031)	–	–
X_7	–	–	–	–
	–	–	–	–

https://doi.org/10.1371/journal.pone.0236067.t008

Table 9. Estimates of regression coefficients for compressive strength (Y_3), with standard error in parentheses.

Variable	IWC		UWC	
	SCAD	LASSO	SCAD	LASSO
X_1	0.1017	0.1032	0.0972	0.0972
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
X_2	0.0322	0.0337	0.0229	0.0299
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
X_3	0.0920	0.0931	0.0871	0.0871
	(0.0004)	(0.0003)	(0.0007)	(0.0007)
X_4	-0.0866	-0.0802	-0.0494	-0.0494
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
X_5	–	0.0173	–	–
	–	(0.0000)	–	–
X_6	–	–	–	–
	–	–	–	–
X_7	0.0165	0.0174	0.0119	0.0119
	(0.0000)	(0.0000)	(0.0000)	(0.0000)

https://doi.org/10.1371/journal.pone.0236067.t009

Conclusion

We have considered the selection of significant variables in multivariate multiple-response regression problems. We developed an extended GEE approach to take into account the correlation among the response variables. Our approach automatically and simultaneously selects the significant variables in high-dimensional models. We also proposed an efficient algorithm to implement the method. We performed many Monte Carlo simulations to assess the performance of the method for different sample sizes. The results showed that the methodology works well, especially when the SCAD penalty function is used together with the BIC tuning criterion. The estimates of $\boldsymbol{\beta}$ are unbiased regardless of the choice of correlation structure. We demonstrated the approach in a case study.

Table 10. Estimates of regression coefficients for slump (Y_1), with standard error in parentheses.

Variable	IWC		UWC	
	SCAD	LASSO	SCAD	LASSO
X_1	–	–	–	–
	–	–	–	–
X_2	-0.0298	-0.0375	–	-0.0173
	(0.0017)	(0.0017)	–	(0.0000)
X_3	-0.0061	-0.0098	-0.0042	-0.0071
	(0.0001)	(0.0016)	(0.0002)	(0.0002)
X_4	0.0869	0.1222	0.0494	0.0753
	(0.0003)	(0.0041)	(0.0014)	(0.0097)
X_5	–	–	–	–
	–	–	–	–
X_6	-0.0011	-0.0017	–	–
	(0.0000)	(0.000)	–	–
X_7	0.0070	–	0.0113	0.0073
	(0.0000)	–	(0.0001)	(0.0006)

https://doi.org/10.1371/journal.pone.0236067.t010

Table 11. Estimates of regression coefficients for flow (Y_2), with standard error in parentheses.

Variable	IWC		UWC	
	SCAD	LASSO	SCAD	LASSO
X_1	–	–	–	–
	–	–	–	–
X_2	-0.0529	-0.0715	-0.0192	-0.0514
	(0.0032)	(0.0672)	(0.0030)	(0.0013)
X_3	–	–	–	–
	–	–	–	–
X_4	0.2868	0.3341	0.2725	0.3171
	(0.0005)	(0.0086)	(0.0005)	(0.0088)
X_5	–	–	–	–
	–	–	–	–
X_6	-0.0034	-0.0121	-0.0041	-0.0104
	(0.0000)	(0.0011)	(0.0000)	(0.0005)
X_7	–	–	–	–
	–	–	–	–

https://doi.org/10.1371/journal.pone.0236067.t011

Table 12. Estimates of regression coefficients for binary compressive strength (Y_3), with standard error in parentheses.

Variable	IWC		UWC	
	SCAD	LASSO	SCAD	LASSO
X_1	0.0378	0.0448	0.0336	0.0431
	(0.0108)	(0.0463)	(0.0004)	(0.0039)
X_2	0.0055	0.0077	0.0018	0.0057
	(0.0045)	(0.0108)	(0.0000)	(0.0016)
X_3	0.0403	0.0471	0.0356	0.0451
	(0.0097)	(0.0430)	(0.0003)	(0.0037)
X_4	-0.0361	-0.0483	-0.0292	-0.0416
	(0.0277)	(0.0410)	(0.0007)	(0.0091)
X_5	–	–	–	–
	–	–	–	–
X_6	-0.0089	-0.0104	-0.0082	-0.0098
	(0.0002)	(0.0008)	(0.0000)	(0.0001)
X_7	–	0.0012	–	–
	–	(0.0009)	–	–

https://doi.org/10.1371/journal.pone.0236067.t012

Acknowledgments

The authors thank the editor and the two referees for their valuable comments that greatly improved the manuscript. The authors' research was supported by a grant from Natural Sciences & Engineering Research Council of Canada.

Author Contributions

Formal analysis: Anita Brobbey.

Investigation: Anita Brobbey.

Methodology: Asokan Mulayath Variyath.

Software: Asokan Mulayath Variyath, Anita Brobbey.

Validation: Anita Brobbey.

Visualization: Asokan Mulayath Variyath.

Writing – original draft: Asokan Mulayath Variyath, Anita Brobbey.

Writing – review & editing: Asokan Mulayath Variyath.

References

1. Yeh I-C. (2006). Exploring concrete slump model using artificial neural networks. *Journal of Computing in Civil Engineering*, 20, 217–221. https://doi.org/10.1061/(ASCE)0887-3801(2006)20:3(217)
2. Yeh I-C. (2007). Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, 28, 474–480. https://doi.org/10.1016/j.cemconcomp.2007.02.001
3. Yeh I-C. (2008). Modeling slump of concrete with fly ash and super plasticizer. *Computers and Concrete*, 5, 559–572. https://doi.org/10.12989/cac.2008.5.6.559
4. Breiman L. and Friedman J. H. (1997). Predicting multivariate responses in multiple regression. *Journal of Royal Statistics Society* B, 1, 3–54. https://doi.org/10.1111/1467-9868.00054
5. Chen F, Song M and Ma X. (2019) Investigation on the injury severity of serivers in rear-end collisions between cars using a random parameters bivariate ordered probit model, *International Journal of Environmental Research and Public Health*. 16(14), 2632. https://doi.org/10.3390/ijerph16142632
6. Dong B, Ma X, Chen F and Chen S (2018). Investigating the differences of single- and multi-vehicle accident probability using mixed logit model. *Journal of Advanced Transportation*. Article ID 2702360, 9 pages.
7. Sun J, Li T. N, Li F and Chen F. (2016). Analysis of safety factors for urban expressways considering the effect of congestion in Shanghai, China, *Accident Analysis and Prevention*, 95, 503–511. https://doi.org/10.1016/j.aap.2015.12.011 PMID: 26721569
8. Zeng Q., Guo Q., Wong S. C. Wen H, Huang H and Pei X. (2019). Jointly modeling area-level crash rates by severity: A Bayesian multivariate random-parameters spatio-temporal Tobit regression *Transportmetrica A: Transport Science*, 15(2), 1867–1884. https://doi.org/10.1080/23249935.2019.1652867
9. Zeng Q, Hao W, Lee J and Chen F (2020). Investigating the impacts of real-time weather conditions on freeway crash severity: A Bayesian spatial analysis. *International Journal of Environmental Research and Public Health*, 2020, 17(8), 2768. https://doi.org/10.3390/ijerph17082768
10. Moser V. C., Casey M., Hamm A., Carter W. H. Jr., Simmons J. E., and Gennings C. (2005). Neurotoxicological and statistical analyses of a yeah I learnedmixture of five organophosphorus pesticides using a ray design, *Toxicological Sciences*, 86, 101–115. https://doi.org/10.1093/toxsci/kfi163 PMID: 15800032
11. Coffey T., Gennings C.(2007a). The Simultaneous Analysis of Mixed Discrete and Continuous Outcomes Using Nonlinear Threshold Models. *Journal of Agricultural*, *Biological*, *and Environmental Statistics*. 12, 55–77. https://doi.org/10.1198/108571107X177735
12. Sammel, M. D. and Landis, J. R. (1998). Summarizing mixed outcomes for pain in intestinal cystitis: A latent variable approach, In Proceedings of the international biometric conference, 21-30.
13. Von Korff M., Ormel J., Keefe F.J., and Dworkin DS. F. (2012). Grading the severity of chronic pain,. *Pain*, 50, 133–149. https://doi.org/10.1016/0304-3959(92)90154-4
14. Fitzmaurice G. M., and Laird N. M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, 53, 110–122. https://doi.org/10.2307/2533101 PMID: 9147588
15. Liu C., and Rubin D. B. (1998). Ellipsoidally symmetric extensions of the general location models for mixed categorical and continuous data, *Biometrika*, 85, 673–688. https://doi.org/10.1093/biomet/85.3.673
16. Sammel M. D., Ryan L. M., and Legler J. M. (1997).Latent variables models for mixed discrete and continuous outcomes., *Journal of the American Statistical Association*, 90, 862–870.
17. Muthen B., and Shedden K. (1999). Finite mixture modeling with mixture out- comes using the EM algorithm, *Biometrics*, 55, 463–469. https://doi.org/10.1111/j.0006-341X.1999.00463.x PMID: 11318201
18. Liang K.Y., Zeger S.L.: Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22 https://doi.org/10.1093/biomet/73.1.13

19. Coffey T., Gennings C.(2007b). D-Optimal designs for mixed discrete and continuous outcomes analyzed with nonlinear models. *Journal of Agricultural*, *Biological*, *and Environmental Statistics*. 12, 78–95.

20. Fan J. and Li R. (2001). Variable selection via non concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96 1348–1360. https://doi.org/10.1198/016214501753382273

21. Nadarajah, T, Variyath, A.M. and Loredo-Osti, J. C, (2015). Penalized Generalized Quasi-Likelihood based Variable Selection for Longitudinal Data (with *Advances and Challenges in Parametric and Semi-parametric Analysis for Correlated Data*, Volume 218 of the series Lecture Notes in Statistics pp 233-250.

22. Tibshirani R. (1996) Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society* B 58, 267–288 (1996)

23. Sutradhar B.C., Das K.: On the efficiency of regression estimators in generalized linear models for longitudinal data. Biometrika 86, 459–465 (1999) https://doi.org/10.1093/biomet/86.2.459

24. Wang Y.G. and Carey V. (2003) Working Correlation Structure Misspecification, Estimation and Covariate Design: Implications for Generalized Estimating Equations Performance. *Biometrika*, 90, 29–41. https://doi.org/10.1093/biomet/90.1.29

25. Shults J., Mazurick C., and Landis J.R. (2006), Analysis of repeated bouts of measurements in the framework of generalized estimating equations. *Statistics in Medicine* 25, 4114–4128. https://doi.org/10.1002/sim.2515

26. Craven P. and Wahba G. (1979). Smoothing noise data with spline functions: validation. *Numerische Mathematika*. 31, 377–403. https://doi.org/10.1007/BF01404567

27. Schwarz G.: Estimating the dimension of a model. 1978, *Annals of Statistics* 6, 461–464 https://doi.org/10.1214/aos/1176344136

28. Dziak, J. J., (2006). Penalized quadratic inference functions for variable selection in longitudinal research. Phd thesis, Pennsylvania State University.

29. Dziak, J. J., Li, R., (2007). An overview on variable selection for longitudinal data. *Quantitative Medical Data Analysis. Singapore: World Sciences*.

30. Amatya A, Demirtas H. (2015). OrdNor: An R Package for Concurrent Generation of Correlated Ordinal and Normal Data, *Journal of Statistical Software*, *Code Snippets*, 68(2), 1–14.

31. Demirtas H, Doganay B (2012). Simultaneous Generation of Binary and Normal Data with Specified Marginal and Association Structures, *Journal of Biopharmaceutical Statistics*, 22 (2), 223–236. https://doi.org/10.1080/10543406.2010.521874 PMID: 22251171