Variable Selection in Measurement Error Models\footnote{AMS 2000 subject classifications. Primary 62G08, 62G10; secondary 62G20. Key Words: Errors in variables; Estimating equations; Kernel regression; Instrumental variables; Latent variables; Local efficiency; Measurement error; Nonconcave penalized likelihood; SCAD; Semiparametric methods. Li’s research was supported by NSF grants DMS-0348869.}

YANYUAN MA  
Institut de Statistique  
Université de Neuchâtel  
Pierre à Mazel 7, 2000 Neuchâtel

RUNZE LI  
Department of Statistics  
Pennsylvania State University  
University Park, PA16802-2111

Abstract

Compared with ordinary regression models, the computational cost for estimating parameters in general measurement error models is often much more expensive because the estimation procedures typically require solving integral equations. In addition, natural criteria functions are often unavailable for general measurement error models. Thus, the traditionally best variable selection procedures become infeasible in the measurement error models context. In this paper, we develop a new framework for variable selection in measurement error models via penalized estimating equations. We first propose a new class of variable selection procedures for general parametric measurement error models and for general semiparametric measurement error models, and study the asymptotic properties of the proposed procedures. Then, under certain regularity conditions and with a properly chosen regularization parameter, we demonstrate that the proposed procedure performs as well as an oracle procedure. We assess the finite sample performance by Monte Carlo simulation studies and illustrate the proposed methodology through the empirical analysis of a real data set.

1 Introduction

Variable selection is fundamental in high-dimensional statistical modeling and has received much attention in the recent literature. Frank and Friedman (1993) proposed bridge regression for variable selection in linear models. Tibshirani (1996) proposed the LASSO method for variable selection in linear models. Fan and Li (2001) proposed a nonconcave penalized likelihood method for variable selection in likelihood-based models. The nonconcave penalized likelihood approach has been
Y. Ma and R. Li

further extended for Cox’s model for survival data (Fan and Li, 2002) and partially linear models for longitudinal data (Fan and Li, 2004). The LASSO method becomes very popular in the literature of machine learning. The whole solution path of LASSO can be obtained by the least angle regression (LARS) algorithm (Efron, et al., 2004). Yuan and Lin (2005) provided insights of the LASSO method from empirical Bayes point of view, while Yuan and Lin (2006) extended the LASSO method for grouped variable selection in linear models. Zou (2006) proposed adaptive LASSO for linear models to reduce the estimation bias in the original LASSO. Zhang and Lu (2007) studied the adaptive LASSO for the Cox model. Fan and Li (2006) presented a review on feature selection in high-dimensional modeling.

In many situations, covariates can only be measured imprecisely or indirectly, thus result in so called measurement error models or errors-in-variable models in the literature. Various statistical procedure have been developed for analyzing measurement error models (Fuller, 1987; Carroll et al., 2006). Tsiatis and Ma (2004) proposed the locally efficient estimator for parametric measurement error models, and Ma and Carroll (2006) further extend the locally efficient estimation procedure for semiparametric measurement error models. To our best knowledge, variable selection for measurement error model has not been systematically studied yet. This paper intends to develop a class of variable selection procedures for both parametric measurement error models and semiparametric measurement error models. Liang and Li (2005) extended the nonconcave penalized likelihood method for linear and partially linear measurement error models, but their method is not applicable for measurement error models beyond linear or partially linear setting, such as logistic measurement error model and generalized partially linear measurement error models. Variable selection for general parametric or semiparametric measurement error models is much more challenging than that for the linear or partially linear models. One major difficulty associated with this topic is the lack of availability of a likelihood function. In fact, classical model selection methods such as BIC, AIC, as well as the aforementioned penalized likelihood methods all require computing the likelihood, which is typically formidable in measurement error models due to the difficulty in obtaining the distribution of the error-prone covariates. Although a reasonable criterion function can be used in place of the likelihood, the difficulty persists in that except for very special models such as in linear or partially linear cases, even a criterion function is unavailable.

To develop variable selection procedures for both parametric and semiparametric measurement error models, we propose a penalized estimating equation method. We systematically study the asymptotic properties of the proposed procedures. In this paper, we first consider two scenarios: one is when the number of regression coefficients is fixed and finite, and the other is when the number of regression coefficients diverges as the sample size increases. We demonstrate how the convergence
rate of the resulting estimator depends on the regularization parameter. We further show that with proper choice of the regularization parameters and the penalty function, the penalized estimating equation estimator possesses the oracle property (Fan and Li, 2001). It is desirable to have an automatic, data-driven method to select the regularization parameters. We propose GCV-type and BIC-type tuning parameter selector for the proposed penalized estimating equation method. Monte Carlo simulation studies are conducted to assess finite sample performance in terms of model complexity and model error.

The rest of the paper is organized as the following. We propose a new class of variable selection procedure for parametric measurement error model and study asymptotic properties of the proposed procedures in Section 2. We develop a new variable selection procedure for semiparametric measurement error model in Section 3. Implementation issues and numerical examples are presented in Section 4, where we describe a data driven automatic tuning parameter selection method (Section 4.1), define the concept of approximate model error to evaluate the selected model (Section 4.2), carry out a simulation study to assess the finite sample performance of the proposed procedures (Section 4.3), and illustrate our method in an example (Section 4.4). Technical details are collected in Section 5. We give some discussions in Section 6.

2 Parametric measurement error models

A general parametric measurement error model has two parts and can be written as

\[ p_{Y|X,Z}(Y|X, Z, \beta) \quad \text{and} \quad p_{W|X,Z}(W|X, Z, \xi). \]  

(1)

The main model is \( p_{Y|X,Z}(Y|X, Z, \beta) \), denoting the conditional probability density function (pdf) of the response variable \( Y \) on the covariates measured with error \( X \) and the covariates measured without error \( Z \). The error model is denoted \( p_{W|X,Z}(W|X, Z, \xi) \), where \( W \) is an observable surrogate of \( X \). Parameter \( \beta \) is a \( d \)-dimensional regression coefficient, \( \xi \) is a finite dimensional parameter, and our main interest is in selecting the relevant subset of covariates in \( X \) and \( Z \) and estimating the subsequent parameters contained in \( \beta \). Typically, \( \xi \) is a nuisance parameter, and its estimation usually requires multiple observations or instrumental variables. As in the literature, for simplicity, we assume in the main context of this paper that the error model \( p_{W|X,Z}(W|X, Z) \) is completely known and hence \( \xi \) is suppressed. We discuss the treatment of the unknown \( \xi \) case in Section 6. The observed data is of the form \( \{(W_i, Z_i, Y_i), i = 1, \ldots, n\} \).

Denote \( S^*_\beta \) the purported score function. That is,

\[ S^*_\beta(W, Z, Y) = \frac{\partial \log \int p_{W|X,Z}(W|X, Z)p_{Y|X,Z}(Y|X, Z)p_{X|Z}(X|Z)d\mu(X)}{\partial \beta}. \]
where $p^*_{X|Z}(X|Z)$ is a conditional pdf that one posits, which can be equal or not equal to the true pdf $p_{X|Z}(X|Z)$. The function $a(X, Z)$ satisfies


where $E^*$ indicates that the expectation is calculated using the posited $p^*_{X|Z}(X|Z)$. Let


Define penalized estimating equations for model (1) to be

$$\sum_{i=1}^n S^*_\text{eff}(W_i, Z_i, Y_i, \beta) - n\dot{p}_\lambda(\beta) = 0. \tag{2}$$

where $\dot{p}_\lambda(\beta) = \{p'_\lambda(\beta_1), \cdots, p'_\lambda(\beta_d)\}^T$ and $p'_\lambda(\cdot)$ is the first order derivative of a penalty function $p_\lambda(\cdot)$. In practice, we may allow different coefficients to have penalty functions with different regularization parameters. For example, we may want to keep some variables in the model and do not penalize their coefficients. For ease of presentation, we assume that the penalty functions and the regularization parameters are the same for all the coefficients in this paper.

The choice of the penalty functions has been studied in Fan and Li (2001) in depth. The penalties in the classic variable selection criteria, such as the AIC and BIC, cannot be applied for the penalized estimating equations. The $L_q$ penalty, namely $p_\lambda(\theta) = q^{-1}\lambda|\theta|^q$, has been proposed for bridge regression in Frank and Friedman (1993). Fan and Li (2001) advocated the use of the SCAD penalty, whose first order derivative is defined as

$$p'_\lambda(\theta) = \lambda \left\{ I(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)}{(a-1)\lambda} I(|\theta| > \lambda) \right\} \text{sign}(\theta), \tag{3}$$

where $a > 2$ and $\text{sign}(\cdot)$ is the sign function, i.e., $\text{sign}(\theta) = -1$, 0, and 1 when $\theta < 0$, $= 0$ and $> 0$ respectively. Fan and Li (2001) further suggested using $a = 3.7$ from a Bayesian point of view. With a proper choice of penalty function, such as the SCAD penalty, the resulting estimate contains some exact zero coefficients. This is equivalent to excluding the corresponding variables from the final selected model, thus achieving the purpose of variable selection for parametric measurement error models.

We now study the asymptotic properties of the resulting estimator from the penalized estimating equations. Denote $\beta_0 = (\beta_{10}, \cdots, \beta_{00})^T$ the true value of $\beta$. Let

$$a_n = \max\{|p'_\lambda(\beta_{00})| : \beta_{00} \neq 0\}, \tag{4}$$

and

$$b_n = \max\{|p''_\lambda(\beta_{00})| : \beta_{00} \neq 0\}. \tag{5}$$

$$4$$
where we write $\lambda$ as $\lambda_n$ to emphasize that $\lambda$ depends on the sample size $n$. We first establish the convergence rate of the penalized estimating equation estimator for the settings in which the dimension of $\beta$ is fixed and finite.

**Theorem 1** Suppose that the regularity conditions (A1)-(A4) in Section 5 hold. If both $a_n$ and $b_n$ tends to 0 as $n \to \infty$, then, with probability tending to one, there exists a root of (2), denoted $\hat{\beta}$, such that $||\hat{\beta} - \beta_0|| = O_p(n^{-1/2} + a_n)$.

Before we demonstrate that the resulting estimator possesses the oracle property, let us introduce some notation. Without loss of generality, it is assumed $\beta_0 = (\beta_{I0}^T, \beta_{II}^T)^T$, and in the true model, any element in $\beta_{I0}$ is not equal to 0 while $\beta_{II} = 0$. Throughout this paper, denote the dimension of $\beta_I$ as $d_1$ and that of $\beta_{II}$ as $d_2$. Let $d = d_1 + d_2$. Furthermore, denote

$$b = \{p_{\lambda_n}'(\beta_{10}), \cdots, p_{\lambda_n}'(\beta_{d_10})\}^T \quad (6)$$

$$\Sigma = \text{diag}\{p_{\lambda_n}''(\beta_{10}), \cdots, p_{\lambda_n}''(\beta_{d_10})\}, \quad (7)$$

and the first $d_1$ components of $S_{eff}^*(W, Z, Y, \beta)$ is denoted $S_{eff,I}^*(\beta)$.

**Theorem 2** Under regularity conditions of Theorem 1, if

$$\liminf_{n \to \infty} \liminf_{\theta \to 0^+} \sqrt{n} p_{\lambda_n}'(\theta) \to \infty, \quad (8)$$

then with probability tending to one, any root-$n$ convergent solution $\hat{\beta} = (\hat{\beta}_I^T, \hat{\beta}_{II}^T)^T$ of (2) must satisfy that

(i) $\hat{\beta}_{II} = 0$;

(ii)

$$\sqrt{n} \left[ \hat{\beta}_I - \beta_{I0} - \left\{ E \frac{\partial S_{eff,I}^*(\beta_{I0})}{\partial \beta_I^T} - \Sigma \right\}^{-1} b \right] \xrightarrow{D} N \left[ 0, \left\{ E \frac{\partial S_{eff,I}^*(\beta_{I0})}{\partial \beta_I^T} - \Sigma \right\}^{-1} E \left\{ S_{eff,I}^*(\beta_{I0}) S_{eff,I}^{*T}(\beta_{I0}) \right\} \left\{ E \frac{\partial S_{eff,I}^*(\beta_{I0})}{\partial \beta_I^T} - \Sigma \right\}^{-T} \right] ,$$

where the notation $\xrightarrow{D}$ stands for convergence in distribution, and we use the notation $M^{-T}$ to denote $(M^{-1})^T$ for a matrix $M$. 

5
For some penalty functions, including the SCAD penalty, $b$ and $\Sigma$ are zero when $\lambda_n$ is sufficiently small. So we actually have

$$\sqrt{n}(\hat{\beta}_I - \beta_{I0}) \rightarrow N(0, E(\partial S_{I, I}^*/\partial \beta_I^T)\{E(S_{I, I}^* S_{I, I}^T)\}^{-1})$$

in distribution. In other words, with probability tending to 1, the penalized estimator performs the same as the locally efficient estimator under the correct model. Hence it has the celebrated oracle property.

In Theorems 1 and 2, both $d_1$ and $d$ are fixed, as $n \rightarrow \infty$. Concerns about model bias often prompts us to build models that contain many variables, especially so when the sample size becomes large. A reasonable way to capture such tendency is to consider the situation where the dimension of the parameter $\beta$ increases along with the sample size $n$. Obviously, the methods in (2) (and also in (11) later on for semiparametric model case) can still be used in this case; the operation does not involve any modification from the fixed number of parameters case. However, the large sample properties of the same procedures will behave inevitably differently under the increasing parameter number case. We therefore study the sampling properties of the penalized estimating equation estimator under the setting in which both $d_1$ and $d$ tend to infinity as $n$ goes to infinity.

To reflect that $d$ and $d_1$ change with $n$, we use the notation $d_{n1}$, $d_n$. The issue of a diverging number of parameters has also been considered in Fan and Peng (2004) in the context of penalized likelihood. Because semiparametric efficiency is not defined under this setting, our result will no longer imply any local efficiency. However, when $\lambda_n$ is sufficiently small, both $b$ and $\Sigma$ are zero, so the oracle property still holds. The main results are given in the following two theorems, whose proofs are given in Section 5.

**Theorem 3** Under the regularity conditions (B1)-(B6) in Section 5, and if $d_{n1}^4 n^{-1} \rightarrow 0$, $\lambda_n \rightarrow 0$ when $n \rightarrow \infty$, then with probability tending to one, there exists a root of (2), denoted $\hat{\beta}_n$, such that $||\hat{\beta}_n - \beta_{n0}|| = O_p\{\sqrt{n}(n^{-1/2} + a_n)\}$.

**Theorem 4** Under regularity conditions (B1)-(B6) in Section 5, assume $\lambda_n \rightarrow 0$ and $d_{n1}^5 / n \rightarrow 0$ when $n \rightarrow \infty$. If

$$\lim \inf_{n \rightarrow \infty} \lim \inf_{\theta \rightarrow 0^+} \sqrt{n/d_{n1}^5} p_{n, n}^*(\theta) \rightarrow \infty,$$  \hspace{1cm} (9)

then with probability tending to one, any root-$n$ consistent solution $\hat{\beta}_n = (\hat{\beta}_I^T, \hat{\beta}_{II}^T)^T$ of (2) must satisfy that

(i) $\hat{\beta}_{II} = 0$, \hspace{1cm} 6
(ii) For any \( d_1 \times 1 \) vector \( v \), s.t. \( v^Tv = 1 \),

\[
\sqrt{n}v^T \left[ E \left\{ S^*_\text{eff,1}(\beta_0)S^*_{\text{eff,1}}(\beta_0) \right\} \right]^{-1/2} \left\{ E \frac{\partial S^*_\text{eff,1}(\beta_0)}{\partial \beta_1^T} - \Sigma \right\} \\
\left[ \hat{\beta}_I - \beta_I - \left\{ E \frac{\partial S^*_\text{eff,1}(\beta_0)}{\partial \beta_1^T} - \Sigma \right\}^{-1} b \right] \overset{D}{\rightarrow} \mathcal{N}(0,1)
\]

3 Semiparametric measurement error models

The semiparametric measurement error model we consider here also has two parts. The major difference from its parametric counter part is that the main model contains an unknown function of one of the observable covariates. For notational convenience, we rewrite the observable covariates \( Z \) into two parts, \( Z \) and \( S \), where now \( Z \) denotes an observable covariate that enters the model nonparametrically, say through \( \theta(Z) \), and \( S \) denotes the observable covariates that enter the model parametrically. The resulting semiparametric measurement error model can be summarized as

\[
p_{Y|X,Z,S}(Y|X,Z,S,\beta,\theta(Z)) \quad \text{and} \quad p_{W|X,Z,S}(W|X,Z,S).
\]

The method we propose in this situation admits a similar simple form

\[
\sum_{i=1}^{n} \mathcal{L}(W_i, Z_i, S_i, Y_i, \beta, \hat{\theta}_i) - n \hat{p}_{\lambda}(\beta) = 0,
\]

where \( \hat{p}_{\lambda}(\beta) \) has the same form as in (3). However, the computation of \( \mathcal{L} \) is more involved as we now describe. If we replace \( \theta(Z) \) with a single unknown parameter \( \alpha \), and append \( \alpha \) to \( \beta \), we obtain a parametric measurement error model with parameters \((\beta^T, \alpha)^T\). For this parametric model, we can compute its corresponding \( S^*_{\text{eff}} \) as we did in Section 2. Decompose \( S^*_{\text{eff}} \) into \( \mathcal{L}(X, Z, S, Y, \beta, \alpha) \) and \( \Psi(X, Z, S, Y, \beta, \alpha) \), where \( \mathcal{L} \) has the same dimension as \( \beta \) and \( \Psi \) is a one-dimensional function.

We solve for \( \hat{\theta}_i, i = 1, \ldots, n \), from

\[
\sum_{i=1}^{n} K_h(s_i - s_1)\Psi(w_i, z_i, s_i, y_i; \beta, \hat{\theta}_1) = 0 \\
\vdots \\
\sum_{i=1}^{n} K_h(s_i - s_n)\Psi(w_i, z_i, s_i, y_i; \beta, \hat{\theta}_n) = 0,
\]

where \( K_h(s) = h^{-1}K(s/h), K \) is a kernel function, \( h \) is a bandwidth. Insert the \( \hat{\theta}_i \)'s into \( \mathcal{L} \) in (11) and we obtain a complete description of the estimator. Note that since \( \hat{\theta}_i \) depends on \( \beta \), so a more precise notation for \( \hat{\theta}_i \) is \( \hat{\theta}_i(\beta) \).
Similar results hold for the semiparametric penalized estimator. In the semiparametric model setting, make the definitions that $L_I$ is the first $d_1$ components of $L$, $L_{I\beta_I}$ is the partial derivative of $L_I$ with respect to $\beta_I$, $L_{I\theta}$ is the partial derivative of $L_I$ with respect to $\theta$, $\Psi_\theta$ is the partial derivative of $\Psi$ with respect to $\theta$ and $\Psi_{\beta_I}$ is the partial derivative of $\Psi$ with respect to $\beta_I$. Also define $\Omega(Z) = E(\Psi_{\theta}|Z)$, $U_I(Z) = E(L_{I\theta}|Z)/\Omega(Z)$ and $\theta_{\beta_I}(Z) = -E(\Psi_{\beta_I}|Z)/\Omega(Z)$. Note that the dimensions of $\beta_1$ and $\beta$ here, $d_1$ and $d$, are fixed even when the sample size $n$ increases.

**Theorem 5** Under the regularity conditions (C1)-(C6) in Section 5, if $\lambda_n \to 0$ when $n \to \infty$, then with probability tending to one, there exists a root of (11), denoted $\hat{\beta}$, such that $||\hat{\beta} - \beta_0|| = O_p(n^{-1/2} + a_n)$.

Further define

$$A = E \left[ L_{I\beta_I}\{W, Z, S, Y, \beta_0, \theta_0(Z)\} + L_{I\theta}\{W, Z, S, Y, \beta_0, \theta_0(Z)\}\theta^T_{\beta_I}(Z, \beta_0) \right] ,$$

$$B = \text{cov} \{ L_I\{W, Z, S, Y, \beta_0, \theta_0(Z)\} - \Psi\{W, Z, S, Y, \beta_0, \theta_0(Z)\}\}U_I(Z)\},$$

we obtain the following result.

**Theorem 6** Under the regularity conditions (C1)-(C6) in Section 5, if $\lambda_n \to 0$ as $n \to \infty$, and (8) holds, then with probability tending to one, any root-n consistent solution $\hat{\beta} = (\hat{\beta}_I^T, \hat{\beta}_{II}^T)^T$ of (11) must satisfy that

(i) $\hat{\beta}_{II} = 0$,

(ii) $\hat{\beta}_I$ has the following asymptotic normality,

$$\sqrt{n}\left\{ \hat{\beta}_I - \beta_{I0} - (A - \Sigma)^{-1}b \right\} \overset{D}{\to} N\left\{ 0, (A - \Sigma)^{-1}B(A - \Sigma)^{-T}\right\} .$$

For some penalty functions, including the SCAD penalty, $b$ and $\Sigma$ are zero when $\lambda_n$ is sufficiently small, so we have $\sqrt{n}(\hat{\beta}_I - \beta_{I0}) \overset{D}{\to} N(0, A^{-1}BA^{-T})$. These are exactly the same first order asymptotic properties of the locally efficient estimator for such semiparametric model in the case when only the first $d_1$ parameters are included in the model, that is, the procedure possesses the oracle property.

Next, in the following two theorems, we present the asymptotic property of the estimator for increasing dimensions $d_n$ and $d_{n1}$.

**Theorem 7** Under the regularity conditions (D1)-(D7) in Section 5, and if $d_{n1}n^{-1} \to 0$, $\lambda_n \to 0$ when $n \to \infty$, then with probability tending to one, there exists a root of (11), denoted $\hat{\beta}_n$, such that $||\hat{\beta}_n - \beta_{n0}|| = O_p\{\sqrt{d_{n1}}(n^{-1/2} + a_n)\}$. 

---

Y. Ma and R. Li
Theorem 8 Under regularity conditions (D1)-(D7) in Section 5, if \( \lambda_n \to 0, d_n^5/n \to 0 \), and (9) holds, then with probability tending to one, any root-\( n \) consistent estimator \( \hat{\beta}_n = (\hat{\beta}_I, \hat{\beta}_I I)^T \) obtained in (11) must satisfy that

(i) \( \hat{\beta}_I I = 0 \),

(ii) for any \( d_{n1} \times 1 \) vector \( v \) such that \( v^T v = 1 \),

\[
\sqrt{n/d_n} v^T B^{-1/2}(A - \Sigma) \left\{ \hat{\beta}_I - \beta_{I0} - (A - \Sigma)^{-1}b \right\} \xrightarrow{D} N(0, 1).
\]

4 Numerical studies and application

In this section, we assess the finite sample performance of the proposed procedure by Monte Carlo simulation, and illustrate the proposed methodology by an empirical analysis of the Framingham heart study data. In our simulation, we concentrate on the performance of the proposed procedure for a quadratic logistic measurement error model and partially linear logistic measurement error model in terms of model complexity and model error.

4.1 Tuning parameter selection

To implement the Newton-Raphson algorithm to solve the penalized estimating equations, we locally approximate the first order derivative of the penalty function by a linear function, following the idea of local quadratic approximation algorithm proposed in Fan and Li (2001). Specifically, suppose that at the \( k \)th step in the course of the iteration, we obtain the value \( \beta^{(k)} \). Then, for \( \beta^{(k+1)}_j \) not very close to zero,

\[
p_\lambda'(\beta_j) = p_\lambda'(|\beta_j|) \text{sign}(\beta_j) \approx \frac{p_\lambda'(|\beta_j^{(k)}|)}{|\beta_j^{(k)}|} \beta_j,
\]

otherwise, we set \( \beta_j^{(k+1)} = 0 \), and exclude the corresponding covariate from the model. This approximation is updated in every step in the course of the Newton-Raphson algorithm iteration. Following the results in Theorem 2 and 6, we can further approximate the estimation variance of the resulting estimate. That is

\[
\hat{\text{cov}}(\hat{\beta}) = \frac{1}{n} (E - \Sigma_{\lambda})^{-1} F(E - \Sigma_{\lambda})^{-T},
\]

where \( \Sigma_{\lambda} \) is a diagonal matrix with elements equal to \( p_\lambda'(|\hat{\beta}_j|)/|\hat{\beta}_j| \) for nonvanishing \( \hat{\beta}_j \), a linear approximation of \( \Sigma \) defined in (7). We use \( E \) to denote the sample approximation of \( E\partial S_{\text{eff},I}^*(W, Z, Y, \beta_I)/\partial \beta_I \) evaluated at \( \hat{\beta} \) for the parametric model (1) and the sample approximation of the matrix \( A \) evaluated at \( \hat{\beta} \) for the semiparametric model (10). Similarly, we use \( F \) to denote the sample approximation of
var($S^{*}_{eff,i}$) evaluated at $\hat{\beta}$ for the parametric model and the sample approximation of the matrix $B$ evaluated at $\hat{\beta}$ for the semiparametric model, respectively. The accuracy of this sandwich formula will be tested in our simulation studies.

It is desirable to have automatic, data-driven method to select the tuning parameter $\lambda$. Here we will consider two tuning parameter selectors, the GCV and BIC tuning parameter selectors. To define the GCV statistic and the BIC statistic, we need to define the degrees of freedom and goodness of fit measure for the final selected model. Similar to the nonconcave penalty likelihood approach (Fan and Li, 2001), we may define effective number of parameter or degrees of freedom to be

$$df_\lambda = \text{trace}\{I(I + \Sigma_\lambda)^{-1}\},$$

where $I$ stands for the Fisher Information matrix. For the logistic regression models that we employ in this section, a natural approximation of $I$, ignoring the measurement error effect, is $C^TQC$, where $C$ represents the covariates included in the model, and $Q$ is a diagonal matrix with the $i$th element equal to $\hat{\mu}_{\lambda,i}(1 - \hat{\mu}_{\lambda,i})$. Here, $\hat{\mu}_{\lambda,i} = P(Y_i = 1|C_i)$.

In the logistic regression model context of this section, we may employ its deviance as goodness of fit measure. Specifically, let $\mu_i$ be the conditional expectation of $Y_i$ given its covariates for $i = 1, \cdots, n$. The deviance of a model fit $\hat{\mu}_\lambda = (\hat{\mu}_{\lambda,1}, \cdots, \hat{\mu}_{\lambda,n})$ is defined to be

$$D(\hat{\mu}_\lambda) = 2 \sum_{i=1}^n \left[ Y_i \log(Y_i/\hat{\mu}_{\lambda,i}) + (1 - Y_i) \log((1 - Y_i)/(1 - \hat{\mu}_{\lambda,i})) \right].$$

Define GCV statistic to be

$$GCV(\lambda) = \frac{D(\hat{\mu}_\lambda)}{n(1 - df_\lambda/n)^2},$$

and the BIC statistic to be

$$BIC(\lambda) = D(\hat{\mu}_\lambda) + 2\log(n)df_\lambda.$$
4.2 Model error

Let us simplify the model error for logistic partially linear measurement error model. Denote 
\( \mu(S, X, Z) = E(Y|S, X, Z) \), and model error for a model \( \hat{\mu}(S, X, Z) \) is defined

\[
ME(\hat{\mu}) = E\{\hat{\mu}(S^*, X^*, Z^*) - \mu(S^*, X^*, Z^*)\}^2,
\]

where the expectation is taken over the new observation \( S^*, X^* \) and \( Z^* \). Let \( g(\cdot) \) be the logit link. For logistic partially linear model, the mean function has the form \( \mu(S, X, Z) = g^{-1}\{\theta(Z) + \beta^T V\} \), where \( V = (S^T, X^T)^T \). If \( \hat{\theta}(\cdot) \) and \( \hat{\beta} \) are consistent estimator for \( \theta(\cdot) \) and \( \beta \), respectively, then by a Taylor expansion, the model error can be approximated by

\[
ME(\hat{\mu}) \approx E\left[g^{-1}\{\theta(Z^*) + \beta^T V^*\}^2(\hat{\beta}(Z^*) - \theta(Z^*))^2 + \beta^T V^* - \beta^T V^*\right].
\]

The first component is the inherent model error due to \( \hat{\theta}(\cdot) \), the second one is due to lack of fit of \( \hat{\beta} \), and the third one is the cross-product between the first two components. Thus, to assess the performance of the proposed variable selection procedure, we define the approximate model error (AME) for \( \hat{\beta} \) to be

\[
AME(\hat{\beta}) = E\left[g^{-1}\{\theta(Z^*) + \beta^T V^*\}^2(\hat{\beta}(Z^*) - \theta(Z^*))^2 + \beta^T V^* - \beta^T V^*\right].
\]

Furthermore, the AME of \( \hat{\beta} \) can be written as

\[
AME(\hat{\beta}) = (\hat{\beta} - \beta)^T C_X (\hat{\beta} - \beta).
\]  

In our simulation, the matrix \( C_X \) is estimated by 1,000,000 Monte Carlo simulations. For measurement error data, we observe \( W \) instead of \( X \). We also consider an alternative approximate model error

\[
AME_W(\hat{\beta}) = (\hat{\beta} - \beta)^T C_W (\hat{\beta} - \beta),
\]

where \( C_W \) is obtained by replacing \( X \) with \( W \) in the definition of \( C_X \). The \( AME(\hat{\beta}) \) and \( AME_W(\hat{\beta}) \) are defined for parametric model case, specifically the quadratic logistic model, by setting \( \theta(\cdot) = 0 \) and append \( X^2 \) (or \( W^2 \)) in \( V^* \). Note that although we defined the model error in the context with a logistic link function, it is certainly not restricted to such case. The general approach for calculating \( AME \) is to approximate the probability density function evaluated at the estimated parameters around the true parameter value, and extract the linear term of the parameter of interest. \( AME_W \) is calculated through replacing \( X \) with \( W \).
4.3 Simulation examples

To demonstrate the performance of the method in both parametric and semiparametric measurement error models, we conduct two simulation studies.

**Example 1.** In this example, we generate data from a logistic model where the covariate measured with error enters the model through a quadratic function, and the covariates measured without error enter linearly. The measurement error follows a normal additive pattern. Specifically, we have

\[
\text{logit}\left\{p(Y = 1|X, Z)\right\} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 Z_1 + \beta_4 Z_2 + \beta_5 Z_3 + \beta_6 Z_4 + \beta_7 Z_5 + \beta_8 Z_6 + \beta_9 Z_7,
\]

and

\[W = X + U,\]

where \(\beta = (0, 1.5, 2, 0, 3, 0, 1.5, 0, 0, 0)^T\), the covariate \(X\) is generated from a normal distribution with mean 0 and variance 1, \((Z_1, \ldots, Z_6)^T\) is generated from a normal distribution with mean 0 and covariance between \(Z_i\) and \(Z_j\) is \(0.5^{|i-j|}\). The last component of the \(Z\) covariates, \(Z_7\), is a binary variable with equal probability to be 0 or 1. \(U \sim N(0, 0.1^2)\). In our simulation, the sample size is taken to be either \(n = 500\) or \(n = 1000\).

The model complexity of the selected model is summarized in terms of the number of zero coefficients, and the model error of the selected model is summarized in terms of relative approximation model error (RAME), defined to be the ratio of model error of the selected model to that of the full model. In Table 1, the RAME column corresponds to the sample median and median absolute deviation (MAD) divided by a factor of 0.6745 of the AME values defined in (13) over 1000 simulations. Similarly, \(\text{RAME}_W\) corresponds to that of the AME\(_W\) values defined in (14) over 1000 simulations. From Table 1, it can be seen that the values of RAME and \(\text{RAME}_W\) are very close. The average count of zero coefficients is also reported in Table 1, in which the column labeled “C” presents the average count restricted only to the true zero coefficients, while the column labeled “E” displays the average count of the coefficients erroneously set to 0.

<table>
<thead>
<tr>
<th>(n)</th>
<th>GCV</th>
<th>BIC</th>
<th>500</th>
<th>1000</th>
<th>500</th>
<th>1000</th>
<th>4.574</th>
<th>0.074</th>
<th>5.857</th>
<th>0.010</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RAME</td>
<td>RAME(_W)</td>
<td>Median(MAD)</td>
<td>Median(MAD)</td>
<td># of Zero Coefficients</td>
<td>C</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCV</td>
<td>0.694(0.231)</td>
<td>0.698(0.228)</td>
<td>4.574</td>
<td>0.006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>0.396(0.188)</td>
<td>0.396(0.187)</td>
<td>5.857</td>
<td>0.074</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCV</td>
<td>0.766(0.187)</td>
<td>0.770(0.185)</td>
<td>4.456</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>0.390(0.157)</td>
<td>0.401(0.158)</td>
<td>5.758</td>
<td>0.010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We next verify the consistency of the estimators and test the accuracy of the proposed standard error formula. Table 2 displays the bias and sample standard deviation (SD) of the estimates for two nonzero coefficients, \((\hat{\beta}_1, \hat{\beta}_2)\), over 1000 simulations, and the sample average and the sample standard deviations of the 1000 standard errors obtained by using the sandwich formula. The row labeled ‘EE’ corresponds to the unpenalized estimating equation estimator. Overall, the sandwich formula works well. It is worth pointing out that a variable selection procedures effectively reduces the finite sample estimation bias and variance, even though asymptotically, the estimator without variable selection is also consistent.

**Example 2.** In this example, we illustrate the performance of the method for a semiparametric model.

<table>
<thead>
<tr>
<th></th>
<th>(\hat{\beta}_1)</th>
<th></th>
<th>(\hat{\beta}_2)</th>
<th></th>
<th>(\hat{\beta}_4)</th>
<th></th>
<th>(\hat{\beta}_6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bias(SD)</td>
<td>SE(Std(SE))</td>
<td>bias(SD)</td>
<td>SE(Std(SE))</td>
<td>bias(SD)</td>
<td>SE(Std(SE))</td>
<td>bias(SD)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 500)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EE</td>
<td>0.212 (0.609)</td>
<td>0.463 (0.418)</td>
<td>0.336 (0.922)</td>
<td>0.585 (0.647)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCV</td>
<td>0.097 (0.426)</td>
<td>0.382 (0.159)</td>
<td>0.081 (0.418)</td>
<td>0.388 (0.178)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>0.014 (0.593)</td>
<td>0.358 (0.189)</td>
<td>0.092 (0.462)</td>
<td>0.367 (0.176)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 1000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EE</td>
<td>0.072 (0.273)</td>
<td>0.268 (0.062)</td>
<td>0.124 (0.332)</td>
<td>0.321 (0.088)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCV</td>
<td>0.029 (0.254)</td>
<td>0.250 (0.048)</td>
<td>0.009 (0.258)</td>
<td>0.253 (0.057)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>0.024 (0.290)</td>
<td>0.249 (0.054)</td>
<td>0.052 (0.255)</td>
<td>0.244 (0.052)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 500)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EE</td>
<td>0.297 (0.618)</td>
<td>0.504 (0.397)</td>
<td>0.164 (0.465)</td>
<td>0.332 (0.256)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCV</td>
<td>0.164 (0.453)</td>
<td>0.403 (0.105)</td>
<td>0.092 (0.302)</td>
<td>0.264 (0.066)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>0.100 (0.462)</td>
<td>0.386 (0.107)</td>
<td>0.057 (0.286)</td>
<td>0.249 (0.065)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 1000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EE</td>
<td>0.128 (0.314)</td>
<td>0.295 (0.046)</td>
<td>0.051 (0.207)</td>
<td>0.203 (0.026)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCV</td>
<td>0.073 (0.282)</td>
<td>0.267 (0.038)</td>
<td>0.028 (0.184)</td>
<td>0.177 (0.022)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>0.055 (0.275)</td>
<td>0.262 (0.037)</td>
<td>0.014 (0.166)</td>
<td>0.170 (0.020)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
measurement error model. Simulation data are generated from

\[
\text{logit}(Y) = \beta_1 X + \beta_2 S_1 + \cdots + \beta_{10} S_9 + \theta(Z)
\]

\[W = X + U\]

where \(\beta, X \) and \(W\) are the same as in the previous simulation. We generate \(S\)'s in a similar fashion as the \(Z\)'s in simulation 1. That is, \((S_1, \ldots, S_8)\) is generated from normal distribution with mean zero and covariance between \(S_i\) and \(S_j\) is \(0.5^{|i-j|}\). \(S_9\) is a binary variable with equal probability to be zero or one. The random variable \(Z\) is generated from a uniform distribution in \([-\pi/2, \pi/2]\). The true function \(\theta(Z) = 0.5 \cos(Z)\). The parameter takes values \(\beta = (1.5, 2.0, 0, 3.0, 1.5, 0, 0, 0)\).

The simulation results are summarized in Table 3, in which the notation is the same as that in Table 1. From Table 3, we can see that the penalized estimating equation estimators can significantly reduce model complexity. Overall, the BIC tuning parameter selectors performs better, while GCV is too conservative. We have further tested the consistency and the accuracy of the standard error formula derived from the sandwich formula. The result is summarized in Table 4, in which the notation is the same as that in table 2, and from which we can see the consistency of the estimator and that the standard error formula performs rather well when \(n = 1000\).

<table>
<thead>
<tr>
<th>Method</th>
<th>n</th>
<th>RME_(X) Median(MAD)</th>
<th>RME_(W) Median(MAD)</th>
<th># of Zero Coefficients</th>
<th>C</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCV</td>
<td>500</td>
<td>0.878(0.161)</td>
<td>0.880(0.158)</td>
<td>4.060</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>500</td>
<td>0.381(0.158)</td>
<td>0.387(0.155)</td>
<td>5.713</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>GCV</td>
<td>1000</td>
<td>0.868(0.164)</td>
<td>0.873(0.160)</td>
<td>4.061</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>1000</td>
<td>0.386(0.162)</td>
<td>0.392(0.161)</td>
<td>5.694</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

4.4 Data Analysis

Framingham heart study data set (Kannel et al., 1986) is a well known data set where it is generally accepted that there exists measurement error on the long term systolic blood pressure (SBP). In addition to SBP, other measurements include age, smoking status and serum cholesterol. In literature, there has been speculation that a second order term involving age might be needed in analyzing the dependence of the occurrence of heart disease. In addition, it is unclear if the interaction (product) between the various covariates plays a role in influencing the heart disease rate. The data set includes 1615 observations.
### Table 4: Bias and Standard Errors for Example 2

<table>
<thead>
<tr>
<th></th>
<th>Bias(SD)</th>
<th>SE(Std(SE))</th>
<th>Bias(SD)</th>
<th>SE(Std(SE))</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_1$</td>
<td></td>
<td></td>
<td>$\hat{\beta}_2$</td>
<td></td>
</tr>
<tr>
<td>$n = 500$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EE</td>
<td>0.101 (0.267)</td>
<td>0.247 (0.043)</td>
<td>0.113 (0.295)</td>
<td>0.281 (0.045)</td>
</tr>
<tr>
<td>GCV</td>
<td>0.112 (0.275)</td>
<td>0.259 (0.047)</td>
<td>0.129 (0.297)</td>
<td>0.285 (0.050)</td>
</tr>
<tr>
<td>BIC</td>
<td>0.078 (0.262)</td>
<td>0.252 (0.044)</td>
<td>0.085 (0.273)</td>
<td>0.272 (0.045)</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.039 (0.170)</td>
<td>0.166 (0.018)</td>
<td>0.057 (0.194)</td>
<td>0.190 (0.018)</td>
</tr>
<tr>
<td>EE</td>
<td>0.047 (0.174)</td>
<td>0.172 (0.020)</td>
<td>0.069 (0.196)</td>
<td>0.191 (0.021)</td>
</tr>
<tr>
<td>GCV</td>
<td>0.031 (0.169)</td>
<td>0.170 (0.019)</td>
<td>0.044 (0.179)</td>
<td>0.185 (0.019)</td>
</tr>
<tr>
<td>$n = 500$</td>
<td></td>
<td></td>
<td>$\hat{\beta}_5$</td>
<td></td>
</tr>
<tr>
<td>EE</td>
<td>0.179 (0.415)</td>
<td>0.376 (0.063)</td>
<td>0.091 (0.288)</td>
<td>0.266 (0.038)</td>
</tr>
<tr>
<td>GCV</td>
<td>0.203 (0.418)</td>
<td>0.384 (0.071)</td>
<td>0.101 (0.285)</td>
<td>0.255 (0.044)</td>
</tr>
<tr>
<td>BIC</td>
<td>0.139 (0.378)</td>
<td>0.365 (0.064)</td>
<td>0.068 (0.249)</td>
<td>0.238 (0.037)</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>0.074 (0.258)</td>
<td>0.251 (0.026)</td>
<td>0.046 (0.181)</td>
<td>0.180 (0.016)</td>
</tr>
<tr>
<td>EE</td>
<td>0.091 (0.265)</td>
<td>0.254 (0.031)</td>
<td>0.056 (0.181)</td>
<td>0.171 (0.020)</td>
</tr>
<tr>
<td>BIC</td>
<td>0.056 (0.242)</td>
<td>0.245 (0.028)</td>
<td>0.038 (0.159)</td>
<td>0.161 (0.016)</td>
</tr>
</tbody>
</table>

With the method developed here, it is possible to perform a variable selection to address these issues. Following the literature, we adopt the measurement error model of $\log(\text{MSBP} - 50) = \log(\text{SBP} - 50) + U$, where $U$ is a mean zero normal random variable with variance $\sigma^2_u = 0.0126$, MSBP is the measured SBP. We denote the standardized $\log(\text{MSBP} - 50)$ as $W$. The standardization using the same parameters on $\log(\text{SBP} - 50)$ is denoted $X$. The standardized serum cholesterol, age are denoted $Z_1, Z_2$, and we use $Z_3$ to denote the binary variable smoking status. Using $Y$ to denote the occurrence of heart disease, the saturated model which includes all the interaction terms and
also the square of age term is of the form

\[
\logit\{p(Y = 1|X, Z's)\} = \beta_1 X + \beta_2 XZ_1 + \beta_3 XZ_2 + \beta_4 XZ_3 + \beta_5 + \beta_6 Z_1 + \beta_7 Z_2 + \beta_8 Z_3 + \beta_9 Z_2^2 + \beta_{10} Z_1Z_2 + \beta_{11} Z_1Z_3 + \beta_{12} Z_2Z_3
\]

\[W = X + U.\]

Figure 1: Tuning parameters and their corresponding BIC and GCV scores for the Framingham data. The scores are normalized to the range [0, 1].

We first used both GCV and BIC tuning parameter selectors to choose \(\lambda\). We present the tuning parameters and the corresponding GCV and BIC scores in Figure 1. The final chosen \(\lambda\) is 0.073 and 0.172 by the GCV and BIC selectors, respectively. The selected model is depicted in Table 5. The GCV criterion selects the covariates \(X, XZ_1, 1, Z_1, Z_2, Z_3, Z_2^2, Z_2Z_3\) into the model, while the BIC criterion selects the covariates \(X, 1, Z_1, Z_2\) into the model. We report the selection and estimation results in Table 5, as well as the semiparametric estimation results without variable selection.

As we can see, the terms \(X, 1, Z_1, Z_2\) are selected by both criteria, while \(Z_3, Z_2^2\) and some of the interaction terms are selected by GCV. The BIC criterion is very aggressive and it results in a very simple final model while the GCV criterion is much more conservative, hence the resulting model is more complex. This agrees with the simulation results we have obtained. Since both criteria have included the covariate \(X\), we see that the measurement error feature and its treatment in the Framingham data is unavoidable.
Table 5: Results for the Framingham data set.

<table>
<thead>
<tr>
<th></th>
<th>EE</th>
<th>GCV</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}$ (SE)</td>
<td>$\hat{\beta}$ (SE)</td>
<td>$\hat{\beta}$ (SE)</td>
</tr>
<tr>
<td>$X$</td>
<td>0.643(0.248)</td>
<td>0.416(0.093)</td>
<td>0.179(0.039)</td>
</tr>
<tr>
<td>$XZ_1$</td>
<td>-0.167(0.097)</td>
<td>-0.072(0.041)</td>
<td>0 (NA)</td>
</tr>
<tr>
<td>$XZ_2$</td>
<td>-0.059(0.111)</td>
<td>0 (NA)</td>
<td>0 (NA)</td>
</tr>
<tr>
<td>$XZ_3$</td>
<td>-0.214(0.249)</td>
<td>0 (NA)</td>
<td>0 (NA)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.415(0.428)</td>
<td>-3.255(0.356)</td>
<td>-2.555(0.092)</td>
</tr>
<tr>
<td>$Z_1$</td>
<td>0.516(0.212)</td>
<td>0.332(0.085)</td>
<td>0.124(0.033)</td>
</tr>
<tr>
<td>$Z_2$</td>
<td>1.048(0.341)</td>
<td>1.044(0.329)</td>
<td>0.398(0.067)</td>
</tr>
<tr>
<td>$Z_3$</td>
<td>1.060(0.443)</td>
<td>0.907(0.373)</td>
<td>0 (NA)</td>
</tr>
<tr>
<td>$Z_1^2$</td>
<td>-0.253(0.125)</td>
<td>-0.262(0.121)</td>
<td>0 (NA)</td>
</tr>
<tr>
<td>$Z_1Z_2$</td>
<td>-0.072(0.103)</td>
<td>0 (NA)</td>
<td>0 (NA)</td>
</tr>
<tr>
<td>$Z_1Z_3$</td>
<td>-0.161(0.225)</td>
<td>0 (NA)</td>
<td>0 (NA)</td>
</tr>
<tr>
<td>$Z_2Z_3$</td>
<td>-0.442(0.336)</td>
<td>-0.473(0.326)</td>
<td>0 (NA)</td>
</tr>
</tbody>
</table>

5 Proofs

List of Regularity Conditions for Theorems 1 and 2

(A1) The model is identifiable.

(A2) The expectation of the first derivative of $S_{eff}^*$ with respect to $\beta$ exists and is not singular at $\beta_0$.

(A3) The second derivative of $S_{eff}^*$ with respect to $\beta$ exists and is continuous and bounded in a neighborhood of $\beta_0$.

(A4) The variance-covariance matrix of the first $d_1$ components of $S_{eff}^*$, $S_{eff,i}^*$, is positive definite at $\beta_0$.

Proof of Theorem 1

From condition (A2), we can denote

$$J = \left\{ E \left( \frac{\partial S_{eff}^*}{\partial \beta^T} | \beta_0 \right) \right\}^{-1}, \quad \phi_{eff}^* = JS_{eff}^* \quad \text{and} \quad q'_{\lambda_n}(\beta) = Jp'_{\lambda_n}(\beta).$$

Let $\alpha_n = n^{-1/2} + a_n$, $\phi_{eff,i}^*(\beta) = \phi_{eff}^*(W_i, Z_i, Y_i, \beta)$. It is sufficient to show that

$$n^{-1/2} \sum_{i=1}^{n} \phi_{eff,i}^*(\beta) - n^{1/2}q'_{\lambda_n}(\beta) = 0 \quad (15)$$
has a solution \( \hat{\beta} \) that satisfies \( ||\hat{\beta} - \beta_0|| = O_p(\alpha_n) \). Consider

\[
(\beta - \beta_0)^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi_{\text{eff},i}^*(\beta) - n^{1/2} q_{\lambda_n}'(\beta) \right\}.
\]

For any \( \beta \) such that \( ||\beta - \beta_0|| = C\alpha_n \) for some constant \( C \), because of condition (A3) and \( a_n \rightarrow 0 \), it follows by the Taylor expansion that

\[
\text{The first term in the above display is of order } O_p.C\sqrt{n}\alpha_n^2, \text{ which dominates the first term as long as } C \text{ is large enough. The remaining terms are dominated by the first two terms. Thus, for any } \epsilon > 0, \text{ as long as } C \text{ is large enough, the probability for the above display to be larger than zero is at least } 1 - \epsilon. \text{ From Brouwer fixed-point theorem, we know there with probability at least } 1 - \epsilon, \text{ there exists at least one solution for (15) in the region } ||\beta - \beta_0|| \leq C\alpha_n. \]

To show Theorem 2, we first prove the following lemma.

**Lemma 1** If conditions in Theorem 2 hold, then with probability tending to 1, for any given \( \beta \) that satisfies \( ||\beta - \beta_0|| = O_p(n^{-1/2}) \), \( \beta_{11} = 0 \) is a solution to the last \( d_2 \) equations of (2).

Proof: Denote the \( k \)th equation in \( \sum_{i=1}^{n} S_{\text{eff}}(W_i, Z_i, Y_i, \beta) \) as \( L_k(\beta), k = d_1 + 1, \ldots, d \), then

\[
L_k(\beta) - np_{\lambda_n}'(\beta_k) = L_k(\beta_0) + \frac{1}{2} \sum_{j=1}^{d} \sum_{l=k+1}^{d} \frac{\partial^2 L_k(\beta^*)}{\partial \beta_l \partial \beta_j} (\beta_l - \beta_0)(\beta_j - \beta_0) - np_{\lambda_n,k}'(\|\beta_k\|) \text{sign}(\beta_k).
\]

The first three terms of the above display are all of order at most \( O_p(n^{1/2}) \), hence we have

\[
L_k(\beta) - np_{\lambda_n}'(\beta_k) = -\sqrt{n} \{ \sqrt{n} p_{\lambda_n}'(\|\beta_k\|) \text{sign}(\beta_k) + O_p(1) \}.
\]
Using (8), the sign of \( L_k(\beta) - n \eta'_{\lambda_n}(\beta_k) \) is decided by \( \text{sign}(\beta_k) \) completely when \( n \) is large enough. From the continuity of \( L_k(\beta) - n \eta'_{\lambda_n}(\beta_k) \), we obtain that it is zero at \( \beta_k = 0 \). \qed

**Proof of Theorem 2** We let \( \lambda_n \) be sufficiently small so that \( a_n = o(n^{-1/2}) \). From Lemma 1, there is a root-\( n \) consistent estimator \( \hat{\beta} \). From Lemma 1, \( \hat{\beta} = (\hat{\beta}_1^T, 0^T)^T \), so (i) is shown. Denote the first \( d_1 \) equations in \( \sum_{i=1}^n S_{\epsilon_{\theta}}^* \{ W_i, Z_i, Y_i, (\hat{\beta}_1^T, 0^T)^T \} \) as \( L(\beta_1) \), then

\[
0 = L(\hat{\beta}_1) - n \eta'_{\lambda_n,1}(\hat{\beta}_1) \\
= L(\beta_{10}) + \left\{ \frac{\partial L(\beta_{10})}{\partial \beta_{10}} + o_p(n) \right\} (\hat{\beta}_1 - \beta_{10}) - nb - n \{ \Sigma + o_p(1) \} (\hat{\beta}_1 - \beta_{10}) \\
= L(\beta_{10}) + n \left\{ \frac{\partial S_{\epsilon_{\theta}}^* (\beta_{10})}{\partial \beta_{10}} - \Sigma \right\}^{-1} \left\{ \hat{\beta}_1 - \beta_{10} - \left\{ \frac{\partial S_{\epsilon_{\theta}}^* (\beta_{10})}{\partial \beta_{10}} - \Sigma \right\}^{-1} b \right\} + o_p(n^{1/2}).
\]

We thus obtain

\[
\sqrt{n} \left[ \hat{\beta}_1 - \beta_{10} - \left\{ \frac{\partial S_{\epsilon_{\theta}}^* (\beta_{10})}{\partial \beta_{10}} - \Sigma \right\}^{-1} b \right] = -n^{-1/2} \left\{ \frac{\partial S_{\epsilon_{\theta}}^* (\beta_{10})}{\partial \beta_{10}} - \Sigma \right\}^{-1} L(\beta_{10}) + o_p(1).
\]

Because of condition (A4), the results in (ii) now follow. \qed

**List of Regularity Conditions for Theorems 3 and 4**

(B1) The model is identifiable.

(B2) The expectation of the first derivative of \( S_{\epsilon_{\theta}}^* \) with respect to \( \beta \) exists at \( \beta_0 \) and its left eigenvalues are bounded away from zero and infinity uniformly for all \( n \). For any entry \( S_{jk} \) in \( \partial S_{\epsilon_{\theta}}^* (\beta_0) / \partial \beta^T \), \( E(S_{jk}^2) < C_3 < \infty \).

(B3) The eigenvalues of the matrix \( E(S_{\epsilon_{\theta}}^* S_{\epsilon_{\theta}}^T) \), satisfy \( 0 < C_1 < \lambda_{\min} < \cdots < \lambda_{\max} < C_2 < \infty \) for all \( n \); For any entries, \( S_k, S_j \) in \( S_{\epsilon_{\theta}}^* (\beta_0) \), \( E(S_k^2 S_j^2) < C_4 < \infty \).

(B4) The second derivatives of \( S_{\epsilon_{\theta}}^* \) with respect to \( \beta \) exists and the entries are uniformly bounded by a function \( M(W_i, Z_i, Y_i) \) in a large enough neighborhood of \( \beta_0 \). In addition, \( E(M^2) < c_5 < \infty \) for all \( c_5 < \infty \).

(B5) \( \min_{1 \leq j \leq d_1} |\beta_{j0}| / \lambda_n \to \infty \) as \( n \to \infty \).

(B6) Let \( c_n = \max_{1 \leq j \leq d_1} \{ p''_{\lambda_j}(|\beta_{j0}|) : \beta_{j0} \neq 0 \} \). Assume that \( \lambda_n \to 0 \), \( a_n = O(n^{-1/2}) \) and \( c_n \to 0 \) as \( n \to \infty \). In addition, there exists constants \( C \) and \( D \) such that when \( \theta_1, \theta_2 \geq C \lambda \), \( \| p''_{\lambda} (\theta_1) - p''_{\lambda} (\theta_2) \| \leq D |\theta_1 - \theta_2| \).
To emphasize the dependence of $v$, $\lambda$, $d$, $d_1$, $d_2$, $S^*_\text{eff}$, $b$, $\Sigma$, $\beta$, $\beta_I$, $\beta_{II}$, $\beta_j$, for $j = 1, \ldots, d$, on $n$, we write them as $v_n$, $\lambda_n$, $d_n$, $d_{n1}$, $d_{n2}$, $S^*_\text{eff}_n$, $b_n$, $\Sigma_n$, $\beta_n$, $\beta_{nI}$, $\beta_{nII}$, $\beta_{nj}$ respectively throughout the proof.

**Proof of Theorem 3**

From condition (B2), we can denote

$$J_n = \left\{ E \left( \frac{\partial S^*_\text{eff}}{\partial \beta^*_n} \big| \beta_0 \right) \right\}^{-1}, \quad \phi^*_n\text{eff} = J_nS^*_n\text{eff} \quad \text{and} \quad q'_\lambda(\beta_n) = J_n p'_\lambda(\beta_n).$$

Let $\alpha_n = n^{-1/2} + a_n$, $\phi^*_n\text{eff,i}(\beta_n) = \phi^*_n\text{eff}(W_i, Z_i, Y_i, \beta_n)$. It suffices to show that

$$n^{-1/2} \sum_{i=1}^{n} \phi^*_n\text{eff,i}(\beta_n) - n^{1/2} q'_\lambda(\beta_n) = 0 \quad (16)$$

has a solution $\hat{\beta}_n$ that satisfies $||\hat{\beta}_n - \beta_{0n}|| = O_p(d_n^{1/2} \alpha_n)$. Consider

$$(\beta_n - \beta_{0n})^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi^*_n\text{eff,i}(\beta_n) - n^{1/2} q'_\lambda(\beta_n) \right\}.$$ 

For any $\beta_n$ such that $||\beta_n - \beta_{0n}|| = C \sqrt{\alpha_n}$ for some constant $C$, because of condition (B2), (B3), (B4) and (B6), we have the expansion

$$n^{-1/2} \sum_{i=1}^{n} \phi^*_n\text{eff,i}(\beta_n) - n^{1/2} q'_\lambda(\beta_n) = 0 + n^{1/2} \frac{\partial \phi^*_n\text{eff,i}}{\partial \beta_n} (\beta_n - \beta_{0n})$$

$$+ 2 n^{-1/2} \sum_{i=1}^{n} (\beta_n - \beta_{0n})^T \frac{\partial^2 \phi^*_n\text{eff,i}}{\partial \beta_n \partial \beta_n^T} (\beta_n - \beta_{0n}) - n^{1/2} q'_\lambda(\beta_{0n}) \{1 + o_p(1)\}$$

$$= (\beta_n - \beta_{0n})^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi^*_n\text{eff,i}(\beta_{0n}) - n^{1/2} q'_\lambda(\beta_{0n}) + n^{1/2}\{1 + o_p(1)\}(\beta_n - \beta_{0n}) + o_p(1) + n^{1/2} O(\alpha_n)(\beta_n - \beta_{0n}) \right\}$$

$$+ 2 n^{-1/2} \sum_{i=1}^{n} (\beta_n - \beta_{0n})^T \frac{\partial^2 \phi^*_n\text{eff,i}}{\partial \beta_n \partial \beta_n^T} (\beta_n - \beta_{0n}) - n^{1/2} q'_\lambda(\beta_{0n}) \{1 + o_p(1)\}$$

$$= (\beta_n - \beta_{0n})^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi^*_n\text{eff,i}(\beta_{0n}) - n^{1/2} q'_\lambda(\beta_{0n}) + n^{1/2}\{1 + o_p(1)\}(\beta_n - \beta_{0n}) + o_p(1) + n^{1/2} ||\beta_n - \beta_{0n}||^2 \right\}$$

$$+ o_p(n^{1/2} ||\beta_n - \beta_{0n}||^2).$$

In the above derivation, $\beta^*_n$ is in between $\beta_n$ and $\beta_{0n}$, and we used

$$||n^{-1/2} \sum_{i=1}^{n} (\beta_n - \beta_{0n})^T \frac{\partial^2 \phi^*_n\text{eff,i}}{\partial \beta_n \partial \beta_n^T} (\beta_n - \beta_{0n})|| = o_p(1).$$
which is shown in detail in (18), hence the proof is omitted here. The first term in the above display is bounded by

\[ ||\beta_n - \beta_0||_O \left\{ n^{-1/2} \sum_{i=1}^n \phi_{n,eff,i}^* (\beta_0) - n^{1/2} \varphi_{n\beta} (\beta_0) || \right\} \]

\[ = ||\beta_n - \beta_0||_O (d_n + d_n n_n^2)^{1/2} = O_p(C n^{1/2} d_n \alpha_n^2), \]

the second term equals \( C^2 n^{1/2} d_n \alpha_n^2 \), which dominates the first term with probability \( 1 - \epsilon \) for any \( \epsilon > 0 \) as long as \( C \) is large enough. The remaining terms are dominated by the first two terms. Thus, for any \( \epsilon > 0 \), for large enough \( C \), the probability for the above display to be larger than zero is at least \( 1 - \epsilon \). From Brouwer fixed-point theorem, we know that with probability at least \( 1 - \epsilon \), there exists at least one solution for (16) in the region \( ||\beta - \beta_0|| \leq C \sqrt{d_n} \alpha_n \).

We first prove the following lemma, then give the proof of Theorem 4.

**Lemma 2** If the conditions in Theorem 4 hold, then with probability tending to 1, for any given \( \beta_n \) that satisfies \( ||\beta_n - \beta_0|| = O_p(\sqrt{d_n / n}) \), \( \beta_{n,0} = 0 \) is a solution to the last \( d_{n,2} \) equations of (2).

**Proof:** Denote the \( k \)th equation in \( \sum_{i=1}^n S_{n,eff}^* (W_i, Z_i, Y_i, \beta_n) \) as \( L_{nk}(\beta_n) \), \( k = d_{n,1} + 1, \ldots, d_n \), then

\[ n \pi_{n,\beta} (\beta_n) - \frac{\partial L_{nk}(\beta_n)}{\partial \beta_{nj}} \]

\[ = L_{nk}(\beta_n) + \sum_{j=1}^{d_n} \frac{\partial L_{nk}(\beta_n)}{\partial \beta_{nj}} (\beta_{nj} - \beta_{n0}) + 2^{-1} \sum_{j=1}^{d_n} \sum_{j=1}^{d_n} \frac{\partial^2 L_{nk}(\beta_n)}{\partial \beta_{nl} \partial \beta_{nj}} (\beta_{nl} - \beta_{n0}) (\beta_{nj} - \beta_{n0}) \]

\[ - np_{nk} (\beta_n) \text{sign}(\beta_n), \]  

(17)

where \( \beta_n^* \) is in between \( \hat{\beta}_n \) and \( \beta_0 \). Because of condition (B3), the first term of (17) is of order \( O_p(n^{1/2}) = o_p(n^{1/2} d_n^{-1/2}) \). The second term (17) can be further written as

\[ \sum_{j=1}^{d_n} \left\{ \frac{\partial L_{nk}(\beta_n)}{\partial \beta_{nj}} - E \frac{\partial L_{nk}(\beta_n)}{\partial \beta_{nj}} \right\} (\beta_{nj} - \beta_{n0}) + \sum_{j=1}^{d_n} E \frac{\partial L_{nk}(\beta_n)}{\partial \beta_{nj}} (\beta_{nj} - \beta_{n0}), \]

where the first term is controlled by

\[ \left[ \sum_{j=1}^{d_n} \left\{ \frac{\partial L_{nk}(\beta_n)}{\partial \beta_{nj}} - E \frac{\partial L_{nk}(\beta_0)}{\partial \beta_{nj}} \right\} \right]^{1/2} \]

\[ = O(d_n^{1/2} n^{-1/2}) O_p \left\{ \text{var} \left( \frac{\partial S_{n,eff,k}^*}{\partial \beta_{n1}} \right) \right\} O_p(d_n^{-1/2} n^1/2) \]

\[ \leq O(d_n^{1/2} n^{-1/2}) O_p \left\{ E(S_{k1}^2) \right\} O_p(d_n^{1/2} n^{-1/2}) = o_p(n^{1/2} d_n^{1/2}), \]
due to condition (B2), and the second term is controlled by

$$n \left\{ \frac{d_n}{\sqrt{2}} \left( E \frac{\partial S_{n,eff,k}}{\partial \beta_n^2} \right)^2 \right\}^{1/2} ||\beta_n - \beta_{n0}|| \leq n\lambda_{\text{max}} \left\{ E \frac{\partial S_{n,eff,k}}{\partial \beta_n^2} \right\}^{2} ||\beta_n - \beta_{n0}|| = O_p(n^{1/2} d_n^{1/2}).$$

The third term of (17) can be further written as

$$\sum_{l=1}^{d_n} \sum_{j=1}^{d_n} \left( E \frac{\partial^2 L_{nk}(\beta_n^*)}{\partial \beta_n \partial \beta_{n,0}} \right) (\beta_{nl} - \beta_{n0})(\beta_{n,j} - \beta_{n0}) + \sum_{l=1}^{d_n} \sum_{j=1}^{d_n} \left( \frac{\partial^2 L_{nk}(\beta_n^*)}{\partial \beta_n \partial \beta_{n,0}} \right) - E \left( \frac{\partial^2 L_{nk}(\beta_n^*)}{\partial \beta_n \partial \beta_{n,0}} \right) (\beta_{nl} - \beta_{n0})(\beta_{n,j} - \beta_{n0})\right),$$

where the first term is bounded by

$$n \left( \sum_{j,l=1}^{d_n} \left( E \left( \frac{\partial^2 L_{nk}(\beta_n^*)}{\partial \beta_n \partial \beta_{n,0}} \right) \right)^2 \right)^{1/2} ||\beta_n - \beta_{n0}||^2 = \left( \sum_{j,l=1}^{d_n} nE \left( \frac{\partial^2 S_{n,eff,k}(\beta_n^*)}{\partial \beta_n \partial \beta_{n,0}} \right) \right)^{2} O_p(n^{-1} d_n)$$

due to condition (B4), and the second term is bounded by

$$\left( \sum_{j,l=1}^{d_n} \left[ \frac{\partial^2 L_{nk}(\beta_n^*)}{\partial \beta_n \partial \beta_{n,0}} - E \left( \frac{\partial^2 L_{nk}(\beta_n^*)}{\partial \beta_n \partial \beta_{n,0}} \right) \right] \right)^{2} ||\beta_n - \beta_{n0}||^2 \leq \left( \sum_{j,l=1}^{d_n} O_p \left( \frac{\partial^2 S_{n,eff,k}(\beta_n^*)}{\partial \beta_n \partial \beta_{n,0}} \right) \right)^{1/2} O_p(n^{-1} d_n)$$

$$\leq \left( \sum_{j,l=1}^{d_n} O_p \left( E \left( \frac{\partial^2 S_{n,eff,k}(\beta_n^*)}{\partial \beta_n \partial \beta_{n,0}} \right) \right)^2 \right)^{1/2} O_p(n^{-1/2} d_n) \leq \left( \sum_{j,l=1}^{d_n} O_p \left( E(M^2) \right) \right)^{1/2} O_p(n^{-1/2} d_n) = O_p(n^{-1/2} d_n^2) = O_p(n^{1/2} d_n^{1/2}),$$

due to condition (B4) and (B5). Hence we have

$$L_{nk}(\beta_n) - np_{\lambda_n,k}^*(\beta_n) = -\sqrt{n} \{ \sqrt{n/d_n} p_{\lambda_n,k}^*(\beta_n) \text{sign}(\beta_n) + O_p(1) \}.$$
Proof of Theorem 4: From Theorem 3 and condition (B6), there is a root-\((n/d_n)\) consistent estimator \(\hat{\beta}_n\). From Lemma 2, \(\hat{\beta}_n = (\hat{\beta}_n^T, 0^T)^T\), so (i) is shown. Denote the first \(d_{n1}\) equations in \(\sum_{i=1}^n S^*_n,eff (W_i, Z_i, Y_i, (\beta_{nI}^*, 0^T)^T)\) as \(L_n(\beta_{nI})\). From Lemma 2, if we select \(\beta_{nI} - \beta_{nI0} = O_p(d_n^{1/2} n^{-1/2})\), then (i) is shown. Now consider solving the first \(d_{n1}\) equations in (2) for \(\beta_{nI}\), while \(\beta_{nII} = 0\). Obviously, when \(\lambda_n\) is sufficiently small, \(a_n = 0\), hence from Theorem 3, there is a root-\(n/d_n\) consistent root, denote it \(\hat{\beta}_I\). Denote the first \(d_{n1}\) equations in \(\sum_{i=1}^n S^*_n,eff (W_i, Z_i, Y_i, \beta_n)\) as \(L_n(\beta_n)\), then

\[
0 = L_n(\hat{\beta}_{nI}) - np'_{\lambda_{nI}}(\hat{\beta}_{nI}) = L_n(\beta_{10}) + \frac{\partial L_n(\beta_{10})}{\partial \beta_{nI}^T}(\hat{\beta}_{nI} - \beta_{10}) + 2^{-1}(\hat{\beta}_{nI} - \beta_{10})^T \frac{\partial^2 L_n(\beta_{nI}^*)}{\partial \beta_{nI} \partial \beta_{nI}^T}(\hat{\beta}_{nI} - \beta_{10}) - nb_n - np''_{\lambda_{nI}}(\hat{\beta}_{nI}) (\hat{\beta}_{nI} - \beta_{10}),
\]

where \(\beta_{nI}^*\) is between \(\beta_{10}\) and \(\hat{\beta}_{nI}\). We also have

\[
||n^{-1}(\hat{\beta}_{nI} - \beta_{10})^T \frac{\partial^2 L_n(\beta_{nI}^*)}{\partial \beta_{nI} \partial \beta_{nI}^T}(\hat{\beta}_{nI} - \beta_{10})||^2 \leq ||\hat{\beta}_{nI} - \beta_{10}||^4 ||n^{-1} \frac{\partial^2 L_n(\beta_{nI}^*)}{\partial \beta_{nI} \partial \beta_{nI}^T}||^2
\]

\[
= O_p(d_n^2 n^{-2} n^{-1} d_{n1}^3) O_p \left( n \left[ E \left\{ \frac{\partial S_{n,eff,I}(\beta_{nI}^*)}{\partial \beta_{nI}^T} \right\} \right] + \text{var} \left( \frac{\partial S_{n,eff,I}(\beta_{nI}^*)}{\partial \beta_{nI}^T} \right) \right)
\]

\[
\leq O_p(d_n^5 n^{-3}) O_p \left( n \left\{ E(M) \right\}^2 + EM^2 \right) \leq O_p(d_n^5 n^{-3}) O_p(nEM^2) = o_p(n^{-1}),
\]

due to condition (B4). In addition,

\[
||n^{-1} \frac{\partial L_n(\beta_{nI0})}{\partial \beta_{nI}^T} - p'_{\lambda_{nI}}(\beta_{nI0}) - E \frac{\partial L_n(\beta_{nI0})}{\partial \beta_{nI}^T} + p''_{\lambda_{nI}}(\beta_{nI0})||^2 \leq 2 ||n^{-1} \frac{\partial L_n(\beta_{nI0})}{\partial \beta_{nI}^T} - E \frac{\partial L_n(\beta_{nI0})}{\partial \beta_{nI}^T}||^2 + O_p(n^{-1} d_{n1})
\]

due to condition (B5) and (B6). For any fixed \(\epsilon > 0\), we have

\[
P_r \left\{ ||n^{-1} \frac{\partial L_n(\beta_{nI0})}{\partial \beta_{nI}^T} - E \frac{\partial L_n(\beta_{nI0})}{\partial \beta_{nI}^T}|| \geq \epsilon d_{n1}^{-1} \right\} \leq \frac{d_{n1}^2}{n^2 \epsilon^2} E \left\| \frac{\partial L_n(\beta_{nI0})}{\partial \beta_{nI}^T} \right\|^2 = O(d_n^2 n^{-2} d_{n1}^2) = o(1),
\]

due to condition (B2), we have

\[
||n^{-1} \frac{\partial L_n(\beta_{nI0})}{\partial \beta_{nI}^T} - E \frac{\partial L_n(\beta_{nI0})}{\partial \beta_{nI}^T}|| = o_p(d_{n1}^{-1}).
\]

Therefore,

\[
||n^{-1} \frac{\partial L_n(\beta_{nI0})}{\partial \beta_{nI}^T} - p'_{\lambda_{nI}}(\beta_{nI0}) - E \frac{\partial L_n(\beta_{nI0})}{\partial \beta_{nI}^T} + p''_{\lambda_{nI}}(\beta_{nI0})||^2 = o_p(d_{n1}^{-2}),
\]

23
and subsequently,
\[
\| \{ n^{-1} \frac{\partial L_n(\beta_{n0})}{\partial \beta_{n0}^T} - p_{\beta_{n0}}^* - E \frac{\partial L_n(\beta_{n0})}{\partial \beta_{n0}^T} + p_{\beta_{n0}}^* \} (\hat{\beta}_{n0} - \beta_{n0}) \| \\
\leq o_p(d_n^{-1}) o_p(n^{-1/2}d_n^{1/2}) = o_p(n^{-1/2}).
\]

We thus obtain
\[
\left\{ -E \frac{\partial L_n(\beta_{n0})}{\partial \beta_{n0}^T} + \Sigma_n \right\} (\hat{\beta}_{n0} - \beta_{n0}) + b_n = n^{-1}L_n(\beta_{n0}) + o_p(n^{-1/2}).
\]

Denoting \( I^* = E \left\{ S^*_{n,e ff,l}(\beta_{n0}) S^T_{n,e ff,l}(\beta_{n0}) \right\} \), using condition (B3), we have
\[
n^{1/2} v_n I^{*-1/2} \left\{ -E \frac{\partial L_n(\beta_{n0})}{\partial \beta_{n0}^T} + \Sigma_n \right\} (\hat{\beta}_{n0} - \beta_{n0}) + b_n \]
\[
= n^{-1/2} v_n I^{*-1/2} L_n(\beta_{n0}) + o_p(v_n I^{*-1/2}) = n^{-1/2} v_n I^{*-1/2} L_n(\beta_{n0}) + o_p(1).
\]

Let
\[
Y_i = n^{-1/2} v_n I^{*-1/2} S_{n,e ff,l}(W_i, Z_i, Y_i, \beta_{n0}), \quad i = 1, \ldots, n.
\]

It follows that for any \( \epsilon > 0 \),
\[
\sum_{i=1}^{n} E[|Y_i|^2 1(|Y_i| > \epsilon)] = n E[|Y_1|^2 1(|Y_1| > \epsilon)] \leq n (E[|Y_1|^4])^{1/2} P_r(|Y_1| > \epsilon)^{1/2}.
\]

We have
\[
P_r(|Y_1| > \epsilon) = P_r(|Y_1|^2 > \epsilon^2) \leq \frac{E[|Y_1|^2]}{\epsilon^2} = \frac{E[v_n I^{*-1/2} S_{n,e ff,l}(W_1, Z_1, Y_1, \beta_{n0})]^2}{n \epsilon^2} = \frac{v_n v_n^T}{n \epsilon^2} = O(n^{-1}),
\]

and
\[
E(|Y_1|^4) = n^{-2} E[|v_n I^{*-1/2} S_{n,e ff,l}(W_1, Z_1, Y_1, \beta_{n0})|^4] = n^{-2} E \{ S_{n,e ff,l}(W_1, Z_1, Y_1, \beta_{n0})^T I^{*-1/2} v_n I^{*-1/2} S_{n,e ff,l}(W_1, Z_1, Y_1, \beta_{n0}) \}^2
\]
\[
\leq n^{-2} \chi^2_{max}(v_n^T v_n) E \{ (S_{n,e ff,l}(W_1, Z_1, Y_1, \beta_{n0})^T I^{*-1} S_{n,e ff,l}(W_1, Z_1, Y_1, \beta_{n0}))^2 \}
\]
\[
\leq n^{-2} \chi^2_{max}(v_n^T v_n) \chi^2_{max}(I^{*-1}) E \{ S_{n,e ff,l}(W_1, Z_1, Y_1, \beta_{n0})^T S_{n,e ff,l}(W_1, Z_1, Y_1, \beta_{n0}) \}^2
\]
\[
= n^{-2} \chi^2_{max}(I^{*-1}) E \{ S_{n,e ff,l}(W_1, Z_1, Y_1, \beta_{n0}) \}^4 = O(d_n^2 n^{-2}),
\]
due to condition (B3). Hence,
\[
\sum_{i=1}^{n} E[|Y_i|^2 1(|Y_i| > \epsilon)] = O(nd_n n^{-1/2} n^{-1/2}) = o(1).
\]
On the other hand,
\[ \sum_{i=1}^{n} \text{cov}(Y_i) = n \text{cov}\{n^{-1/2}v_n I^{*-1/2}S_{n,eff,I}(W_1, Z_1, Y_1, \beta_0)\} = v_n I^{*-1/2}E\{S_{n,eff,I}(W_1, Z_1, Y_1, \beta_0)S_{n,eff,I}(W_1, Z_1, Y_1, \beta_0)^T\}I^{*-1/2}v_n^T = 1. \]

Following Lindeberg-Feller central limit theorem, the results in (ii) now follow. \[ \square \]

**List of Regularity Conditions for Theorems 5 and 6**

(C1) The model is identifiable.

(C2) The first derivatives of \( L \) with respect to \( \beta \) and \( \theta \) exist and are denoted as \( L_\beta \) and \( L_\theta \) respectively. The first derivative of \( \theta \) with respect to \( \beta \) exists and is denoted as \( \theta_\beta \). Then \( E(L_\beta + L_\theta^T) \) exists and is not singular at \( \beta_0 \) and the true function \( \theta_0(Z) \).

(C3) The second derivatives of \( L \) with respect to \( \beta \) and \( \theta \), \( \partial^2 L/\partial \beta \partial \beta^T \), \( \partial^2 L/\partial \theta \partial \theta^T \), and \( \partial^2 L/\partial \beta \partial \theta \), exist and are uniformly bounded in a neighborhood of \( \beta_0, \theta_0 \). The second derivative of \( \theta \) with respect to \( \beta \) exists and is uniformly bounded in a neighborhood of \( \beta_0 \) for all \( \theta_0(Z) \) in a neighborhood of \( \theta_0 \).

(C4) The variance-covariance matrix of \( L_I - \Psi_{U_1}(Z) \) is positive definite at \( \beta_0, \theta_0 \).

(C5) The random variable \( Z \) has compact support and its density \( f_Z(z) \) is positive on that support. The bandwidth \( h \) satisfies \( nh^4 \to 0 \) and \( nh^2 \to \infty \). \( \theta(z) \) has bounded second derivative.

**Proof of Theorem 5**

From condition (C2), we can denote
\[ J = \left[ E \left\{ L_\beta + L_\theta^T \right|_{\beta_0, \theta_0} \right\}^{-1}, \quad \phi_{\text{eff}, i}(\beta, \theta) = JL_\beta(\beta, \theta) \] and \( q_{\lambda_n}^*(\beta) = Jp_{\lambda_n}^*(\beta) \).

Let \( \alpha_n = n^{-1/2} + a_n \), \( \phi_{\text{eff}, i}(\beta, \hat{\theta}) = \phi_{\text{eff}}^*\{W_i, Z_i, Y_i, \beta, \hat{\theta}_i(\beta)\} \), and \( \Psi_{\text{eff}, i}(\beta, \hat{\theta}) = \Psi_{\text{eff}}^*\{\beta, \hat{\theta}_i(\beta)\} \), where \( \hat{\theta}_i \) are obtained from (12). We will prove that
\[ n^{-1/2} \sum_{i=1}^{n} \phi_{\text{eff}, i}^*(\beta, \hat{\theta}) - n^{1/2}q_{\lambda_n}^*(\beta) = 0 \] (19)
has a solution \( \hat{\beta} \) that satisfies \( ||\hat{\beta} - \beta_0|| = O_p(\alpha_n) \). Consider
\[ (\beta - \beta_0)^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi_{\text{eff}, i}^*(\beta, \hat{\theta}) - n^{1/2}q_{\lambda_n}^*(\beta) \right\}. \]
For any $\beta$ such that $||\beta - \beta_0|| = C_{\alpha_n}$ for some constant $C$, because of condition (C3) and (C6), using the expansion (A.3) in Ma and Carroll (2006), we have

$$
(\beta - \beta_0)^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi^*_e f_{i,1}(\beta, \hat{\theta}) - n^{1/2} q^*_{\lambda_n}(\beta) \right\}
$$

$$
= (\beta - \beta_0)^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi^*_e f_{i,1}(\beta_0, \theta_0) - n^{1/2} q^*_{\lambda_n}(\beta_0) 
+ \mathbb{E} \{ \phi^*_e f_{i,1}(\beta_0, \theta_0) + \phi^*_e f_{i,1}(\beta_0, \theta_0) \theta_T^T \{ Z(s) - \beta_0 \} \} n^{1/2}(\beta - \beta_0) 
+ n^{-1/2} \sum_{i=1}^{n} \frac{\partial \phi^*_e f_{i,1}(\beta_0, \theta_0)}{\partial \theta} \{ \hat{\theta}(\beta_0) - \theta_0 \} + o_p(1) - n^{1/2} \frac{\partial q^*_{\lambda_n}(\beta_0)}{\partial \beta_T} (\beta - \beta_0) \{ 1 + o_p(1) \} \right\}
$$

$$
= (\beta - \beta_0)^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi^*_e f_{i,1}(\beta_0, \theta_0) - n^{1/2} q^*_{\lambda_n}(\beta_0) + n^{1/2}(\beta - \beta_0) 
+ n^{-1/2} \sum_{i=1}^{n} \frac{\partial \phi^*_e f_{i,1}(\beta_0, \theta_0)}{\partial \theta} \{ \hat{\theta}(\beta_0) - \theta_0 \} + n^{1/2} O(b_n)(\beta - \beta_0) \{ 1 + o_p(1) \} \right\}
$$

$$
= (\beta - \beta_0)^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi^*_e f_{i,1}(\beta_0, \theta_0) - n^{1/2} q^*_{\lambda_n}(\beta_0) + n^{1/2}(\beta - \beta_0) 
+ (\beta - \beta_0)^T n^{-1/2} \sum_{i=1}^{n} \frac{\partial \phi^*_e f_{i,1}(\beta_0, \theta_0)}{\partial \theta} \{ \hat{\theta}(\beta_0) - \theta_0 \} + o_p(1) \right\} + n^{1/2} ||\beta - \beta_0||^2. \tag{20}
$$

Due to the usual local estimating equation expansion

$$
\hat{\theta}(z, \beta_0) - \theta_0(z)
$$

$$
= \left( h^2/2 \right) \theta_0^T(z) - n^{-1} \sum_{j=1}^{n} K_h(Z_j - z) \Psi_j(\beta_0, \theta_0) / \{ f_Z(z) \Omega(z) \} + o_p(n^{-1/2}), \tag{21}
$$

substituting into the third term in (20), making use of condition (C5), it can be easily seen that

$$
(\beta - \beta_0)^T n^{-1/2} \sum_{i=1}^{n} \frac{\partial \phi^*_e f_{i,1}(\beta_0, \theta_0)}{\partial \theta} \{ \hat{\theta}(\beta_0) - \theta_0 \} = - (\beta - \beta_0)^T \left\{ n^{-1/2} \sum_{i=1}^{n} \Psi_i(\beta_0, \theta_0) \mathcal{U}(Z_i) + o_p(1) \right\},
$$

where $\mathcal{U}(Z) = E(\partial \mathcal{L} / \partial \theta | Z) / \Omega(Z)$. Continuing from (20), we have

$$
(\beta - \beta_0)^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi^*_e f_{i,1}(\beta, \hat{\theta}) - n^{1/2} q^*_{\lambda_n}(\beta) \right\}
$$

$$
= (\beta - \beta_0)^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi^*_e f_{i,1}(\beta_0, \theta_0) - n^{1/2} q^*_{\lambda_n}(\beta_0) - n^{-1/2} \sum_{i=1}^{n} \Psi_i(\beta_0, \theta_0) \mathcal{U}(Z_i) \right\}
$$

$$
+ n^{1/2} ||\beta - \beta_0||^2 + o_p(1) \right\} + n^{1/2} ||\beta - \beta_0||^2.
$$

The first term in the above display is of order $O_p(C n^{1/2} \alpha_n^2)$, the second term equals $C^2 n^{1/2} \alpha_n^2$, which dominates the first term as long as $C$ is large enough. The last term is dominated by the first two

26
terms. Thus, for any \( \epsilon > 0 \), as long as \( C \) is large enough, the probability for the above display to be larger than zero is at least \( 1 - \epsilon \). From Brouwer fixed-point theorem, we know there with probability at least \( 1 - \epsilon \), there exists at least one solution for (19) in the region \( \|\beta - \beta_0\| \leq C\alpha_n \). □

We first prove the following lemma, then give the proof of Theorem 6.

**Lemma 3** If conditions in Theorem 6 hold, then with probability tending to 1, for any given \( \beta \) that satisfies \( \|\beta - \beta_0\| = O_p(n^{-1/2}) \), \( \beta_{11} = 0 \) is a solution to the last \( d_2 \) equations of (11).

**Proof:** Denote the kth equation in \( \sum_{i=1}^{n} \mathcal{L}_i \{ \beta, \hat{\theta}(\beta) \} = L_k(\beta, \hat{\theta}) \), and in \( \sum_{i=1}^{n} \Psi_i(\beta_0, \theta_0) \mathcal{U}(Z_i) \) as \( G_k(\beta_0, \theta_0) \), \( k = d_1 + 1, \ldots, d \), then the expansion in Lemma 5 leads to

\[
L_k(\beta, \hat{\theta}) - np'_{A_n,k}(\beta_k) = L_k(\beta_0, \theta_0) - G_k(\beta_0, \theta_0) + n \sum_{j=1}^{d} (J^{-1})_{kj} (\beta_j - \beta_{j0}) - np'_{A_n,k}(|\beta_k|) \text{sign}(\beta_k) + o_p(n^{1/2}).
\]

The first three terms of the above display are all of order \( O_p(n^{1/2}) \), hence we have

\[
L_k(\beta, \hat{\theta}) - np'_{A_n,k}(\beta_k) = -\sqrt{n}\{\sqrt{n}p'_{A_n}(|\beta_k|) \text{sign}(\beta_k) + O_p(1)\}.
\]

Due to (8), the sign of \( L_k(\beta) - np'_{A_n}(\beta_k) \) is decided by \( \text{sign}(\beta_k) \) completely. From the continuity of \( L_k(\beta) - np'_{A_n}(\beta_k) \), we obtain that it is zero at \( \beta_k = 0 \) with probability larger than any \( 1 - \epsilon \). □

**Proof of Theorem 6:** We let \( \lambda_n \) be sufficiently small so that \( a_n = o(n^{-1/2}) \). From Theorem 5, there is a root-\( n \) consistent estimator \( \hat{\beta} \). From Lemma 3, \( \hat{\beta} = (\hat{\beta}_I^T, 0^T)^T \), so (i) is shown. Denote the first \( d_1 \) equations in \( \sum_{i=1}^{n} \mathcal{L}_i \{ (\beta_I^T, 0^T)^T, \hat{\theta} \} = L\{ \hat{\beta}_I, \hat{\theta}(\hat{\beta}_I) \} \), and in \( \sum_{i=1}^{n} \Psi_i(\beta_0, \theta_0) \mathcal{U}(Z_i) \) as \( G(\beta_{10}, \theta_0) \). Note that the \( d_1 \times d_1 \) upper left block of \( J^{-1} \) is the matrix \( A \) defined in Theorem 6. Use the expansion in Theorem 5 at \( \beta = (\hat{\beta}_I^T, 0^T)^T \), the first \( d_1 \) equations yield

\[
0 = L\{ \hat{\beta}_I, \hat{\theta}(\hat{\beta}_I) \} - np'_{A_n,k}(\hat{\beta}_I) = L(\beta_{10}, \theta_0) - G(\beta_{10}, \theta_0) + nA(\hat{\beta}_I - \beta_{10}) - nb - n\{\Sigma + o_p(1)\}(\hat{\beta}_I - \beta_{10}) + o_p(n^{1/2}) = L(\beta_{10}, \theta_0) - G(\beta_{10}, \theta_0) + n(A - \Sigma)[\hat{\beta}_I - \beta_{10} - (A - \Sigma)^{-1}b] + o_p(n^{1/2}).
\]

We thus obtain

\[
\sqrt{n}\{\hat{\beta}_I - \beta_{10} - (A - \Sigma)^{-1}b\} = -n^{-1/2}(A - \Sigma)^{-1}\{L(\beta_{10}, \theta_0) - G(\beta_{10}, \theta_0)\} + o_p(1).
\]

Because of condition (C4), the results in (ii) now follow. □

**List of Regularity Conditions for Theorem 7 and 8**
(D1) The model is identifiable.

(D2) The first derivatives of $\mathcal{L}$ with respect to $\beta$ and $\theta$ exist and are denoted as $\mathcal{L}_\beta$ and $\mathcal{L}_\theta$ respectively. The first derivative of $\theta$ with respect to $\beta$ exists and is denoted as $\theta_\beta$. Then $E(\mathcal{L}_\beta + \mathcal{L}_\theta\theta^T_\beta)$ exists and its left eigenvalues are bounded away from zero and infinity uniformly for all $n$ at $\beta_0$ and the true function $\theta_0(Z)$. For any entry $S_{jk}$ of the matrix $d(\mathcal{L}_\beta + \mathcal{L}_\theta\theta^T_\beta)/d\beta$, $E(S_{jk}^2) < C_3 < \infty$.

(D3) The eigenvalues of the matrix $E\{\mathcal{L}_I - \Psi \mathcal{U}_I(Z)\} \{\mathcal{L}_I - \Psi \mathcal{U}_I(Z)\}^T$ satisfy $0 < C_1 < \lambda_{\min} < \cdots < \lambda_{\max} < C_2 < \infty$ for all $n$; for any entries $S_k, S_j$ in $\{\mathcal{L}_\beta + \mathcal{L}_\theta\theta^T_\beta\}$, $E(S_k^2S_j^2) < C_4 < \infty$.

(D4) The second derivatives of $\mathcal{L}$ with respect to $\beta$ and $\theta$ exist, the second derivative of $\theta$ with respect to $\beta$ exist, and the entries are uniformly bounded by a function $M(W_i, Z_i, S_i, Y_i)$ in a neighborhood of $\beta_0, \theta_0$. In addition, $E(M^2) < c_5 < \infty$ for all $n, d$.

(D5) The random variable $Z$ has compact support and its density $f_Z(z)$ is positive on that support. The bandwidth $h$ satisfies $nh^4 \to 0$ and $nh^2 \to \infty$. $\phi(z)$ has bounded second derivative.

(D6) $\min_{1 \leq j \leq d_1} |\beta_{j0}|/\lambda \to \infty$ as $n \to \infty$.

(D7) Let $c_n = \max_{1 \leq j \leq d_1} \{p_n^\prime(|\beta_{j0}|) : \beta_{j0} \neq 0\}$. Assume that $\lambda \to 0$, $a_n = O(n^{-1/2})$ and $c_n \to 0$ as $n \to \infty$. In addition, there exists constants $C$ and $D$ such that when $\theta_1, \theta_2 > C\lambda$, $|p_n^\prime(\theta_1) - p_n^\prime(\theta_2)| \leq D|\theta_1 - \theta_2|$.

To emphasize the dependence of $v, \lambda, d, d_1, d_2, \mathcal{L}, \mathcal{U}, A, B, \Sigma, \beta, \beta_1, \beta_{II}, \beta_j$, for $j = 1, \ldots, d$, on $n$, we write them as $v_n, \lambda_n, d_n, d_{n1}, d_{n2}, \mathcal{L}_n, \mathcal{U}_n, A_n, B_n, \Sigma_n, \beta_n, \beta_{n1}, \beta_{nII}, \beta_{nj}$ respectively throughout the proof.

Proof of Theorem 7

From condition (D2), we can denote

$$J_n = \left[ E \left\{ \left( \mathcal{L}_{n\beta_n} + \mathcal{L}_{n\theta_\beta_{n\theta_0}} \right) |_{\beta_{n0} = \theta_0} \right\} \right]^{-1}, \quad \phi_{n,eff}^\star(\beta_n, \theta) = J_n \mathcal{L}_n(\beta_n, \theta) \quad \text{and} \quad q_{n,\lambda_n}^\prime(\beta_n) = J_n p_{n,\lambda_n}^\prime(\beta_n).$$

Let $\alpha_n = n^{-1/2} + a_n$, $\phi_{n,eff,i}^\star(\beta_n, \hat{\theta}) = \phi_{n,eff}^\star\{W_i, Z_i, S_i, Y_i, \beta_n, \hat{\theta}(\beta_n)\}$. It will be enough to show that

$$n^{-1/2} \sum_{i=1}^n \phi_{n,eff,i}^\star(\beta_n, \hat{\theta}) - n^{1/2} q_{n,\lambda_n}^\prime(\beta_n) = 0 \quad (22)$$

has a solution $\hat{\beta}_n$ that satisfies $||\hat{\beta}_n - \beta_{n0}|| = O_P(a_n^{1/2} \alpha_n)$.
Note that since (21) concerns only the true fixed value \( \beta_0 \), it is still valid in this setting, which implies that \( \hat{\theta}(z, \beta_0) - \theta_0(z) = O_p(h^2 + n^{-1/2}h^{-1/2}) \). For any \( \beta_n \) such that \( ||\beta_n - \beta_0|| \leq Cd_n^{1/2} \alpha_n \) for some constant \( C \), we obtain the expansion

\[
\begin{align*}
&n^{-1/2} \sum_{i=1}^{n} \phi_{n, eff, i}^* \{ \beta_n, \hat{\theta}(\beta_n) \} \\
= &n^{-1/2} \sum_{i=1}^{n} \phi_{n, eff, i}^* \{ \beta_{n0}, \hat{\theta}(\beta_{n0}) \} \\
+ &n^{-1/2} \sum_{i=1}^{n} \left[ \frac{\partial \phi_{n, eff, i}^* \{ \beta_{n0} + \hat{\theta}(\beta_{n0}) \}}{\partial \beta_n^T} \right] \left( \frac{\partial \hat{\theta}(\beta_{n0})}{\partial \beta_n} \right) (\beta_n - \beta_{n0}) \\
+ &\frac{1}{2\sqrt{n}} \sum_{i=1}^{n} (\hat{\beta} - \beta_0)^T \frac{d}{d \beta_n^T} \left[ \phi_{n, eff, i}^* \{ \beta_{n0} + \hat{\theta}(\beta_{n0}) \} + \phi_{n, eff, i}^* \{ \beta_{n0} + \hat{\theta}(\beta_{n0}) \} \frac{\partial \hat{\theta}(\beta_{n0})}{\partial \beta_n} \right] |\beta_n - \beta_{n0}|,
\end{align*}
\]

where \( \beta_n^* \) is in between \( \beta_n \) and \( \beta_{n0} \). Because of condition (D4), each component of the last term is uniformly of order \( O_p(n^{1/2} ||\beta_n - \beta_{n0}||^2) \). The second term can be written as \( n^{1/2} \{1 + o_p(1)\} (\beta_n - \beta_{n0}) \) under condition (D2), (D4) and (D5). The first term can be further expanded as

\[
\begin{align*}
&n^{-1/2} \sum_{i=1}^{n} \phi_{n, eff, i}^* \{ \beta_{n0}, \hat{\theta}(\beta_{n0}) \} \\
= &n^{-1/2} \sum_{i=1}^{n} \left[ \frac{\partial \phi_{n, eff, i}^* \{ \beta_{n0}, \hat{\theta}(\beta_{n0}) \}}{\partial \theta} \right] \{ \hat{\theta}(\beta_{n0}) - \theta_0 \} + O_p(n^{1/2}) (\hat{\theta}(\beta_{n0}) - \theta_0)^2 \\
= &n^{-1/2} \sum_{i=1}^{n} \left[ \frac{\partial \phi_{n, eff, i}^* \{ \beta_{n0}, \hat{\theta}(\beta_{n0}) \}}{\partial \theta} \right] \{ \hat{\theta}(\beta_{n0}) - \theta_0 \} + o_p(1)
\end{align*}
\]

under condition (D4) and (D5). Summarizing the above results, making use of (21), we obtain

\[
\begin{align*}
&n^{-1/2} \sum_{i=1}^{n} \phi_{n, eff, i}^* \{ \beta_n, \hat{\theta}(\beta_n) \} \\
= &n^{-1/2} \sum_{i=1}^{n} \phi_{n, eff, i}(\beta_{n0}, \theta_0) + n^{1/2}(\beta_n - \beta_{n0}) \\
&- n^{-3/2} \sum_{j,i=1}^{n} \left[ \frac{\partial \phi_{n, eff, i}(\beta_{n0}, \theta_0)}{\partial \theta} \right] \frac{K_h(Z_j - Z_i)}{f(Z_j)\Omega(Z_i)} \Psi_j(\beta_{n0}, \theta_0) + o_p(1) \\
= &n^{-1/2} \sum_{i=1}^{n} \phi_{n, eff, i}(\beta_{n0}, \theta_0) + n^{1/2}(\beta_n - \beta_{n0}) - n^{-1/2} \sum_{i=1}^{n} \Psi_j(\beta_{n0}, \theta_0)J\Omega(Z_i) + o_p(1)
\end{align*}
\]
under condition (D5). Similar to the situation in Theorem 3, under condition (D6) and (D7), we further obtain

\[
(\beta_n - \beta_{n0})^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi_{n,eff,i}^*(\beta_n, \hat{\theta}) - n^{1/2} q_{\lambda_n}^*(\beta_n) \right\}
\]

\[
= (\beta_n - \beta_{n0})^T \left\{ n^{-1/2} \sum_{i=1}^{n} \phi_{n,eff,i}^*(\beta_{n0}, \theta_0) - n^{1/2} q_{\lambda_n}^*(\beta_{n0}) - n^{-1/2} \sum_{i=1}^{n} \Psi_i(\beta_{n0}, \theta_0) \mathcal{U}_n(Z_i) \right\}
\]

\[+ n^{1/2} |\beta_n - \beta_{n0}|^2 + o_p(n^{1/2} |\beta_n - \beta_{n0}|^2). \tag{23}
\]

The first term in the above display is of order \(O_p(Cn^{1/2}d_n\alpha_n^2)\), the second term equals \(C^2n^{1/2}d_n\alpha_n^2\), which dominates the first term as long as \(C\) is large enough. The last term is dominated by the first two terms. Thus, for any \(\epsilon > 0\), as long as \(C\) is large enough, the probability for the above display to be larger than zero is at least \(1 - \epsilon\). From Brouwer fixed-point theorem, we know there with probability at least \(1 - \epsilon\), there exists at least one solution for (22) in the region \(|\beta_n - \beta_{n0}| \leq C\alpha_n^{1/2}\).

We first prove the following lemma, then give the proof of Theorem 8.

**Lemma 4** If conditions in Theorem 8 hold, then with probability tending to 1, for any given \(\beta_n\) that satisfies \(|\beta_n - \beta_{n0}| = O_p(d_n^{1/2}n^{-1/2})\), \(\beta_{nII} = 0\) is a solution to the last \(d_n\) equations of (11).

**Proof:** Denote the \(k\)th equation in \(\sum_{i=1}^{n} \mathcal{C}_n \{ \beta_n, \hat{\theta}(\beta_n) \} = L_{nk}(\beta_n, \hat{\theta})\), and that in \(\sum_{i=1}^{n} \Psi_i(\beta_{n0}, \theta_0) \mathcal{U}_n(Z_i)\) as \(G_{nk}(\beta_{n0}, \theta_0)\), \(k = d_n + 1, \ldots, d_n\), then the expansion in Theorem 7 leads to

\[
L_{nk}(\beta_n, \hat{\theta}) - np_{\lambda_n,k}(\beta_n)
\]

\[
= L_{nk}(\beta_{n0}, \theta_0) - G_{nk}(\beta_{n0}, \theta_0) + n \sum_{j=1}^{d} (J_{n}^{-1})_{kj} (\beta_{nj} - \beta_{n0j}) - np_{\lambda_n,k}(|\beta_{nk}|) \text{sign}(\beta_{nk}) + o_p(d_n^{1/2}n^{1/2}).
\]

Similar to the derivation in Lemma 2, the first three terms of the above display are all of order \(O_p(n^{1/2}d_n^{1/2})\), hence we have

\[
L_{nk}(\beta_n, \hat{\theta}) - np_{\lambda_n,k}(\beta_{nk}) = -\sqrt{n}d_n \{ \sqrt{n}d_n p_{\lambda_n,k}(|\beta_{nk}|) \text{sign}(\beta_{nk}) + O_p(1) \}.
\]

Because of (9), the sign of \(L_{nk}(\beta_n) - np_{\lambda_n}(\beta_{nk})\) is decided by \(\text{sign}(\beta_{nk})\) completely. From the continuity of \(L_{nk}(\beta_n) - np_{\lambda_n}(\beta_{nk})\), we obtain that it is zero at \(\beta_{nk} = 0\) with probability larger than any \(1 - \epsilon\). \(\square\)

**Proof of Theorem 8** We let \(\lambda_n\) be sufficiently small so that \(a_n = o(n^{-1/2})\). From Theorem 7, there is a root-\(n/d_n\) consistent estimator \(\hat{\beta}_n\). From Lemma 4, \(\hat{\beta}_n = (\hat{\beta}_n^T, 0^T)^T\), so (i) is
shown. Denote the first $d_{n1}$ equations in $\sum_{i=1}^{n} L_{ni}\{(\beta_{ni}^T, 0^T)^T\} = L_{n}\{\beta_{ni}, \hat{\theta}(\beta_{ni})\}$, and that in $\sum_{i=1}^{n} \Psi_i(\beta_{ni0}, \theta_0) U_{ni}(Z_i)$ as $G_n(\beta_{ni0}, \theta_0)$. Note that the $d_{n1} \times d_{n1}$ upper left block of $J_{n}^{-1}$ is the matrix $A_n$ defined in Theorem 8. Use the expansion in Lemma 7 at $\beta_n = (\beta_{n1}^T, 0)^T$, the first $d_{n1}$ equations yield

\begin{align*}
0 &= L_n\{\beta_{ni}, \hat{\theta}(\beta_{ni})\} - np_{\hat{\lambda}_{ni}}(\beta_{ni}) \\
&= L_n(\beta_{ni0}, \theta_0) - G_n(\beta_{ni0}, \theta_0) + n A_n(\hat{\beta}_{ni} - \beta_{ni0}) - nb_n - n \{\Sigma_n + o_p(1)\} (\hat{\beta}_{ni} - \beta_{ni0}) + o_p(d_n^{1/2}n^{1/2}) \\
&= L_n(\beta_{ni0}, \theta_0) - G_n(\beta_{ni0}, \theta_0) + n (A_n - \Sigma_n) (\hat{\beta}_{ni} - \beta_{ni0}) - (A_n - \Sigma_n)^{-1}bn + o_p(d_n^{1/2}n^{1/2}).
\end{align*}

Using condition (D3), we have

\begin{align*}
n^{1/2}v_nB^{-1/2} \left\{ (-A_n + \Sigma_n)(\hat{\beta}_{ni} - \beta_{ni0}) + bn \right\} \\
= n^{-1/2}v_nB_n^{-1/2}\{L_n(\beta_{ni0}, \theta_0) - G_n(\beta_{ni0}, \theta_0)\} + o_p(v_nB^{-1/2}) \\
= n^{-1/2}v_nB_n^{-1/2}\{L_n(\beta_{ni0}, \theta_0) - G_n(\beta_{ni0}, \theta_0)\} + o_p(1).
\end{align*}

Let

$$Y_i = n^{-1/2}v_nB^{-1/2}\{L_{ni1}(\beta_{ni0}, \theta_0) - \Psi_1(\beta_{ni0}, \theta_0)U_{ni1}(Z_i)\}, \quad i = 1, \ldots, n$$

it follows that for any $\epsilon > 0$,

$$\sum_{i=1}^{n} E||Y_i||^2 1(||Y_i|| > \epsilon) = nE||Y_i||^2 1(||Y_i|| > \epsilon) \leq n(\epsilon E||Y_i||^4)^{1/2}(P_r(||Y_i|| > \epsilon))^{1/2}.$$

We have

$$P_r(||Y_i|| > \epsilon) = P_r(||Y_i||^2 > \epsilon^2) \leq \frac{E||Y_i||^2}{\epsilon^2} = \frac{v_nv_n^2}{\epsilon^2} = O(n^{-1}),$$

and

$$E(||Y_i||^4) = n^{-2}E[\sum_{i=1}^{n} v_nB^{-1/2}\{L_{ni1}(\beta_{ni0}, \theta_0) - \Psi_1(\beta_{ni0}, \theta_0)U_{ni1}(Z_i)\}]^4$$

$$= n^{-2}E[(\{L_{ni1}(\beta_{ni0}, \theta_0) - \Psi_1(\beta_{ni0}, \theta_0)U_{ni1}(Z_i)\})^T B^{-1/2}v_nB^{-1/2}$$

$$\leq n^{-2}\lambda_{\max}^2(v_n^T v_n)E[(\{L_{ni1}(\beta_{ni0}, \theta_0) - \Psi_1(\beta_{ni0}, \theta_0)U_{ni1}(Z_i)\})^T B^{-1}$$

$$\leq n^{-2}\lambda_{\max}^2(v_n^T v_n)^2 \lambda_{\max}^2(B^{-1})E[(\{L_{ni1}(\beta_{ni0}, \theta_0) - \Psi_1(\beta_{ni0}, \theta_0)U_{ni1}(Z_i)\})^T$$

$$\leq n^{-2}\lambda_{\max}^2(B^{-1})E[\sum_{i=1}^{n} v_nB^{-1/2}\{L_{ni1}(\beta_{ni0}, \theta_0) - \Psi_1(\beta_{ni0}, \theta_0)U_{ni1}(Z_i)\}]^4 = O(d_n^2n^{-2}),$$

31
due to condition (D3). Hence,
\[ \sum_{i=1}^{n} E[||Y_i||^21(||Y_i|| > \epsilon)] = O(nd_{n1}n^{-1}n^{-1/2}) = o(1). \]

On the other hand,
\[ \sum_{i=1}^{n} \text{cov}(Y_i) = n\text{cov}[n^{-1/2}v_nB^{-1/2}\{L_{n1}(\beta_{nI0}, \theta_0) - \Psi_{1}(\beta_{nI0}, \theta_0)U_{nI}(Z_1)\}] = 1. \]

Following Lindeberg-Feller central limit theorem, the results in (ii) now follow.

6 Discussion

In this paper, we have proposed a new variable selection procedure in the framework of measurement error models. The procedure is proposed in a completely general functional measurement error model setting, and it is suitable for both parametric models and semiparametric models that contain an unspecified smooth function of observable covariates. We have assumed the error model \( p_{W|X,Z}(W|X, Z) \) to be completely known for ease of presentation. In the situation when the error model contains an unknown parameter \( \xi \), the identifiability of the problem requires additional information such as multiple measurements or instruments. These information should be incorporated to estimate \( \xi \), as illustrated in Ma and Carroll (2006). In the variable selection context, we can simply append the estimating equation with these additional estimating equations, while appending the penalty function \( p'_{\lambda} \) with zeros. The same asymptotic convergence rates and oracle properties hold as in the known \( \xi \) case, without any efficiency loss. In the situation where the error model \( p_{W|X,Z}(W|X, Z) \) is totally unspecified, a nonparametric estimation of the measurement error distribution has to be carried out first, then the result can be plugged into the proposed variable selection and estimation procedure. In this case, the asymptotic convergence rate of the parameters and the oracle property still remain the same, but the asymptotic variance will increase, see Hall and Ma (2007) for details.

Finally, we would like to point out that in the special case of generalized linear models and normal additive error with possible heteroscedasticity, the procedure of solving linear integral equations can be spared and the estimating equations are simplified significantly (Ma and Tsiatis, 2006). In such situations, the computation complexity of the proposed procedure will be reduced to about the same level as the regressions without errors in the variables.
References


