# Variable Selection in Multivariate Multiple Regression

by

**©Anita Brobbey**

*A thesis submitted to the School of Graduate Studies*

*in partial fulfillment of the requirement for the Degree of*

*Master of Science*

**Department of Mathematics and Statistics**

**Memorial University**

St. John's Newfoundland and Labrador, Canada September 2015

# Abstract

Multivariate analysis is a common statistical tool for assessing covariate effects when only one response or multiple response variables of the same type are collected in experimental studies. However with mixed continuous and discrete outcomes, traditional modeling approaches are no longer appropriate. The common approach used to make inference is to model each outcome separately ignoring the potential correlation among the responses. However a statistical analysis that incorporates association may result in improved precision. Coffey and Gennings (2007a) proposed an extension of the generalized estimating equations (GEE) methodology to simultaneously analyze binary, count and continuous outcomes with nonlinear functions. Variable selection plays a pivotal role in modeling correlated responses due to large number of covariate variables involved. Thus a parsimonious model is always desirable to enhance model predictability and interpretation. To perform parameter estimation and variable selection simultaneously in the presence of mixed discrete and continuous outcomes, we propose a penalized based approach of the extended generalized estimating equations. This approach only require to specify the first two marginal moments and a working correlation structure. An advantageous feature of the penalized GEE is that the consistency of the model holds even if the working correlation is misspecified. However it is important to use appropriate working correlation structure in small samples since it improves the statistical efficiency of the regression parameters. We develop a computational algorithm for estimating the parameters using local quadratic approximation (LQA) algorithm proposed by Fan and Li (2001). For tuning parameter selection, we explore the performance of unweighted Bayesian information criterion(BIC) and generalized cross validation (GCV) for least absolute shrinkage and selection operator(LASSO) and smoothly clipped absolute deviation (SCAD). We discuss the asymptotic properties for the penalized GEE estimator when the number of subjects $n$ goes to infinity. Our simulation studies reveal that when correlated mixed outcomes are available, estimates of regression parameters are unbiased regardless of the choice of correlation structure. However, estimates obtained from the unstructured working correlation (UWC) have reduced standard errors. SCAD with BIC tuning criteria works well in selecting important variables. Our approach is applied to concrete slump test data set.

# Contents

# List of Tables

# List of Figures

# Acknowledgements

I thank the Almighty God for his blessings throughout my education. I would like to express my deepest gratitude to my supervisor, Dr. Asokan Mulayath Variyath, for his patience, invaluable assistance, constructive criticisms and commitment which brought my thesis to a successful completion.

I gratefully acknowledge the funding received towards my MSc. from School of Graduate Studies, the Department of Mathematics & Statistics, and my supervisor in the form of graduate fellowships and teaching assistantships.

I also appreciate the assistance I received from the faculty members and staff of our department over the last two years, I am very grateful.

My warmest gratitude goes to my dear parents and guardians for their love, encouragement and constant support.

Finally, I extend my sincere thanks to graduate students in the department, friends and loved ones who contributed in various ways to make this work a success.

# Dedication

This thesis is dedicated to my father, Kofi Brobbey

and to the memory of my mother, Agartha Opoku

in appreciation for their unrelenting love and immense support.

# Chapter 1

# Introduction

In many applied science or public health studies, researchers are interested in modeling the relationship between response variable(s) and explanatory variables (independent variables). For example, in the well known Framingham Heart Study (Kannel et al., 1961), many covariates including age, sex, smoking status, cholesterol level, blood pressure were recorded on the participants over the years to identify risk factors for coronary heart disease. Despite the large number of covariates, some of them have no influence on the response variable. In some studies, the number of explanatory variables can be considerably large due to addition of interaction effects of covariates. If there are more than one response of interest, then the number of model parameters to be estimated will be much higher. Moreover, a model with all covariates may lead to an over-fitting problem. Thus, parameter estimation and variable selection are two important problems in multivariate regression analysis. Selecting a smaller number of important variables results in a simpler and interpretable model. In this thesis, we address the variable selection problem in multivariate multiple regression models.

## 1.1 Modelling Multiple Outcomes

Multivariate multiple regression analysis is a common statistical tool for assessing covariate effects when only one response or multiple response variables are collected in observational or experimental studies. Many multivariate regression techniques are designed for univariate response cases. A common approach to dealing with multiple response variables is to apply the univariate response regression technique separately on each response variable ignoring the joint information among the responses. To solve this multi-response regression problem, several methodologies in generalized linear model (GLM) framework have been proposed in literature.

### 1.1.1 Literature Review

Breiman and Friedman (1997) proposed the curd and whey method that uses the correlation among response variables to improve predictive accuracy. They showed that their method can outperform separate univariate regression approaches but did not address variable selection. In general, using multivariate multiple linear regression is more appropriate in investigating relations between multiple response (Goldwasser & Fitzmaurice 2006). The analysis of multivariate outcomes is especially challenging when multiple types of outcomes are observed, the methodology is comparatively scarce when each response is to be modeled with a nonlinear function. However, multivariate outcomes of mixed types occur frequently in many research areas including dose-response experiment in toxicology (Moser et al 2005; Coffey & Gennings, 2007a, 2007b), birth defects in teratology (Sammel, Ryan & Legler, 1997) and pain in public health research (Von Korff et al 1992; Sammel & Landis, 1998). In the last three decades, methodologies for mixed-type outcomes includes using factorization approaches based on extensions of the general location model proposed by Fitzmaurice and Laird (1997) and Liu and Rubin (1998). These likelihood based methodologies factor the joint distribution of the random variables as the product of marginal and conditional distributions, but can be unattractive because of their dependence on parametric distributional assumptions. Sammel et al. (1997) proposed a latent variable model for cross-sectional mixed outcomes using generalized linear mixed model with continuous latent variables, allowing covariate effects on both the outcomes and the latent variables. Muthen and Shedden (1999) proposed a general latent variable modeling framework that incorporates both continuous and categorical outcomes and associates separate latent variables for outcomes of each type. Miglioretti (2003) also developed a methodology based on latent transition regression model for mixed outcomes. Other authors (Prentice and Zhao 1991; Rochon 1996; Bull 1998; Gray and Brookmeyer 2000; Rochon and Gillespie 2001) handled mixed outcomes through modification of generalized estimating equation (GEE) of Liang and Zeger (1986). Although some of these approaches (Lefkopoulou, Moore, and Ryan 1989; Contreras and Ryan 2000) may incorporate the use of GEEs for nonlinear models, none of the methodologies have formally extended the modeling of mixed discrete and continuous outcomes to nonlinear functions. Coffey and Gennings (2007a) proposed an extension of the generalized estimating equation (GEE) methodology to simultaneously analyze binary, count, and continuous outcomes with nonlinear models that incorporates the intra-subject correlation. The methodology uses a quasi-likelihood framework and a working correlation matrix. The incorporation of the intra-subject correlation resulted in decreased standard errors for the parameters. In addition, Coffey and Gennings (2007b) developed a new application to the traditional D-optimality criterion to create an optimal design for experiments measuring mixed discrete and continuous outcomes that are analyzed with nonlinear models. These designs are to choose the location of the dose groups and proportion of total sample size that result in a minimized generalized variance. The designs were generally robust to different correlation structures. Coffey and Gennings (2007b) observed a substantial gain in efficiency compared to optimal designs created for each outcome separately when the expected correlation was moderate or large. In this thesis, we use the GEE approach (Coffey and Gennings, 2007a) and also conduct a series of simulations to investigate the performance of their method. Since we use GEE approach, we briefly review it in the next section.

## 1.2 Generalized Estimating Equation (GEE)

### 1.2.1 Generalized Linear Models (GLM)

Nelder and Wedderburn (1972) introduced the class of generalized linear models (GLMs) which extends ordinary model to encompass non-normal response distributions and modeling of the mean. The distribution of $\boldsymbol{y}$ is a member of an exponential family such as the Gaussian, binomial, Poisson or inverse-Gaussian. For a GLM, let $E(\boldsymbol{y}|\boldsymbol{X})$ denote the conditional expectation of the response variable, $\boldsymbol{y}$ given the covariates, $\boldsymbol{X}$ and $g(\cdot)$ denote a known link function then

$$\boldsymbol{\mu} = E(\boldsymbol{y}|\boldsymbol{X}) = g(\boldsymbol{X}\boldsymbol{\beta}) \tag{1.1}$$

where $\boldsymbol{\beta}$ is the vector of unknown regression coefficients to be estimated. Generalized linear models consists of three components, the random, systematic and link component.

- A *random component* specifying the conditional distribution of the response variable $Y$ given the explanatory variables $X$. The densities of the random component can be written in the form,

  $$f(y \mid \theta, \phi) = exp\Big(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\Big)$$

  where $a(\cdot), b(\cdot)$ and $c(\cdot)$ are arbitrary known functions, $\phi$ is the dispersion parameter and $\theta$ is the canonical parameter of the distribution.

- A *systematic component* specifying a linear predictor function. For each subject $i$,

  $$\eta_i(\beta) = x_i^T\beta.$$

- A *link function*, $g(\cdot)$ defines the relationship between the linear predictor $\eta_i$ and the mean $\mu_i$ of $Y_i$.

  $$g(\mu_i) = \eta_i(\beta) = x_i^T\beta.$$

For GLMs, estimation starts by defining a measure of the goodness of fit between the observed data and the fitted values generated by the model. The parameter estimates are values that minimize the goodness-of-fit criterion, we obtain the parameter estimates by maximizing the likelihood of the observed data. The log-likelihood based on a set of independent observations $y_1, y_2, y_3, ..., y_n$ with density $f(y_i; \beta)$ is

$$\ell(\mu; y) = \sum_{i=1}^{n} \log f(y_i; \beta).$$

The goodness-of-fit criterion is

$$D(y; \mu) = 2\ell(y; y) - 2\ell(\mu; y).$$

This is called the *scaled deviance*. Deviance is one of the methods used for model checking and inferential comparisons. The greater the scaled deviance, the poorer the fit.

### 1.2.2 Quasi-Likelihood (QL) Functions

Wedderburn (1974) proposed to use the quasi-score function, which assumes only a mean-variance relationship to estimate regression coefficients, $\beta$ without fully specifying the distribution of the observed data, $y_i$. The score equation is of the form,

$$S(\beta) = \sum_{j=1}^{n} S_i(\beta) = \sum_{j=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T Var^{-1}(y_i; \beta, \phi)(y_i - \mu_i(\beta)) = 0. \tag{1.2}$$

To obtain the score function, the random component in the generalized estimating equations was replaced by the following assumptions:

$$E[Y_i] = \mu_i(\beta),$$

$$\text{Var}[Y_i] = V_i = a(\phi)V(\mu_i).$$

Consider independent vector of responses $Y_1, Y_2, ..., Y_n$ with common mean $\mu$ and covariance matrix $a(\phi)V(\mu)$.

The quasi-likelihood function is

$$Q(\mu; y) = \sum_{j=1}^{n} \int_{y}^{\mu} \frac{y - t}{a(\phi)V(t)} dt. \tag{1.3}$$

The quasi-score function is

$$S(\beta; y) = \frac{\partial Q}{\partial \beta} = \sum_{j=1}^{n} \frac{y - \mu}{a(\phi)V(\mu)},$$

where $S(\beta; y)$ possess the following properties: replaced by the following assumptions:

$$E[S] = 0$$

$$\text{Var}[S] = -E\left(\frac{\partial S}{\partial \mu}\right)$$

These properties form the basis of most asymptotic theory for likelihood-based inference. Thus in general, $S$ behaves like a score function and $Q$ like a log-likelihood function. The quasi-score function, $S(\beta; y)$ would be the true score function of $\beta$ if $Y_i$'s have a distribution in the exponential family. We find the value $\beta_{QL}$ that maximizes $Q$ by setting $S(\beta_{QL}; y) = 0$,this is called QL estimating equations. In matrix form, we can express the score equation as;

$$S(\beta; y) = \frac{D^T V^{-1}(y - \mu)}{\phi},$$

where $\mathbf{D}$ is the $n \times p$ matrix with $(i,j)th$ entry $\partial\mu_i/\partial\beta_j$,$\mathbf{V}$ is the $n \times n$ diagonal matrix with $i^{th}$ diagonal entry $V(\mu_i)$, $\mathbf{y} = (y_1, y_2, ..., y_n)$, and $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_n)$. The covariance matrix of $S(\beta)$ plays the same role as Fisher information matrix in the asymptotic variance of $\beta$;

$$I_n = D^T V^{-1} D,$$

$$Var(\hat{\beta}) = I_n^{-1}.$$

These properties are based only on the correct specification of the mean and variance of $Y_i$.

Method of moments is used for the estimation of $a(\phi)$.

$$a(\hat{\phi}) = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu_i})^2}{V(\hat{\mu_i})} = \frac{\chi^2}{n-p},$$

where $\chi^2$ is the generalized Pearson statistics. As in the case of GLM, the quasi-deviance function corresponding to a single observation is

$$D(y;u) = -2\sigma^2 Q(\mu;y) = 2\int_{\mu}^{y} \frac{y-t}{V(t)}dt. \tag{1.4}$$

The deviance function for the complete observation $y$ when the observations are independent is defined as $D(y:\mu) = \sum_{i=1}^{n} D(y_i:\mu_i)$

### 1.2.3 Generalized Estimating Equation (GEE)

The GEE approach was first developed by Liang and Zeger (1986) for longitudinal data. Suppose we have a random sample of observations from $n$ individuals. For each individual $i$ we have a vector of responses $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ and corresponding covariates $X_i = (X'_{i1}, X'_{i2}, \ldots, X'_{in_i})'$, where each $Y_{ij}$ is a scalar and $X'_{ij}$ a $p$-vector. In general, the components of $Y_i$ are correlated but $Y_i$ and $Y_k$ are independent for any $i \neq k$ given the covariates. To model the relationship between the response and covariates one can use a regression model similar to the generalized linear model(GLM): (see equation (1.1)). The GEE approach suggests estimating $\boldsymbol{\beta}$ by solving the following estimating equation.(Liang and Zeger,1986)

$$S(\beta) = \sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i) = 0, \tag{1.5}$$

where $D_i = \partial\mu_i\beta/\partial\beta'$ and $V_i$ is a working covariance matrix of $Y_i$ and $V_i = A_i^{1/2}R(\alpha)A_i^{1/2}$ where $R(\alpha)$ is a working correlation matrix and $A_i$ is a diagonal matrix with elements $\text{var}(Y_{ij}) = \phi V(\mu_{ij})$ which is specified as a function of the mean $\mu_{ij}$. The correlation parameter $\alpha$ can be estimated through the method of moments or another set of estimating equations. The GEE can be regarded as a quasi-likelihood (QL) score equation.

## 1.3 Variable Selection

The problem of predicting the response using high-dimensional covariates has always been an important problem in statistical modeling. Researchers are often interested in selecting a smaller number of important variables to adequately represent the relationship and obtain a more interpretable model. To select the best and simplest model, several model selection techniques have been developed in recent years especially for linear models and generalized linear models(GLM). In this section, we discuss existing variable selection approaches as well as their advantages and disadvantages.

### 1.3.1 Sequential Approaches

Sequential model selection methods include forward selection, backward elimination and stepwise regression. Forward selection starts with intercept alone model and sequentially adds the most significant variable that improves the model fit. The problem with forward selection is that, the addition of a new variable may render one or more of the already included variables redundant. Alternately, the backward elimination starts with the full model with all the variables in the model, then sequentially eliminates the least significant variable. The final model is obtained when either no variables remain in the model or the criteria for removal is not met. Backward elimination has drawbacks, for example a variable dropped in the process may be significant when added to the final reduced models. Thus, stepwise regression has been proposed as a technique that combines advantages of forward selection and backward elimination. In this approach, we consider both forward selection and backward elimination at each step and uses the thresholds to determine if the variable needs to be added or dropped or the selection should stop. Stepwise regression evaluates more subsets than the other two techniques, so in practice it tends to produce better subsets (Miller, 1990). However, there is no strong theoretical results for comparing the effectiveness of stepwise regression against forward selection or backward elimination.

### 1.3.2 Information Criteria

Information criterion selects the best model from all possible subset models. Akaike's information criterion (AIC)(Akaike, 1973) and Schwarz's bayesian information criterion (BIC)(Schwartz, 1978) are the most widely used information criteria. The criteria consist of a measure of model fit based on the log-likelihood, $\ell(X(s), y, \beta(s))$ of submodel $s$ and a penalty term, $q(k, n)$ with $k$ being the number of parameters for model complexity and $n$, the number of observations that contributes to the likelihood. The general form of an information criteria of submodel $s$ is defined to be

$$-2\ell(X(s), y, \beta(s)) + q(k, n).$$

Typical choices of the penalty term for AIC and BIC include:

- **Akaike's information criterion (AIC)**

$$q(k, n) = 2k.$$

- **Bayesian information criterion (BIC)**

$$q(k, n) = klog(n).$$

For linear regression with Gaussian assumption, Mallow's $C_p$ (Mallows, 1973) is equivalent to AIC. Under information criteria, the first step is to calculate the chosen information criterion for all possible models and the model with the minimum value for the information criterion is then declared optimal. Information criteria approaches are computationally inefficient due to evaluation of all possible models.

### 1.3.3 Penalized Likelihood Methods

Traditional approaches such as $C_p$ (Mallow's, 1973), Akaike's information criterion (Akaike, 1974) and Bayesian information criterion (Schwarz, 1978) cease to be useful due to computational infeasibility and model non-identifiability. Recently developed approaches based on penalized likelihood methods have been proved to be an attractive approach both theoretically and empirically for dealing with these problems. In addition, all variables are considered at the same time which may lead to better global submodel. Penalized regression estimates a sparse vector of regression coefficients by minimizing an objective function that is composed of a loss function subject to a constraint on the coefficients. A general form proposed by Fan and Li (2001) is defined by

$$\ell_p(\beta) = \ell(\beta \mid y, X) - n \sum_{j=1}^{p} P_{\lambda_n}(|\beta|), \tag{1.6}$$

where $X$ is the matrix of covariates, $y$ is the response vector, $\beta$ is the regression coefficient vector, $P_{\lambda_n}$ is a penalty function and $\lambda_n$ is the tuning parameter which controls the degree of penalization. Maximizing (1.6) leads to simultaneous estimation and variable selection of the regression model. The mostly used penalty functions includes the least shrinkage and selection operator (LASSO; Tibshirani 1996, 1997), Bridge (Fu, 1998), the smoothly clipped absolute deviation (SCAD; Fan and Li 2001), Elastic Net (Zou and Hastie, 2003) and other extended forms.

## 1.4 Motivation and Proposed Approach

Generalized estimating equation (GEE) is playing an increasingly important role in the analysis of correlated outcomes. Recently, Coffey and Gennings (2007a) proposed an extension of the GEE methodology to simultaneously analyze binary, count and continuous outcomes with nonlinear function. However, the joint model for all responses results in high dimension of covariates therefore selecting significant variables become necessary in model building. Several model selection methods have been developed to select the best submodel. Sequential approaches have been found to be unstable in the selection process: a small change in the data could cause a very different selection. This is partially because once a covariate has been added to (dropped from) the model at any step, it is never removed from (added to) the final model. Information Criteria approaches such as AIC and BIC are computationally inefficient due to evaluation of all possible models. Penalization based methods such as LASSO and SCAD have continuous selection procedure and hence it provides more robust selection results. Penalized likelihood methods are computationally efficient and have been proved to be attractive both theoretically and empirically. In order to deal with high dimensionality in the mixed continuous and discrete outcomes model, it is preferred to use penalization based variable selection approach of the extended GEE approach (Coffey & Gennings, 2007a). In our study, we have developed a penalized GEE approach to multi-response regression problem using LASSO and SCAD penalty functions. We conduct a series of simulations to investigate the performance of our proposed approach using both independent working correlation (IWC) and unstructured working correlation (UWC). Our simulation studies showed that the proposed methodology work well and helps improve precision.

The remaining part of the thesis is organized as follows. In Chapter 2, we briefly review properties of the LASSO and SCAD penalty functions and discuss local quadratic approximation (LQA) algorithm and estimation of standard error of parameters proposed by Fan and Li (2001). We introduce generalized estimating equations (GEE) for mixed outcomes and then discuss our proposed penalization based approach, the computational algorithm, the tuning parameter selection problem and asymptotic properties. In Chapter 3, we investigate the performance of our approach with LASSO and SCAD penalty functions through simulation, in the context of continuous, binary and count outcomes with both unstructured working correlation (UWC) and independent working correlation (IWC). In Chapter 4, we apply our method to concrete slump test data set. Our concluding remarks are provided in Chapter 5.

# Chapter 2

# Penalized Generalized Estimating Equations(GEE)

In this chapter, we review properties of the LASSO and SCAD penalty functions as members of the penalized likelihood family and introduce the local quadratic approximation (LQA) algorithm proposed by Fan and Li (2001). We briefly introduce generalized estimating equations (GEE) for mixed outcomes. We then discuss our proposed penalized based approach.

Suppose $\ell(y_i; \boldsymbol{X_i \beta})$ denote the loss function (log-likelihood or log-quasi-likelihood) of $\boldsymbol{\beta}$ then a general form of the penalized likelihood is defined by

$$\sum_{i=1}^{n} \ell(y_i; \boldsymbol{X_i}\boldsymbol{\beta}) - n \sum_{i=1}^{p} p_\lambda(|\beta_j|), \tag{2.1}$$

where $p_\lambda(|\beta_j|)$ is a penalty function, and $\lambda$ is the tuning parameter.

## 2.1 Penalty Functions and Optimization

### 2.1.1 LASSO

The least absolute shrinkage and selection operator (LASSO) was proposed by Tibshirani (1996) which performs parameter estimation and shrinkage there by variable selection automatically. The LASSO penalty function is the $L_1$ penalty, $p_{\lambda_n}(|\beta|) = \lambda_n|\beta|$. We obtain the penalized estimates of the LASSO regression by maximizing the function:

$$\ell_p(\beta) = \sum_{i=1}^{n} \ell(y_i; \boldsymbol{X_i}\boldsymbol{\beta}) - n\lambda_n \sum_{j=1}^{p} |\beta_j|, \tag{2.2}$$

where $\lambda_n$ controls the variable selection as $\lambda_n$ increases model parsimony increases as more variables are selected out of the model. This is known as soft thresholding. LASSO is closely related with ridge regression. Ridge regression is a popular regularization technique proposed by Hoerl and Kennard (1970). Equation (2.1) results in a ridge penalized regression model when $p_{\lambda_n}(|\beta|) = \lambda_n |\beta|^2$, called $L_2$ penalty. Equivalently the solution $\hat{\beta}^{ridge}$ can be written as follows

$$\hat{\beta}^{ridge}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left[ \sum_{i=1}^{n} \ell(y_i; \boldsymbol{X_i \beta}) - n\lambda_n \sum_{j=1}^{p} \beta_j^2 \right]. \tag{2.3}$$

Efficient ways to compute the analytic solution for $\hat{\beta}^{ridge}$ along with its properties are presented in Hastie et al. (2001). Ridge ($L_2$) and the LASSO ($L_1$) are special cases of $L_\lambda(\lambda > 0)$ penalties. Zou and Hastie (2005) proposed Elastic Net which combines the Ridge and LASSO constraints to allow both stability with highly correlated variables and variable selection.

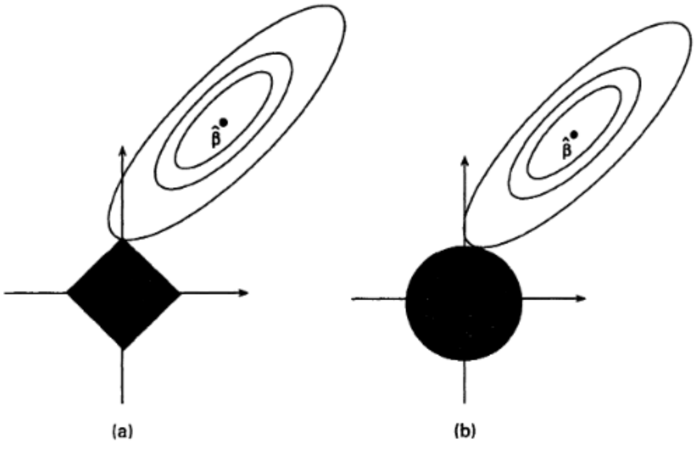

Figure 2.1: Geometry of LASSO vs Ridge
Estimation picture for (a) the LASSO and (b) ridge regression

Figure 2.1 (a) results in sparsity, the LASSO solution is the first place that the contours touch the square and this sometimes occur at a corner corresponding to a zero coefficient. On the contrary, Figure 2.1 (b) depicting ridge regression solution has no corners for the contours to hit hence zero solutions will rarely result.

### 2.1.2 SCAD

Fan and Li (2001) argued that an ideal penalty function should yield an estimator with the following three properties;

1. Unbiasedness: The estimator is nearly unbiased when the true unknown parameter is large to reduce model bias.
2. Sparsity: The estimator is a thresholding rule which automatically sets small estimated coefficients to zero to reduce model complexity.
3. Continuity: The estimator is continuous in the data to reduce instability in model prediction.

In contrast, the convex LASSO penalty ($L_1$ penalty) does not satisfy the unbiasedness condition, the convex $L_q$ penalty with $q > 1$ does not satisfy the sparsity condition and the concave $L_q$ penalty with $0 \leq q < 1$ does not satisfy the continuity condition. Thus, Fan and Li (2001) proposed a non-concave penalty function referred to as the

Smoothly Clipped Absolute Deviation (SCAD) which simultaneously achieves the three desirable properties: unbiasedness, sparsity and continuity. The SCAD penalty function is continuous and the first derivative for some $a > 2$ and $\beta > 0$ is

$$p'_\lambda(\beta) = \lambda\left\{ I(\beta > \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\}. \tag{2.4}$$

The SCAD function is given by

$$p_\lambda(\beta_j) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| \leq \lambda \\ -\left(\dfrac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}\right) & \text{if } \lambda < |\beta_j| \leq a\lambda \\ \dfrac{(a-1)\lambda^2}{2} & \text{if } |\beta_j| > a\lambda. \end{cases} \tag{2.5}$$

The SCAD penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$, but not differentiable at zero. Its derivative vanishes outside $[-a\lambda, a\lambda]$. As a consequence, SCAD penalized regression can produce sparse set of solution and approximately unbiased coefficients for large coefficients.

In Figure 2.2, we sketch the LASSO penalty along with the SCAD. Both penalty functions are equal to zero when the regression coefficient is equal to zero. It is seen that for small values SCAD is similar to the LASSO penalty whereas for larger values SCAD levels off. The SCAD improves the LASSO by reducing the estimation bias. Following Fan and Li (2001), let the parameter vector $\boldsymbol{\beta}$ be partitioned into $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$ and assume $\boldsymbol{\beta}_2 = 0$, with $\boldsymbol{J_1}(\boldsymbol{\beta_1})$ denoting Fisher information matrix

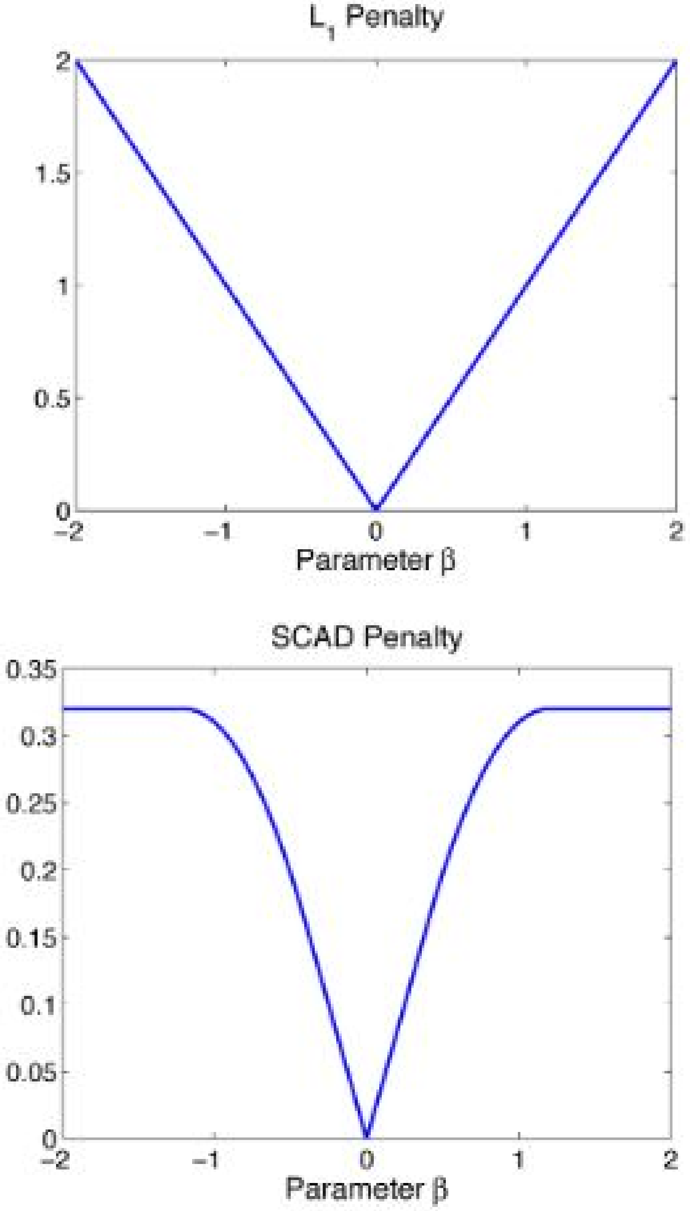

Figure 2.2: LASSO (top) and SCAD (down) penalty functions

given $\boldsymbol{\beta} = 0$ . Under some regularity conditions, it may be shown that $\hat{\boldsymbol{\beta}}^T = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)$ satisfies the oracle properties, since $\hat{\boldsymbol{\beta}}_2 \xrightarrow{P} 0$ and $\hat{\boldsymbol{\beta}}_1$ is asymptotic normal with covariance matrix $\boldsymbol{J_1(\beta_1)^{-1}}$ if $n^{-1/2}\lambda_n \to \infty$. To obtain a penalized maximum likelihood estimator of $\boldsymbol{\beta}$, we maximize (2.1) with respect to $\boldsymbol{\beta}$ for some thresholding parameter $\lambda$. For computational purposes, Fan and Li (2001) used quadratic functions to locally approximate the penalty function.

### 2.1.3 Local Quadratic Approximation (LQA) Algorithm

Suppose we choose an initial value $\boldsymbol{\beta}_0$ near the maximizer of (2.1). If the $j$th component of $\boldsymbol{\beta}_0$, $\beta_{j0}$ is very close to zero, then set $\hat{\beta}_{j0} = 0$, otherwise, the penalty $P_\lambda(|\beta_j|)$ can be approximated as

$$P_\lambda(|\beta_j|) \approx P_\lambda(|\beta_{j0}|) + \frac{1}{2}\Big\{P'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\Big\}(\beta_j^2 - \beta_{j0}^2),$$

for $\beta_j \approx \beta_{j0}$. In other words,

$$[P_\lambda(|\beta_j|)]' = P'_\lambda(|\beta_j|)sgn(\beta_j) \approx \{P'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}\beta_j, \text{ when } \beta_j \neq 0.$$

This method significantly reduces the computational burden. However, a drawback of this approximation is that once a coefficient is shrunken to zero, it will be excluded from the final selected model. The maximization problem (2.1) can be reduced to a quadratic maximization problem assuming that the log-likelihood function is smooth with respect to $\boldsymbol{\beta}$ so that its first two partial derivatives are continuous. Thus using Taylor expansion, the first term in (2.1) can be locally approximated by

$$\ell(\boldsymbol{\beta}_0)+\nabla\ell(\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\nabla^2\ell(\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+\frac{1}{2}n\boldsymbol{\beta}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta} \quad (2.6)$$

with $\nabla\ell(\boldsymbol{\beta}_0)=\dfrac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}$, $\nabla^2\ell(\boldsymbol{\beta}_0)=\dfrac{\partial^2\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$ and

$$\boldsymbol{\Sigma_\lambda(\beta_0)}=\mathrm{diag}(P'_\lambda(|\beta_{10}|)/|\beta_{10}|,\ldots,P'_\lambda(|\beta_{p0}|)/|\beta_{p0}|)$$

With the aid of this local quadratic approximation, Newton-Raphson (N-R) algorithm can be used to maximize (2.1) iteratively. The estimate of $\hat{\boldsymbol{\beta}}$ at the $(r+1)$th iterative step is

$$\boldsymbol{\hat{\beta}_{r+1}}=\boldsymbol{\hat{\beta}_r}-\left\{\nabla^2\ell(\boldsymbol{\beta}_r)-n\boldsymbol{\Sigma_\lambda(\hat{\beta}_r)}\right\}^{-1}\left\{\nabla\ell(\boldsymbol{\beta}_r)-n\boldsymbol{U_\lambda(\hat{\beta}_r)}\right\}, \quad (2.7)$$

with $\boldsymbol{U_\lambda(\hat{\beta}_r)}=\boldsymbol{\Sigma_\lambda(\beta_r)\beta_r}$. We iterate this algorithm until convergence.

A perturbed version of the LQA, the Minorization-Maximization (MM) (Hunter and Li (2005)) algorithms have been introduced which alleviates a drawback of backward stepwise variable selection in LQA, but it is difficult to choose the size of perturbation. LQA and MM share the convergence properties of the modified N-R algorithm, using a robust local quadratic approximation. In both cases, the Hessian matrix is guaranteed to be positive definite, driving convergence at least to a local maximum.

### 2.1.4 Standard Error

Fan and Li (2001) recommend estimating the covariance matrix of the non-vanishing (non-zero) component of $\hat{\boldsymbol{\beta}}$ via sandwich formula:

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \left[\nabla^2\ell(\hat{\boldsymbol{\beta}}) - n\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}})\right]^{-1}\widehat{\text{cov}}\left\{\nabla\ell(\hat{\boldsymbol{\beta}})\right\}\left[\nabla^2\ell(\hat{\boldsymbol{\beta}}) - n\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}})\right]^{-1}. \tag{2.8}$$

Fan and Li (2001) showed that the LASSO penalty proposed by Tibshirani (1996) has good performance when the signal to noise ratio is large, but creates excessive biases compared to using the SCAD penalty.

## 2.2 GEE for Mixed Outcomes

In a longitudinal study of $n$ subjects, if the investigators are mainly interested in the covariate effect on the response variable, Liang and Zeger (1986) proposed the GEE model based on the marginal distributions of the response. In a cross-sectional study with multiple responses, Coffey and Gennings (2007a) used GEE approach to estimate the parameters. Let the observations $(y_i^m, x_i^m)$ denote the response and covariate respectively for the $m$th response ($m = 1, 2, \dots, M_i$ ) measured on subject $i = 1, \dots, n$. The $M_i \times 1$ vector of responses for the $i$th subject is $\mathbf{y} = (y_i^{(1)}, y_i^{(2)}, ..., y_i^{(M_i)})$.

To apply quasi-likelihood method to the analysis, we define the first two moments of $y_i^{(m)}$;

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

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

where $h^{(m)}(\cdot)$ is a known function, $s^{(m)}$ is a scaling parameter, $f^{(m)}(\cdot)$ is the nonlinear function of the coefficients and $\boldsymbol{\beta}^{(m)}$ is a $p^{(m)} \times 1$ vector of model coefficients for the $m$th response variable. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)^T}, \boldsymbol{\beta}^{(2)^T}, \dots, \boldsymbol{\beta}^{(M)^T})^T$ be the $p \times 1$ vector of model parameters for all M outcomes, where $p = (p^{(1)} + p^{(2)} + \dots + p^{(M)})$. In the quasi - likelihood framework with multiple outcomes, the regression coefficients $\boldsymbol{\beta}$ can be estimated by solving the Generalized Estimating Equations (GEEs)

$$\boldsymbol{S(\beta)} = \sum_{i=1}^{n} \boldsymbol{D_i^T V_i^{-1} r_i} = \mathbf{0}. \tag{2.9}$$

For each subject $i$, let

$$\mathbf{D}_i = \begin{pmatrix} \dfrac{\partial \mu_i^{(m)}}{\partial \boldsymbol{\beta}^{(1)^T}} & \mathbf{0}^{\boldsymbol{T}} & \cdots & \mathbf{0}^{\boldsymbol{T}} \\ \mathbf{0}^{\boldsymbol{T}} & \dfrac{\partial \mu_i^{(m)}}{\partial \boldsymbol{\beta}^{(2)^T}} & \cdots & \mathbf{0}^{\boldsymbol{T}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^{\boldsymbol{T}} & \mathbf{0}^{\boldsymbol{T}} & \cdots & \dfrac{\partial \mu_i^{(m)}}{\partial \boldsymbol{\beta}^{(M_i)^T}} \end{pmatrix},$$

be a $M_i \times p$ full-rank derivative matrix, $\boldsymbol{r_i = (y_i - \mu_i)}$ be a $M_i \times 1$ vector of residuals and $\boldsymbol{V_i} = \boldsymbol{A_i^{1/2}R_i(\alpha)A_i^{1/2}}$, the $M_i \times M_i$ working covariance matrix of $\boldsymbol{y_i}$. Here, $\boldsymbol{A_i}$ $= \mathrm{diag}(\sigma_i^{2(1)}, \sigma_i^{2(2)}, \ldots, \sigma_i^{2(M_i)})$ is a $M_i \times M_i$ diagonal matrix of $\mathrm{var}(y_i^{(m)})$ and $\boldsymbol{R_i(\alpha)}$ is a $M_i \times M_i$ working correlation matrix parameterized with parameter vector $\boldsymbol{\alpha}$. The GEE estimator $\hat{\boldsymbol{\beta}}$ is asymptotically consistent as n goes to infinity. In the presence of high dimensional covariates we extend (2.9) to penalized estimating equations. Thus a penalty term can be incorporated with the aim of adjusting the model to facilitate the estimation of unbiased parameter estimates.

## 2.3 Variable Selection via Penalized GEE

Fu (2003) proposed a generalization of the bridge and LASSO penalties to GEE models, which minimizes the penalized deviance criterion

$$D(\boldsymbol{\beta; X, y}) + P(\boldsymbol{\beta}), \tag{2.10}$$

where $D(\boldsymbol{\beta; X, y}) = 2\ell(\boldsymbol{y}; \boldsymbol{y}) - 2\ell(\boldsymbol{\mu}; \boldsymbol{y})$ (McCullagh and Nelder, 1989) with log-likelihood $\ell(\boldsymbol{\mu}; \boldsymbol{y})$ and $P(\boldsymbol{\beta}) = \lambda \sum_j |\beta_j|^q$, given $q > 0$. The LASSO estimator is defined to be a special case with $q = 1$ (Tibshirani, 1996). This leads to solving penalized equations;

$$\begin{cases} F_1(\beta, X, y) + \dot{P}_1 = 0 \\ \dots \\ F_p(\beta, X, y) + \dot{P}_p = 0 \end{cases}, \tag{2.11}$$

where $F_j(\beta, X, y)$ is the $j$th score of the likelihood and $\dot{P}_j = \lambda \sum_j q|\beta_j|^{q-1} sgn(|\beta_j|)$. This could be generalized to GEE quasi-score function equations.

$$\sum_{i=1}^{n} \boldsymbol{D_i^T V_i^{-1} r_i - n\dot{P}_\lambda(\beta) = 0}, \tag{2.12}$$

where $\dot{\mathbf{P}}_\lambda(\boldsymbol{\beta}) = \partial P_\lambda(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is the vector derivative of the penalty function. Fu (2003) proposed a method by adjusting the iteratively reweighted least squares method for the penalty function which is equivalent to LQA algorithm. Dziak and Li (2006) proposed using SCAD for GEE models and showed that SCAD may provide better estimation and selection performance than LASSO. Although different penalty functions can be adopted, in this research we consider only two important penalty functions: LASSO and SCAD. The first possesses the sparsity function and the second simultaneously achieves the three desirable properties of variable selection: sparsity, unbiasedness and continuity, Fan and Li (2001).

### 2.3.1 Correlation Structure

An attractive feature of the penalized GEE is that the consistency of the estimated parameters hold even if the working correlation, $\boldsymbol{R(\alpha)}$ is misspecified. There are several choices for the working correlation structure - independent, exchangeable, and first-order autoregressive (AR(1)) must be specified. However, Sutradhar and Das (1999), Wang and Carey (2003), and Shults et al (2006) showed that an incorrectly specified correlation structure leads to substantial loss in estimation efficiency. The correlation pattern in analyses of different types of responses is rarely known and difficult to specify. Thus, we suggest using unstructured correlation structure, $\boldsymbol{R_u(\alpha)}$ to prevent misspecification and loss of efficiency. Liang and Zeger (1986) suggested simply using the moment estimators based on Pearson residuals to estimate the correlation. 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. Specifically,

$$\hat{R}_u = \frac{1}{n}\sum_{i=1}^{n} \hat{A}_i^{-1/2} r_i r_i^T \hat{A}_i^{-1/2}, \tag{2.13}$$

where $\hat{R}_u$ is obtained without any assumption on the specific structure on the true correlation matrix.

### 2.3.2 Computational Algorithm

To compute $\hat{\boldsymbol{\beta}}$, we use the local quadratic approximation (LQA) algorithm suggested by Fan and Li (2001). With the aid of the LQA, the optimization of (2.12) can be carried out using a modified Newton-Raphson (MNR) algorithm. Let $\beta_r = (\beta_{1r}, \ldots, \beta_{pr})$ be the parameter estimate at the $r$th iteration.

- We start with an initial $\boldsymbol{\beta}_0$ ordinary least squares estimate.
- For each iteration $r$, if $\beta_{jr}$ is very close to 0 then set $\hat{\beta}_{jr} = 0$.
- Otherwise the penalty can be locally approximated by the quadratic function. The derivative of the penalty can be approximated as

  $$[P_\lambda(|\beta_j|)]' = P'_\lambda(|\beta_j|)sgn(\beta_j) \approx \{P'_\lambda(|\beta_j|)/|\beta_j|\}\beta_j.$$

  Thus using Taylor expansions, we can locally approximate equation (2.12) by

  $$S(\beta_r) + \frac{\partial S(\beta_r)}{\partial \beta}(\beta - \beta_r) - nU_\lambda(\beta_r) - n\Sigma_\lambda(\beta_r)(\beta - \beta_r) + \cdots = 0 \qquad (2.14)$$

  where

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

- Applying Newton-Raphson method to equation (2.14), we obtain the following iteration for solving the penalized generalized estimating equation. The estimate of $\hat{\boldsymbol{\beta}}$ in the $(r+1)$th iteration is,

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{r+1}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{r}} - \left\{ \frac{\partial S(\hat{\boldsymbol{\beta}}_{\boldsymbol{r}})}{\partial \boldsymbol{\beta}} - n\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{r}}) \right\}^{-1} \left\{ S(\hat{\boldsymbol{\beta}}_{\boldsymbol{r}}) - n\boldsymbol{U}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{r}}) \right\}. \tag{2.15}$$

- Given a selected tuning parameter $\lambda$, we repeat the above algorithm to update $\hat{\boldsymbol{\beta}}_{\boldsymbol{r}}$ until convergence. The convergence criterion is

$$\|\hat{\boldsymbol{\beta}}_{\boldsymbol{r}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{r-1}}\|_2 < \epsilon.$$

  for a pre-specified small constant, $\epsilon$.

### 2.3.3 Tuning Parameter Selection

The numerical performance and the asymptotic behaviour of the penalized regression models rely on the appropriate choice of the tuning parameter. The tuning parameters are often employed to balance model sparsity and goodness-of-fit. To optimize the thresholding parameters $\boldsymbol{\theta} = (\lambda, a)$ for SCAD, we fix $a = 3.7$ as suggested by Fan and Li (2001) in practice and only tune $\lambda$ for SCAD and $\boldsymbol{\theta} = \lambda$ for other penalty functions (LASSO). Here we discuss two methods of estimating $\lambda$: Generalized Cross-Validation (GCV) (Craven and Wahba, 1979) and Bayesian Information Criterion (BIC) (Schwarz, 1978).

### Generalized Cross-Validation (GCV)

Generalized Cross-Validation (GCV) proposed by Craven and Wahba (1979) aims to approximate the leave- one-out cross validation criterion. In the GCV approach, the value of $\lambda$ that achieves the minimum of the GCV is the optimal tuning parameter. The minimization can be carried out by searching over a predetermined grid of points for $\lambda$. For linear smoothers ($\hat{y} = Ly$), the GCV is defined by

$$GCV(\lambda) = \frac{1}{n}\frac{RSS(\beta(\lambda))}{(1 - n^{-1}df(\lambda))^2}, \tag{2.16}$$

where $RSS(\beta(\lambda)) = (y - X\beta)^T(y - X\beta)$ and $df(\lambda) = tr(X(X^TX + n\Sigma_\lambda)^{-1}X^T)$ is the trace of the smoothing matrix $L$, often called effective number of parameters; Hastie & Tibshirani (1990), Tibshirani (1996) and Fan & Li (2001). The GCV is computationally convenient and remains as one popular criterion is selecting smoothing parameter. The nonlinear GCV for the generalized linear model is defined as

$$GCV(\lambda) = \frac{Dev}{n(1 - n^{-1}df(\lambda))^2}, \tag{2.17}$$

where $Dev = 2\ell(y, y) - 2\ell(\mu, y)$ is the model deviance (McCullagh and Nelder, 1989). The model deviance replaces the $RSS$ in the GCV for non-Gaussian distributions in the exponential family. Fu (2003) also recommended an adaptation of GCV where

$RSS$ is generalized to the weighted deviance,

$$WDev = \sum_{i=1}^{n} r_i^T R_i^{-1} r_i, \tag{2.18}$$

where $\boldsymbol{r_i} = (\boldsymbol{y_i} - \boldsymbol{\mu_i})$ are the deviance residuals and $\boldsymbol{R_i}$ is the working correlation matrix.

**Bayesian Information Criterion(BIC)**

In model selection, Wang et al.(2007) showed that the tuning parameter that is selected by the BIC can identify the true model consistently. Several researchers use BIC for selecting the optimal $\lambda$ by minimizing

$$BIC(\lambda) = log\Big(\frac{RSS(\lambda)}{N}\Big) + \Big(\frac{log(N)}{N}\Big)df(\lambda) \tag{2.19}$$

where $df(\lambda)$ is estimated as the number of nonzero variables in $\hat{\boldsymbol{\beta}}(\lambda)$ (Zou et al., 2007). The resulting optimal regularization parameter $\hat{\lambda}_{BIC}$ is then selected as the one that minimizes the $BIC(\lambda)$. The BIC criteria can also be extended beyond linear models by replacing $RSS(\lambda)$ with a weighted sum of squares or model deviance (Poisson, binomial, etc).

### 2.3.4 Asymptotic Properties

In this section, we discuss the asymptotic properties for penalized GEE estimator as the number of subjects goes to infinity. Let

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D_i^T V_i^{-1} r_i}, \tag{2.20}$$

$$\boldsymbol{K}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{D_i^T V_i^{-1} D_i}, \tag{2.21}$$

where $V_i$ is the working covariance and is not assumed to be the same as the true covariance. For analysis of longitudinal data using penalized estimating equations, Dziak (2006) showed that the asymptotic consistency and normality of $\hat{\boldsymbol{\beta}}$ depends on the following regularity conditions:

(1) $\boldsymbol{S(\beta)}$ and $\boldsymbol{K(\beta)}$ have continuous third derivative in $\boldsymbol{\beta}$.

(2) $\boldsymbol{K(\beta)}$ is positive definite with probability approaching one andthere exist a non-random function $\boldsymbol{K}_0(\boldsymbol{\beta})$ such that $\|\boldsymbol{K}(\boldsymbol{\beta}) - \boldsymbol{K}_0(\boldsymbol{\beta})\| \xrightarrow{p} 0$ uniformly, $\boldsymbol{K}_0(\boldsymbol{\beta}) > 0$ for all $\boldsymbol{\beta}$.

(3) $\boldsymbol{S_i} = \sum_{i=1} \boldsymbol{D_i^T V_i^{-1} r_i}$ have finite covariance for all $\boldsymbol{\beta}$.

(4) The derivative of $\boldsymbol{K}_0(\boldsymbol{\beta})$ in $\boldsymbol{\beta}$ are $O_p(n^{-1/2})$ for all $\boldsymbol{\beta}$.

Liang and Zeger (1986) proposed GEEs for the analysis of longitudinal data with a generalized linear model. The GEEs are multivariate extensions of quasi-likelihood.

Modeling the same way as longitudinal data, the following theorems from Dziak (2006) can be easily extended to apply to GEE for the multivariate multiple regression case. Assuming $\mu_i^{(m)}$ is correctly specified by $f^{(m)}(x_i^{(m)}, \boldsymbol{\beta}^{(m)})$, $\hat{\boldsymbol{\alpha}}$ and scaling parameters are appropriately chosen, then smooth nonlinear models with continuous derivatives have been shown to satisfy these regularity conditions for penalized estimating equation with multiple outcomes. Dziak (2006) states the following theorems;

**Theorem 2.1.** *Under regularity conditions* $(1)-(4)$*, for LASSO with* $\lambda = O_p(n^{-1/2})$ *or for SCAD penalty with* $\lambda = o_p(1)$ *there exists a sequence* $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ *of solutions such that* $\|\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}\| = O_p(n^{-1/2})$.

Following Dziak (2006), Theorem 2.1 shows model consistency of the penalized estimating equation 2.12 with LASSO ($\lambda = O_p(n^{-1/2})$) and SCAD penalty ( $\lambda = o_p(1)$) when the number of subjects n goes to infinity. If $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}, \boldsymbol{\beta}_{\mathcal{N}})$ is the true vector of regression coefficients with two subsets: $\mathcal{A} = \{j : \beta_j \neq 0\}$ as the active (non-zero) coefficients and $\mathcal{N} = \{j : \beta_j = 0\}$ as the inactive (zero) coefficients then for selection consistency, we require both sparsity (deleting zero coefficients) and sensitivity (retaining non-zero coefficients) properties (Fan and Li, 2001).

**Theorem 2.2.** ***(Asymptotic Normality)*** *Under the conditions of Theorem 2.1,* *there exist a sequence $\hat{\boldsymbol{\beta}}$ of solutions to equation 2.12 such that*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{L} N(\mathbf{0}, \boldsymbol{\Phi}) \tag{2.22}$$

Again, following Dziak (2006) Theorem 2.2 indicates that the parameter estimates for 2.12 are asymptotically normal, i.e,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{L} N(\mathbf{0}, \boldsymbol{\Phi}) \tag{2.23}$$

where $\boldsymbol{\Phi}$ is the limit in probability of

$$\boldsymbol{\Phi}_n = \Big[\sum_{i=1}^{n} \boldsymbol{D_i^T V_i^{-1} D_i} - n\boldsymbol{\Sigma_\lambda(\hat{\beta})}\Big]^{-1} \Big\{\sum_{i=1}^{n} \boldsymbol{D_i^T V_i^{-1}} \mathrm{cov(y_i)} \boldsymbol{V_i^{-1} D_i}\Big\} \Big[\sum_{i=1}^{n} \boldsymbol{D_i^T V_i^{-1} D_i} - n\boldsymbol{\Sigma_\lambda(\hat{\beta})}\Big]^{-1}$$

Since the proofs of Theorem 2.1 and 2.2 are similar to that of Dziak (2006) by replacing quasi-likelihood based on multiple responses (Coffey and Gennings, 2007a, 2007b), we ignore the proofs here.

It should be noted that the variable selection methods in general does not guarantee the consistency property there by does not guarantee classical inference theory in some situations. Post-selection inference procedure is one of the option to overcome the problem by utilizing the cross-validation approach to part of the data.

# Chapter 3

# Simulation Studies

We conducted a series of simulation studies to investigate the performance of our proposed variable selection approach on continuous, binary and count response outcomes using the LASSO and SCAD penalty functions. Simulations were conducted using the R software. For faster computations in optimization of tuning parameter $\lambda$, we used the "warm-starting" principle, where the initial value of $\boldsymbol{\beta}$ is replaced by $\hat{\boldsymbol{\beta}}_{(\lambda+\delta\lambda)}$ for the modified N-R algorithm in each simulation. The model that has minimum BIC($\lambda$) or GCV($\lambda$) is identified as the best model. The model performance is assessed using model error (ME, Fan and Li 2001) and their standard error, correct deletions and incorrect deletions. Model error is due to lack of fit of an underlying model and is denoted by $ME(\hat{\beta})$. The size of the model error reflects how well the model fits the data.

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

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

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

where $ME_{full}$ is the model error calculated by fitting the data with the full model. Correct deletions are the average number of true zero coefficients correctly estimated as zero and incorrect deletions are the average number of true nonzero coefficients erroneously set to zero. Estimated values for correct and incorrect deletions are reported in the columns "Correct" and "Incorrect", respectively. For comparison purposes, we estimated the covariance matrix of the response variables based on both unstructured working correlation (UWC) and independent working correlation (IWC) to investigate the performance of the GEE methodology (Coffey & Gennings, 2007a). We simulated 1000 data sets consisting of $n = 50$ and $n = 100$ observations from the response model

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

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

## 3.1 Simulation for Normal and Binary responses

### 3.1.1 Case 1: Correlated Three Normal Responses

We consider correlated normal responses ($m = 3$) with AR(1) true correlation with parameter $\rho_y = 0.7$ and two covariates ($k = 2$) with $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)}) = ((3, 1.5), (0, 0), (2, 0))$. Simulation results are summarized in Tables 3.1 and 3.2 for IWC and UWC respectively. From Table 3.1 & 3.2, we see 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 $\beta_1^{(3)} = 2$ but the standard errors of the estimates in Table 3.2 decreases which can be attributed to the correlation between the responses. For both $n = 50$ and $n = 100$, the mean model error and its standard error for SCAD are smaller than LASSO. The average number of zero coefficients increases as $n$ increases in Table 3.2 especially for SCAD.

| 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) |

Table 3.1: Simulations results for correlated normal responses (Case 1) with IWC.

This indicates that SCAD performs well compared to LASSO.

### 3.1.2 Case 2: Correlated Two Normal and One Independent Binary Responses

We simulated 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 with parameter $\rho_y = 0.7$ and the binary outcome from an independent binary observation and two covariates ($k = 2$) with $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)}) =$

| 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) |

Table 3.2: Simulations results for correlated normal responses (Case 1) with UWC.

$((3, 1.5), (0, 0), (2, 0))$. Simulation results are summarized in Tables 3.3 and 3.4 for IWC and UWC respectively. We see from Tables 3.3 & 3.4 that, the nonzero estimates for IWC remained similar to those in UWC. However because of the large correlation (0.7) between the continuous responses, the standard errors of $\beta_1^{(1)} = 3$, $\beta_2^{(1)} = 1.5$ decreases for UWC. Again, the average number of zero coefficients increases for UWC compared to IWC. As the sample size of SCAD is increased, the mean model error and its standard error decreases for both GCV and BIC. LASSO estimates for $\beta_3^{(1)}$ are not close to the true value but the estimates of the nonzero coefficients are all

| 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) |

Table 3.3: Simulations results for correlated normal and independent binary responses (Case 2) with IWC.

close to the true values for SCAD. Thus, SCAD performs well compared to LASSO.

### 3.1.3 Case 3 : Correlated Two Normal and One Binary Responses

We simulated three outcomes ($m = 3$) - two continuous and one binary generated using unstructured correlation structure with parameters $\rho_{12} = 0.3, \rho_{13} = 0.4$ and $\rho_{23} = 0.6$ and two covariates ($k = 2$) with $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)}) = ((3, 1.5), (0, 0), (2/3, 0))$.The $\boldsymbol{\beta}$ values for the binary outcome had to be smaller than before to avoid numerical

| 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) |

Table 3.4: Simulations results for correlated normal and independent binary responses (Case 2) with UWC.

instability. Correlated normal and binary outcomes were generated in R using the **BinNor** package of Anup Amatya and Hakan Demirtas for generating multiple binary and normal variables simultaneously given marginal characteristics and association structure based on the methodology proposed by Demirtas and Doganay (2012). Simulation results are summarized in Tables 3.5 and 3.6 for IWC and UWC respectively.

From Tables 3.5 & 3.6., we see that if the sample size is increased, the mean model error and its standard error are reduced. Again, the standard error of the nonzero parameter estimates for UWC are reduced compared to IWC. The average number of

| 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) |

Table 3.5: Simulations results for correlated normal and binary responses (Case 3) with IWC.

zero coefficients using SCAD with BIC for all sample size are close the target value of three and the nonzero estimated coefficients are close to the true values for $n = 50$ and $n = 100$ for SCAD with GCV.

### 3.1.4 Case 4 : Correlated Two Normal and One Independent Count Responses

We simulated three outcomes ($m = 3$) - two continuous and one count. The continuous outcomes were generated from a normal distribution and were correlated with

| 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) |

Table 3.6: Simulations results for correlated normal and binary responses (Case 3) with UWC.

AR(1) true correlation with parameter $\rho_y = 0.7$ and the count outcome from an independent Poisson observations and two covariates ($k = 2$) with $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)}) = ((3, 1.5), (0, 0), (2, 0))$. Simulation results are summarized in Table 3.7 and 3.8 for IWC and UWC respectively. From Tables 3.7 & 3.8., we see that the nonzero parameter estimates are close to the true values. The incorporation of the correlation resulted in decreased standard errors of nonzero parameters.

Overall, from Tables 3.1-3.8, we see that the nonzero estimates are unbiased regardless of the correlation structure. However the unstructured correlation resulted in

| Selection | Penalty | MRME | Correct | Incorrect |
|---|---|---|---|---|
| $n = 50$ | | | | |
| $\hat{\lambda}_{GCV}$ | SCAD | 0.214 | 0.932 | 0.000 |
| | LASSO | 0.254 | 0.957 | 0.010 |
| $\hat{\lambda}_{BIC}$ | SCAD | 0.188 | 1.110 | 0.000 |
| | LASSO | 0.911 | 1.178 | 0.014 |
| $n = 100$ | | | | |
| $\hat{\lambda}_{GCV}$ | SCAD | 0.906 | 0.988 | 0.000 |
| | LASSO | 0.871 | 1.013 | 0.002 |
| $\hat{\lambda}_{BIC}$ | SCAD | 0.834 | 1.096 | 0.000 |
| | LASSO | 0.864 | 1.225 | 0.000 |

| Selection | Penalty | $\hat{\beta}_1^{(1)}$ | $\hat{\beta}_2^{(1)}$ | $\hat{\beta}_1^{(3)}$ |
|---|---|---|---|---|
| $n = 50$ | | | | |
| $\hat{\lambda}_{GCV}$ | SCAD | 3.000(0.172) | 1.502(0.174) | 1.999(0.057) |
| | LASSO | 2.873(0.291) | 1.379(0.243) | 1.978(0.073) |
| $\hat{\lambda}_{BIC}$ | SCAD | 3.000(0.172) | 1.502(0.174) | 1.990(0.052) |
| | LASSO | 2.831(0.340) | 1.338(0.268) | 1.972(0.085) |
| $n = 100$ | | | | |
| $\hat{\lambda}_{GCV}$ | SCAD | 3.004(0.117) | 1.497(0.119) | 1.999(0.032) |
| | LASSO | 2.910(0.185) | 1.405(0.173) | 1.989(0.042) |
| $\hat{\lambda}_{BIC}$ | SCAD | 3.003(0.118) | 1.497(0.119) | 1.999(0.030) |
| | LASSO | 2.876(0.181) | 1.366(0.198) | 1.987(0.033) |

Table 3.7: Simulations results for correlated normal and independent count responses (Case 4) with IWC.

decreased standard errors of estimates compared to independent working correlation based estimates. The average number of zero coefficients increases in unstructured correlation tables compared to independent. We notice a decrease in mean model error when the sample size increases from 50 to 100 for both LASSO and SCAD. SCAD has smaller mean model error than LASSO in all cases. Specifically, SCAD with BIC perform well.

| Selection | Penalty | MRME | Correct | Incorrect |
|---|---|---|---|---|
| $n = 50$ | | | | |
| $\hat{\lambda}_{GCV}$ | SCAD | 0.209 | 1.183 | 0.000 |
| | LASSO | 0.244 | 1.076 | 0.001 |
| $\hat{\lambda}_{BIC}$ | SCAD | 0.185 | 1.303 | 0.000 |
| | LASSO | 0.851 | 1.173 | 0.012 |
| $n = 100$ | | | | |
| $\hat{\lambda}_{GCV}$ | SCAD | 0.894 | 1.260 | 0.000 |
| | LASSO | 0.866 | 1.205 | 0.001 |
| $\hat{\lambda}_{BIC}$ | SCAD | 0.818 | 1.372 | 0.000 |
| | LASSO | 0.832 | 1.264 | 0.000 |

| Selection | Penalty | $\hat{\beta}_1^{(1)}$ | $\hat{\beta}_2^{(1)}$ | $\hat{\beta}_1^{(3)}$ |
|---|---|---|---|---|
| $n = 50$ | | | | |
| $\hat{\lambda}_{GCV}$ | SCAD | 3.003(0.162) | 1.496(0.169) | 2.000(0.055) |
| | LASSO | 2.934(0.178) | 1.426(0.171) | 1.981(0.060) |
| $\hat{\lambda}_{BIC}$ | SCAD | 3.000(0.157) | 1.498(0.164) | 2.000(0.051) |
| | LASSO | 2.909(0.271) | 1.408(0.203) | 1.976(0.101) |
| $n = 100$ | | | | |
| $\hat{\lambda}_{GCV}$ | SCAD | 3.005(0.109) | 1.495(0.110) | 2.998(0.032) |
| | LASSO | 2.951(0.140) | 1.443(0.117) | 1.991(0.034) |
| $\hat{\lambda}_{BIC}$ | SCAD | 3.003(0.104) | 1.495(0.103) | 2.000(0.030) |
| | LASSO | 2.948(0.105) | 1.439(0.114) | 1.990(0.033) |

Table 3.8: Simulations results for correlated normal and independent count responses (Case 4) with UWC.

# Chapter 4

# Case Studies

## 4.1 Concrete Slump Test Data

In this section, we apply variable selection to concrete slump test data set. The data comes from a study by Yeh, I-Cheng (2006, 2007, 2008, 2009) to model the slump-flow of fly ash and slag concrete as a function of seven concrete ingredients measured in $kg/m^3$, including cement ($X_1$), fly ash ($X_2$), blast furnace slag ($X_3$), water ($X_4$), superplasticizer ($X_5$), and coarse aggregate ($X_6$) and fine aggregate ($X_7$). The data set report some results about two kinds of tests executed on concrete. Concrete is a highly complex material, which makes modeling its behavior a very difficult task. The workability of concrete can be measured by the "concrete slump test", a simplistic measure of the plasticity of a fresh batch of concrete. The concrete slump test is in essence, a method of quality control. For a particular mix, the slump should be consistent. A change in slump height would demonstrate an undesired change in the ratio of the concrete ingredients; the proportions of the ingredients are then adjusted to keep a concrete batch consistent. This homogeneity improves the quality and structural integrity of the concrete. The second test considered is "compressive strength test" where this test measure the capacity of a material to withstand axially directed pushing forces. The variance of slump and flow was observed. The slump is the difference of height of the concrete mix after being placed in the slump cone and the cone. It differs from one sample to another. Samples with lower heights are predominantly used in construction, with samples having high slumps commonly used to construct roadway pavements. The flowability is measured in terms of spread, hence the flow correspond to the width of the patty. The three output variables include slump (cm), flow (cm) and 28-day compressive strength (CS) (Mpa). The data comprises 103 samples and it is available at *http://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test*. A more detailed description of the data set can be found in Yeh, I-Cheng (2006, 2007, 2008, 2009). Figure 4.1 and Table 4.1 show the association strength among the three responses. It is shown that slump ($Y_1$) and flow ($Y_2$) are highly correlated, with a positive correlation of 0.9061. We can use penalized GEE to utilize that additional

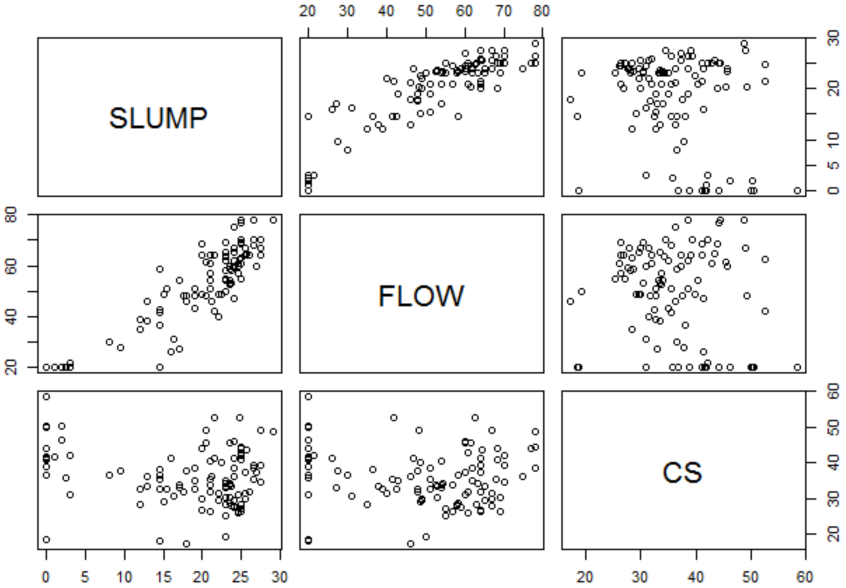

Figure 4.1: Scatter plot demonstrating visually the relationship between slump ($Y_1$), flow ($Y_2$) and compressive strength (CS) ($Y_3$)

| | SLUMP | FLOW | CS |
|---|---|---|---|
| SLUMP | 1 | 0.9061 | -0.2233 |
| FLOW | 0.9061 | 1 | -0.1240 |
| CS | -0.2233 | -0.1240 | 1 |

Table 4.1: Correlation matrix for the responses

information in the selection of significant variables for this data set. The estimates are given in Table 4.2-4.4.

The second and third columns of Tables 4.2-4.4 represent performance using penalized GEE with IWC for SCAD and LASSO. The fourth and fifth columns of the

| | IWC | | UWC | |
|---|---|---|---|---|
| Variable | 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) |

Table 4.2: Estimates of regression coefficients for slump ($Y_1$), with standard error in parentheses

tables represent performance using penalized GEE with UWC. For the model selection procedures, both unweighted BIC and GCV were used to estimate regression coefficients. However, their performance was similar. Therefore, we present only the results based on the unweighted BIC for both SCAD and LASSO. We see from Table 4.2 that, SCAD with IWC identified 5 out of the 7 covariates as important for slump($Y_1$) whereas LASSO with IWC identified 4 covariates. The difference between them is that SCAD kept fine aggregate ($X_7$). SCAD and LASSO with UWC obtained the same estimates for all variables, they retained fine aggregate ($X_7$) but forced fly

| | IWC | | UWC | |
|---|---|---|---|---|
| Variable | 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) | – | – |
| $X_7$ | – | – | – | – |
| | – | – | – | – |

Table 4.3: Estimates of regression coefficients for flow ($Y_2$), with standard error in parentheses

ash ($X_2$) and coarse aggregate ($X_6$) to zero. From Table 4.3 we see that, both SCAD and LASSO with IWC chose fly ash ($X_2$), water ($X_4$) and coarse aggregate ($X_6$) as significant ingredients for flow ($Y_2$) but SCAD and LASSO with UWC identified only fly ash ($X_2$) and water ($X_4$) as significant variables. The standard errors of estimates with UWC decreases. From Table 4.4 we see that, LASSO with IWC chose all covariates as important ingredients for CS ($Y_3$) except coarse aggregate ($X_6$) whereas the others dropped coarse aggregate ($X_6$) as well as superplasticizer ($X_5$).

| | IWC | | UWC | |
|---|---|---|---|---|
| Variable | 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) |

Table 4.4: Estimates of regression coefficients for compressive strength ($Y_3$), with standard error in parentheses

### 4.1.1 Concrete Slump Test Data With Artificial Binary Response

For illustration purposes, we create an artificial binary response variable to indicate whether a specimen can sustain a heavy load before distortion. For this analysis, we consider that concrete with compressive strength less than 35 is of poor quality. So for illustration purpose, we convert this continuous variable response to binary based on the quality. Let $Y_3 = 1$ if the compressive strength is more than 35, and

$Y_3 = 0$ otherwise. The goal is to apply variable selection method to model correlated continuous and binary outcomes. The estimates are given in Tables 4.5-4.7. The description of Tables 4.5-4.7 is the same as Tables 4.2-4.4.

| | IWC | | UWC | |
|---|---|---|---|---|
| Variable | 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) |

Table 4.5: Estimates of regression coefficients for slump ($Y_1$), with standard error in parentheses

From Table 4.5 we see that, SCAD with IWC identified 5 out of the 7 covariates as important for slump ($Y_1$) whereas LASSO with IWC identified 4 covariates. The difference between them is that SCAD kept fine aggregate ($X_7$). These results are similar to independent results in Table 4.2, which confirms the use of IWC. SCAD with UWC forced fly ash ($X_2$) to zero compared to SCAD with IWC. LASSO with

| | IWC | | UWC | |
|---|---|---|---|---|
| Variable | 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) |
| $X_7$ | – | – | – | – |
| | – | – | – | – |

Table 4.6: Estimates of regression coefficients for flow ($Y_2$), with standard error in parentheses

UWC maintained the same important variables as LASSO with IWC. From Table 4.6, we see that all methods identified fly ash ($X_2$), water ($X_4$) and aggregate ($X_6$) as significant variables for flow ($Y_2$). From Table 4.7, we see that all methods chose 5 covariates as important ingredients for binary CS ($Y_3$) except LASSO with IWC. Estimates obtained with UWC have reduced standard errors.

| | IWC | | UWC | |
|---|---|---|---|---|
| Variable | 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) | – | – |

Table 4.7: Estimates of regression coefficients for binary compressive strength ($Y_3$), with standard error in parentheses

# Chapter 5

# Conclusion

Variable selection plays a pivotal role in modeling correlated responses due to large number of covariate variables involved. Thus a parsimonious model is always desirable to enhance model predictability and interpretation especially in multi-response regression models. To automatically and simultaneously select significant variables, we proposed penalized GEE approach to multi-response regression problem using LASSO and SCAD penalty functions. To implement the proposed approach, one need to estimate the covariance matrix of the response variables and we recommend covariance matrix based on the estimate of the unstructured correlation matrix. For model selection, the performance of unweighted BIC and GCV were explored for LASSO and SCAD through series of simulation studies. In each case, we performed the entire analysis with both unstructured working correlation (UWC) and independent working correlation (IWC) for comparison purpose. We discussed the computational algorithm and asymptotic properties of our approach. Simulation studies showed that SCAD with BIC tuning criteria works well compared to the other pairs. The estimates of $\boldsymbol{\beta}$ are unbiased (Liang and Zeger, 1986) regardless of the choice of correlation structure. However, estimates obtained from the UWC have reduced standard errors. We also applied our method to concrete slump test data to investigate variable selection in continuous and binary multi-response framework. Future research are warranted to gain more insights on their properties including their strengths and weakness. In conclusion, we hope our methodology may prove useful and support variable selection and estimation of coefficients in multivariate multi-response regression problem.

# Bibliography

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrove, B. N. and Csaki, F. (eds.) *Second Symposium of Information Theory*, Akademial Kiado, Budapest, 267-282.

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

[3] Breiman, L. & Friedman, J. H. (1997). Predicting multivariate responses in multiple regression. *Journal of Royal Statistics Society B.* **1**, 3-54.

[4] Bull, S. B. (1998). Regression models of multiple outcomes in large epidemiologic studies. *Statistics in Medicine.* **17**, 2179-2197.

[5] 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.

[6] 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.

[7] Contreras, M., and Ryan, L. M. (2000). Fitting nonlinear and constrained generalized estimating equations with optimization software. *Biometrics.* **56**, 1268-1271.

[8] Craven, P. and Wahba, G. (1979). Smoothing noise data with spline functions: validation. *Numerische Mathematika.* **31**, 377-403.

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

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

[11] 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.

[12] Fitzmaurice, G. M., and Laird, N. M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, **53**, 110-122.

[13] Fu, W. J. J., (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7**, 397-416.

[14] Fu, W. J. J., (2003). Penalized Estimating Equations. *Biometrics*, **59**, 126-132.

[15] Fu, W. J.,(2005). Nonlinear GCV and quasi-GCV for shrinkage models. *Journal of Statistical Planning and Inference*, **131**, 333-347.

[16] Goldwasser M. A. & Fitzmaurice G. M. (2006). Multivariate linear regresion analysis of child psychopathology using multiple informant data, *International Journal of Methods in Psychiatric Research*, **10**, 1-10.

[17] Gray, S. M., and Brookmeyer, R. (2000). Multidimensional longitudinal data: Estimating a treatment effect from continuous, discrete, or time-to-event response variables, *Journal of the American Statistical Association*, **95**, 396-406.

[18] Hastie, T. J & Tibshirani, R. J. (1990). *Generalized Additive Models*, New York: Chapman and Hall.

[19] Hoerl, A. E., & Kennard, R. W., (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55-67.

[20] Kannel, W.B., et al,(1961). Factors of risk in the D=development of coronary heart diseasesix-year Follow-up experience: The framingham study. *Annals of of Internal Medicine*, **55**, 33-50.

[21] Hunter, D. R., Li, R. Z.,(2005). Variable selection using MM algorithms. *Annals of Statistics*, **33**, 1617-1642.

[22] Lee, W., and Liu, Y. (2012), Simultaneous Multiple Response Regression and Inverse Covariance Matrix Estimation via Penalized Gaussian Maximum Likelihood, *Biometrika*, **111**, 241-255.

[23] Lefkopoulou, M., Moore, D., and Ryan, L. (1989), The analysis of multiple correlated binary outcomes: Application of rodent teratology experiments, *Journal of the American Statistical Association* , **84**, 810-815.

[24] Liang, K. Y., and Zeger, S.L. (1986). Longitudinal data Analysis using Generalized Linear Models, *Biometrika*, **73**, 13-22.

[25] Liu, C., and Rubin, D. B. (1998). Ellipsoidally symmeteric extensions of the general location models for mixed categorical and continuous data, *Biometrika*, **85**, 673-688.

[26] Mallows, C. L. (1973). Some comments on $C_p$. *Technometrics*, **15**, 661-675.

[27] McCullagh, P.(1983). Quasi-Likelihood Functions *Annals of Statistics*, **11**, 59-67.

[28] McCullagh, P., and Neldar, J. A. (1989). *Generalized Linear Models* (2nd ed.), London: Chapman and hall.

[29] Miglioretti, D. L. (2003). Latent Transition Regression for Mixed Outcomes, *Biometrics*, **59**, 710-720.

[30] Moser, V. C., Casey, M., Hamm, A., Carter, Jr. W. H., Simmons, J. E., and Gennings, C. (2005). Neurotoxicological and statistical analyses of a mixture of five organophosphorus pesticides using a ray design, *Toxicological Sciences*, **86**, 101-115.

[31] Muthen, B., and Shedden, K.(1999). Finite mixture modeling with mixture outcomes using the EM algorithm, *Biometrics*, **55**, 463-469.

[32] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society A*, **135**, 370-384.

[33] Prentice, R. L., and Zhao, L. P. (1991). Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses, *Biometrics*, **47**, 825-839.

[34] Rochon, J. (1996). Analyzing bivariate repeated measures of discrete and continuous outcomes, *Biometrics*, **52**, 740-750.

[35] Rochon, J., and Gillespie, B. W. (2001). A methodology for analyzing a repeated measures and survival outcome simultaneously, *Statistics in Medicine*, **20**, 1173-1184.

[36] 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.

[37] 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.

[38] Schwarz, G. (1973). Estimating likelihood and general estimating equations., *Annals of Statistics*, **22**, 416-464.

[39] Shults, J., C. A. Mazurick, and J. R. Landis (2006). Analysis of repeated bouts of measurements in the framework of generalized estimating equations., *Statistics in Medicine*, **25**, 4114-4128.

[40] Sutradhar, B. C. and Das, K. (1999). On the efficiency of regression estimators in generalised linear models for longitudinal data, *Biometrika*, **86**, 459-465.

[41] Tibshirani, R. (1996). Regression shrinkage and variable selection via the lasso, *Journal of Royal Statistical society*, **58**, 267 - 288.

[42] Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems*, Washington: Winston & Sons.

[43] Von. Korff, M., Ormel, J., Keefe F. J., and Dworkin, DS. F. (2012). Grading the severity of chronic pain,. *Pain*, **50**, 133-149.

[44] Wang, Y. G., and Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. *Biometrics*, **90**, 29-41.

[45] Wang, Li, B., and Leng, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553-568.

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

[47] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, Generalized linear models, and the Gauss- Newton method. *Biometrika*, **61**, 439-447.

[48] Yeh, I-Cheng. (2006). Exploring concrete slump model using artificial neural networks. *J. of Computing in Civil Engineering*, **20**, 217-221.

[49] Yeh, I-Cheng. (2007). Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, **29**, 474-480.

[50] Yeh, I-Cheng. (2008). Modeling slump of concrete with fly ash and superplasticizer. *Computers and Concrete*, **5**, 559-572.

[51] Yeh, I-Cheng. (2008). Prediction of workability of concrete using design of experiments for mixtures. *Computers and Concrete*, **5**, 1-20.

[52] Yeh, I-Cheng. (2009). Simulation of concrete slump using neural networks. *Construction Materials*, **162**, 11-18.

[53] Zeger, S. L., Liang, K. Y., (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121-130.

[54] Zou, H., Hastie, T., (2005), Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **67**, 301-320.

[55] Zou, H., Hastie, T., and Tibshirani. (2007), On the Degree of Freedom of the Lasso. *Annals of Statistics*, **35**, 2173-2192.

[56] Zou, H., Li, R. (2008), One-step Sparse Estimates in Nonconcave Penalized Likelihood Models. *Ann. Stat*, **36**, 1509-1533.