Regularized Regression in Generalized Linear Measurement Error Models with Instrumental Variables -Variable Selection and Parameter Estimation

by

Lin Xue

A Thesis Submitted to the Faculty of Graduate Studies of
the University of Manitoba
in Partial Fulfillment of the Requirements of the Degree of

Doctor of Philosophy

Department of Statistics
University of Manitoba
Winnipeg

Abstract

Regularization method is a commonly used technique in high dimensional data analysis. With properly chosen tuning parameter for certain penalty functions, the resulting estimator is consistent in both variable selection and parameter estimation. Most regularization methods assume that the data can be observed and precisely measured. However, it is well-known that the measurement error (ME) is ubiquitous in real-world datasets. In many situations some or all covariates cannot be observed directly or are measured with errors. For example, in cardiovascular disease related studies, the goal is to identify important risk factors such as blood pressure, cholesterol level and body mass index, which cannot be measured precisely. Instead, the corresponding proxies are employed for analysis. If the ME is ignored in regularized regression, the resulting naive estimator can have high selection and estimation bias. Accordingly, the important covariates are falsely dropped from the model and the redundant covariates are retained in the model incorrectly. We illustrate how ME affects the variable selection and parameter estimation through theoretical analysis and several numerical examples.

To correct for the ME effects, we propose the instrumental variable assisted regularization method for linear and generalized linear models. We showed that the proposed estimator has the oracle property such that it is consistent in both variable selection and parameter estimation. The asymptotic distribution of the estimator is derived. In addition, we showed that the implementation of the proposed method is equivalent to the plug-in approach under linear models, and the asymptotic variance-covariance matrix has a compact form. Extensive simulation studies in linear, logistic and poisson log-linear regression showed that the proposed estimator outperforms the naive estimator in both linear and generalized linear models. Although the focus of this study is the classical ME, we also discussed the variable selection and estimation in the setting of Berkson ME. In particular, our finite sample simulation studies show that in contrast to the estimation in linear regression, the Berkson ME may cause bias in variable selection and estimation. Finally, the proposed method is applied to real datasets of diabetes and Framingham heart study.

Acknowledgements

I would like to express my deepest appreciation and thankfulness to my advisor Dr. Liqun Wang for his guidance, support and patience over the years. His unique insights into statistical science encouraged me to pursue my PhD research. His advice and guidance are helpful to both my research and personal life. This dissertation would not have been possible without his supervision and help.

I would also like to thank my committee members Dr. Mahmoud Torabi, Dr. Xikui Wang and Dr. Wenqing He for providing thoughtful comments and constructive feedback. I greatly appreciate your help and time commitment serving on my examining committee.

I am grateful for all the help and support received from the faculty and staff members at Department of Statistics, where I had enjoyable teaching and research experience. I also thank Faculty of Graduate Studies for the financial support through University of Manitoba Graduate Fellowship.

I would also like to thank my colleagues at Department of Statistics for providing support and interesting discussions. I want to thank my fellow student Zhiyong Jin for his help and pleasant collaboration over many projects. I thank Han Yu and Yu Zhang for their help and interesting conversations.

Last but not the least, I want to thank my family for their love and support. I want to thank Xuan Chen for her love and continuous encouragement. I couldn't be more grateful having her in my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Variable selection is an important data analysis technique in high-dimensional problems. Regularization methods achieve the goal of variable selection and parameter estimation simultaneously, which has been a popular research area recently. On the other hand, the measurement error is ubiquitous in real data applications. It is known that ignoring ME can result in biased estimation in conventional regression methods. In a similar way, ignoring the ME in high-dimensional can result in estimation and selection bias. In this thesis we illustrate how ME affects the variable selection through several heuristic examples, and propose a new method correcting for ME effect in regularization methods. This thesis is organized as follows. The background information of regularization methods and measurement error models, including different theories and methodologies are reviewed in Chapter 1. Regularized regression in linear ME model is presented in Chapter 2. Specifically, a motivating example is introduced at first, followed by theoretical results, numerical examples, real data application and theorem proofs. The estimation performance with different penalty functions and model selection criteria are also discussed. Chapter 3 covers the topic of regularized regression in generalized linear ME model, which consists of theories, numerical examples, proofs and a real data application. Summaries and conclusions of the thesis are presented in Chapter 4, along with the discussions and future research. Technique details are relegated to the Appendix.

1.1. Motivating Example

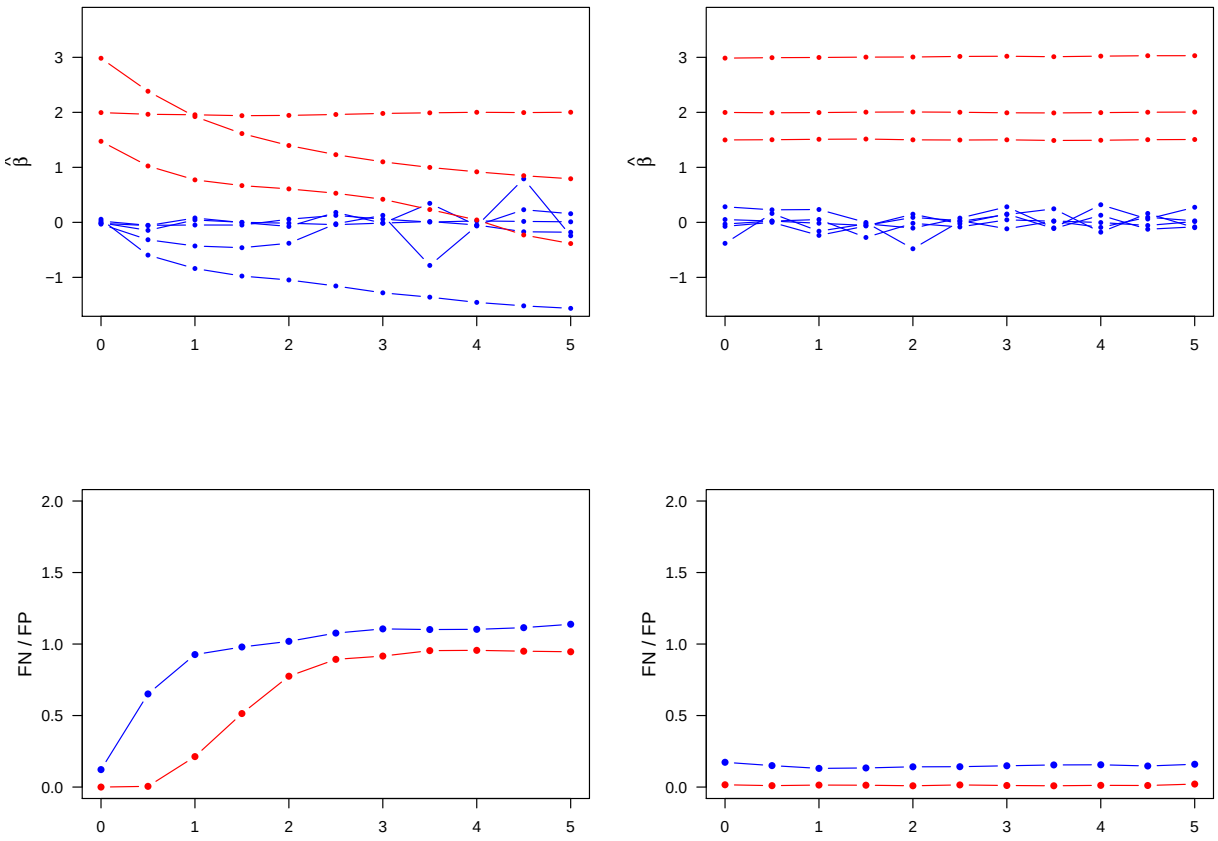
1.1.1 Variable Selection under Classical ME Model

Consider the following linear model with classical ME,

$$\begin{aligned} y &= \beta_x x + \beta_z^T z + \epsilon, \\ x &= 1.5w + u, \\ x^* &= x + \delta, \end{aligned}$$

where w is an instrumental variable and $(z_1, -w, z_2, \ldots, z_7)^T$ are jointly generated from $N(0, \Sigma)$ with $\Sigma_{ij} = 0.7^{|i-j|}$. The coefficients $(\beta_x, \beta_z^T) = (3, 1.5, 0, 0, 2, 0, 0, 0)$, ϵ and u are standard normal, whence the correlation between w and x is around 0.83. The random ME δ follows normal distribution with mean zero and variance σ_δ^2. The details can be found in example (2.2.2) of this chapter. Figure (1.1) shows the estimation and selection results of the naive estimator ignoring the ME and the proposed estimator (denoted as RIV for regularized IV estimator). It can be observed that the naive estimator is biased away from the true value. The values of false positive (FP) and false negative (FN) are both nonzero meaning that redundant features are falsely retained in the model and some important features are removed incorrectly. As a comparison, the estimation is stable and close to the true value of β for the proposed RIV estimator. In addition, the values of FP and FN are both close to zero across different values of $\sigma_\delta^2/\sigma_x^2$. Since the ME causes unpredictable estimation and selection results, it is of interest to develop new methods correcting for ME effects and recovering the underlying true model.

Figure 1.1: Estimation and selection eesults for Example (2.2.2) with $n = 200$. The x axis represents $\sigma_\delta^2/\sigma_x^2$. Top left: estimation of naive estimator; top right: estimation of RIV estimator (blue dotted-line corresponds to zero true coefficient; red dotted-line corresponds to nonzero true coefficient); bottom left: selection results of naive estimator; bottom right: selection results of RIV estimator (blue dotted-line: FP; red dotted-line: FN).



1.1.2 Variable Selection under Berkson ME Model

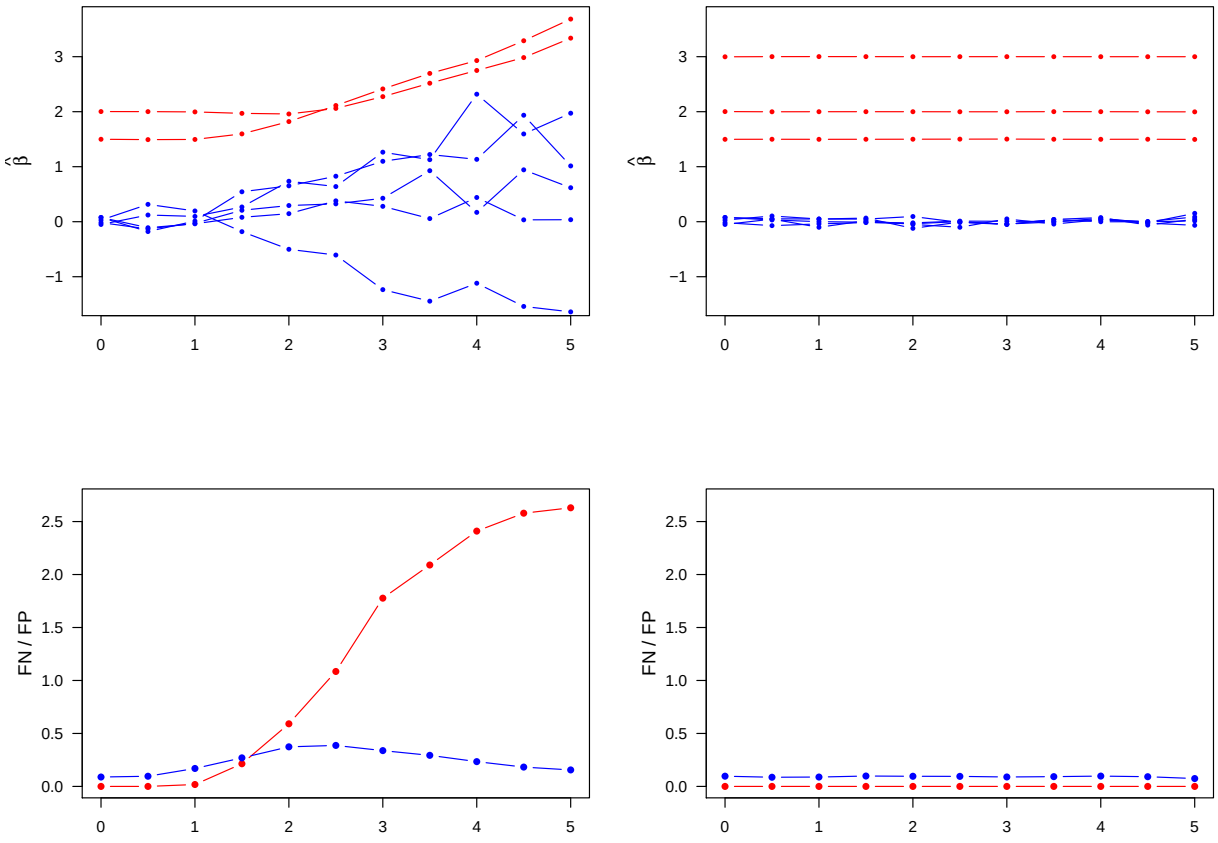
Now consider the following Berkson ME model, where the datasets are simulated from the linear model

$$y = \beta_x x + \beta_z^T z + \epsilon,$$

where $(z_1, x^*, z_2, \ldots, z_7)^T$ are jointly generated from $N(0, \Sigma)$ with $\Sigma_{ij} = 0.7^{|i-j|}$, $(\beta_x, \beta_z^T) = (3, 1.5, 0, 0, 2, 0, 0, 0)$. The covariate x is generated as $x = x^* + \delta$ with the random errors ϵ and δ being standard normal. It is known that the naive estimator for Berkson ME model is consistent. We examine finite sample performance and compare the naive estimator with

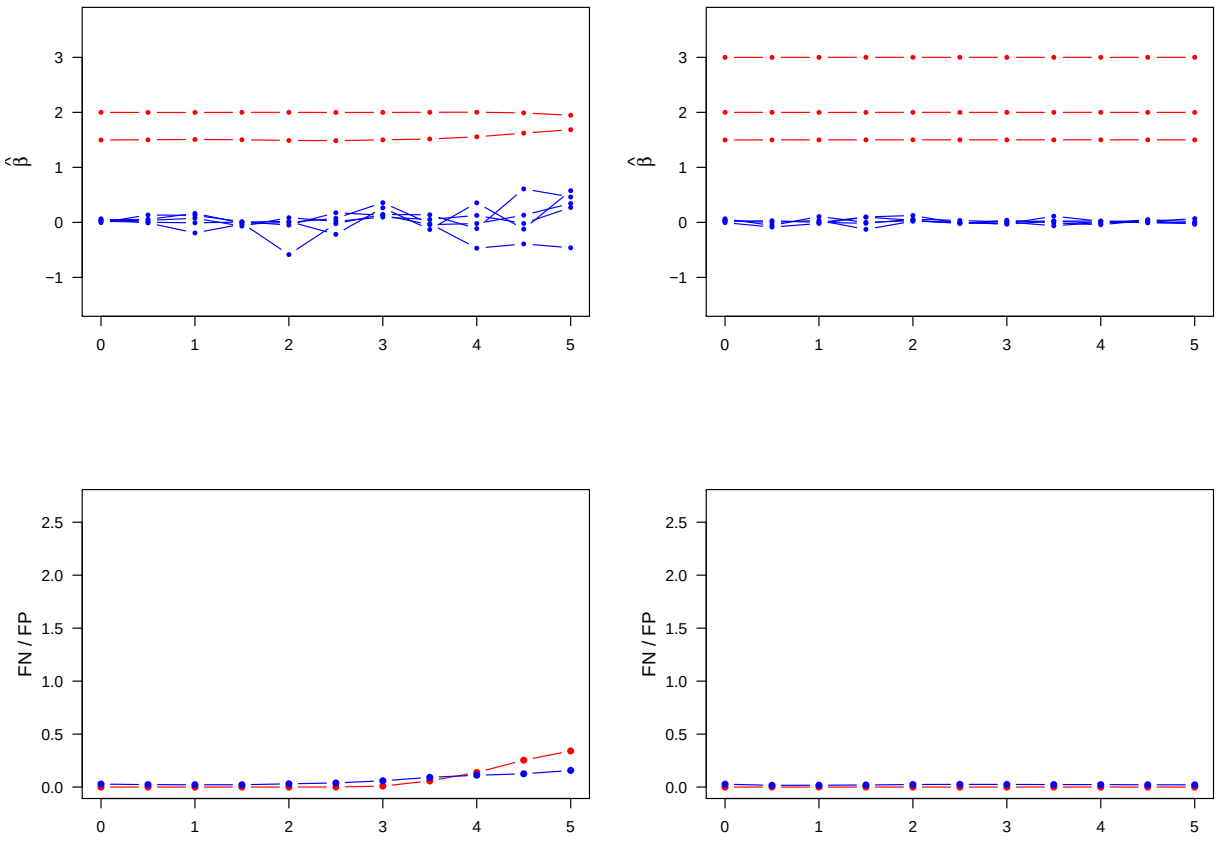
the estimator obtained using the true unobserved data (TR). From the Figure (1.2) it can be seen that for a relatively small sample size, the estimation of naive method under Berkson ME model is unstable as the ratio $\sigma_\delta^2/\sigma_x^2$ increases, compared with the TR method, which is centered around the true value of β and remains stable across different values of the ratio.

Figure 1.2: Estimation and selection results with $n = 200$. The x axis represents $\sigma_\delta^2/\sigma_x^2$. Top left: estimation of naive estimator; top right: estimation of TR estimator (blue dotted-line corresponds to zero true coefficient; red dotted-line corresponds to nonzero true coefficient); bottom left: selection results of naive estimator; bottom right: selection results of TR estimator (blue dotted-line: FP; red dotted-line: FN).



The simulation results for a larger sample size ($n = 2000$) is reported in Figure (1.3). It can be seen that for large sample size, the naive estimator performs much better compared with small sample size. It is as expected since we know that the naive estimator under Berkson ME model is consistent.

Figure 1.3: Estimation and selection results with $n = 2000$. The x axis represents $\sigma_\delta^2/\sigma_x^2$. Top left: estimation of NA estimator; top right: estimation of TR estimator (blue dotted-line corresponds to zero true coefficient; red dotted-line corresponds to nonzero true coefficient); bottom left: selection results of NA estimator; bottom right: selection results of TR estimator (blue dotted-line: FP; red dotted-line: FN).



1.2. Objectives of the Thesis

High-dimensional variable selection has been an active research area in statistics, economics, genomics, computer sciences and health sciences. The high dimensionality and enormous data size make the conventional statistical methods infeasible theoretically and computationally. In most cases, only a subset of covariates are important and the rest are redundant. To address this problem, various regularization methods are proposed for variable selection. For example, the bridge regression (Frank and Friedman, 1993), least absolute shrinkage and selection operator (Lasso, Tibshirani 1996), smoothly clipped absolute deviation (SCAD, Fan and Li 2001), adaptive Lasso (Zou, 2006), minimax concave penalty (MCP, Zhang et al. 2010), Elastic net (Zou and Hastie, 2005) and Dantzig selector (Candes et al., 2007). More detailed review of regularization methods can be found in Negahban et al. (2009) and Fan and Lv (2010).

In real data analysis, some covariates cannot be measured precisely or observed directly. For example, in cardiovascular disease studies, we are interested in identifying important risk factors of cardiovascular heart disease (CHD) such as long-term average systolic blood pressure, cholesterol level and body mass index. Those factors are either unobservable or measured with errors. In lung cancer risk studies, we are interested in the relationship between lung cancer incidence and the individual exposure to the air pollutants. The actual amount of pollutant inhaled by each individual cannot be measured directly. Instead, the pollution level are measured by several monitoring stations in a certain area. In pharmacokinetic study, the goal is to examine the efficacy of a drug. The actual absorption of medical substance in bloodstream is unobservable. Instead, the predetermined dosage of a drug is used in analysis. In the agriculture study, we are interested in the relationship between yield of a crop and the amount of fertilizer. The actual amount of fertilizer absorbed in the crop is unobservable and the predetermined dose of the fertilizer is used instead. More theoretical methods, examples and applications in ME models can be found in Carroll et al. (2006).

Applying the regularization methods naively on proxy or mismeasured covariates can lead to biased estimates and possible omission of important variables. As a consequence, methods for variable selection in ME models are proposed. For example, Liang and Li (2009) applied the correction-for-attenuation and orthogonal regression approach on the penalized least squares and quantile regression, respectively. Ma and Li (2010) proposed the variable selection technique for general parametric and semi-parametric ME models. Zhang et al. (2017) developed a model selection criterion based on minimizing prediction errors for linear model. Instrumental variable methods are also proposed for high-dimensional problems. For example, Caner and Fan (2010) and Caner and Fan (2015) suggested selecting

relevant instrumental variables at the first stage, before the other procedures that are applied afterwards. Fan and Liao (2014) proposed the focused generalized method of moments, which applies the instrumental method for variable selection in high dimensions. Lin et al. (2015) proposed two-stage regularization method for instruments and covariates selection under the joint normality assumption of random errors for linear regression model. Huang and Zhang (2013) proposed the penalized score functions for variable selection in linear ME models assuming the variance covariance matrix of ME to be known.

In this thesis, we study the effects of ME on variable selection and parameter estimation by developing theoretical results and conducting numerical examples. We propose the regularized regression method in generalized linear ME models based on instrumental variables (IV). The simulation studies are conducted comparing the performance of the naive estimator ignoring the ME with the proposed estimator. The ME introduces selection and estimation bias for naive estimator, especially with high ME variance. Whereas the regularized instrumental variables (RIV) estimator is robust to the magnitude of the ME variance. In addition, we show that the proposed estimator has the oracle property such that it is consistent in both variable selection and parameter estimation, and the estimators corresponding to nonzero coefficients follow normal distribution asymptotically. In this chapter, we introduce some notations used in the thesis, and review several regularization methods and ME models.

1.3. Regularization Methods

Massive and high-dimensional data are becoming available in many areas, such as astronomy, physics, genome and health sciences, business and finance, social media, signal processing and imaging, etc. In regression settings, it is common that there is a large number of predictors for a given response variable. Therefore, it is of interest to identify a relatively small set of important predictors. Specifically, in a linear model

$$y_i = \beta^T x_i + \epsilon_i, i = 1, 2, \ldots, n$$

where $x_i, \beta \in \mathbb{R}^p$ are a vector of predictors and parameters respectively, and ϵ_i is a random error. Both the sample size n and the number of predictors p can be very large and it is typical that $p > n$. In this case, conventional inference methods either fail or become inefficient. To overcome this difficulty, various regularized regression methods have been developed in the literature. Suppose the true model is sparse, i.e. the number of nonzero elements in β_0 is bounded by some positive integer $s < n$. For example, the constrained

least squares estimator is the solution of the problem

$$\text{minimize } \sum_{i=1}^{n}(y_i - \beta^T x_i)^2 \text{ subject to } ||\beta||_0 \leq s,$$

where $||\beta||_0 \leq s$ denotes the number of nonzero elements in β.

The estimator can be computed by the regularization method of the following form

$$\min_{\beta} \sum_{i=1}^{n}(y_i - \beta^T x_i)^2 + \sum_{j=1}^{p} p_{\lambda_n}(\beta_j), \tag{1.3.1}$$

where $p(\cdot)$ is a penalty function with tuning parameter λ_n. A commonly used penalty function is least absolute shrinkage and selection operator (Lasso) proposed by Tibshirani (1996), which takes the form $p_{\lambda_n}(\beta_j) = \lambda_n|\beta_j|$. Besides parameter estimation, the Lasso achieves the goal of variable selection, which sets some of the estimates to be exactly zero. To gain an insight of this mechanism, suppose the columns of n by p matrix X is orthonormal such that $X^T X = I_p$. Rewrite the penalized least squares objective function as

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - \beta^T x_i)^2 + \sum_{j=1}^{p} p_{\lambda_n}(\beta_j) \\
=&(y - XX^T y)^T(y - XX^T y) + (X^T y - \beta)^T(X^T y - \beta) + \sum_{j=1}^{p} p_{\lambda_n}(\beta_j) \\
=&(y - XX^T y)^T(y - XX^T y) + \sum_{j=1}^{p}(z_j - \beta_j)^2 + \sum_{j=1}^{p} p_{\lambda_n}(\beta_j),
\end{aligned}$$

where z_j is the jth element of $X^T y$. Then (1.3.1) becomes a componentwise minimization problem. For Lasso penalty, the univariate version of penalized least squares problem

$$\frac{1}{2}(z_j - \beta_j)^2 + \lambda|\beta_j|$$

has the following closed-form solution

$$\hat{\beta}_j^{\text{Lasso}} = \text{sign}(z_j)(|z_j| - \lambda)_+,$$

where a_+ denotes the positive part of a. If the magnitude of ordinary least squares estimator is less than λ_n, the Lasso estimator shrinks it to zero. Besides Lasso, there are many other penalty functions with good properties. Fan and Li (2001) suggested that a good estimator should have the oracle property. That is, the zero coefficients are estimated as zero with

probability approaching 1, and the nonzero coefficients are estimated as if the subset of covariates under true model is known. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty function which is defined as

$$p_\lambda(\beta_j) = \begin{cases} \lambda|\beta_j|, & \text{if } |\beta_j| \leq \lambda; \\ -\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}, & \text{if } \lambda < |\beta_j| \leq a\lambda; \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\beta_j| > a\lambda; \end{cases}$$

for some $a > 2$ and $\lambda > 0$. Similarly, the solution of SCAD estimator in univariate case is

$$\hat{\beta}_j^{\text{SCAD}} = \begin{cases} \text{sgn}(z_j)(|z_j| - \lambda)_+, & \text{if } |z_j| \leq 2\lambda \\ \{(a-1)z_j - \text{sgn}(z_j)a\lambda\}/(a-2), & \text{if } 2\lambda < |z_j| \leq a\lambda \\ z_j, & \text{if } |z_j| > a\lambda. \end{cases}$$

Compared with the univariate Lasso solution, it can be observed that except for a short interval, $\hat{\beta}_j^{\text{SCAD}}$ is consistent in estimation for large values of parameters. Taking $a = 3.7$ as suggested by Fan and Li (2001), the 3D plot and the corresponding heat map of $p_\lambda(\beta)$ with respect to the parameters λ and β is given in Figure (1.4) and (1.5). It is easy to see that the value of penalty function stays close to zero if β or λ is close to zero. In addition, the first-order derivative of SCAD is given by

$$p'_\lambda(|\beta_j|) = \lambda \left\{ I(|\beta_j| \leq \lambda) + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda} I(|\beta_j| > \lambda) \right\}.$$

From the Figure (1.6) and (1.7) it can be observed that for a given value of β_j, $p'_\lambda(|\beta_j|)$ is zero with a sufficiently small value of λ (especially when $\lambda < |\beta_j|/a$).

There are several regularization methods that also process the oracle property. For example, the minimax concave penalty (MCP, Zhang et al. 2010) is defined as

$$p_\lambda(\beta_j) = \begin{cases} \lambda|\beta_j| - \frac{\beta_j^2}{2a}, & \text{if } |\beta_j| \leq a\lambda, \\ \frac{1}{2}a\lambda^2, & \text{if } |\beta_j| > a\lambda, \end{cases}$$

which has the oracle property with properly chosen parameters a and λ. The adaptive Lasso (Zou, 2006; Zhang and Lu, 2007) is defined as weighted version of Lasso which is given by $\sum_{j=1}^p w_j|\beta_j|$. It was shown that with properly chosen tuning parameter and data-driven weights, the adaptive Lasso estimator performs as well as an oracle procedure.

Figure 1.4: 3D Plot of SCAD penalty

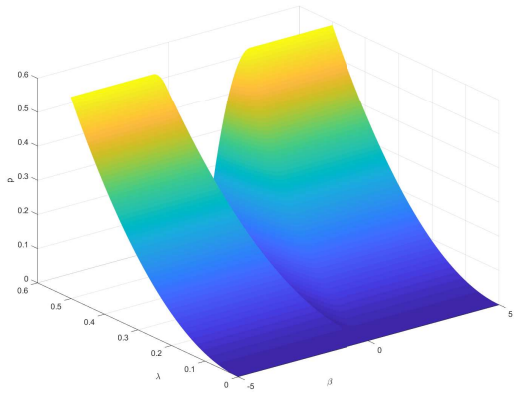


Figure 1.5: Heat Map of SCAD penalty

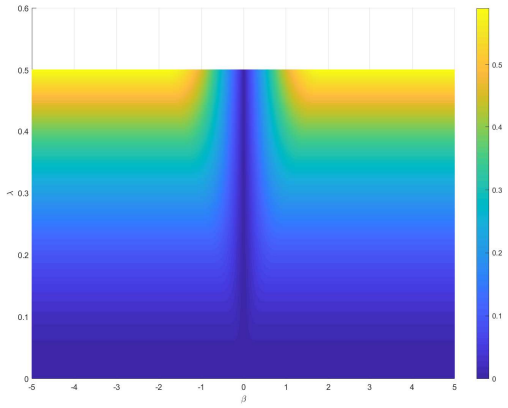


Figure 1.6: 3D Plot of SCAD gradient

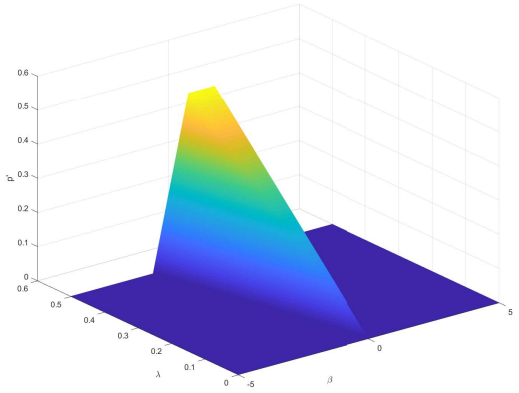
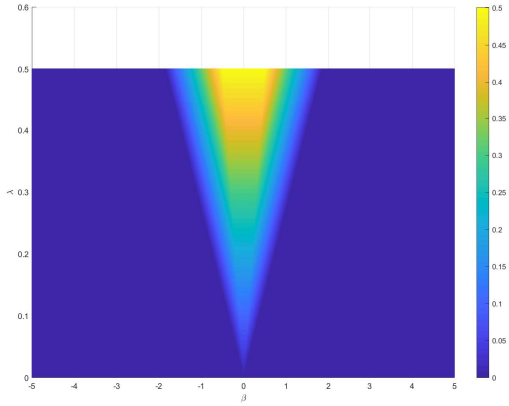


Figure 1.7: Heat Map of SCAD gradient



The elastic net method proposed by Zou and Hastie (2005) encourages the grouping effect where highly correlated covariates are retained in/dropped out from the model together, while retaining the sparsity property like Lasso. The elastic net penalty function is given by

$$(1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2$$

where $\alpha \in [0,1]$. It can be observed that the elastic net is a convex combination of Lasso and ridge penalty since it becomes ridge penalty when $\alpha = 1$ and to Lasso penalty when $\alpha = 0$. Some other regularization methods include the bridge regression (Frank and Friedman, 1993), group Lasso Yuan and Lin (2006) and Dantzig selector (Candes et al., 2007). More detailed

review of regularization methods can be found in Negahban et al. (2009) and Fan and Lv (2010).

1.4. Model Selection Criteria

Conventional variable selection methods like stepwise regression, best subset selection often involve choosing the best model among a sequence of candidate models with different complexities. Similarly, there is a sequence of candidate models in regularized regression indexed by the tuning parameter. The choice of tuning parameter is important for selecting the best model. We discuss several model selection criteria in this section for choosing the optimal tuning parameter.

Cross-Validation

Cross-validation is a simple, intuitive and commonly used method to estimate the prediction error by splitting the data into training and testing sets. The following algorithm illustrates the way that cross-validation works. For choosing the optimal tuning parameter λ, grid search over λ or $\log(\lambda)$ is routinely performed over the interval $[\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\min}$ is zero or some small positive number and $\lambda_{\max}$ is often set as the smallest number such that all coefficients are shrunken as zero. In practice, there may exist more than two candidate models yielding small cross-validation errors, which do not differ much with each other. In this case we use the one-standard-error rule to select the tuning parameter. Specifically, the standard error of cross-validation error is calculated for each λ, then choose the sparsest model whose cross-validation error is within the one standard error of the model with lowest cross-validation error.

Algorithm 1: K-fold cross-validation

randomly divide the dataset into K folds $S_1, S_2, \ldots, S_K$;
for $k = 1, 2, \ldots, K$ **do**
    fit the model on training set $S_{[-k]}$ (the complete dataset without the S_k);
    calculate the prediction error e_k on test set S_k;
end
calculate the cross-validation error $\sum_{k=1}^{K} e_k/K$

Information Criteria

Basically, the information criteria are measures for model selection based on likelihood. A typical information criterion consists of two parts: measure of model fitting with some penalty on the model complexity which can be written as

$$IC = -2 \log L(\theta) + kd,$$

where k is the number of estimated parameters in the model and d is some coefficient. The Akaike information criterion (AIC, Akaike 1974) corresponds to the case where $d = 2$. Founded on information theory, the AIC is obtained by minimizing the information lost by a given model, measured by Kullback–Leibler distance of the likelihood function of the candidate model from the unknown true likelihood function. In the literature, the AIC is mostly criticized for not yielding the consistent estimator. However, in situations where statistical models are used to approximate complex systems for certain objectives, a "good" model is preferred over the "true" model, as suggested in Konishi and Kitagawa (2008). The Bayesian information criterion (BIC, Schwarz et al. 1978) corresponds to the case where $d = \log(n)$. The BIC is obtained by maximizing the posterior probability, from the Bayesian point of view. Compared with AIC, the BIC is shown to be consistent in model selection. Variants of BIC were also developed in statistical literature accounting for the high dimensionality, for example in Chen and Chen (2012).

Besides the criteria discussed above, there are other model selection criteria in the literature, for example, Mallow's C_p (Mallows, 1973), the risk inflation criterion (RIC, Foster and George 1994) and generalized cross-validation (GCV, Golub et al. 1979).

1.5. Measurement Error Models

Measurement error is ubiquitous in real data analysis. In medical and clinical studies, it is often the case that some or all of the variables cannot be precisely or directly measured. Instead, indirect or proxy measurements are used. Several examples of ME models are listed as follows. (1) In Framingham Heart Study, a cohort of residents are followed for the development of coronary heart disease. Important risk factors such as long-term average systolic blood pressure, cholesterol level and body mass index cannot be observed directly. The observed values are the measurements during clinic visits on a given day. (2) In lung cancer study, the outcome of interest is the incidence of lung cancer. The actual amount of pollutants inhaled by individual is an important factor, which is unobservable. Instead, the observed exposure is measured by some monitoring station for the air pollution level at

certain area. (3) When evaluating the efficacy of a drug in pharmacokinetics study, the actual absorption of the medical substance in bloodstream cannot be measured. The predetermined dosage of the drug are used for analysis instead. (4) Fertilizers contain essential elements that are needed by plants, such as nitrogen, phosphorus and potassium. The predetermined dose of the fertilizer serves as proxy for the absorbed amount, in modeling the relationship between the yield of crop and the amount of fertilizer used. (5) In biology, the biomass cannot be measured precisely. Indirect measurements like estimation with satellite image is used to approximate the true value of biomass. (6) In health-care research, it is commonly known that ME is manifest and masking the relationship between diet and health status. The long-term nutrition, fat, energy, carbohydrates intake, alcohol and/or tobacco consumption are unobservable, which are approximated by the self-report food questionnaire or a 24 hour recall interview. (7) In econometrics, a classical example is to model the relationship between wage and factors like education, experience, age, gender, race. Obviously the factors like education and experience are unobservable and are approximated by measures like schooling and number of work years. A commonly used ME model for the true predictor x and its proxy x^* is defined as

$$x^* = x + \delta, \tag{1.5.1}$$

where δ is a random measurement error satisfying $E(\delta|x) = 0$. Model (1.5.1) is also called the classical additive ME model. In other situations like the lung cancer example, the relationship is modeled as

$$x = x^* + \delta, \tag{1.5.2}$$

where $E(\delta|x^*) = 0$. Model (1.5.2) is called Berkson ME model. To gain some insights about the two ME models, consider the following two numerical examples. For the classical ME example, the data is generated as follows.

$$\begin{aligned}
y_1 &= \sin(x_1) + \epsilon_1, \\
x_1 &\sim \text{Uniform}(-\pi, \pi), \\
\epsilon_1 &\sim N(0, 0.2), \\
x_1^* &= x_1 + \delta_1, \\
\delta_1 &\sim N(0, \pi)
\end{aligned}$$

such that $\sigma^2_{\delta_1}/\sigma^2_{x_1} \approx 0.95$. For the Berkson ME example, generate the data as below.

$$
\begin{aligned}
y_2 &= \sin(x_2) + \epsilon_2, \\
x_2 &= x_2^* + \delta_2, \\
\epsilon_2 &\sim N(0, 0.2), \\
x_2^* &\sim \text{Uniform}(-\pi, \pi), \\
\delta_2 &\sim N(0, \pi).
\end{aligned}
$$

The sample size is 500 in both examples. The scatter plot of both examples are shown in the Figure (1.8) and (1.9), where black circle represents sample (x, y) and red cross represents sample (x^*, y). It can be observed that the ME masks the relationship between x and y in both ME models.

Figure 1.8: Scatterplot with classical ME

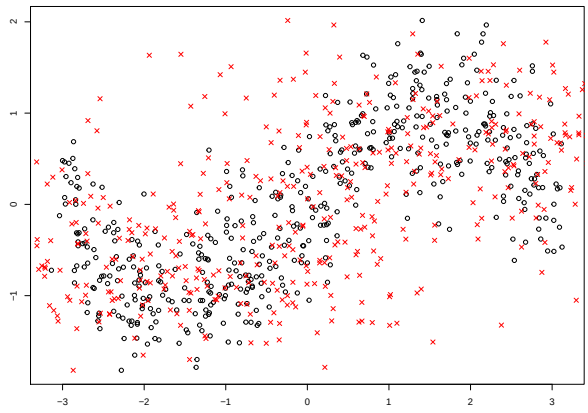
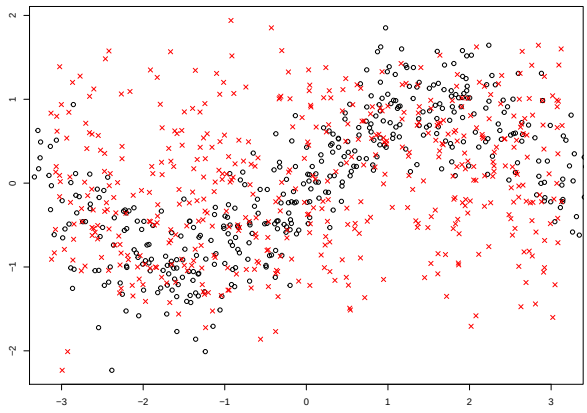


Figure 1.9: Scatterplot with Berkson ME



Consider another two simple linear regression examples. The model setting is the same as above except that the response y_j is generated by $y_j = 1.5x_j + \epsilon_j$, $j = 1, 2$ in both classical and Berkson ME models. The sample size is $n = 200$. The fitted least squares regression line is shown in the graphs below. The black points and lines represent the scatterplots and regression lines fitted with true datasets (x, y). The red color corresponds to the datasets with ME (x^*, y). It can be observed that the slope of regression line is attenuated towards zero in classical ME model whereas the Berkson ME does not cause bias in estimation but with increased variance.

Figure 1.10: Fitted line with classical ME

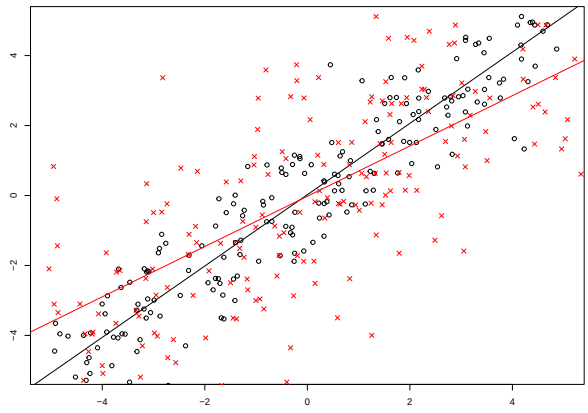
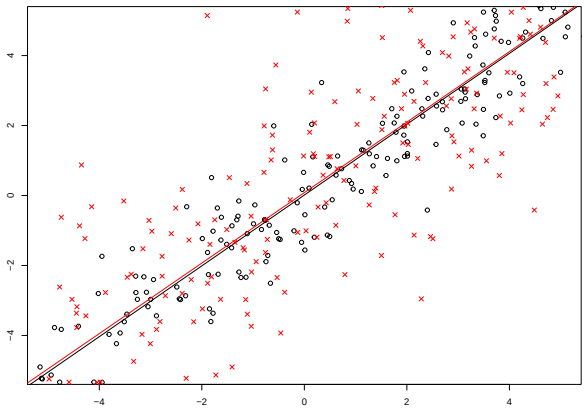


Figure 1.11: Fitted line with Berkson ME



The true patterns of the datasets can be masked by ME and the estimation and statistical inference become invalid in the presence of ME. Different methods were proposed to correct for the ME effects. We briefly introduce some ME model analysis techniques in this section.

Method of Moments

As illustrated in Figure (1.10), the ME causes estimation bias for least squares estimator in linear regression. Thus the resulting naive estimator is inconsistent. To quantify this effect, consider a univariate simple linear ME model

$$y = \alpha + \beta x + \epsilon, \tag{1.5.3}$$

$$x^* = x + \delta, \tag{1.5.4}$$

where $E(\epsilon|x, x^*) = 0$ and $E(\delta|x) = 0$. It is known that if the ME is ignored, the naive estimator $\hat{\beta}_{NA} \xrightarrow{p} \kappa\beta$, where $\kappa = \sigma_x^2/(\sigma_x^2 + \sigma_\delta^2)$ is often referred as attenuation factor or reliability ratio. Therefore, we can obtain an bias-corrected estimator $\hat{\beta} = \hat{\beta}_{NA}/\kappa$ if κ is known or can be consistently estimated. Since it can be observed that $\sigma_{x^*}^2 = \sigma_x^2 + \sigma_\delta^2$ from (1.5.4), κ can be estimated if one of the following three terms σ_x^2, ME variance σ_δ^2 or the ratio $\sigma_x^2/\sigma_\delta^2$ is known. The extension of method of moments to multivariate linear model is straightforward and can be found in Fuller (2009).

Regression Calibration

The regression calibration is one of the statistical methods that correct for the covariates ME effects. The basic idea is to replace the unobserved covariates x with the regression of x on the observed x^* along with other ME free covariates. Consider the model (1.5.3) and (1.5.4). Taking the conditional expectation on x^* we obtain the following equation

$$E(y|x^*) = \alpha + \beta E(x|x^*).$$

This equation suggests that the unbiased estimator of (α, β) can be obtained by regressing y on $E(x|x^*)$. Since x is not observed, extra information is required for estimating $E(x|x^*)$. Usually we need a validation set where the true value of x can be observed. Or there exists an unbiased instrument w such that the regression of w on x^* yields the same results as if we are regressing x on x^* according to Carroll et al. (2006). The estimation can also be obtained with replicate data. For example, consider the model (1.5.4) where $x^*_{i1} = x+\delta_{i1}$ with a second measurement $x^*_{i2} = x + \delta_{i2}$, where $i = 1, 2, \ldots, n$, $x \sim N(\mu_x, \sigma_x^2)$ and $\delta_1, \delta_2 \sim N(0, \sigma_\delta^2)$. The parameters μ_x, σ_x^2 and σ_δ^2 can be estimated with standard analysis of variance technique. Then the conditional expectation of x on x^* is given by

$$E(x|x^*) = (\sigma_\delta^2 \mu_x + \sigma_x^2 x^*)/(\sigma_\delta^2 + \sigma_x^2).$$

The generalization to multiple covariates with ME and repeated measurements is straightforward. Note that the regression calibration is an approximation method such that the resulting estimator is approximately consistent in generalized linear models. It should be taken with caution when working with nonlinear models where the performance of regression calibration can be worse. A comprehensive and detailed review of regression calibration method can be found in Chapter 4 of Carroll et al. (2006).

Simulation Extrapolation

The simulation extrapolation (SIMEX) proposed by Cook and Stefanski (1994) is a simulation-based approach correcting for ME effect. Given the simple linear model (1.5.3) with classical additive ME (1.5.4), it is known that the naive estimator

$$\hat{\beta}_{NA} \xrightarrow{p} \beta\sigma_x^2/(\sigma_x^2 + \sigma_\delta^2).$$

The SIMEX procedure consists of two stages, simulation and extrapolation. In the simulation stage, pseudo errors δ_{bi} are generated and added to x^* such that

$$x^*_{bi}(\zeta) = x^*_i + \sqrt{\zeta}\delta_{bi},\ i = 1, 2, \ldots, n,\ b = 1, 2, \ldots, B, \tag{1.5.5}$$

where δ_{bi} is independently and identically distributed from $N(0, \sigma^2_\delta)$. For each simulated dataset indexed by b, the estimator

$$\hat{\beta}_b(\zeta) \xrightarrow{p} \beta\sigma^2_x/(\sigma^2_x + (1+\zeta)\sigma^2_\delta).$$

Averaging over b we obtain the estimator $\hat{\beta}_{SIM}(\zeta) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b(\zeta)$. According to the asymptotic form of $\hat{\beta}_b(\zeta)$, the estimator is consistent if $\zeta = -1$. However, since the variance does not decrease as seen from the mechanism of (1.5.5), the consistent estimator when $\zeta = -1$ can only be extrapolated in this case. To this end, let ζ_m be a sequence of positive numbers with increasing order $0 < \zeta_1 < \zeta_2 < \cdots < \zeta_M$. Then a sequence of estimators are obtained and denoted as $\hat{\beta}_{SIM}(\zeta_1), \hat{\beta}_{SIM}(\zeta_2), \ldots, \hat{\beta}_{SIM}(\zeta_M)$. As a function of ζ_m, the function $\hat{\beta}_{SIM}(\zeta_m)$ is then extrapolated to $\zeta = -1$ to obtain the proposed SIMEX estimator. The extension of SIMEX method to more complicated cases where the ME is non-additive or σ_δ is unknown can be found in Chapter 5 of Carroll et al. (2006).

Instrumental Variable Method

Compared with the methods that require the variance of ME σ^2_δ to be known or can be estimated, the instrumental variable (IV) method is applicable provided that there exists some instrumental variable that satisfies certain conditions. Specifically, an instrument w for x is said to be valid if it is correlated with x, uncorrelated with the ME δ and does not contain any information of the response variable after accounting for all other possible covariates. In this section, the IV method is illustrated under the setting of simple linear model (1.5.3) and (1.5.4). Now assume there exists an instrumental variable w such that it is related with x through

$$x = \gamma w + u, \tag{1.5.6}$$

where u is independent of w with $E(u) = 0$. Note that γ can be consistently estimated from (1.5.4) and (1.5.6) using the regular least squares method. Further, the equation (1.5.3) can be rewritten as

$$y = \alpha + \beta\gamma w + \epsilon^*,$$

where $\epsilon^* = \beta u + \epsilon$. Thus β can be consistently estimated by regressing y on γw, in which case the computation procedure is also called the two-stage least squares method.

1.6. Variable Selection with Measurement Error

In this section we review some existing literature on variable selection problems with measurement errors, under the settings of regularized ME models.

Yi et al. (2015) proposed the estimation and model selection procedure for longitudinal data which is subject to missingness and measurement errors. The algorithm is briefly described as below. Firstly the SIMEX procedure is employed and a sequence of estimators $\hat{\beta}(\zeta)$ (see (1.5.5)) from the corresponding unbiased estimating equations. Then the SIMEX estimator $\tilde{\beta}$ is obtained by extrapolating ζ to the value $\zeta = -1$, namely $\tilde{\beta} = \hat{\beta}(-1)$. Note that in Yi et al. (2015) the SIMEX estimator $\tilde{\beta}$ is assumed to be normally distributed asymptotically. Then the final step is to solve the following minimization problem of some penalized quadratic loss function

$$l(\beta) = \frac{1}{2}(\beta - \tilde{\beta})^T V_n (\beta - \tilde{\beta}) - n \sum_{j=1}^{d} p_\lambda(|\beta_j|)$$

for some positive definite weight matrix V_n. The estimator $\tilde{\beta}$ is shown to have oracle property under some standard conditions in the settings of regularized regression.

Ma and Li (2010) proposed penalized estimating equation for variable selection in ME models. For a general parametric model, denote $p_{y|xz}(y|x, z; \beta)$ as the conditional probability density function of the response variable y on the covariates (x, z), where x is unobservable or measured with errors, and z is ME free. Suppose instead of x, we observe x^* where $x^* = x + \Delta$. The unbiased estimation equation is defined as

$$S(x^*, z, y) = S^*_\beta(x^*, z, y) - E^*(a(x, z)|x^*, z, y),$$

where

$$S^*_\beta(x^*, z, y) = \partial \log \int p_{x^*|x,z}(x^*|x, z) p_{y|x,z}(y|x, z) p^*_{x|z}(x|z) \mathrm{d}\mu(x) / \partial\beta,$$

and $a(x, z)$ is some function that satisfies

$$E[E^*\{a(x, z)|x^*, z, y\}|x, z] = E\left\{S^*_\beta(x^*, z, y)|x, z\right\}$$

with the expectation E^* calculated with respect to some posited density function $p^*_{x|z}(x|z)$. The penalized estimating equation is then defined as

$$\sum_{i=1}^{n} S(x_i^*, z_i, y_i; \beta) - np'_\lambda(\beta) = 0,$$

where $p'_\lambda(\beta) = (p'_\lambda(\beta_1), p'_\lambda(\beta_2), \ldots, p'_\lambda(\beta_d))^T$. For certain types of penalty functions and properly chosen tuning parameters, the resulting estimator is shown to have properties like consistency and asymptotic normality.

Liang and Li (2009) proposed two variable selection approaches for partially linear ME models. For simplicity, we illustrate the idea under the settings of simple linear regression model. The method can be generalized to partially linear ME model by using partial residual-based loss function. Consider the linear ME model

$$\begin{aligned} y &= \beta_x^T x + \beta_z^T z + \epsilon \\ x^* &= x + \delta \end{aligned} \tag{1.6.1}$$

where $E(\epsilon|x, z) = 0$, δ is independent of (x, z, ϵ) with $E(\delta) = 0$. Denote $\beta = (\beta_x^T, \beta_z^T)^T$, Σ_δ as the covariance matrix of the random error δ. The first estimator based on correction for attenuation method is then given by minimizing the following penalized least squares function

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_x^T x_i^* - \beta_z^T z_i)^2 - \frac{n}{2}\beta_x^T \Sigma_\delta \beta_x + n\sum_{j=1}^{d} p_\lambda(|\beta_j|).$$

The second method, penalized quantile function is given by

$$\sum_{i=1}^{n} \rho_\tau \left((y_i - \beta_x^T x_i^* - \beta_z^T z_i)/\sqrt{1 + \beta_x^T C_\delta \beta_x} \right) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|),$$

where $\rho_\tau(r) = \tau \max(r, 0) + (1 - \tau)\max(-r, 0)$ and C_δ is some matrix assumed to be known. The estimators of the two methods are shown to have good properties as in Liang and Li (2009).

Huang and Zhang (2013) proposed to perform variable selection in linear ME models via penalized score functions. Under the linear ME model (1.6.1) with the assumption that the random error $\epsilon \sim N(0, \sigma_\epsilon^2)$, the quantity

$$\Delta = x^* + y\Sigma_\delta \beta_x / \sigma_\epsilon^2$$

is a sufficient statistic for x as shown in (Stefanski and Carroll, 1987). The conditional score function is given by

$$S_1(\beta, \sigma_\epsilon^2) = \begin{bmatrix} (y - E(y|\Delta, z)) \begin{pmatrix} \Delta \\ z \end{pmatrix} \\ \frac{n-d}{n}\sigma_\epsilon^2 - \frac{(y - E(y|\Delta, z))^2}{V(y|\Delta, z)/\sigma_\epsilon^2} \end{bmatrix},$$

where

$$E(y|\Delta, z) = (\beta_x^T \Delta + \beta_z^T z)/(1 + \beta_x^t \Sigma_\delta \beta_x/\sigma_\epsilon^2)$$

and

$$V(y|\Delta, z) = \sigma_\epsilon^2/(1 + \beta_x^t \Sigma_\delta \beta_x/\sigma_\epsilon^2).$$

Another unbiased estimating equation called corrected score function is given by

$$S_2(\beta) = \begin{bmatrix} (y - \beta_x^T x^* - \beta_z^T z)x^* + \Sigma_\delta \beta_x \\ (y - \beta_x^T x^* - \beta_z^T z)z \end{bmatrix}.$$

Then the penalized score equations are defined as

$$S_k(\beta) - np'_\lambda(\beta) = 0,$$

where $k = 1, 2$ refers to penalized conditional score and penalized corrected score method, respectively. Specifically, under the score-based information criteria, Huang and Zhang (2013) showed that the estimation procedure is consistent in model selection.

The regularized regression for linear ME model with IV is presented in next chapter. The linear ME model settings, methods and theoretical results are presented in Section 1. Numerical examples comparing the proposed estimator with naive estimator using different model selection criteria are given in Section 2, followed by a real data application on diabetes example in Section 3. The proofs of theorem are relegated to the Section 4.

Chapter 2

Regularized Regression in Linear ME Model

Massive and high-dimensional data are becoming available in many areas, such as astronomy and physics, genome and health science, business and finance, social media, signal processing and imaging, etc. For a given response variable of interest, the number of potential predictors can be very large. A subset of important covariates can improve the prediction accuracy and the interpretability of the model, which is usually done through regularized regression. However, most of the existing literature assume the data are measured precisely, which is not the case in many real applications. When there is ME in the dataset, the so-called oracle property does not exist anymore, which results in estimation and selection bias. Hence, under the ME model settings, how to reduce the dimension by removing the redundant features, without losing those important ones, is of primary interest. We introduce the proposed regularized instrumental variable method for linear ME model in this chapter.

2.1. Regularized Linear ME Model

In this section we discuss the instrumental variable method under the settings of linear ME model. In particular, we compare the regularized naive estimator with the proposed regularized instrumental variable estimator through different scenarios of numerical examples. Different model selection criteria are also discussed. In addition, the proposed RIV estimator is shown to have the oracle property, which is consistent in both variable selection and parameter estimation. Specifically, consider a linear regression model

$$y = \alpha + \beta_x^T x + \beta_z^T z + \epsilon, \tag{2.1.1}$$

where $x \in \mathbb{R}^p$ is a vector of covariates that are unobservable or measured with errors, $z \in \mathbb{R}^q$ is a vector of error-free covariates, the coefficients $\beta = (\beta_x^T, \beta_z^T)^T \in \mathbb{R}^d$ is assumed to be sparse. Without loss of generality, assume the intercept α is zero. If the covariates x are observable and are measured precisely, the coefficients can be estimated through the regularized least squares method

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_x^T x_i - \beta_z^T z_i)^2 + n \sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|) \right) \tag{2.1.2}$$

for a random sample (y_i, x_i, z_i), where $p_{\lambda_n}(\cdot)$ is some penalty function with tuning parameter λ_n. However, in real applications, some or all of the predictors are usually unobservable or are measured with errors. For example, long-term average systolic blood pressure is an important factor affecting the cardiovascular heart disease and is generally accepted in the literature to have ME issues. Suppose the covariates x are unobservable, instead we observe

$$x^* = x + \delta, \tag{2.1.3}$$

where δ is a random ME. Further assume that there exists an instrument variable (IV) w that is related with x through the equation

$$x = \Gamma w + u, \tag{2.1.4}$$

where Γ is a $p \times l$ matrix with rank p, u is independent of w with $E(u) = 0$ and $E(uu^T|z) = \Sigma_u$. The random errors in (2.1.1) and (2.1.3) are assumed to satisfy $E(\epsilon|x, z, w) = 0$ with constant variance and $E(\delta|x, z, w) = 0$. For an independently and identically distributed random sample (y_i, x_i^*, z_i, w_i), let $\tilde{w}_i = (w_i^T, z_i^T)^T$, $\tilde{\Gamma} = \operatorname{diag}(\hat{\Gamma}, I_q)$, $\tilde{x}_i = (\hat{x}_i^T, z_i^T)^T$, where $\hat{x}_i = \hat{\Gamma} w_i$, $\hat{\Gamma}$ is a consistent estimator of Γ and can be estimated by multivariate least squares fitting of x^* on w, i.e.

$$\hat{\Gamma} = (\sum_{i=1}^{n} x_i^* w_i^T)(\sum_{i=1}^{n} w_i w_i^T)^{-1}.$$

In addition, denote β_0 as the true model parameter,

$$\begin{aligned} \beta_J &= \{\beta_j, j \in J\}, \ J = \{j : \beta_{0j} \neq 0\}, \\ \beta_{J^c} &= \{\beta_j, j \in J^c\}, \ J^c = \{j : \beta_{0j} = 0\}, \end{aligned}$$

$s = |J|$ the cardinality of J, and $\tilde{\Gamma}_J$ the matrix consisting of rows of $\tilde{\Gamma}$ corresponding to the index set J. Furthermore, write

$$\begin{aligned} a_n &= \max\{p'_{\lambda_n}(|\beta_{0J}|), j \in J\}, \\ b_n &= \max\{p''_{\lambda_n}(|\beta_{0J}|), j \in J\}, \\ \boldsymbol{b} &= p'_{\lambda_n}(|\beta_{0J}|) \circ \text{sign}(|\beta_{0J}|), \\ \Sigma &= \text{diag}(p''_{\lambda_n}(|\beta_{0J}|)), \end{aligned}$$

where $\circ$ is Hadamard product.

Similar as (2.1.2), the regularized instrumental variable estimator is defined as the minimizer of the following objective function

$$Q_n(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \tilde{x}_i^T\beta)^2 + n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|). \tag{2.1.5}$$

Remark 1. Since the naive estimator is inconsistent in estimation and selection generally, the observed covariates are replaced by its corrected version $\tilde{x}$ based on instruments. Furthermore, since the objective function in (2.1.5) involves the non-independence of random sample $(y_i, \tilde{x}_i)$ due to the involvement of $\hat{\Gamma}$, the standard results for regularized linear regression cannot be applied.

Remark 2. In general, a larger value of λ_n imposes more weights on the penalty and produces a sparser model. The tuning parameter λ_n can be chosen in different ways. For example, the Akaike information criterion (AIC), Bayesian information criterion (BIC), k-fold cross validation and generalized cross validation (GCV). With a properly chosen tuning parameter λ_n, the proposed estimator is shown to have the following properties.

Theorem 1. *If $a_n = O(n^{-1/2})$, $b_n = o(1)$ and $E(\tilde{w}\tilde{w}^T)$ is positive definite, then there exists a local minimizer $\hat{\beta}$ of $Q_n(\beta)$ such that $||\hat{\beta} - \beta_0|| = O_p(n^{-1/2})$.*

Theorem 2. *If $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ and $\liminf_{n\to\infty} \liminf_{\xi\to 0^+} p'_{\lambda_n}(\xi)/\lambda_n > 0$, then with probability approaching 1, the root n consistent estimator $\hat{\beta}$ in (2.1.5) satisfies*

(a) $\hat{\beta}_{J^c} = 0$,

(b) $\hat{\beta}_J$ has the following asymptotic normal distribution

$$\sqrt{n}(H + \Sigma)(\hat{\beta}_J - \beta_{0J} + (H + \Sigma)^{-1}\boldsymbol{b}) \xrightarrow{d} N(0, DCD^T),$$

where

$$H = \tilde{\Gamma}_{0J} E\tilde{w}\tilde{w}^T\tilde{\Gamma}_{0J}^T,$$

$$D = (I_s, \tilde{\Gamma}_{0J}(\beta_{0J}^T \otimes I_{l+q})),$$

$$C = E(KK^T)$$

and

$$K = \begin{pmatrix} \tilde{\Gamma}_{0J}\tilde{w}(y - \beta_{0J}^T\tilde{\Gamma}_{0J}\tilde{w}) \\ (\tilde{\Gamma}_{0J}\tilde{w} - \tilde{x}_J^*) \otimes \tilde{w}) \end{pmatrix}.$$

Remark 3. According to Theorem 2, the asymptotic covariance matrix of $\hat{\beta}_J$ can be estimated with the following consistent estimator

$$n^{-1}(\hat{H}_n + \Sigma(\hat{\beta}_J))^{-1}(\hat{D}_n\hat{C}_n\hat{D}_n^T)(\hat{H}_n + \Sigma(\hat{\beta}_J))^{-1},$$

where $\hat{H}_n$, $\hat{D}_n$, $\hat{C}_n$ are the sample counterparts of H, D, C evaluated at $\hat{\beta}_J$ and $\Sigma(\hat{\beta}_J) = \text{diag}(p''_{\lambda_n}(|\hat{\beta}_J|))$.

Remark 4. For some penalty functions (e.g. SCAD and MCP), $\boldsymbol{b}$ and Σ are both zero when the tuning parameter λ_n is sufficiently small. Hence the resulting estimator has the oracle performance such that $\hat{\beta}_{J^c} = 0$ and the asymptotic distribution of $\hat{\beta}_J$ is given by

$$\sqrt{n}(\hat{\beta}_J - \beta_{0J}) \xrightarrow{d} N(0, H^{-1}DCD^TH^{-1}).$$

2.2. Simulation Studies

Finite sample simulations are conducted to assess the performance of the proposed estimator in this section. We compare the variable selection and parameter estimation results from regularized linear regression models using the following three random samples: the precisely measured datasets (x_i, z_i, y_i) without ME, the observed sample (x_i^*, z_i, y_i) ignoring ME, and the predicted sample $(\hat{x}_i, z_i, y_i)$ using instrumental variables. The results corresponding to the three methods are denoted as TR, NA and IV, respectively. The proposed method is implemented with SCAD penalty function. Other penalty functions are also included in example (2.2.4). According to Wang et al. (2007), the tuning parameter selected by BIC has the property of recovering the true model consistently for SCAD penalty. We use the BIC criteria to select the tuning parameter in simulation examples. The simulation examples using other model selection criteria are also discussed. In this chapter, the optimization is conducted using R package called *ncvreg* developed by Breheny and Huang (2011). The false positive (FP) reported in the table represents the average number of zero coefficients incorrectly estimated as nonzero. Similarly, the false negative (FN) represents the average

number of nonzero coefficients incorrectly estimated as zero. The MCC stands for Matthews correlation coefficient, which is a general measure of describing the confusion matrix of true/false positives/negatives and is defined as

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

The MCC ranges from -1 to 1, where large value indicates good prediction. Finally, the mean of $||\hat{\beta} - \beta_0||^2$ is denoted as the mean squared error (MSE).

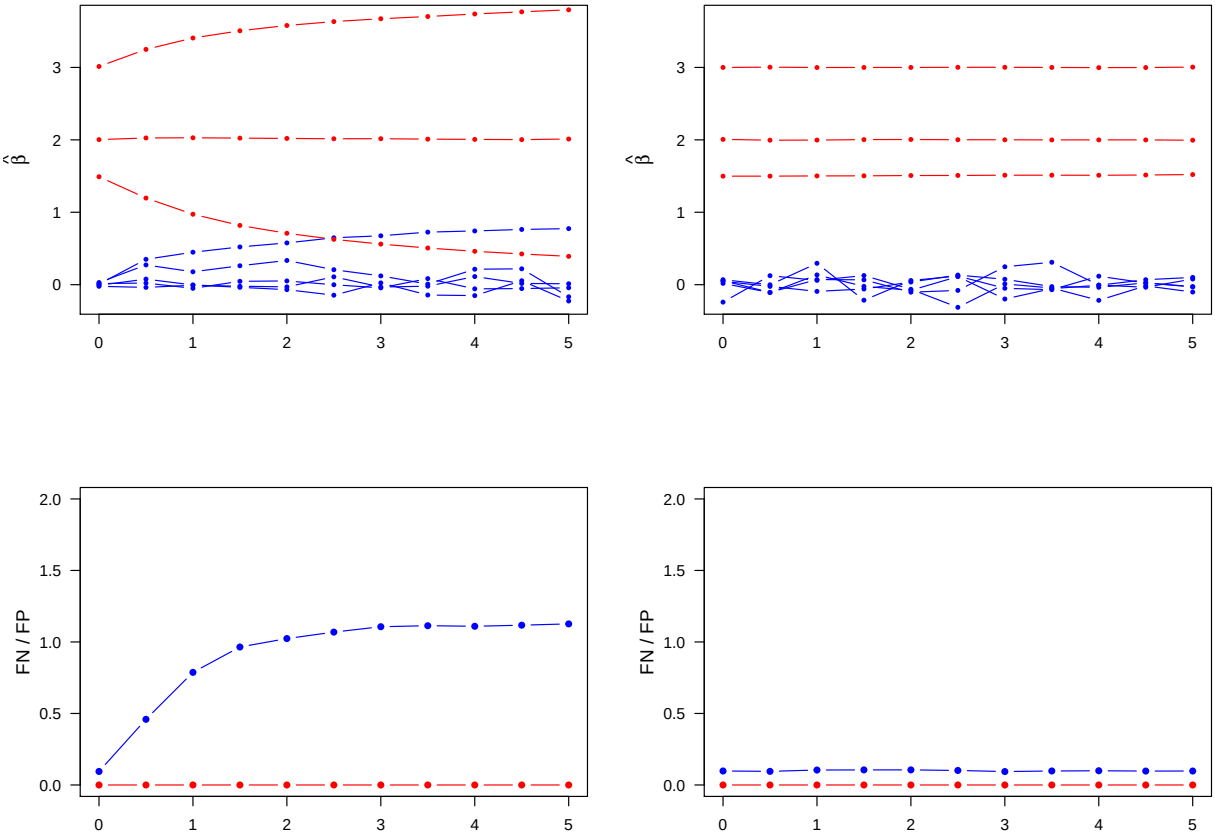
2.2.1 Numerical Example for Linear Model 1

In this example we simulate 1000 datasets consisting of 200 observations from the following linear regression model

$$y = \beta_x x + \beta_z^T z + \epsilon,$$

where $(\beta_x, \beta_z^T) = (1.5, 3, 0, 0, 2, 0, 0, 0)$ and ϵ is standard normal. In addition, the covariate x is generated as $x = 1.5w + u$ where $(z_1, w, z_2, \ldots, z_7)^T$ are jointly generated from $N(0, \Sigma)$ with $\Sigma_{ij} = 0.7^{|i-j|}$ and u is standard normal. In this example the correlation between w and x is around 0.83. The unobserved covariate is generated as $x^* = x + \delta$, where δ follows normal distribution with mean zero and variance σ_δ^2. Figure (2.1) shows the estimation mean, FP and FN across different values of σ_δ^2. The results from naive method are on the left hand side and the results of RIV method are on the right. For the naive method, the estimation is biased and FP is inflated. The bias and FP get larger as $\sigma_\delta^2/\sigma_x^2$ increases. As a comparison, the RIV estimator is robust against the magnitude of $\sigma_\delta^2/\sigma_x^2$ in terms of estimation, FP and FN.

Figure 2.1: Estimation and selection results of Example 2.2.1 with $n = 200$



The simulation results for $\sigma_\delta^2 = 2$ are reported in Table (2.1). It can be observed that the estimation is biased for naive method due to ME effects. Specifically, β_x is biased towards zero due to the attenuation effect and the estimation of β_1, β_2 are inflated due to the correlation structure of the covariates. In contrast, the estimation for TR and RIV method is close to the true value of model parameter. The selection results are shown in Table (2.2). Specifically, the table on the left shows the four summarized measures FP, FN, MCC and MSE, and the percentage of correct specification is shown in the table on the right. The results from the TR sample have the lowest FP, FN, MSE and highest MCC among all three methods. The IV method performs similarly as the TR model with respect to all four measures. In contrast, the selection bias is high when ME is ignored, as shown from the results of NA method. For NA method, it can also be observed that the false inclusion of z_2 contributes to the majority of FP, although the covariate z_2 is not affected by ME directly.

Table 2.1: Estimation eesults of Example 2.2.1 with $n = 200$, $\sigma_\delta^2 = 2$

	β_1= 3	β_x= 1.5	β_2= 0	β_3= 0	β_4= 2	β_5= 0	β_6= 0	β_7= 0
TR	3.00	1.50	-0.00	0.00	2.00	-0.00	0.00	-0.00
IV	3.01	1.51	0.00	-0.00	2.00	-0.00	-0.00	0.00
NA	3.66	0.59	0.61	0.01	2.01	0.00	0.00	-0.00

Table 2.2: Selection eesults of Example 2.2.1 with $n = 200$, $\sigma_\delta^2 = 2$

	FP	FN	MCC	MSE
TR	0.1	0.0	0.97	0.03
IV	0.1	0.0	0.97	0.07
NA	1.1	0.0	0.76	1.77

	z_1	x	z_2	z_3	z_4	z_5	z_6	z_7
TR	100	100	98	98	100	98	98	98
IV	100	100	98	98	100	98	98	98
NA	100	100	8	97	100	98	96	94

The estimation results of mean and standard deviations are shown in Table (2.3). The numbers in parentheses are standard errors. The standard error formula performs satisfactorily. The boxplots corresponding to Table (2.1) for all three methods are shown in Figure (2.2). The boxplot in the first, second and third row corresponds to results for TR, IV and NA, respectively. It can be observed that the estimation mean center around the true value of the coefficients for TR and IV methods, compared with the NA method which has substantial bias. In addition, the spread for all coefficients estimates is small. Note that there are a few outliers for the estimates of β_2 in IV method, which is due to the prediction error in the estimation procedure. The number of points marked as outliers is negligible compared to the total number of points in that column. The performance of variable selection among three methods ($\sigma_\delta^2 = 1$) with sample sizes $n = 50, 100, 200$ are reported in Table (2.4). As the sample size increases, it can be seen that both FP and FN decrease for TR and RIV methods, whereas the FP increases for naive method. In addition, the performance of MCC and MSE is better for TR and RIV methods than that of the naive method. The selection is biased for naive method no matter how large the sample size is.

Figure 2.2: Boxplots of coefficient estimates in Example 2.2.1 with $n = 200$, $\sigma_\delta^2 = 2$; First row: TR, second row: IV, third row: NA

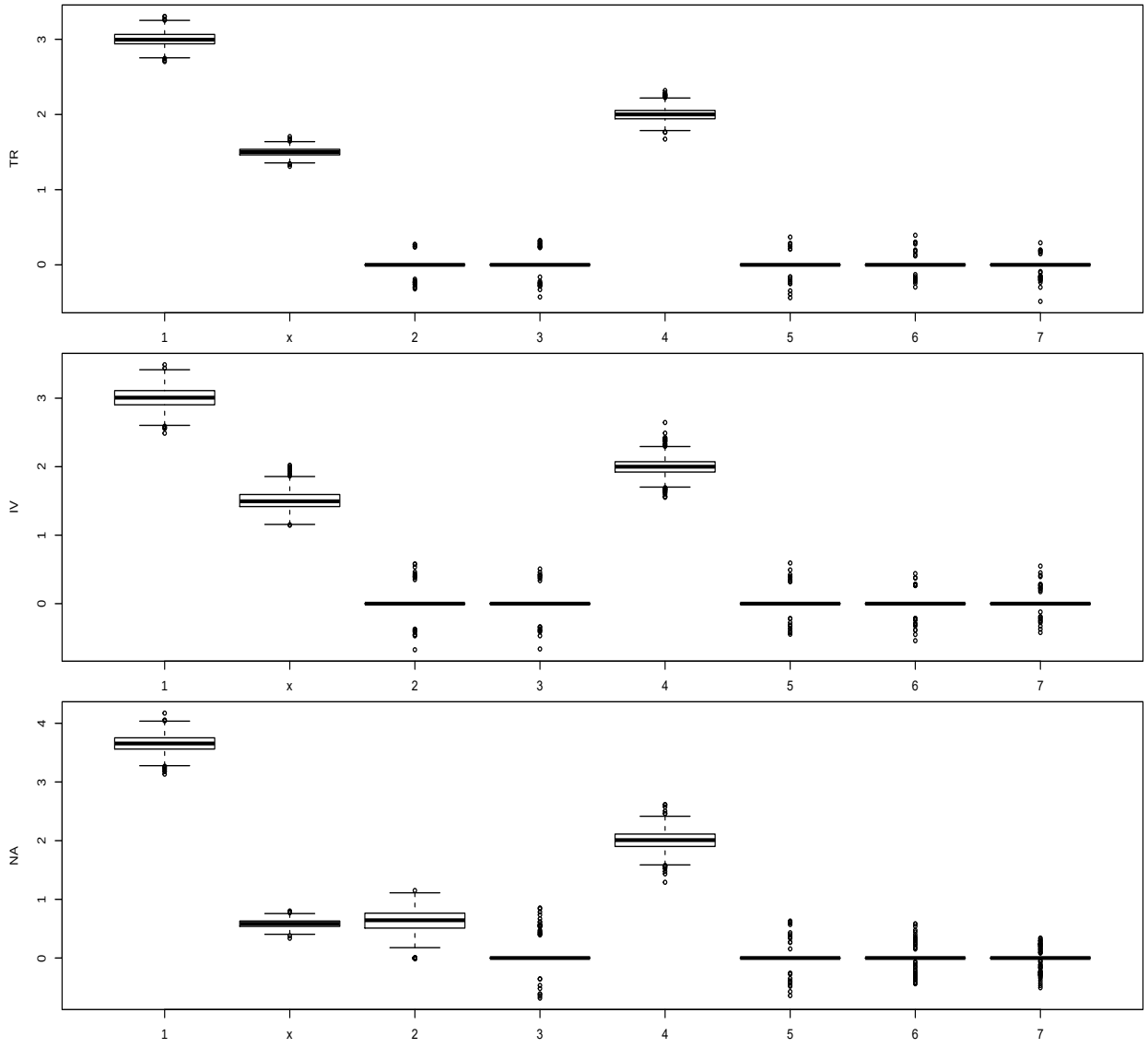


Table 2.3: Mean and standard errors for nonzero coefficients of Example 2.2.1 with $n = 200$, $\sigma_\delta^2 = 2$

	Mean	SD
$\hat{\beta}_1$	3.01 (0.180)	0.178 (0.018)
$\hat{\beta}_x$	1.50 (0.146)	0.138 (0.014)
$\hat{\beta}_5$	1.99 (0.151)	0.140 (0.032)

Table 2.4: Selection results with different sample size of Example 2.2.1 with $\sigma_\delta^2 = 1$

	n=50				n=100				n=200			
	FP	FN	MCC	MSE	FP	FN	MCC	MSE	FP	FN	MCC	MSE
TR	0.3	0	0.92	0.08	0.1	0	0.96	0.03	0.1	0	0.98	0.01
IV	0.3	0	0.92	0.52	0.2	0	0.96	0.19	0.1	0	0.98	0.07
NA	0.5	0	0.87	0.87	0.6	0	0.85	0.71	0.8	0	0.82	0.64

2.2.2 Numerical Example for Linear Model 2

In this example we simulated 1000 datasets consisting of 200 observations from the following linear regression model

$$y = \beta_x x + \beta_z^T z + \epsilon,$$

where $(\beta_x, \beta_z^T) = (3, 1.5, 0, 0, 2, 0, 0, 0)$, and ϵ is standard normal. In addition, the covariate x is generated as $x = 1.5w + u$ where $(z_1, -w, z_2, \ldots, z_7)^T$ are jointly generated from $N(0, \Sigma)$ with $\Sigma_{ij} = 0.7^{|i-j|}$ and u is standard normal. Note that in this example the covariate x is negatively correlated with all other covariates and the values of the first two coefficients are interchanged. The estimation, FP and FN are shown in Figure (1.1). In this example both the values of FP and FN increase with σ_δ^2 for naive method, as seen from the bottom left graph. Similarly as in Example (2.2.1), the RIV estimator is robust against the magnitude of σ_δ^2. The simulation results where $\sigma_\delta^2 = 2$ are reported in Table (2.5) and (2.6). Besides the similar patterns that are observed in Example 1, it can be seen that the NA method have both high FP and high FN in selection results. The increase of FN is due to the fact that z_1 is dropped from the model incorrectly, as shown in Table (2.6). On the other hand the TR and IV methods perform well in recovering the true model. The selection results for $\sigma_\delta = 2$ with sample size $n = 50, 100, 200$ are reported in Table (2.7). The TR and IV method perform like the oracle procedure as both values of FP and FN are decreasing towards zero as the sample size increases. For the NA method, the FP and FN remain at a high level regardless of the sample size. The boxplots for all coefficients of three methods are shown in Figure (2.3). There are two points worth noting here. First, the mean estimates of β_2 center around the true values for TR and IV methods. Whereas it is attenuated towards zero due to the ME effect for NA method, of which the spread is small from the interquartile range. Second, the estimate for β_1 of NA method is centered around zero, which shows the ME effect on other nonzero coefficients from another point of view.

Table 2.5: Estimation results of Example 2.2.2 with $n = 200$, $\sigma_\delta^2 = 2$

	β_1= 1.5	β_x= 3	β_2= 0	β_3= 0	β_4= 2	β_5= 0	β_6= 0	β_7= 0
TR	1.50	3.00	0.00	0.00	2.00	0.00	0.00	-0.00
IV	1.49	3.00	-0.00	-0.00	2.00	-0.00	0.01	0.00
NA	0.05	1.15	-1.20	-0.00	1.99	0.00	0.00	-0.00

Table 2.6: Selection results of Example 2.2.2 with $n = 200$, $\sigma_\delta^2 = 2$

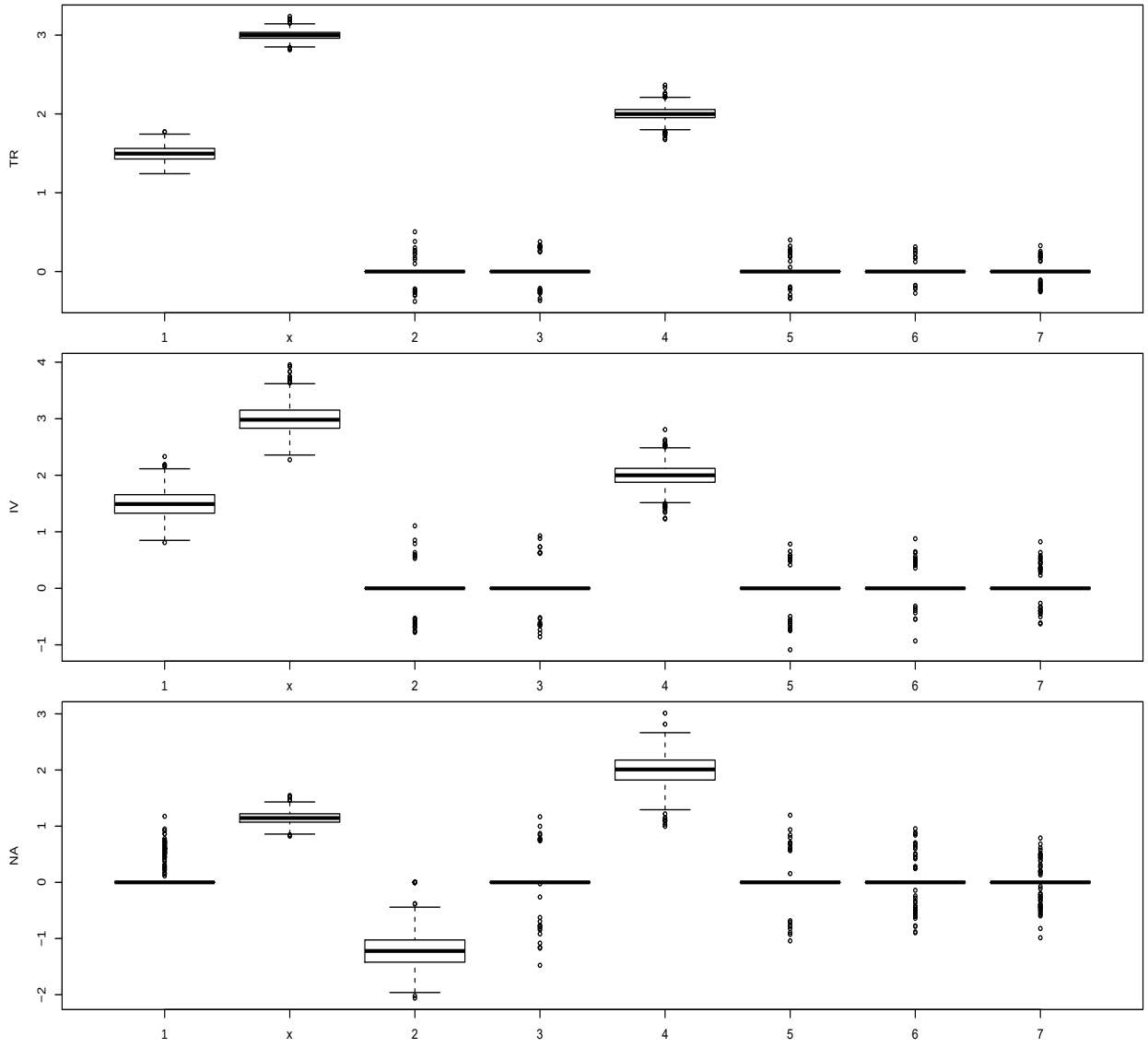
	FP	FN	MCC	MSE
TR	0.1	0.0	0.97	0.03
IV	0.1	0.0	0.97	0.20
NA	1.1	0.9	0.46	7.26

	z_1	x	z_2	z_3	z_4	z_5	z_6	z_7
TR	100	100	98	98	100	98	99	97
IV	100	100	98	98	100	98	97	97
NA	8	100	2	98	100	98	96	94

Table 2.7: Selection results with different sample size of Example 2.2.2 with $\sigma_\delta^2 = 2$

	n=50				n=100				n=200			
	FP	FN	MCC	MSE	FP	FN	MCC	MSE	FP	FN	MCC	MSE
TR	0.3	0	0.93	0.14	0.2	0	0.96	0.05	0.1	0	0.98	0.02
IV	0.5	0.2	0.77	3.34	0.3	0	0.88	0.91	0.2	0	0.96	0.34
NA	1	0.9	0.51	7.46	0.9	0.9	0.53	6.17	1	0.8	0.54	5.69

Figure 2.3: Boxplots of Example 2.2.2 with $n = 200$, $\sigma_\delta^2 = 2$; First row: TR, second row: IV, third row: NA



2.2.3 Numerical Example for Linear Model 3

In this example we show that in some special cases the ME can cause the nonzero coefficients to be incorrectly shrunken to zero asymptotically, even if the corresponding covariates are error free. Let (w, z) be generated the same way as in Example (2.2.1) and $x = 0.707w + u$, $u \sim N(0, 0.5)$ so that $\text{cor}(w, x) \approx 0.7$. Further, let $x^* = x + \delta$, $\delta \sim N(0, 1)$ and $\beta_0 = (0.7, -1^*, 0.2, 0, 0, 0.7)$. It is known that the naive estimator

$$\hat{\beta}_{NA} \xrightarrow{p} (\Sigma_x + \Sigma_\delta)^{-1}\Sigma_x\beta.$$

In this example, we have

$$\hat{\beta}_{NA} \xrightarrow{p} (0.5, -0.4, 0, 0.02, -0.02, 0.7).$$

Therefore, the naive method will result in nonzero values of FN asymptotically. The selection results with sample sizes $n = 100, 500, 1000$ are reported in Table (2.8). The TR and RIV estimators perform like the oracle procedure, where the values of FP, FN and MSE decrease as sample size gets larger. On the other hand, the FN remains at one for naive estimator regardless of sample size.

Table 2.8: Selection results of Example 2.2.3 with $\sigma_\delta^2 = 1$

	n=100				n=500				n=1000			
	FP	FN	MCC	MSE	FP	FN	MCC	MSE	FP	FN	MCC	MSE
TR	0.1	0.5	0.77	0.06	0.1	0.2	0.90	0.02	0	0	0.99	0.01
IV	0.2	0.8	0.67	0.24	0.1	0.5	0.80	0.07	0	0.1	0.96	0.02
NA	0.1	0.9	0.67	0.49	0	1	0.7	0.45	0	1	0.7	0.44

2.2.4 Numerical Example for Linear Model 4

In this example we examine the effects of ME on all the covariates (with high/medium/low correlations). Notations are changed a bit in this example. Specifically, the datasets are generated from the linear model

$$y = \beta^T t + \epsilon$$

where $\beta = (1, 0, 0.7, 0.6, 0, 0.5, 0.4, 0, 0, \ldots, 0)$, the covariates $t = (t_1, t_2, \ldots, t_{20})$ are jointly generated from multivariate normal distribution $N(0, \Sigma)$ with $\Sigma_{ij} = 0.7^{|i-j|}$. In addition, t_k is measured with errors $\delta_k \sim N(0, \sigma_\delta^2)$, $(k = 1, 4, 7)$. The penalty functions are chosen as SCAD, MCP and Lasso. BIC, AIC and Cross-validation are used as model selection criteria. The simulation results are reported in Table (2.9), (2.10) and (2.11). Each table consists of the selection results from penalized regression with SCAD, MCP and Lasso penalty functions, respectively. The results from TR, IV and NA methods are compared with respect to four summary statistics FP, FN, MCC and MSE, with different sample sizes $(n = 50, 100, 200)$ and standard deviations of ME $(\sigma_\delta = 1, 2, 5)$.

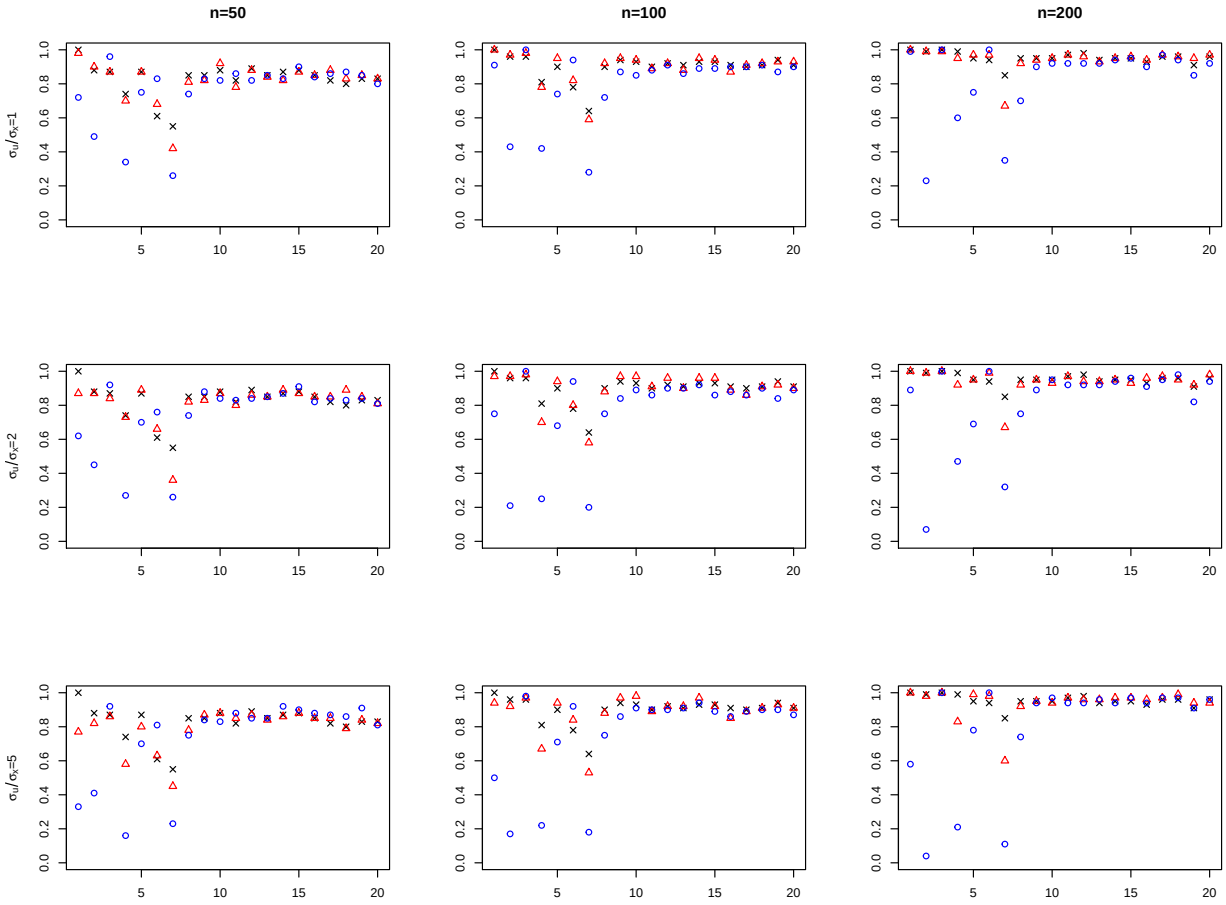
Table (2.9), Figure (2.4), (2.5) and (2.6) show the linear model selection results with the tuning parameter chosen by BIC. The figures show the frequency of correct selection for each covariate. Note that the low values of FP, FN and MSE and high values of MCC indicate a good model fit. First, regardless of the penalty functions chosen, values of σ_δ and sample

sizes, the model fit for TR and IV method is consistently better than the naive method. For SCAD penalty with $n = 50$, the FP remains at 2.3 for TR and IV, with the value for NA being slightly higher. The FN is around 1.4 for TR and IV methods, whereas FN is greater than 2 for naive method and is increasing with the variance of ME. The other two model fit statistics are also better for TR and IV methods with larger MCC and smaller MSE. As the sample size increases, the results from TR and IV are approaching the oracle as both the FP and FN are decreasing towards zero. Also it is worth noting that the IV method is robust against the magnitude of ME variance, which is not the case for naive method. The effect of ME can be observed from the selection results in details. Figure (2.4) shows the frequency plot of the correct selection for the SCAD penalty. The graphs in the same row correspond to the same value of σ_δ/σ_x, and the graphs in the same column correspond to the same value of sample size n.

Table 2.9: Linear model selection results of Example 2.2.4 with BIC

			n=50				n=100				n=200			
			FP	FN	MCC	MSE	FP	FN	MCC	MSE	FP	FN	MCC	MSE
SCAD		True	2.3	1.4	0.54	1.199	1.2	0.7	0.75	0.469	0.8	0.2	0.88	0.155
	$\sigma_\delta = 1$	IV	2.3	1.5	0.53	1.372	1.0	0.9	0.74	0.613	0.8	0.3	0.86	0.197
		Naive	2.5	2.1	0.41	2.441	2.3	1.4	0.53	1.771	2.1	1.2	0.60	1.466
	$\sigma_\delta = 2$	IV	2.4	1.7	0.47	1.777	1.2	1.0	0.72	0.71	0.7	0.5	0.85	0.279
		Naive	2.5	2.3	0.36	3.015	2.5	2.0	0.42	2.389	2.4	1.7	0.48	2.09
	$\sigma_\delta = 5$	IV	2.3	1.4	0.53	2.264	1.2	1.0	0.72	0.851	0.8	0.7	0.80	0.461
		Naive	2.6	2.4	0.35	3.164	2.6	2.2	0.38	2.515	2.3	2.0	0.45	2.218
MCP		True	2.0	1.4	0.57	1.157	0.8	0.7	0.80	0.399	0.5	0.1	0.92	0.118
	$\sigma_\delta = 1$	IV	2.0	1.5	0.55	1.312	0.8	0.9	0.78	0.538	0.4	0.2	0.91	0.155
		Naive	2.3	2.2	0.40	2.419	1.6	1.7	0.56	1.709	1.9	1.1	0.62	1.387
	$\sigma_\delta = 2$	IV	2.0	1.7	0.51	1.683	0.8	0.9	0.76	0.711	0.4	0.4	0.89	0.245
		Naive	2.3	2.5	0.36	2.999	1.9	2.3	0.44	2.309	2.3	1.8	0.47	2.048
	$\sigma_\delta = 5$	IV	1.9	1.5	0.56	2.171	0.9	1.1	0.74	0.832	0.5	0.6	0.86	0.403
		Naive	2.2	2.6	0.34	3.081	2.1	2.3	0.40	2.505	1.9	2.2	0.45	2.209
Lasso		True	2.9	0.4	0.65	0.554	2.0	0.0	0.78	0.196	1.6	0.0	0.82	0.097
	$\sigma_\delta = 1$	IV	2.7	0.6	0.64	0.708	1.9	0.2	0.77	0.294	1.8	0.0	0.80	0.137
		Naive	3.3	1.3	0.47	1.648	3.3	0.8	0.55	1.373	3.3	0.5	0.61	1.265
	$\sigma_\delta = 2$	IV	3.0	0.9	0.57	0.907	2.4	0.4	0.70	0.408	1.7	0.1	0.80	0.21
		Naive	3.2	1.9	0.37	2.096	3.3	1.5	0.44	1.891	3.3	1.3	0.48	1.81
	$\sigma_\delta = 5$	IV	2.9	0.9	0.57	1.005	2.2	0.4	0.71	0.534	2.2	0.2	0.75	0.331
		Naive	3.2	2.2	0.33	2.211	3.1	2.0	0.37	2.024	3.3	1.6	0.42	1.942

Figure 2.4: The frequency of correct selection in Example 2.2.4 for SCAD with BIC; black cross - True, red triangle - IV, blue circle - Naive



The common selection errors for naive method are on x_1, x_2, x_4, x_5, x_7, as more blue circles are appearing close to bottom. From the data correlation structure, and the simulation setting such that x_1, x_4, x_7 are measured with errors, it can be seen that the ME affects the selection of the covariates that are highly/moderately correlated with mis-measured covariates. The weakly correlated covariates are also affected, for example $x_8, x_9, x_{11}, x_{16}, x_{19}, x_{20}$. As the ratio σ_δ/σ_x increases, the correct selection frequency drops for the naive method (eg. x_1, x_4). As comparison, the RIV method is robust against the ratio. For MCP penalty (Figure (2.5)), the FP is slightly lower compared with SCAD. The overall pattern for both penalty function remains similar.

Figure 2.5: Correct selection frequency results of Example 2.2.4 for MCP with BIC; black cross - True, red triangle - IV, blue circle - Naive

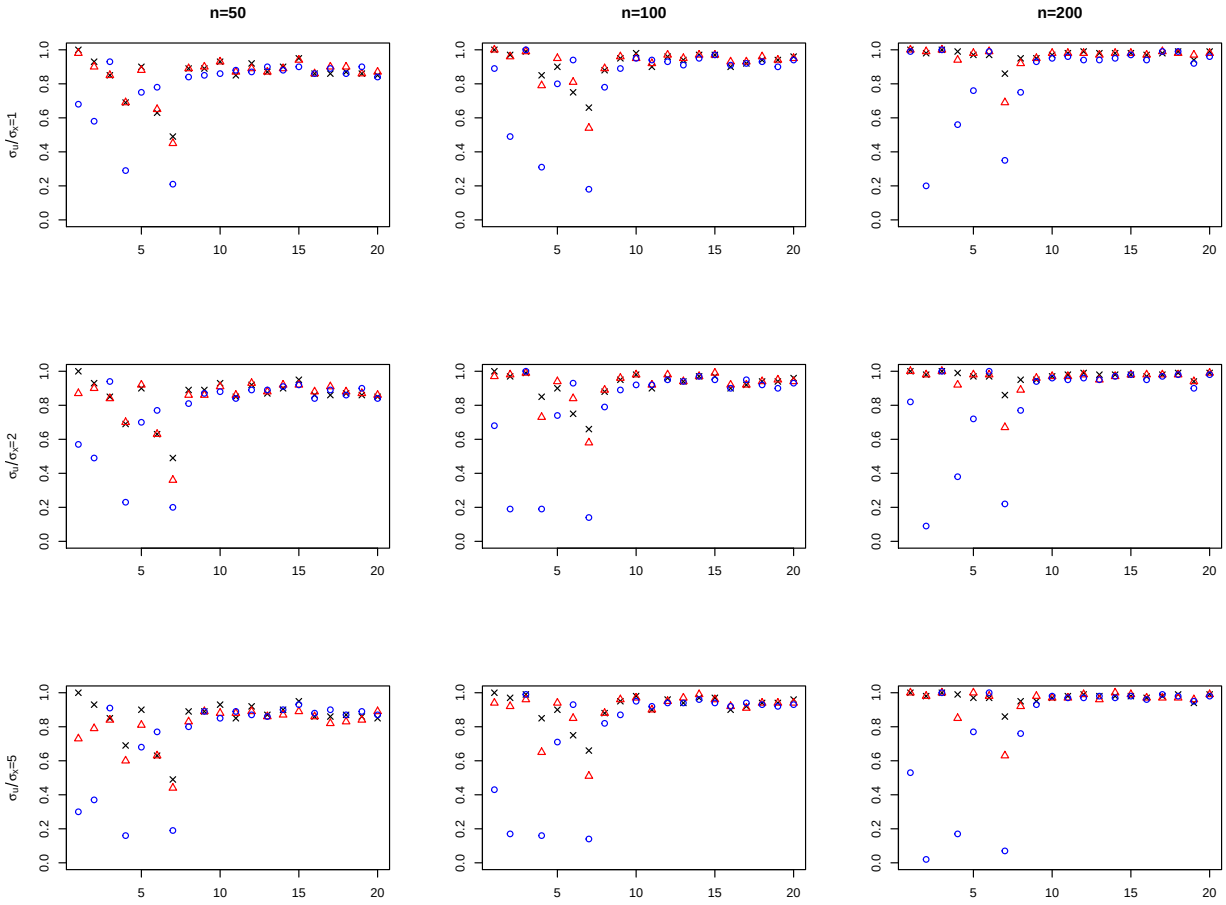


Figure (2.6) shows the frequency plot of correct selection for Lasso. First, the pattern resembles that for the SCAD and MCP cases. The TR method performs the best among the three methods, followed by the RIV method. The naive method cannot select the true model consistently among all scenarios. For Lasso penalty, the FN is lower than that of SCAD and MCP for all TR, IV and NA methods. Whereas the FP is much higher as a trade-off. For example, the values of FP and FN are 0.8, 0.7 for SCAD penalty in naive method ($\sigma_\delta = 5$), compared with 2.2, 0.2 for that of Lasso. In particular, x_2 and x_5 contribute the most to the excessive covariates selected. In this sense Lasso tends to select less sparse model compared with the other two penalty functions.

Figure 2.6: Correct selection frequency results of Example 2.2.4 for Lasso with BIC; black cross - True, red triangle - IV, blue circle - Naive

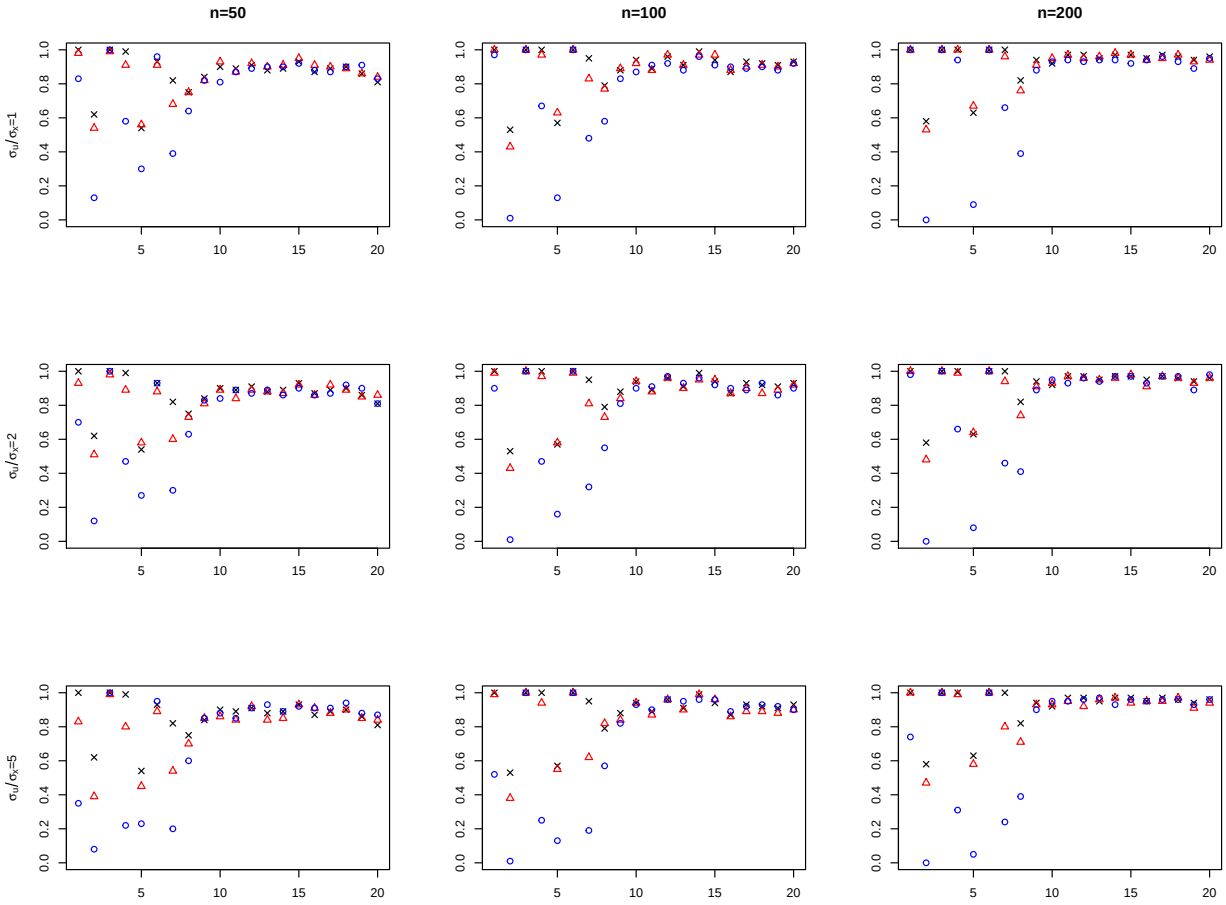
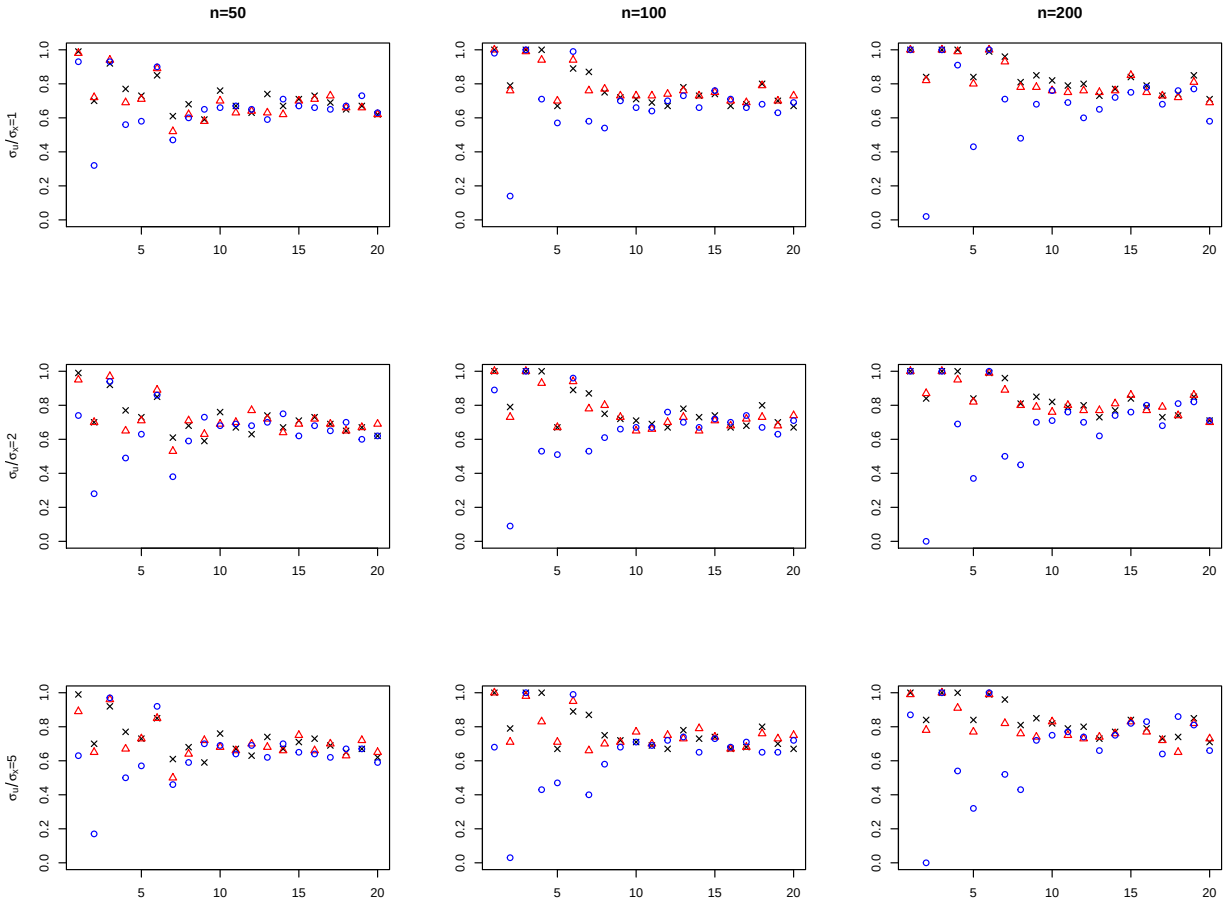


Table (2.10), figure (2.7), (2.8) and (2.9) show the linear model selection results with the tuning parameter chosen by AIC. As discussed in the section of model selection criteria, the AIC always tends to select a bigger model compared with BIC. Hence, the results obtained from using AIC criteria have higher FP and lower FN, which is the case for all three models. This effect improves the FN for all models in some sense. However, the increase of FP is too much compared with the extent of improvement in FN. For example, consider the case where $\sigma_\delta^2 = 2$ and $n = 200$. For the SCAD penalty with BIC, the (FP,FN) for IV and NA are (0.4,0.4) and (2.3,1.8), respectively.

Table 2.10: Linear model selection results of Example 2.2.4 with AIC

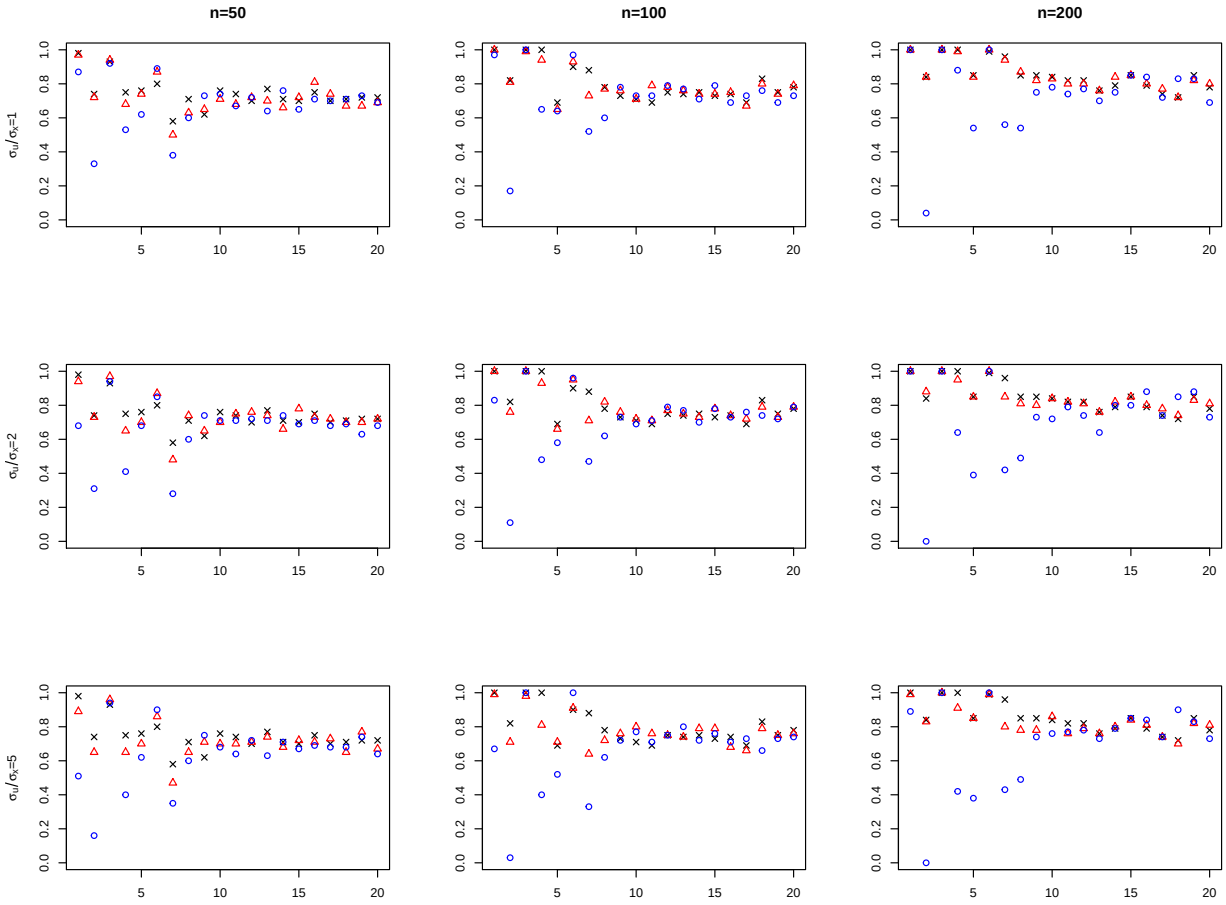
			n=50				n=100				n=200			
			FP	FN	MCC	MSE	FP	FN	MCC	MSE	FP	FN	MCC	MSE
SCAD		True	4.5	0.9	0.46	1.4	4.0	0.2	0.6	0.499	3.4	0.0	0.67	0.188
	$\sigma_\delta = 1$	IV	4.5	0.9	0.45	1.567	4.3	0.4	0.56	0.641	3.4	0.1	0.66	0.231
		Naive	5.7	1.2	0.34	2.771	5.2	0.8	0.43	1.67	5.0	0.6	0.48	1.323
	$\sigma_\delta = 2$	IV	4.6	1.1	0.42	2.407	4.2	0.5	0.54	1	3.0	0.2	0.67	0.333
		Naive	5.1	1.6	0.31	3.285	5.3	1.2	0.36	2.351	5.7	0.9	0.39	1.976
	$\sigma_\delta = 5$	IV	5.0	1.1	0.39	3.309	4.1	0.6	0.53	1.836	3.3	0.3	0.63	0.388
		Naive	5.6	1.6	0.27	3.751	5.1	1.3	0.35	2.533	5.3	1.1	0.37	2.191
MCP		True	4.0	0.9	0.49	1.353	3.4	0.2	0.64	0.482	2.9	0.1	0.7	0.194
	$\sigma_\delta = 1$	IV	4.2	1.0	0.46	1.603	3.6	0.3	0.61	0.611	2.9	0.1	0.69	0.238
		Naive	4.9	1.3	0.36	2.702	4.6	0.9	0.45	1.669	4.3	0.7	0.51	1.311
	$\sigma_\delta = 2$	IV	4.1	1.1	0.45	2.298	3.7	0.5	0.57	0.998	2.9	0.2	0.68	0.336
		Naive	4.6	1.8	0.31	3.24	4.5	1.4	0.38	2.307	4.9	1.0	0.41	1.971
	$\sigma_\delta = 5$	IV	4.4	1.2	0.42	3.294	3.6	0.6	0.57	1.836	2.9	0.3	0.66	0.39
		Naive	4.8	1.8	0.29	3.599	4.7	1.4	0.35	2.518	4.7	1.3	0.37	2.167
Lasso		True	5.6	0.1	0.52	0.731	4.9	0.0	0.58	0.286	4.6	0.0	0.6	0.124
	$\sigma_\delta = 1$	IV	5.4	0.4	0.49	0.917	5.3	0.1	0.54	0.398	4.4	0.0	0.61	0.166
		Naive	6.1	0.7	0.39	1.844	5.8	0.4	0.46	1.42	5.8	0.3	0.47	1.234
	$\sigma_\delta = 2$	IV	5.2	0.6	0.47	1.536	5.3	0.2	0.52	0.616	4.3	0.1	0.61	0.249
		Naive	6.3	1.0	0.33	2.562	5.3	1.0	0.39	1.979	6.1	0.6	0.41	1.84
	$\sigma_\delta = 5$	IV	6.3	0.6	0.39	2.273	5.6	0.3	0.49	1.31	4.6	0.2	0.58	0.314
		Naive	6.2	1.3	0.28	2.734	5.9	1.0	0.35	2.174	6.2	0.9	0.36	2.016

Figure 2.7: Correct selection frequency results of Example 2.2.4 for SCAD with AIC; black cross - True, red triangle - IV, blue circle - Naive



If AIC is used in the same scenario, the (FP,FN) for IV and NA become (3.0,0.2) and (5.7,0.9). It can be seen that the small decrease in FN comes with large increase of FP. On the other hand, consider the same case where $\sigma_\delta^2 = 2$ and $n = 200$. The (FP,FN) for IV and Naive method for Lasso penalty with BIC are (1.7, 0.1) and (3.3, 1.3), respectively. Whereas they become (4.3, 0.1) and (6.1, 0.6) if AIC is used instead. In this case there is no improvement in FN. In other words, the model selected by AIC is getting bigger without any substantial improvement, in terms of the model selection metrics used in the current simulation. Figure (2.7) and (2.8) show the frequency plot of correct selection for SCAD and MCP penalty functions.

Figure 2.8: Correct selection frequency results of Example 2.2.4 for MCP with AIC; black cross - True, red triangle - IV, blue circle - Naive



The performance is similar with each other and we see the overall correct selection frequencies drop due to bigger models are falsely selected. Figure (2.9) shows the frequency plot for Lasso penalty, which amplifies the false selection effect due to the nature of the Lasso penalty.

Figure 2.9: Correct selection frequency results of Example 2.2.4 for Lasso with AIC; black cross - True, red triangle - IV, blue circle - Naive

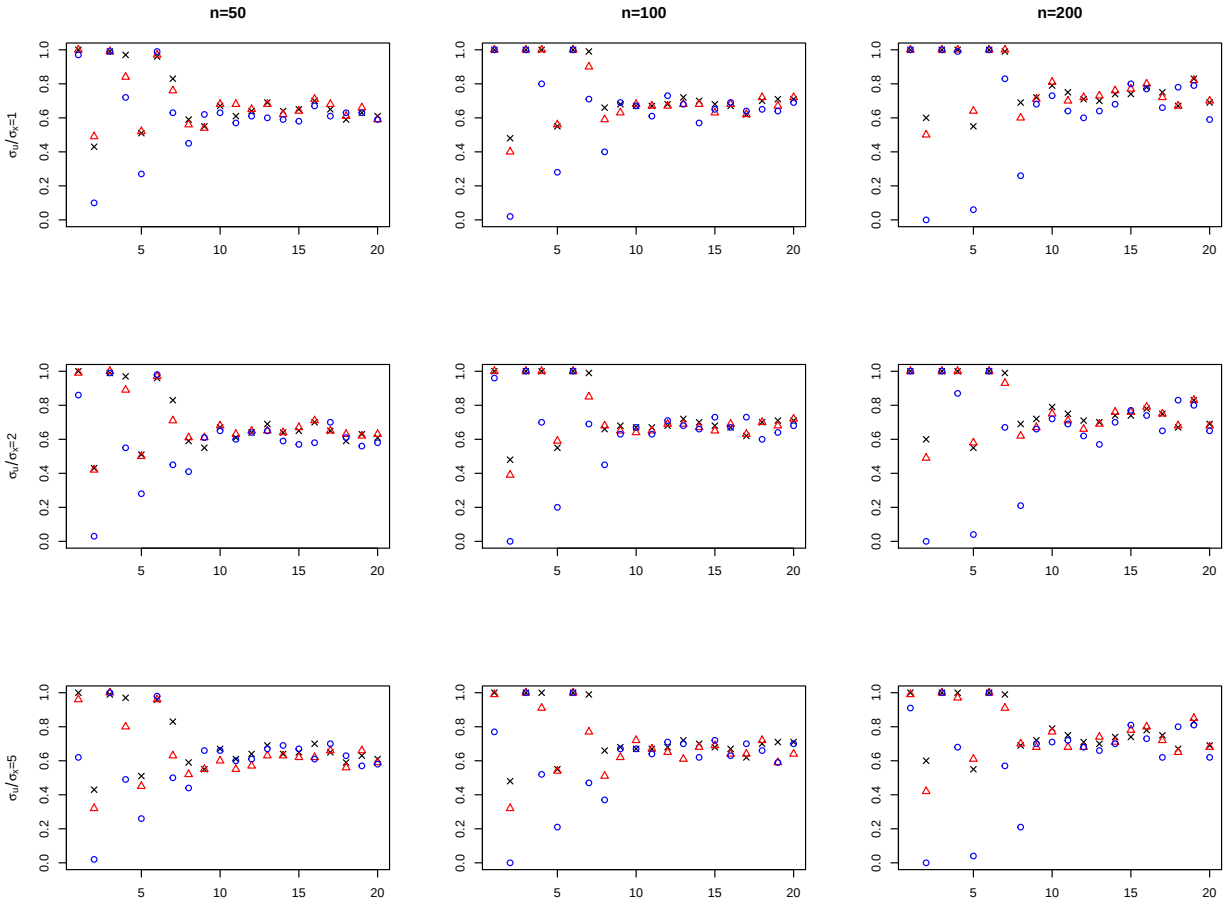


Table (2.11), figure (2.10), (2.11) and (2.12) show the linear model selection results with the tuning parameter chosen by cross-validation. The cross-validation is based on minimizing the prediction error. From this point of view, cross-validation is not guaranteed to choose the true model consistently. It can be observed that the selection results for cross-validation is similar to that of AIC. Note that the overall pattern among the three different methods (TR, IV and NA) remain the same.

Table 2.11: Linear model selection results of Example 2.2.4 with cross-validation

			n=50				n=100				n=200			
			FP	FN	MCC	MSE	FP	FN	MCC	MSE	FP	FN	MCC	MSE
SCAD		True	3.5	1.1	0.48	1.241	2.7	0.5	0.65	0.391	2.7	0.1	0.71	0.148
	$\sigma_\delta = 1$	IV	3.4	1.4	0.45	1.946	3.0	0.7	0.59	0.858	2.6	0.4	0.69	0.29
		Naive	3.3	2.1	0.34	2.97	4.0	1.3	0.42	2.279	4.2	1.2	0.43	1.987
	$\sigma_\delta = 2$	IV	3.6	1.4	0.44	1.567	2.9	0.6	0.63	0.521	2.5	0.2	0.72	0.187
		Naive	3.6	1.7	0.39	2.471	4.3	1.0	0.46	1.604	4.7	0.7	0.48	1.335
	$\sigma_\delta = 5$	IV	3.7	1.4	0.43	2.261	3.0	0.8	0.57	0.935	2.5	0.5	0.67	0.401
		Naive	3.7	2.1	0.32	3.12	3.9	1.8	0.35	2.463	4.1	1.5	0.39	2.177
MCP		True	2.3	1.4	0.54	1.159	2.4	0.4	0.68	0.383	1.9	0.2	0.78	0.145
	$\sigma_\delta = 1$	IV	2.7	1.6	0.47	1.792	2.2	0.8	0.64	0.842	1.5	0.4	0.77	0.285
		Naive	2.8	2.4	0.33	3.151	3.5	1.6	0.40	2.355	3.7	1.4	0.42	2.004
	$\sigma_\delta = 2$	IV	2.2	1.6	0.51	1.552	2.0	0.7	0.67	0.537	1.7	0.2	0.77	0.2
		Naive	2.4	2.0	0.43	2.474	2.8	1.3	0.51	1.601	3.7	0.8	0.53	1.343
	$\sigma_\delta = 5$	IV	2.5	1.7	0.47	1.989	1.9	1.0	0.64	0.943	1.3	0.5	0.77	0.408
		Naive	2.5	2.5	0.33	3.157	2.8	2.2	0.37	2.452	3.6	1.7	0.38	2.187
Lasso		True	4.6	0.2	0.57	0.55	4.0	0.0	0.63	0.205	4.3	0.0	0.62	0.119
	$\sigma_\delta = 1$	IV	5.1	0.6	0.47	1.01	4.4	0.3	0.57	0.513	4.7	0.0	0.59	0.225
		Naive	5.0	1.5	0.33	2.118	5.3	1.0	0.38	1.913	5.8	0.7	0.42	1.82
	$\sigma_\delta = 2$	IV	5.0	0.4	0.51	0.743	3.9	0.1	0.63	0.303	4.4	0.0	0.61	0.151
		Naive	5.0	0.9	0.42	1.652	5.4	0.5	0.46	1.34	5.5	0.3	0.50	1.24
	$\sigma_\delta = 5$	IV	5.2	0.7	0.45	1.291	4.7	0.3	0.54	0.644	5.0	0.2	0.54	0.331
		Naive	4.8	1.7	0.31	2.273	5.1	1.4	0.32	2.072	5.6	0.9	0.39	1.983

Figure 2.10: Correct selection frequency results of Example 2.2.4 for SCAD with cross-validation; black cross - True, red triangle - IV, blue circle - Naive

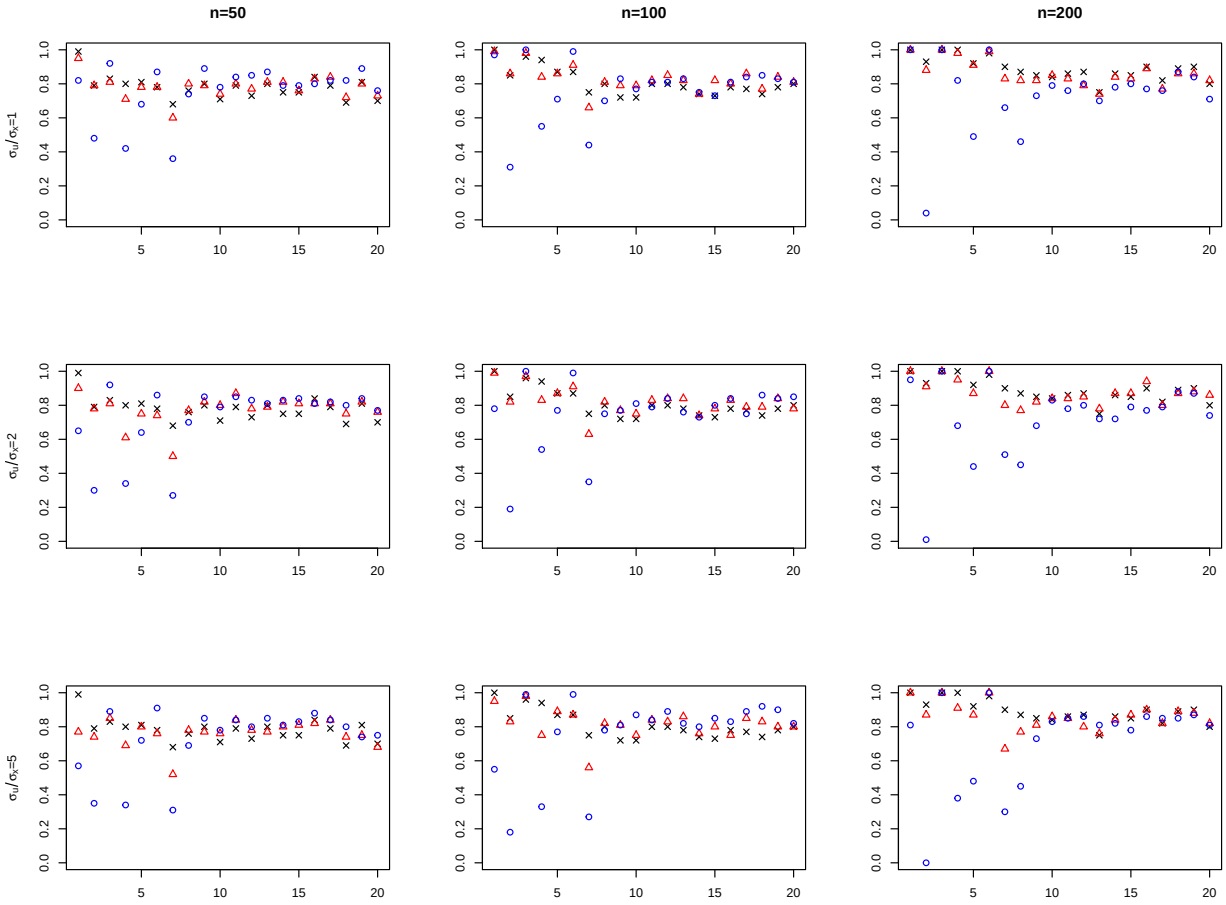


Figure 2.11: Correct selection frequency results of Example 2.2.4 for MCP with Cross-Validation; black cross - True, red triangle - IV, blue circle - Naive

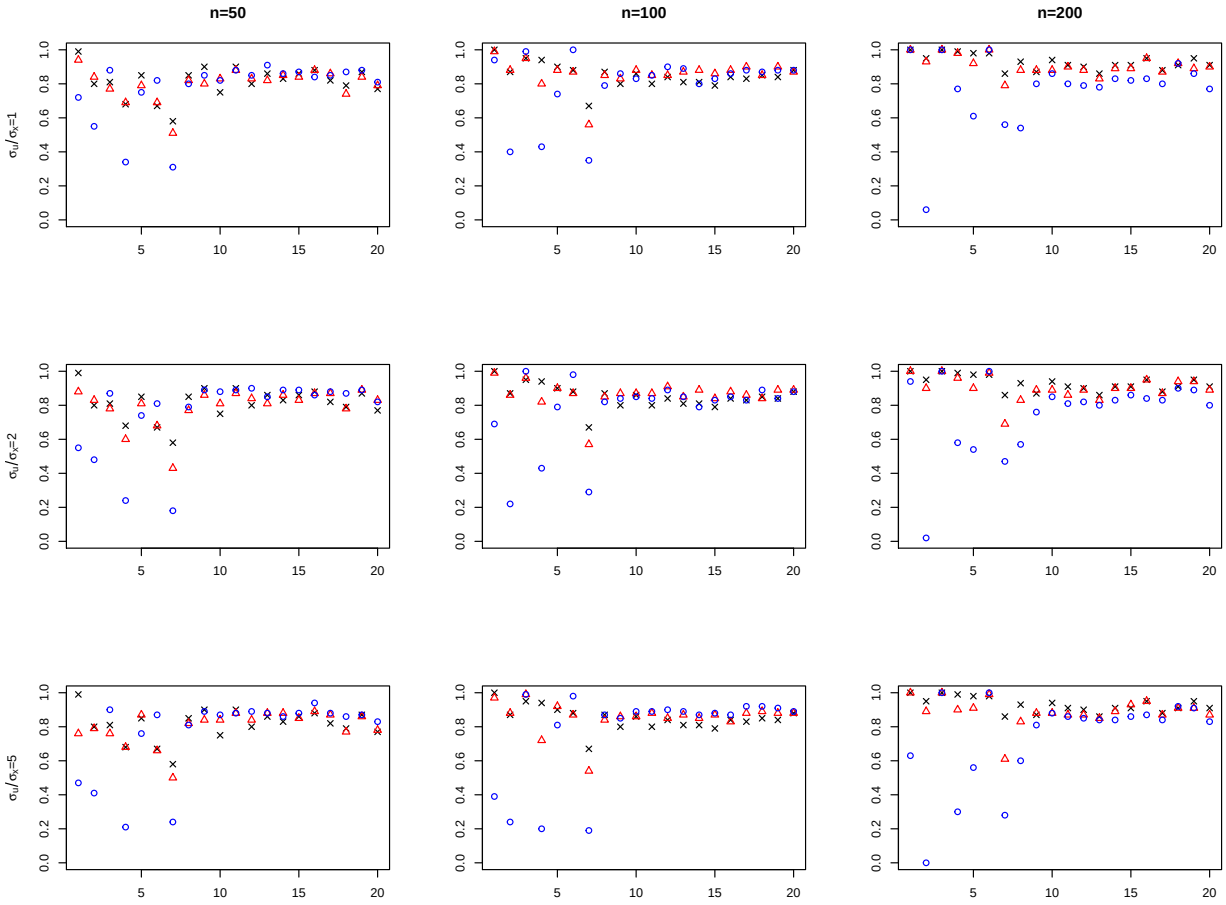
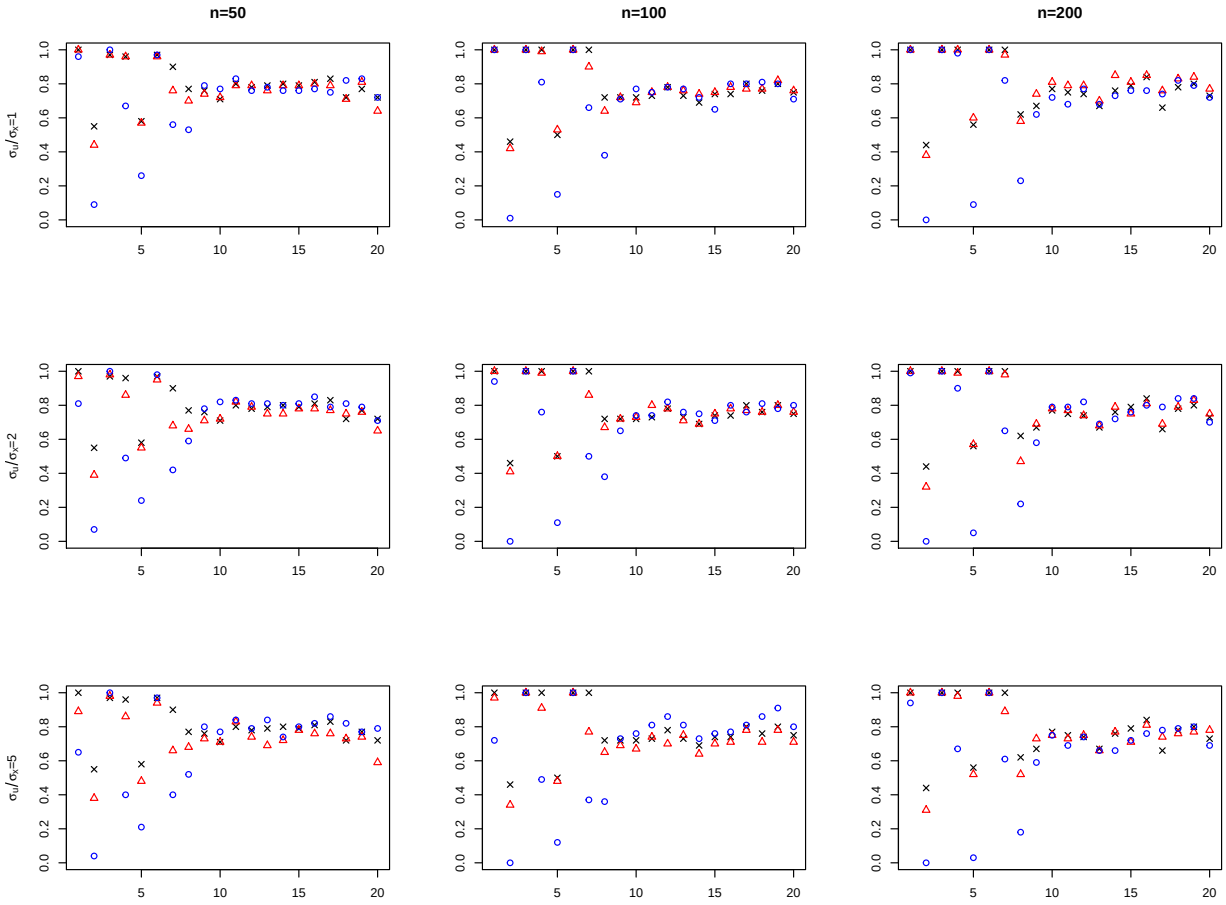


Figure 2.12: Correct selection frequency results of Example 2.2.4 for Lasso with Cross-Validation; black cross - True, red triangle - IV, blue circle - Naive



2.3. Real Data Application

In this section we apply the proposed method on a real dataset. The data is from a community-based study on the prevalence of coronary heart disease risk factors in Virginia by Willems et al. (1997). Risk factors like smoking habits, blood pressure, high-density lipoprotein, cholesterol and glycosylated hemoglobin are collected from 403 among rural blacks in Virginia. In our analysis, the outcome of interest is systolic blood pressure. The covariates being considered are: glycosylated hemoglobin (glyhb), body mass index (bmi). total cholesterol (chol), high-density lipoprotein (hdl), chol/hdl (ratio), age, gender, height in inchs, weight in kilogram. Specifically, the stabilized glucose, waist and hip in inches are used as instrumental variables for glycosylated hemoglobin and bmi. A regularized linear regression model is fitted to the dataset to identify the important factors affecting systolic blood pressure. The tuning parameter is selected by BIC. The results from the naive method ignoring the ME and the proposed RIV method are presented in Table (2.12), as well as the results from the ordinary least squares. Numbers in parentheses are estimated standard errors. From the results in Table (2.12) it can be observed that the naive method selects only age as the important predictor for blood pressure. However, growing evidence in epidemiological studies shows that there is positive relationship between the blood pressure and bmi, eg. Falkner et al. (2006) and Linderman et al. (2018). After correcting for the ME effect, the proposed RIV method selects bmi, cholesterol and age as important covariates. The plots of coefficient estimates against the tuning parameter for both methods are shown in Figure (2.13). It can be observed from Figure (2.13) such that, besides the magnitude of the estimates, the ME also affects the selection results through the model selection criteria.

Figure 2.13: Estimates of regression coefficients for diabetes data. Left panel: estimates of naive method as a function of tuning parameter λ; Right panel: estimates of IV method. The vertical line corresponds to the optimal value of λ selected by BIC.

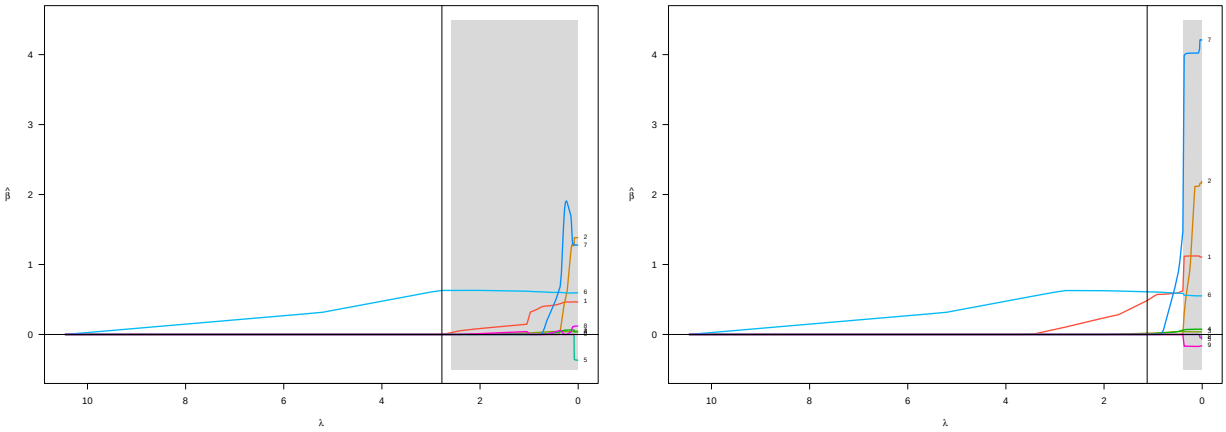


Table 2.12: Estimated regression coefficients and standard errors (SE) of diabetes data

	SCAD-BIC		OLS	
	IV	Naive	IV	Naive
intercept	91.03 (7.031)	107.86 (3.187)	72.68 (31.622)	71.43 (82.255)
bmi	0.48 (0.176)	0 (-)	1.11 (0.478)	0.46 (1.336)
stab.glu	0 (-)	0 (-)	2.19 (4.536)	1.38 (3.274)
chol	0.02 (0.024)	0 (-)	0.04 (0.046)	0.05 (0.046)
hdl	0 (-)	0 (-)	0.07 (0.142)	0.04 (0.143)
ratio	0 (-)	0 (-)	-0.07 (1.579)	-0.37 (1.591)
age	0.61 (0.065)	0.63 (0.064)	0.55 (0.073)	0.59 (0.071)
gender	0 (-)	0 (-)	4.21 (3.265)	1.28 (3.043)
height	0 (-)	0 (-)	-0.04 (0.409)	0.12 (1.220)
weight	0 (-)	0 (-)	-0.16 (0.156)	0.00 (0.483)

2.4. Linear ME Model Proofs

Proof of Theorem 1

The idea of the proof is similar to that of Ma and Li (2010). Define the score function as

$$s_n(\beta) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{x}_i^T\beta)\tilde{x}_i + \sum_{j=1}^{p} p'_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j)$$

It is sufficient to show that $s_n(\beta) = 0$ has a solution β satisfying $||\hat{\beta} - \beta_0|| = O_p(n^{-1/2})$. To this end, we show for any β such that $||\hat{\beta} - \beta_0|| = n^{-1/2}C$ the inequality $(\beta - \beta_0)^T s_n(\beta) > 0$ holds with probability approaching 1. Using the Taylor expansion, we have

$$\begin{aligned}
&(\beta - \beta_0)^T s_n(\beta) \\
=&(\beta - \beta_0)^T \left(-n^{-1}\sum_{i=1}^{n}(y_i - \tilde{x}_i^T\beta_0)\tilde{x}_i + p'_{\lambda_n}(|\beta_0|) \circ \text{sign}(\beta_0) \right) \\
+&(\beta - \beta_0)^T n^{-1}\sum \tilde{x}_i\tilde{x}_i^T(\beta - \beta_0) + (\beta - \beta_0)^T p''_{\lambda_n}(|\beta_0|)(\beta - \beta_0)(1 + o(1))
\end{aligned}$$

The term in first line of last equation is of order $O_p(Cn^{-1})$. The second term is $O_p(C^2n^{-1})$ and the third term is $o_p(C^2n^{-1})$. Hence the second term dominates the other terms with a sufficiently large C. Together with the positive definiteness of $E(\tilde{w}\tilde{w}^T)$, $(\beta - \beta_0)^T s_n(\beta)$ is shown to be positive with probability tending to 1, which completes the proof.

Proof of Theorem 2

It is sufficient to show that for any β satisfying $||\beta - \beta_0|| = O_p(n^{-1/2})$,

$$\begin{cases} \partial Q(\beta)/\partial\beta_j > 0, & 0 < \beta_j < Cn^{-1/2}; \\ \partial Q(\beta)/\partial\beta_j < 0, & -Cn^{-1/2} < \beta_j < 0 \end{cases}$$

for every $\beta_j \in \beta_{J^c}$. Note that

$$\begin{aligned}
\frac{\partial Q(\beta)}{\partial\beta_j} &= -\sum_{i=1}^{n}(y_i - \beta^T\tilde{x}_i)\tilde{x}_{ij} + np'_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) \\
&= n\left\{ -n^{-1}\sum_{i=1}^{n}(y_i - \beta_0^T\tilde{x}_i)\tilde{x}_{ij} + n^{-1}\sum_{i=1}^{n}(\beta - \beta_0)^T\tilde{x}_i\tilde{x}_{ij} + p'_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) \right\} \\
&= n\lambda_n\left\{ O_p(n^{-1/2}/\lambda_n) + \lambda_n^{-1}p'_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) \right\}
\end{aligned}$$

The first term of the last second equation is $O_p(n^{-1/2})$, together with $||\beta - \beta_0|| = O_p(n^{-1/2})$ and the assumption $\liminf_{n\to\infty} \liminf_{\phi\to 0^+} p'_{\lambda_n}(\phi)/\lambda_n > 0$, the sign of the partial derivative is completely determined by the sign of β_j, which completes the proof of (a) in Theorem 2.

The first order differential equation with respect to β_J (given $\hat{\beta}_{J^c} = 0$) is

$$\frac{\partial Q(\hat{\beta}_J)}{\partial \beta_J} = -\sum_{i=1}^{n}(y_i - \hat{\beta}_J^T \tilde{x}_{Ji})\tilde{x}_{Ji} + np'_{\lambda_n}(|\hat{\beta}_J|) \circ \text{sign}(\hat{\beta}_J)$$

where $\circ$ is Hadamard product. Setting the first order derivative as zero, then

$$n^{-1}\sum_{i=1}^{n}(\tilde{x}_{Ji}\tilde{x}_{Ji}^T + \Sigma(\tilde{\beta}_J))n^{1/2}(\hat{\beta}_J - \beta_{0J}) + n^{1/2}\boldsymbol{b} = n^{-1/2}\sum_{i=1}^{n}(y_i - \beta_{0J}^T\tilde{x}_{Ji})\tilde{x}_{Ji}.$$

The RHS can be written as

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}(y_i - \beta_{0J}^T\tilde{x}_{Ji})\tilde{x}_{Ji} \\
=&n^{-1/2}\sum(y_i - \beta_{0J}^T\alpha_i + \beta_{0J}^T\alpha_i - \beta_{0J}^T\tilde{x}_{Ji})\tilde{x}_{Ji} \\
=&n^{-1/2}\sum\left\{(y_i - \beta_{0J}^T\tilde{x}_{Ji})\alpha_i + (y_i - \beta_{0J}^T\alpha_i)(\tilde{x}_{Ji} - \alpha_i) + \beta_{0J}^T(\alpha_i - \tilde{x}_{Ji})(\tilde{x}_{Ji} - \alpha_i)\right\}
\end{aligned}$$

where $\alpha_i = \tilde{\Gamma}_{0J}\tilde{w}_i$. It can be shown that the second and third term is $o_p(1)$ and the first term is:

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}(y_i - \beta_{0J}^T\tilde{x}_{Ji})\alpha_i \\
=&n^{-1/2}\sum\left(y_i - \beta_{0J}^T\tilde{\Gamma}_{0J}\tilde{w}_i + \tilde{w}_i^T I_{l+q}(\tilde{\Gamma}_{0J} - \tilde{\Gamma}_J)^T\beta_{0J}\right)\alpha_i \\
=&n^{-1/2}\sum\left((y_i - \beta_{0J}^T\tilde{\Gamma}_{0J}\tilde{w}_i)\alpha_i + \tilde{\Gamma}_{0J}(\beta_{0J}^T \otimes I_{l+q})(\sum(\tilde{\Gamma}_{0J}\tilde{w}_i - x^*_{Ji}) \otimes \tilde{w}_i)\right) \\
=&n^{-1/2}\left(I_s, \tilde{\Gamma}_{0J}(\beta_{0J}^T \otimes I_{l+q})\right) \\
&\cdot\sum\begin{pmatrix}(y_i - \beta_{0J}^T\alpha_i)\alpha_i \\ (\tilde{\Gamma}_{0J}\tilde{w}_i - \tilde{x}^*_{Ji}) \otimes \tilde{w}_i)\end{pmatrix} \\
\equiv&Dn^{-1/2}\sum K_i \\
\to&D \cdot N(0, C)
\end{aligned}$$

where

$$D = \left(I_s, \tilde{\Gamma}_{0J}(\beta_{0J}^T \otimes I_{l+q})\right),$$

$$C = E(KK^T).$$

Also

$$\begin{aligned} H_n \equiv & n^{-1} \sum \tilde{x}_{Ji} \tilde{x}_{Ji}^T \\ = & n^{-1} \tilde{\Gamma}_J \sum \tilde{w}_i \tilde{w}_i^T \tilde{\Gamma}_J^T \\ \xrightarrow{p} & \tilde{\Gamma}_{0J} E \tilde{w} \tilde{w}^T \tilde{\Gamma}_{0J}^T \\ = & H. \end{aligned}$$

The proof is then completed by applying Slutsky's theorem. Note that the following facts are used in the proof and are listed as below.

Kronecker Product Properties

For matrices A_j, B_j, C_j and D_j, $j = 1, 2$, we have the following facts

$$\begin{aligned} \text{vec}(A_1 B_1 C_1) &= (C_1^T \otimes A_1)\text{vec}(B_1) \\ (A_2 C_2 \otimes B_2 D_2) &= (A_2 \otimes B_2)(C_2 \otimes D_2) \end{aligned}$$

provided that the dimensions of the matrices match. Under the current model setting, we have

$$\begin{aligned} \text{vec}[(\hat{\Gamma} - \Gamma_0)^T] &= \begin{pmatrix} \hat{\gamma}_1 - \gamma_{01} \\ \vdots \\ \hat{\gamma}_p - \gamma_{0p} \end{pmatrix} \\ &= \begin{pmatrix} (W^T W)^{-1} & & \\ & \ddots & \\ & & (W^T W)^{-1} \end{pmatrix} \begin{pmatrix} W^T (X_1^* - W\gamma_{01}) \\ \vdots \\ W^T (X_p^* - W\gamma_{0p}) \end{pmatrix} \\ &= \left(I_p \otimes \sum w_i w_i^T \right)^{-1} \left(\sum (x_i^* - \Gamma_0 w_i) \otimes w_i \right). \end{aligned}$$

Chapter 3

Regularized Regression in Generalized Linear ME Model

In this chapter, we discuss about the regularized instrumental variable method in generalized linear ME models. In particular, the RIV estimator is obtained by minimizing the proposed objective function, which is based on the conditional moments with some penalty function. It is shown that under some conditions, the proposed estimator enjoys the oracle property. The asymptotic distribution is derived and finite sample performance is examined through numerical examples. Finally, the proposed method is applied to Framingham heart study dataset.

3.1. Regularized Generalized Linear ME Model

In generalized linear model, the response variable y has the following density function

$$f(y;\eta,\varphi) = \exp[(y\eta - b(\eta))/\varphi + c(y,\varphi)]$$

where $\eta = \alpha + \beta_x^T x + \beta_z^T z$, $x \in \mathbb{R}^p$ is a vector of covariates that are unobservable or measured with errors, $z \in \mathbb{R}^q$ is a vector of error-free covariates, the coefficients $\beta = (\beta_x^T, \beta_z^T)^T \in \mathbb{R}^d$ is assumed to be sparse. In addition, $b(\cdot)$, $c(\cdot,\cdot)$ are known functions and φ is dispersion parameter which is assumed to be known. If the covariates x are observable and measured precisely, the coefficients can be estimated through the regularized likelihood function

$$\underset{\alpha,\beta}{\operatorname{argmin}} \left(l_n(\alpha,\beta) + n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|) \right) \tag{3.1.1}$$

where $l_n(\alpha, \beta)$ is the log-likelihood function, $p_{\lambda_n}(\cdot)$ is some penalty function with tuning parameter λ_n. When x are measured with errors or cannot be observed directly, the regularization procedure applied on mismeasured proxies of x does not have the oracle property anymore. Hence the true model cannot be identified correctly due to ME effect. We propose to correct the ME effect with instrumental variables. Similar as in the linear model (2.1.4), suppose the covariates x are unobservable and we observe

$$x^* = x + \delta,$$

where δ is a random ME. Further assume that there exists an instrument variable (IV) w that is related with x through the equation

$$x = \Gamma w + u,$$

where Γ is a $p \times l$ matrix with rank p. In contrast to the linear model, the distribution $f_U(u; \phi)$ of the random error u is assumed to be known indexed by some unknown parameter ϕ and is independent of (w, z) with $E(u) = 0$. The random error δ in (2.1.3) is supposed to satisfy $E(\delta|x, z, w) = 0$ and (x^{*T}, w^T) is a surrogate for x.

In generalized linear model, the conditional expectation of the response y on covariates (x^T, z^T) is given by

$$E(y|x, z) = G^{-1}(\alpha + \beta_x^T x + \beta_z^T z),$$

where G is link function. For example, $G(a) = \text{logit}(a)$ for logistic model and $G(a) = \log(a)$ in Poisson model. Note that all expectations are conditional on z throughout this chapter unless stated explicitly. Denoting $\breve{x}^* = (1, x^{*T}, z^T)^T$ and $\breve{x} = (1, x^T, z^T)^T$, we have

$$\begin{aligned} E(\breve{x}^* y|w) &= \int \breve{x}^* G^{-1}(\alpha + \beta_x^T \Gamma w + \beta_z^T z + \beta_x^T u) f_U(u; \phi) du \\ &= \int \breve{x} G^{-1}(\alpha + \beta_x^T x + \beta_z^T z) f_U(x - \Gamma w; \phi) dx \end{aligned} \tag{3.1.2}$$

Let random sample (y_i, x_i^*, z_i, w_i) be independent and identically distributed. Define

$$m(v; \psi) = \int \breve{x} G^{-1}(\alpha + \beta_x^T x + \beta_z^T z) f_U(x - v; \phi) dx,$$

$$\hat{\rho}_i(\psi) = y_i \breve{x}_i^* - m(\hat{\Gamma} w_i; \psi),$$

where $\psi = (\alpha, \beta_x^T, \beta_z^T, \phi)^T$ and $\hat{\Gamma} = (\sum x_i^* w_i^T)(w_i w_i^T)^{-1}$. It is easy to see that $m(\Gamma w; \psi) = E(\breve{x}^* y|w)$. The proposed regularized IV estimator is defined as the minimizer of the following

function

$$Q_n(\psi) + n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|) \tag{3.1.3}$$

where $Q_n(\psi) = \frac{1}{2}\sum_{i=1}^{n} \hat{\rho}_i(\psi)^T A_i \hat{\rho}_i(\psi)$, $A_i = A(w_i)$ is a nonnegative definite matrix which may depend on w_i.

Remark 1. In general, a larger value of λ_n imposes more weights on the penalty and produces a sparser model. The tuning parameter λ_n can be chosen in different ways. For example, the Akaike information criterion (AIC), Bayesian information criterion (BIC), K-fold cross validation and generalized cross validation (GCV). With a properly chosen tuning parameter λ_n ,the proposed estimator is shown to have the following properties.

The notations are slightly different in this chapter. Denote $H_n(\psi)$ as Hessian matrix of $Q_n(\psi)$, $\psi_0 = (\alpha_0, \beta_0^T, \phi_0)^T = (\alpha_0, \beta_{0x}^T, \beta_{0z}^T, \phi_0)^T$ as the true value of model parameters. Define $\psi_J = (\alpha, \beta_J^T, \phi)^T$ where

$$\begin{aligned}
\beta_J &= \{\beta_j, j \in J\}, \; J = \{j : \beta_{0j} \neq 0\}, \\
\beta_{J^c} &= \{\beta_j, j \in J^c\}, \; J^c = \{j : \beta_{0j} = 0\},
\end{aligned}$$

$s = |J|$ the cardinality of J, Γ_J the matrix consisting of rows of Γ corresponding to the index set J, and $\gamma = \mathrm{vec}(\Gamma^T)$ as the vector that is consisting of the columns of Γ^T. In addition, denote

$$\begin{aligned}
a_n &= \max\{p'_{\lambda_n}(|\beta_{0j}|), j \in J\}, \\
b_n &= \max\{p''_{\lambda_n}(|\beta_{0j}|), j \in J\}, \\
\boldsymbol{b} &= (0, p'_{\lambda_n}(|\beta_{0J}^T|), 0)^T \circ \mathrm{sign}(\psi_{0J}), \\
\Sigma &= \mathrm{diag}(0, p''_{\lambda_n}(|\beta_{0J}^T|), 0).
\end{aligned}$$

With the notations defined above, we have the following theorems.

Theorem 3. *Assume* $E||y\breve{x}^*||^2 < \infty$ *and* $E[H(\psi_0)]$ *is positive definite,* $a_n = O(n^{-1/2})$, $b_n = o(1)$, *then there exists a local minimizer* $\hat{\psi}$ *of the objective function* (3.1.3) *such that* $||\hat{\psi} - \psi_0|| = O_p(n^{-1/2})$.

Theorem 4. *If* $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ *and* $\liminf_{n\to\infty} \liminf_{\xi\to 0^+} p'_{\lambda_n}(\xi)/\lambda_n > 0$, *then with probability approaching 1, the root n consistent estimator* $\hat{\psi}$ *satisfies*

(a) $\hat{\beta}_{J^c} = 0$,

(b) $\hat{\psi}_J$ *has the following asymptotic normal distribution*

$$\sqrt{n}(H + \Sigma)(\hat{\psi}_J - \psi_{0J}) + \sqrt{n}\boldsymbol{b} \to_d N(0, DCD^T)$$

where

$$H = E\left[\frac{\partial \rho^T(\psi_{0J})}{\partial \psi_J} A(w) \frac{\partial \rho(\psi_{0J})}{\partial \psi_J^T}\right],$$

$$D = \left[I_{s+2}, E\left(\frac{\partial \rho^T(\psi_{0J})}{\partial \psi_J} A(w) \frac{\partial \rho(\psi_{0J})}{\partial \gamma^T}\right)\left(I_p \otimes E(ww^T)^{-1}\right)\right],$$

$$C = E(KK^T)$$

and

$$K = \begin{pmatrix} \partial \rho^T(\psi_{0J}) / \partial \psi_J \cdot A(w) \rho(\psi_{0J}) \\ (x_J^* - \Gamma_{0J} w) \otimes w \end{pmatrix}$$

Remark 2. The sample counterpart of covariance matrix DCD^T has an alternative expression that eases the calculation and is given by

$$\widehat{DCD^T} = \frac{1}{n} \frac{\partial Q_n(\hat{\psi}_J)}{\partial \psi_J} \frac{\partial Q_n(\hat{\psi}_J)}{\partial \psi_J^T}$$

where

$$\frac{\partial Q_n(\hat{\psi}_J)}{\partial \psi_J} = \sum_{i=1}^{n} \frac{\partial \rho_i^T(\psi_{0J})}{\partial \psi_J} A_i \rho_i(\psi_{0J})$$

Remark 3. Though the estimator is consistent regardless of the choice of $A(w)$, there exists an optimal weight $A(w)$ matrix theoretically for a most efficient estimator. Following Wang and Hsiao (2011), the optimal weight matrix is given by

$$A(w) = E[\rho(\psi_{0J})\rho^T(\psi_{0J})|w].$$

Since the optimal weight matrix involves unknown parameters, the calculation of $A(w)$ can be done via a two-stage estimation procedure. First, minimize the objective function using the identity matrix as weight matrix. In the second stage, the estimators are obtained with the optimal weight matrix which is calculated with the estimates from first stage.

Remark 4 For some penalty functions (e.g. SCAD and MCP), $\boldsymbol{b}$ and Σ are both zero when the tuning parameter λ_n is sufficiently small. Hence the resulting estimator has the oracle performance such that $\hat{\beta}_{J^c} = 0$ and the asymptotic distribution of $\hat{\psi}_J$ is given by

$$\sqrt{n}(\hat{\psi}_J - \psi_{0J}) \xrightarrow{d} N(0, H^{-1} DCD^T H^{-1}).$$

Remark 5 In situations where the integral in (3.1.2) does not have analytical form, Monte Carlo methods (e.g. importance sampling) are used to approximate the integral. Specifically,

we can follow the suggestions in Wang and Hsiao (2011) for calculating the (3.1.2).

- Choose a candidate distribution whose density function $h(x)$ is known;
- Generate i.i.d. random sample $\{x_{is}, s = 1, 2, \ldots, S, S+1, \ldots, 2S; i = 1, 2, \ldots, n\}$ from density function $h(x)$;
- Calculate the Monte Carlo approximation of $m(\Gamma w_i; \psi)$ as

$$m_{S1}(\Gamma w_i; \psi) = \frac{1}{S} \sum_{s=1}^{S} \frac{\tilde{x}_{is} G^{-1}(\alpha + \beta_x^T x_{is} + \beta_z^T z_i) f_U(x_{is} - \Gamma w_i; \phi)}{h(x_{is})},$$

  and

$$m_{S2}(\Gamma w_i; \psi) = \frac{1}{S} \sum_{s=S+1}^{2S} \frac{\tilde{x}_{is} G^{-1}(\alpha + \beta_x^T x_{is} + \beta_z^T z_i) f_U(x_{is} - \Gamma w_i; \phi)}{h(x_{is})};$$

- The approximated loss function is then calculated as

$$Q_n(\psi) = \frac{1}{2} \sum_{i=1}^{n} \hat{\rho}_{i,S1}^T(\psi) A_i \hat{\rho}_{i,S2}(\psi)$$

  where $\hat{\rho}_{i,S1}(\psi) = y_i \breve{x}_i^* - m_{S1}(\hat{\Gamma} w_i; \psi)$ and $\hat{\rho}_{i,S2}(\psi) = y_i \breve{x}_i^* - m_{S2}(\hat{\Gamma} w_i; \psi)$.

Remark 6 As noted in Abarin and Wang (2012), for some models like gamma log-linear and poisson log-liner model, the analytical form of the expectation (3.1.2) can be obtained for some error distribution $f_U(u)$. For example, when the random error u follows an univariate normal distribution $u \sim N(0, \phi)$, the integral in (3.1.2) has the following closed-form expression

$$E(\breve{x}^* y | w) = \breve{\boldsymbol{a}}^T \xi,$$

where $\breve{\boldsymbol{a}} = (1, \Gamma w + \beta_x \phi, z^T)^T$ and $\xi = \exp(\alpha + \beta_x \Gamma w + \beta_z^T z + \frac{1}{2} \beta_x^2 \phi)$. With the closed-form expression, the computation burden is eased a lot.

3.2. Simulation Studies

3.2.1 Numerical Example for Logistic Regression 1

In this example we simulate 1000 datasets from the model $y \sim$ Bernoulli$(p(\alpha + x^T \beta))$, where $p(b) = \exp(b)/(1 + \exp(b))$ and the coefficients $(\alpha, \beta^T) = (1, 3, 1.5, 0, 0, 2, 0, 0, 0)$. The covariate $x = 1.5w + u$ where $(z_1, w, z_2, \ldots, z_7)^T$ are jointly generated from $N(0, \Sigma)$ with $\Sigma_{ij} = 0.7^{|i-j|}$ and u is standard normal. In this example the correlation between w and

x is around 0.83. The unobserved covariate is generated as $x^* = x + \delta$, where δ follows normal distribution with mean zero and variance σ_δ^2. The estimation and selection results with BIC model selection criteria are reported in Tables (3.1) and (3.2). It can be observed that the estimation for TR and IV is close to the true values of coefficients compared with NA method, of which the estimation is biased due to the ME effect. In terms of selection results, the values of FP and FN are both low for TR and IV methods. The boxplots for all coefficients of three methods are shown in Figure (3.1). The pattern of IV results mimics that of TR method, with the mean values centering around the true values of coefficients. The estimation of naive method is biased, with larger spread for some covariates (e.g. β_2, β_4).

Table 3.1: Estimation results of Example 3.2.1 with $n = 200$, $\sigma_\delta^2 = 5$

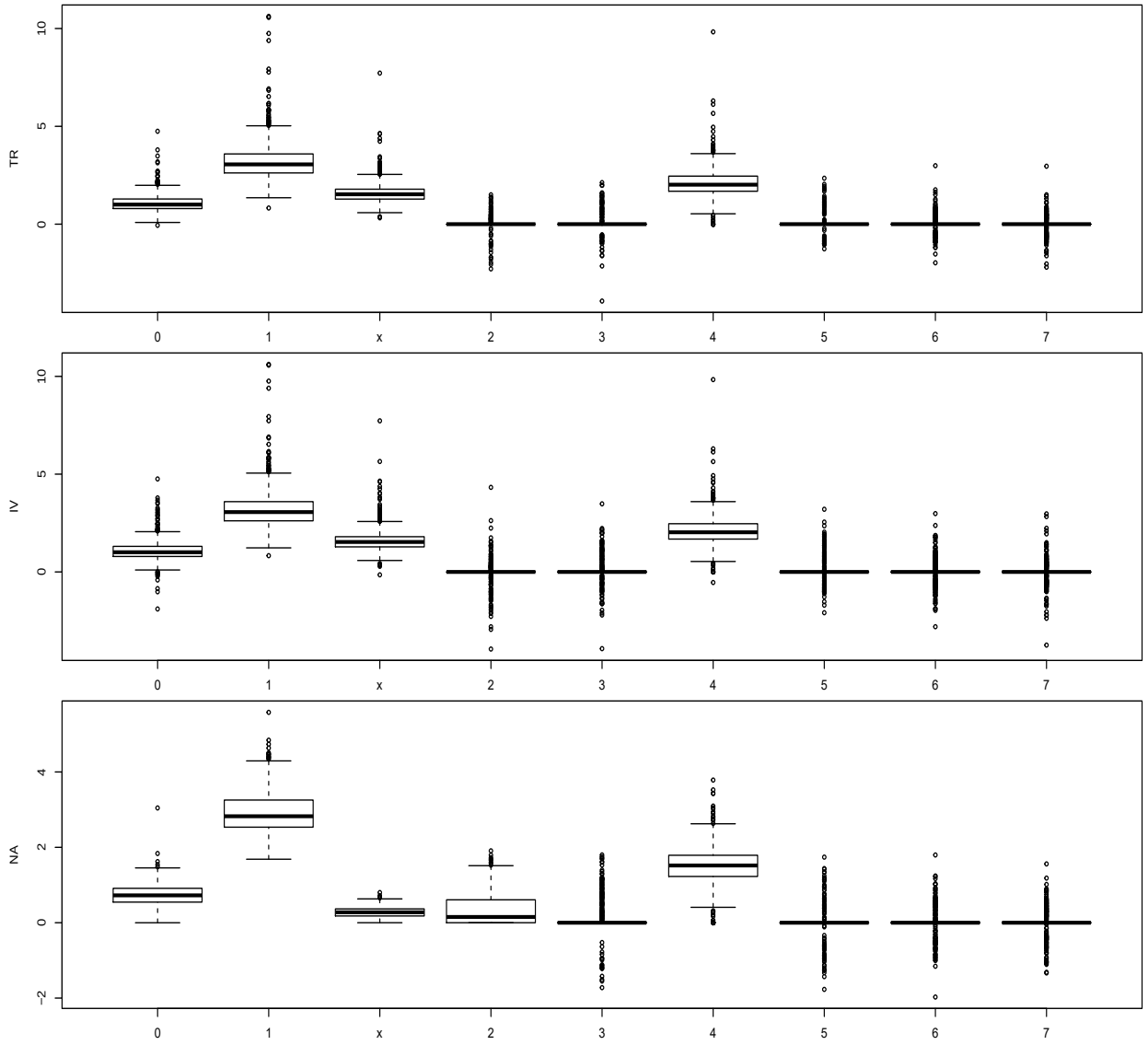
	α=1	β_1= 3	β_x= 1.5	β_2= 0	β_3= 0	β_4= 2	β_5= 0	β_6= 0	β_7= 0
TR	1.04	3.09	1.54	0.02	0.06	1.92	0.04	0.02	0.04
IV	1.03	3.12	1.47	0.01	0.07	1.91	0.05	0.04	0.04
NA	0.72	2.87	0.28	0.38	0.12	1.39	0.03	0.01	0.05

Table 3.2: Selection results of Example 3.2.1 with $n = 200$, $\sigma_\delta^2 = 5$

	FP	FN	MCC	MSE
TR	0.4	0.0	0.90	1.51
IV	0.6	0.0	0.87	1.62
NA	1.1	0.1	0.75	3.20

	z_0	z_1	x	z_2	z_3	z_4	z_5	z_6	z_7
TR	100	100	100	90	90	98	93	93	89
IV	100	100	100	87	87	98	90	90	86
NA	100	100	96	41	82	96	86	94	83

Figure 3.1: Boxplots of Example 3.2.1 with $n = 200$, $\sigma_\delta^2 = 5$; first row: TR, second row: IV, third row: NA; intercept α is denoted as 0



3.2.2 Numerical Example for Logistic Regression 2

In this example we simulate 1000 datasets from the model $y \sim \text{Bernoulli}(p(\alpha + x^T\beta))$, where the coefficients $(\alpha, \beta^T) = (1, 1.5, 3, 0, 0, 2, 0, 0, 0)$. The covariates $(z_1, -w, z_2, \ldots, z_7)^T$ are jointly generated from $N(0, \Sigma)$ with $\Sigma_{ij} = 0.7^{|i-j|}$. In this example the correlation between x and other covariates is negative. The rest of model setting remains the same as in Example (3.2.1). The overall pattern of the results in this example is similar as that in Example (3.2.1). In addition, it is worth noting that the selection of NA method has high values in both FP and FN. In other words, ignoring the ME can result in selecting irrelevant variables

and failing to retain those important ones, as observed in the simulation examples.

Table 3.3: Estimation results of Example 3.2.2 with $n = 200$, $\sigma_\delta^2 = 5$

	α=1	β_1= 1.5	β_x= 3	β_2= 0	β_3= 0	β_4= 2	β_5= 0	β_6= 0	β_7= 0
TR	1.12	1.53	3.24	0.05	0	2.13	0.02	0.03	-0.04
IV	1.12	1.78	3.1	-0.21	-0.25	2.15	0.13	-0.08	-0.1
NA	0.47	-0.03	0.36	-0.59	0	0.82	-0.01	0.02	-0.02

Table 3.4: Selection results of Example 3.2.2 with $n = 200$, $\sigma_\delta^2 = 5$

	FP	FN	MCC	MSE
TR	0.2	0.0	0.94	1.35
IV	0.8	0.0	0.84	5.02
NA	1.2	0.9	0.53	11.68

	z_0	z_1	x	z_2	z_3	z_4	z_5	z_6	z_7
TR	100	98	100	94	100	99	96	93	95
IV	100	99	100	84	87	100	84	83	85
NA	100	14	100	17	97	93	93	89	88

3.2.3 Numerical Example for Poisson Log-linear Regression 1

In this example we simulate 1000 datasets from the model $y \sim \text{Poisson}(\exp(\alpha + x^T\beta))$, where the coefficients $(\alpha, \beta^T) = (1, 1.5, 3, 0, 0, 2, 0, 0, 0)$. The instrumental variable and covariates $(z_1, w, z_2, \ldots, z_7)^T$ are jointly generated from $N(0, \Sigma)$ with $\Sigma_{ij} = 0.7^{|i-j|}$. The generation mechanism of x is the same as in Example (3.2.1). The estimation results are reported in Table (3.5) and the selection resulted are reported in Table (3.6). It can be seen that in poisson log-linear model, the naive method performs the worst among all three methods. The FP and FN remain at high level when $\sigma_\delta^2 = 5$. As comparison, the results from RIV method is similar to that of the TR method, where values of FP, FN and MSE are close to zero and MCC is close to one.

Table 3.5: Estimation results of Example 3.2.3 with $n = 200$, $\sigma_\delta^2 = 5$

	α= 1	β_1= 1.5	β_x= 3	β_2= 0	β_3= 0	β_4= 2	β_5= 0	β_5= 0	β_5= 0
TR	1.0	1.51	3.01	0	0	1.99	0	0	0
IV	1.02	1.50	2.98	0.01	0	2.02	0	0.02	0
NA	4.7	2.82	0.75	1.5	0.69	2.22	0.83	1.22	0.5

Table 3.6: Selection results of Example 3.2.3 with $n = 200$, $\sigma_\delta^2 = 5$

	FP	FN	MCC	MSE
TR	0	0	0.99	0.03
IV	0.3	0	0.94	0.07
NA	2.6	0.7	0.32	25.68

	z_0	z_1	x	z_2	z_3	z_4	z_5	z_6	z_7
TR	100	100	100	100	100	100	100	100	100
IV	100	100	100	93	94	100	95	95	96
NA	74	96	84	22	53	75	58	48	61

3.2.4 Numerical Example for Poisson Log-linear Regression 2

In this example we simulate 1000 datasets from the model $y \sim \text{Poisson}(\exp(\alpha + x^T\beta))$, where the coefficients $(\alpha, \beta^T) = (1, 3, 1.5, 0, 0, 2, 0, 0, 0)$. The instrumental variable and covariates $(z_1, w, z_2, \ldots, z_7)^T$ are jointly generated from $N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$. The generation mechanism of the covariate x is the same as in Example (3.2.1). Note that there are two things that differ from the previous example. First, the values of the first two coefficients are interchanged such that the coefficient corresponding to ME is smaller. Second, the correlation among all covariates also decreases. The estimation results are reported in Table (3.7) and the selection results are reported in Table (3.8). The pattern among the three methods is similar to that in Example (3.2.3). Though the FP and FN improve a bit for naive method as the variance of ME is low compared with the previous example, the results of RIV method is still much better compared with that of naive method.

Table 3.7: Estimation results of Example 3.2.4 with $n = 200$, $\sigma_\delta^2 = 1$

	α= 1	β_1= 3	β_x= 1.5	β_2= 0	β_3= 0	β_4= 2	β_5= 0	β_5= 0	β_5= 0
TR	1	3	1.5	0	0	2	0	0	0
IV	1	3	1.5	0.01	0.05	2	-0.02	0	0
NA	1.49	3.42	1.04	0.4	0.24	2.02	0.29	0.33	0.23

Table 3.8: Selection results of Example 3.2.4 with $n = 200$, $\sigma_\delta^2 = 1$

	FP	FN	MCC	MSE
TR	0.	0	0.98	0.02
IV	1.4	0	0.73	0.07
NA	2.4	0.3	0.49	5.72

	z_0	z_1	x	z_2	z_3	z_4	z_5	z_6	z_7
TR	100	100	100	100	100	100	100	100	100
IV	100	100	100	74	66	100	71	75	73
NA	73	100	100	38	60	100	52	58	55

3.3. Real Data Application

The Framingham heart study (Kannel et al., 1986) is a long-term cohort study monitoring the development of coronary heart disease (CHD). The dataset consists of 1615 observations

with a binary outcome variable indicating the occurrence of CHD. There are 128 CHD cases in the dataset. The covariates include age, systolic blood pressure (SBP), smoking status and serum cholesterol. It is well-known that the long-term SBP cannot be measured precisely hence has the measurement error. For this reason, we apply the proposed method on the Framingham dataset to identify important factors of the occurrence of CHD. For the covariate SBP, we apply the logarithm transformation log(SBP-50) and use the transformed SBP at exam 2 as instrumental variable. The data is standardized prior the analysis. The covariates include transformed SBP (x), serum cholesterol (z_1), age (z_2) and smoking status (z_3). Following Ma and Li (2010), all main factors and interaction terms, as well as the age squared are considered in the model. The results of RIV, naive and the naive plug-in methods are reported in Table (3.9). Specifically, the naive plug-in method refers to the two-stage estimation method, where the analysis is conducted based on the predicted covariate $\hat{x}$. We use the SCAD penalty function with BIC chosen as the model selection criteria. It can be observed that the results for naive and naive plug-in methods are similar, which select more covariates compared with RIV method. However, both methods fail to retain the main effect z_2 in the model. As a comparison, the RIV method selects a sparser model including all main effects only.

Table 3.9: Estimated coefficients and standard errors (SE) of Framingham dataset

	RIV	Plug-in	Naive
intercept	-2.68 (0.276)	-2.74 (0.118)	-2.74 (0.118)
x	0.34 (0.093)	0.34 (0.094)	0.33 (0.090)
xz_1	0 (-)	0.30 (0.087)	0 (-)
xz_2	0 (-)	0.62 (0.135)	0.64 (0.134)
xz_3	0 (-)	0 (-)	0 (-)
z_1	0.32 (0.020)	0 (-)	0.32 (0.088)
z_2	0.46 (0.010)	0 (-)	0 (-)
z_3	0 (-)	0.24 (0.104)	0.23 (0.105)
z_2^2	0 (-)	-0.23 (0.109)	-0.23 (0.109)
z_1z_2	0 (-)	0 (-)	0 (-)
z_1z_3	0 (-)	0 (-)	0 (-)
z_2z_3	0 (-)	0 (-)	0 (-)

3.4. Generalized Linear ME Model Proofs

Proof of Theorem 3

The proof follows from that of Ma and Li (2010). Write the score and hessian function of $Q_n(\psi)$ as $G_n(\psi)$ and $H_n(\psi)$ respectively. Using the Taylor expansion, the score of the objective function $Q_n(\psi) + \sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|)$ can be written as

$$S_n(\psi) = G_n(\psi_0) + \tilde{p}'_{\lambda_n}(|\beta_0|) \circ \text{sign}(\psi_0) + H_n(\psi^*)(\psi - \psi_0) + \tilde{p}''_{\lambda_n}(|\beta_0|)(\psi - \psi_0)(1 + o_p(1))$$

where ψ^* is in between ψ and ψ_0. It is sufficient to show that $S_n(\psi) = 0$ has a solution ψ satisfying $||\hat{\psi} - \psi_0|| = O_p(n^{-1/2})$. To this end, we show for any ψ such that $||\psi - \psi_0|| = n^{-1/2}C$ the inequality $(\psi - \psi_0)^T S_n(\psi) > 0$ holds with probability approaching 1. It follows that

$$(\psi - \psi_0)^T S_n(\psi) = (\psi - \psi_0)^T (G_n(\psi_0) + \tilde{p}'_{\lambda_n}(|\beta_0|) \circ \text{sign}(\psi_0)) + n||\psi - \psi_0||^2(1 + o_p(1))$$

It is can be seen that the first term is of order $O_p(C)$. The second term is of order $O_p(C^2)$. Hence for a sufficiently large C, the second term dominates the others. $(\psi - \psi_0)^T S_n(\psi)$ is shown to be positive with probability tending to 1, which completes the proof.

Proof of Theorem 4a

Taylor expansion of $S_n(\psi)$ around ψ_0 is given by

$$\begin{aligned} S_n(\psi) &= G_n(\psi_0) + H_n(\psi^*)(\psi - \psi_0) + n\tilde{p}'_{\lambda_n}(|\beta|) \circ \text{sign}(\psi) \\ &= n\lambda_n \left\{ \frac{1}{n\lambda_n} G_n(\psi_0) + \frac{1}{n\lambda_n} H_n(\psi^*)(\psi - \psi_0) + \frac{1}{\lambda_n} \tilde{p}'_{\lambda_n}(|\beta|) \circ \text{sign}(\psi) \right\} \end{aligned}$$

For $j \in J^c$ and $\epsilon_n = Cn^{-1/2}$ it can be shown that

$$S_n(\beta_j) = n\lambda_n \left\{ O_p(\frac{1}{\sqrt{n}\lambda_n}) + \frac{p'_{\lambda_n}(|\beta_j|)}{\lambda_n} \text{sign}(\beta_j) \right\}.$$

Together with the condition $\liminf_{n\to\infty} \liminf_{\xi \to 0^+} p'_{\lambda_n}(\xi)/\lambda_n > 0$ we have $G_n(\beta_j) > 0$ if $0 < \beta_j < \epsilon_n$; $G_n(\beta_j) < 0$ if $-\epsilon_n < \beta_j < 0$. Hence $P(\beta_j = 0) \to 1$ for $j \in J^c$.

Proof of Theorem 4b

The Taylor expansion of $S_n(\psi_J)$ around ψ_{0J} is given by

$$S_n(\psi_J) = G_n(\psi_{0J}) + H_n(\psi_J^*)(\psi_J - \psi_{0J}) + n\tilde{p}'_{\lambda_n}(|\beta_{0J}|) \circ \text{sign}(\psi_{0J}) + n\tilde{p}''_{\lambda_n}(|\beta_J^*|)(\psi_J - \psi_{0J})$$

where

$$\tilde{p}'_{\lambda_n}(|\beta_{0J}|) = (0, p'_{\lambda_n}(|\beta_{0J}^T|), 0)^T,$$

$$\tilde{p}''_{\lambda_n}(|\beta_J^*|) = \text{diag}(0, p''_{\lambda_n}(|\beta_J^{*T}|), 0),$$

$$G_n(\psi_{0J}) = \sum_{i=1}^{n} \frac{\partial \hat{\rho}_i^T(\psi_{0J})}{\partial \psi_J} A_i \hat{\rho}_i(\psi_{0J})$$

and

$$H_n(\psi_J^*) = \sum_{i=1}^{n} \left[\frac{\partial \hat{\rho}_i^T(\psi_J^*)}{\partial \psi_J} A_i \frac{\partial \hat{\rho}_i(\psi_J^*)}{\partial \psi_J^T} + \left(\hat{\rho}_i^T(\psi_J^*) A_i \otimes I_{s+2}\right) \frac{\partial \text{vec}\left(\partial \hat{\rho}_i^T(\psi_J^*)/\partial \psi_J\right)}{\partial \psi_J^T} \right].$$

Rearrange the terms we get

$$-\frac{1}{\sqrt{n}} G_n(\psi_{0J}) = \sqrt{n} \left(\frac{1}{n} H(\psi_J^*) + \tilde{p}''_{\lambda_n}(|\beta_J^*|) \right) (\psi_J - \psi_{0J}) + \sqrt{n} \tilde{p}'_{\lambda_n}(|\beta_{0J}|) \circ \text{sign}(\psi_{0J}).$$

Note that

$$\begin{aligned} \frac{1}{n} H(\psi_J^*) &\to_p E\left[\frac{\partial \rho^T(\psi_{0J})}{\partial \psi_J} A(w) \frac{\partial \rho(\psi_{0J})}{\partial \psi_J^T} + \left(\rho^T(\psi_{0J}) A \otimes I_{s+2}\right) \frac{\partial \text{vec}\left(\partial \rho^T(\psi_{0J})/\partial \psi_J\right)}{\partial \psi_J^T} \right] \\ &= B \end{aligned} \tag{3.4.1}$$

since the expectation of the second term is

$$\begin{aligned} & E\left[\rho^T(\psi_J^*) A \otimes I_{s+2} \frac{\partial \text{vec}\left(\partial \rho^T(\psi_J^*)/\partial \psi_J\right)}{\partial \psi_J^T} \right] \\ =& E\left[E(\rho^T(\psi_J^*)|\tilde{w}) A \otimes I_{s+2} \frac{\partial \text{vec}\left(\partial \rho^T(\psi_J^*)/\partial \psi_J\right)}{\partial \psi_J^T} \right] \\ =& 0. \end{aligned}$$

Now consider the first order Taylor expansion of $G_n(\psi_{10})$ around γ_{J0}

$$G_n(\psi_{0J}) = \sum_{i=1}^{n} \frac{\partial \rho_i^T(\psi_{0J})}{\partial \psi_J} A_i \rho_i(\psi_{0J}) + \frac{\partial^2 \tilde{Q}_n(\psi_{0J})}{\partial \psi_J \partial \gamma_J^T} (\hat{\gamma}_J - \gamma_{0J}) \tag{3.4.2}$$

where

$$\begin{aligned}
&\frac{\partial^2 \tilde{Q}_n(\psi_{0J})}{\partial \psi_J \partial \gamma_J^T} \\
=&\sum_{i=1}^{n} \left[\frac{\partial \rho_i^T(\psi_{0J}, \gamma_J^*)}{\partial \psi_J} A_i \frac{\partial \rho_i(\psi_{0J}, \gamma_J^*)}{\partial \gamma_J^T} + (\rho_i^T(\psi_{0J}, \gamma_J^*) A_i \otimes l_{s+2}) \frac{\partial \mathrm{vec}(\partial \rho_i^T(\psi_{0J}, \gamma_J^*)/\partial \psi_J)}{\partial \gamma_J^T} \right].
\end{aligned}$$

Using similar argument of equation (3.4.1), it can be shown that

$$\frac{1}{n} \frac{\partial^2 \tilde{Q}_n(\psi_{0J})}{\partial \psi_J \partial \gamma_J^T} \to_p E\left(\frac{\partial \rho^T(\psi_{0J})}{\partial \psi_J} A \frac{\partial \rho(\psi_{0J})}{\partial \gamma_J^T} \right).$$

In addition, the term $\hat{\gamma}_J - \gamma_{0J}$ in equation (3.4.2) can be written as

$$\hat{\gamma}_J - \gamma_{0J} = (\sum I_p \otimes w_i w_i^T)^{-1} \left(\sum (x_{Ji}^* - \Gamma_{0J} w_i) \otimes w_i \right)$$

Hence the equation (3.4.2) can be written as

$$G_n(\psi_{0J}) = D_n \sum_{i=1}^{n} K_i$$

where

$$\begin{aligned}
D_n &= \left(I_{s+2}, \frac{\partial^2 \tilde{Q}_n(\psi_{0J})}{\partial \psi_J \partial \gamma_J^T} (I_p \otimes (\sum_{i=1}^{n} w_i w_i^T)^{-1}) \right) \\
&\to \left(I_{s+2}, E\Big(\frac{\partial \rho^T(\psi_{0J})}{\partial \psi_J} A \frac{\partial \rho(\psi_{0J})}{\partial \gamma_J^T} \Big) (I_p \otimes E(w w^T)^{-1}) \right) \\
&= D,
\end{aligned}$$

$$K_i = \begin{pmatrix} \partial \rho_i^T(\psi_{0J}) / \partial \psi_J \cdot A_i \rho_i(\psi_{0J}) \\ (x_{Ji}^* - \Gamma_{0J} w_i) \otimes w_i \end{pmatrix}.$$

Then

$$\sqrt{n}(B + \Sigma)(\hat{\psi}_J - \psi_{0J}) + \sqrt{n}\boldsymbol{b} \to_d N(0, DCD^T)$$

Chapter 4

Conclusion and Discussion

In this thesis we illustrated how ME affects the variable selection through several heuristic examples, and proposed regularized instrument variable method correcting for ME effects in both linear ME and generalized linear ME models. Specifically, the proposed estimator is shown to have the oracle property, which is consistent in both variable selection and parameter estimation. The asymptotic distribution is derived for the proposed estimator in both linear and generalized linear ME models. Extensive simulation studies for linear, logistic and poisson log-linear models are conducted examining the performance of the proposed estimator, as well as the naive estimator. Simulation results show that the proposed estimator performs well in various model settings with finite sample size, compared with naive estimator. The ME effect on variable selection in classical and Berkson ME models, as well as different penalty functions and model selection criteria are discussed. Finally the proposed method is applied to real datasets of diabetes and Framingham heart study. For the future research, there are many other models besides the generalized linear ME model, the proposed method can be future generalized to nonlinear models. It is worth noting that computation in high-dimensional nonlinear model is changeling and new algorithms need to be developed.

Appendix

A. Derivation of BIC Model Selection Criteria

In this section we sketch the idea of the derivation of the BIC criteria, the detailed proof can be found in Schwarz et al. (1978), Neath and Cavanaugh (2012), and Bhat and Kumar (2010). Assume the data $(y_1, y_2, \ldots, y_n)$ come from some unknown probability density function $g(y)$, where $g(y)$ is referred as the true model. In addition, assume that the L candidate models are coming from a family of distributions

$$\mathcal{F}(k) = \{f(y|\theta_k) : \theta_k \in \Theta_k\}$$

for $k = 1, 2, \ldots, L$. Note that Θ_k are parameter spaces with different subsets of covariates. Let $\pi(k_1), \pi(k_2), \ldots, \pi(k_L)$ be the prior distribution of the candidate models M_k, $k = k_1, k_2, \cdots, k_L$. Then the posterior distribution of model $M_k(\theta_k)$ is given by

$$p(k, \theta_k|y) = \frac{\pi(k)g(\theta_k|k)f(y|\theta_k)}{m(y)}$$

where $g(\theta_k|k)$ is the prior distribution of θ_k given the candidate model M_k, $f(y|\theta_k)$ is the density function of y and $m(y)$ is the marginal distribution of y. The Bayesian information criteria chooses the model with maximum posteriori, which amounts to maximizing the posterior probability with respect to k

$$\begin{aligned} p(k|y) &= \int p(k, \theta_k|y)d\theta_k \\ &= \int_{\Theta_k} \frac{\pi(k)g(\theta_k|k)f(y|\theta_k)}{m(y)}d\theta_k. \end{aligned}$$

Maximizing $p(k|y)$ is equivalent to minimizing the negative log-likelihood function (with a constant factor 2)

$$-2\log p(k|y) = 2\log(m(y)) - 2\log(\pi(k)) - 2\log\left(\int_{\Theta_k} g(\theta_k|k)f(y|\theta_k)d\theta_k\right). \tag{4.0.1}$$

Now write $L(\theta_k)$ as the likelihood function corresponding to $f(y|\theta_k)$. Use Taylor expansion of $\log L(\theta_k)$ around $\hat{\theta}_k$, the maximizer of $L(\theta_k)$

$$\begin{aligned}
&\log L(\theta_k) \\
=&\log L(\hat{\theta}_k) + (\theta_k - \hat{\theta}_k)^T \frac{\partial \log L(\hat{\theta}_k)}{\partial \theta_k} + \frac{1}{2}(\theta_k - \hat{\theta}_k)^T \left[\frac{\partial^2}{\partial \theta_k \theta_k^T} L(\hat{\theta}_k^*)\right](\theta_k - \hat{\theta}_k) \\
=&\log L(\hat{\theta}_k) - \frac{1}{2}(\theta_k - \hat{\theta}_k)^T nI(\hat{\theta}_k^*)(\theta_k - \hat{\theta}_k)
\end{aligned}$$

where $\hat{\theta}_k^*$ is in between θ_k and $\hat{\theta}_k$. Hence

$$\begin{aligned}
&\int_{\Theta_k} g(\theta_k|k)f(y|\theta_k)d\theta_k \\
=&\int_{\Theta_k} g(\theta_k|k)\left[L(\hat{\theta}_k)\exp\left(-\frac{1}{2}(\theta_k - \hat{\theta}_k)^T nI(\hat{\theta}_k^*)(\theta_k - \hat{\theta}_k)\right)\right]d\theta_k.
\end{aligned}$$

Using the non-informative prior where $g(\theta_k|k) = 1$ and the fact that $\hat{\theta}_k \to_p \theta_k$, we have

$$\int_{\Theta_k} L(\theta_k) \approx L(\hat{\theta}_k)\sqrt{(2\pi)^k n^{-k}|I^{-1}(\hat{\theta}_k)|}.$$

Now the negative log-likelihood function (4.0.1) can be written as follows.

$$\begin{aligned}
&-2\log p(k|y) \\
=&2\log(m(y)) - 2\log(\pi(k)) - 2\log\left(L(\hat{\theta}_k)\sqrt{(2\pi)^k n^{-k}|I^{-1}(\hat{\theta}_k)|}\right) \\
=&2\log(m(y)) - 2\log(\pi(k)) - 2\log L(\hat{\theta}_k) + k\log\frac{n}{2\pi} + \log|I(\hat{\theta}_k)| \\
\propto&-2\log L(\hat{\theta}_k) + k\log(n)
\end{aligned}$$

which is the Bayesian information criteria after ignoring the constant term with respect to n.

B. Score and Hessian Matrix Examples in Simulation

In this section we provide the details of score and hessian expressions used in the numerical examples. Note that the conditional expectation (3.1.2) is given by

$$\begin{aligned} E(\tilde{x}^* y|w) &= \int \tilde{x}^* G^{-1}(\alpha + \beta_x^T \Gamma w + \beta_z^T z + \beta_x^T u) f_U(u;\phi) du \\ &= \int (1, (\Gamma w + u)^T)^T G^{-1}(\alpha + \beta_x^T \Gamma w + \beta_z^T z + \beta_x^T u) f_U(u;\phi) du \end{aligned} \tag{4.0.2}$$

For the first stage estimation where the identity matrix is used as weight matrix, the score and hessian function are given by

$$\frac{\partial Q_n(\psi)}{\partial \psi} = -\frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \int \tilde{x}_i (1, x_i^T) b_i''(\cdot) f(u;\phi) du \\ \int b_i'(\cdot) \partial f(u;\phi)/\partial \phi (1, x_i^T) du \end{pmatrix} (y_i \tilde{x}_i^* - E(\tilde{x}^* y|w))$$

where $b_i(\cdot) = b(\alpha + \beta_x^T \Gamma w_i + \beta_z^T z_i + \beta_x^T u_i)$. When the random error u follows an univariate normal distribution, it can be shown that

$$\frac{\partial f(u;\phi)}{\partial \phi} = f(u;\phi)\frac{u^2 - \phi}{2\phi^2}$$

and

$$\frac{\partial^2 f(u;\phi)}{\partial \phi^2} = f(u;\phi)\frac{u^4 - 6u^2\phi + 3\phi^2}{4\phi^4}$$

In which case we have

$$\frac{\partial Q_n^2(\psi)}{\partial \psi \partial \psi^T} = -\frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} H_i^{(11)} & H_i^{(12)} \\ (H_i^{(12)})^T & H_i^{(22)} \end{pmatrix}$$

where $H_i^{(11)} \in \mathbb{R}^{(d+1)\times(d+1)}$, $H_i^{(12)} \in \mathbb{R}^{d+1}$ and $H_i^{(22)} \in \mathbb{R}$. Write

$$e_i^{(1)} = y_i - \int b_i'(\cdot) f(u;\phi) du$$

, and

$$e_i^{(2)} = y_i x_i^* - \int b_i'(\cdot) x_i f(u;\phi) du,$$

we have

$$H_i^{(11)} = e_i^{(1)} \int b_i^{'''}(\cdot)\tilde{x}_i\tilde{x}_i^T f(u;\phi)du - \int b_i^{''}(\cdot)\tilde{x}_i f(u;\phi)du \int b_i^{''}(\cdot)\tilde{x}_i^T f(u;\phi)du +$$
$$e_i^{(2)} \int b_i^{'''}(\cdot)x_i\tilde{x}_i\tilde{x}_i^T f(u;\phi)du - \int b_i^{''}(\cdot)x_i\tilde{x}_i f(u;\phi)du \int b_i^{''}(\cdot)x_i\tilde{x}_i^T f(u;\phi)du$$

$$H_i^{(12)} = e_i^{(1)} \int b_i^{''}(\cdot)\tilde{x}_i \partial f(u;\phi)/\partial\phi du - \int b_i^{'}(\cdot)\partial f(u;\phi)/\partial\phi du \int b_i^{''}(\cdot)\tilde{x}_i f(u;\phi)du +$$
$$e_i^{(2)} \int b_i^{''}(\cdot)x_i\tilde{x}_i \partial f(u;\phi)/\partial\phi du - \int b_i^{'}(\cdot)x_i \partial f(u;\phi)/\partial\phi du \int b_i^{''}(\cdot)x_i\tilde{x}_i f(u;\phi)du$$

$$H_i^{(22)} = e_i^{(1)} \int b_i^{'}(\cdot)\partial^2 f(u;\phi)/\partial\phi^2 du - (\int b_i^{'}(\cdot)\partial f(u;\phi)/\partial\phi du)^2 +$$
$$e_i^{(2)} \int b_i^{'}(\cdot)x_i\partial^2 f(u;\phi)/\partial\phi^2 du - \int b_i^{'}(\cdot)x_i \partial f(u;\phi)/\partial\phi du$$

where in logistic model $b^{'}(\eta) = 1/(1+exp(-\eta)) = p$, $b^{''}(\eta) = p(1-p)$ and $b^{'''}(\eta) = p(1-p)(1-2p)$; in poisson model $b^{'}(\eta) = b^{''}(\eta) = b^{'''}(\eta) = exp(\eta)$.

Bibliography

Abarin, T. and Wang, L. (2012). Instrumental variable approach to covariate measurement error in generalized linear models. *Annals of the Institute of Statistical Mathematics*, 64(3):475–493.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Bhat, H. S. and Kumar, N. (2010). On the derivation of the bayesian information criterion. *School of Natural Sciences, University of California*.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232.

Candes, E., Tao, T., et al. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Caner, M. and Fan, Q. (2010). The adaptive lasso method for instrumental variable selection. Technical report, Working Paper, North Carolina State University.

Caner, M. and Fan, Q. (2015). Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive lasso. *Journal of Econometrics*, 187(1):256–274.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., and Stefanski, L. A. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Chen, J. and Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, pages 555–574.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association*, 89(428):1314–1328.

Falkner, B., Gidding, S. S., Ramirez-Garnica, G., Wiltrout, S. A., West, D., and Rappaport, E. B. (2006). The relationship of body mass index and blood pressure in primary care pediatric patients. *Journal of Pediatrics*, 148(2):195–200.

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

Fan, J. and Liao, Y. (2014). Endogeneity in high dimensions. *The Annals of Statistics*, 42(3):872.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Fuller, W. A. (2009). *Measurement error models*, volume 305. John Wiley & Sons.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Huang, X. and Zhang, H. (2013). Variable selection in linear measurement error models via penalized score functions. *Journal of Statistical Planning and Inference*, 143(12):2101–2111.

Kannel, W., Neaton, J., Wentworth, D. f., Thomas, H., Stamler, J., Hulley, S., and Kjelsberg, M. (1986). Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for the mrfit. *American heart journal*, 112(4):825–836.

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science & Business Media.

Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, 104(485):234–248.

Lin, W., Feng, R., and Li, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509):270–288.

Linderman, G. C., Lu, J., Lu, Y., Sun, X., Xu, W., Nasir, K., Schulz, W., Jiang, L., and Krumholz, H. M. (2018). Association of body mass index with blood pressure among 1.7 million chinese adults. *JAMA Network Open*, 1(4):e181271–e181271.

Ma, Y. and Li, R. (2010). Variable selection in measurement error models. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 16(1):274.

Mallows, C. L. (1973). Some comments on c p. *Technometrics*, 15(4):661–675.

Neath, A. A. and Cavanaugh, J. E. (2012). The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203.

Negahban, S., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74(4):703–716.

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

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568.

Wang, L. and Hsiao, C. (2011). Method of moments estimation and identifiability of semiparametric nonlinear errors-in-variables models. *Journal of Econometrics*, 165(1):30 – 44. Moment Restriction-Based Econometric Methods.

Willems, J. P., Saunders, J. T., Hunt, D. E., and Schorling, J. B. (1997). Prevalence of coronary heart disease risk factors among rural blacks: a community-based study. *Southern Medical Journal*, 90(8):814–820.

Yi, G. Y., Tan, X., and Li, R. (2015). Variable selection and inference procedures for marginal analysis of longitudinal data with missing observations and covariate measurement error. *Canadian Journal of Statistics*, 43(4):498–518.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zhang, X., Wang, H., Ma, Y., and Carroll, R. J. (2017). Linear model selection when covariates contain errors. *Journal of the American Statistical Association*, 112(520):1553–1561.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67(2):301–320.