Empirical Likelihood-based Dimension Reduction Inference for Linear Error-in-Responses Models with Validation Study

BY Qihua Wang

Academy of Mathematics and System Science, Chinese Academy of Science
Beijing 100080, P.R. China

AND

Wolfgang Härdle
Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin
10178 Berlin, Germany

Abstract

In this paper, linear errors-in-response models are considered in the presence of validation data on the responses. A semiparametric dimension reduction technique is employed to define an estimator of β with asymptotic normality, the estimated empirical loglikelihoods and the adjusted empirical loglikelihoods for the vector of regression coefficients and linear combinations of the regression coefficients, respectively. The estimated empirical log-likelihoods are shown to be asymptotically distributed as weighted sums of independent χ^2_1 and the adjusted empirical loglikelihoods are proved to be asymptotically distributed as standard chi-squares, respectively. A simulation study is conducted to compare the proposed methods in terms of coverage accuracies and average lengths of the confidence intervals.

Key Words. Confidence intervals; Error-in-response; Validation data.

AMS 2000 Subject Classifications. Primary 62J05, Secondary 62E20

1 Introduction

Let Y be a scalar response variable and X be a p-variate explanatory variable in regression. We consider the linear model

$$Y = \mathbf{X}^\top \beta + e, \tag{1.1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a $p \times 1$ vector of unknown regression coefficients, ϵ is a random statistical error, and given X, the errors $e = Y - X^\top \beta$ are identically independently distributed.

Regression problems where some of the predictors are measured with error have been extensively studied. Excellent introductions to the area were provided by Fuller (1987) and Carroll, Ruppert and Stefanski (1995). Here, we consider the problem of error-in-response variables. This is a realistic situation. In a study of factors affecting dietary intake of fat, e.g., sex, race, age, socioeconomic status, etc., true long-term dietary intake is impossible to determine and instead it is necessary to use error-prone measures of long term dietary intakes. Wittes, *et al* (1989) describe another example in which damage to the heart muscle caused by a myocardial infarction can be assessed accurately using arterioscintography, but the procedure is expensive and invasive, and instead it is common practice to use peak cardiac enzyme level in the bloodstream as a proxy for the true response. Generally, the relationship between the surrogate variables $\widetilde{Y}$ and the true variables Y can be rather complicated compared to the classical additive error structure usually assumed in literature The additive error model is often not appropriate, and some authors [e.g., Buonaccorsi (1996); Carroll and Stefenski (1990); Pepe (1992)] have considered more complex measurement error models for either regression or the response. The resulting inferences, however, could be sensitive to the assumed model. Actually, in many practical settings, it is even difficult to specify the relationship between true variables and their surrogated variables. The most realistic situation may be that no model structure between the true variables and surrogate variables or distribution

assumption of the true variables given the surrogate variables is specified. However, this situation presents serious difficulties towards obtaining correct statistical analysis. Biases caused by measurement errors could be difficult to access accurately without extra observations and information. One of solutions is to use the help of validation data. Some examples where validation data are available can be found in Wittes, Lakatos and Probstfied (1989), Duncan and Hill (1985) and Pepe (1992) among others. With help of validation data, some statisticians developed statistical inference techniques based on primary observations without specifying any error structure and the distribution assumption of the true variable given the surrogate variable. See, for example, Stefanski and Carroll (1987), Carroll and Wand (1991), Pepe and Fleming (1991), Pepe (1992), Pepe *et al* (1994), Reilly and Pepe (1995), Sepanski and Lee (1995), Wang (1999,2000) and Wang and Rao (2002) among others.

For model (1.1), we consider settings where no model structure assumption between the true variables and surrogate variables or distribution assumption of Y given $\widetilde{Y}$ is specified, but some validation data are available to relate Y and $\widetilde{Y}$. With help of validation data, we define the estimator of β and develop empirical likelihood inference for β and its linear combinations. To use the surrogate $\widetilde{Y}'s$, let us rewrite the model (1.1) such that $\widetilde{Y}$ is related to X. Notice that $\widetilde{Y}$ and X provide useful information in predicting the unknown Y. We rewrite the model (1.1) as

$$u(Z) = \mathbf{X}^\top \beta + \epsilon \tag{1.2}$$

where $Z = (\widetilde{Y}, \mathbf{X})$, $u(z) = E[Y|Z = z]$ and $\epsilon = e - Y - u(Z)$. If $u(\cdot)$ was a known function , (1.2) is then a standard statistical model and hence standard statistical inference approaches such as the least square and empirical likelihood due to Owen (1991) for linear model can be applied to inference for β or linear combinations of β from the primary data. Usually, $u(\cdot)$ is unknown. Hence, the LSE and empirical log-likelihood functions contain unknown $u(\cdot)$. A natural method is to replace $u(\cdot)$

in the LSE and empirical log-likelihood functions by an estimator of $u(\cdot)$ and define a final estimator of β and estimated empirical log-likelihood functions. Here, we estimate $u(\cdot)$ by kernel regression approach. This method requires a large validation data set, which is difficult or expensive to obtain, in order to be feasible because of the use of kernel regression with $p+1$ dimension explanatory variables Z. That is, "curse of dimension" will limit this approach. We therefore propose a dimension reduction technique by assuming

$$u(z) = m(\alpha^\top z), \tag{1.3}$$

where $m(\cdot)$ is an unknown function and α is a $(p+1) \times 1$ vector of unknown parameter. This assumption is reasonable in many applications. It applies at least to generalized linear models and is conform with the class of single index models. In (1.3), α can be first estimated by sliced inverse regression (SIR) techniques [see, e.g., Li (1991), Duan and Li (1991) and Zhu and Fang (1996)]. Then, we can estimate $u(\cdot)$ by the kernel regression with univariate explanatory variable with validation data. We will prove that the resulting estimator of β is asymptotically normal and the estimated empirical log-likelihood functions for β and its linear combinations are asymptotically weighted sums of independent χ^2_1 variables with unknown weights, respectively. As a result, they cannot be applied directly to construct confidence regions for β. To overcome this difficulty, several different methods may be used. In the first method, the unknown weights are estimated consistently so that the distribution of the estimated weighted sums of chi-squared variables can be calculated from the data. In the second method, the estimated empirical loglikelihood functions are adjusted so that the resulting adjusted empirical loglikelihood functions are asymptotically distributed as standard chi-squares.

This paper is organized as follows. In Section 2, we define a modified LSE with asymptotic normality. In Section 3, we define an estimated empirical loglikelihood and an adjusted empirical loglikelihood for β, and show that the estimated empirical

loglikelihood is asymptotically distributed as a weighted sum of indepedent χ^2 and the adjusted empirical loglikelihood is asymptotically distributed as a standard chi-square. In Section 4, we define an estimated empirical loglikelihood and an adjusted empirical loglikelihood for linear combinations of β and state their asymptotic results similar to those in Section 3. In Section 5, we report some Monte Carlo simulation results for the finite sample performation of the proposed approaches. The appendix presents the proofs of the main results.

2 Estimation

Assume we have a validation data set containing n independent and identically distributed observations of $\{(\widetilde{Y}_i, Y_i, \mathbf{X}_i)_{i=1}^n\}$ and a primary data set containing N independent and identically distributed observations of $\{(\widetilde{Y}_j, \mathbf{X}_j)_{j=n+1}^{n+N}\}$. The primary data set is independent of the validation data set. If $u(\cdot)$ was known in (1.2), the LSE for β with the primary data can be defined to be

$$\widetilde{\beta}_N = \left(\sum_{j=n+1}^{n+N} \mathbf{X}_j \mathbf{X}_j^\top \right)^{-1} \sum_{j=n+1}^{n+N} \mathbf{X}_j u(Z_j),$$

In our setup, $u(\cdot)$ is unknown. We therefore use an estimator for $u(\cdot)$ in the above formula. In what follows, we define the estimator of $u(\cdot)$ based on the dimension reduction model (1.3).

Denote $\mathbf{X} = (X_1, X_2, \cdots, X_p)$, $R(Y) = (R_1(Y), \cdots, R_{p+1}(Y))^T = (E[\widetilde{Y}|Y], E[X_1|Y], \cdots, E[X_p|Y])^T$, $\Lambda = Cov(R(Y)) = Cov(E[Z|Y])$. Denote by Z_{ij} the jth component of Z_i for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, p+1$. Let

$$R_{nj}^*(y) = \frac{1}{nh_{1,n}} \sum_{i=1}^n Z_{ij} K_1 \left(\frac{y - Y_i}{h_{1,n}} \right), j = 1, 2, \cdots, p+1$$

and

$$\widehat{f}_n(y) = \frac{1}{nh_{1,n}} \sum_{i=1}^n K_1 \left(\frac{y - Y_i}{h_{1,n}} \right),$$

where $K_1(\cdot)$ is a kernel function and $h_{1,n}$ is a bandwidth. For each fixed $b > 0$, let

$$\begin{aligned} \widehat{f}_{nb}(y) &= max(\widehat{f}_n(y), b) \\ \widehat{R}_{nb}(y) &= \left(\frac{R^*_{nj}(y)}{\widehat{f}_{nb}(y)}\right)_{(p+1)\times 1} \end{aligned}$$

and

$$\widehat{\Lambda}_n = \frac{1}{n}\sum_{j=1}^{n}(\widehat{R}_{nb}(Y_j))(\widehat{R}_{nb}(Y_j))^\top - \left(\frac{1}{n}\sum_{j=1}^{n}\widehat{R}_{nb}(Y_j)\right)\left(\frac{1}{n}\sum_{j=1}^{n}\widehat{R}_{nb}(Y_j)\right)^\top$$

Let α_n be the eigenvector corresponding to the maximum eigenvalue of $\widehat{\Lambda}_n$. By Zhu and Fang (1996), we can estimate α by α_n. Then, $u(z) = m(\alpha^\top z)$ can be estimated by

$$\widehat{u}_n(z) = \frac{\sum_{i=1}^{n} K_2\left(\frac{\alpha_n^\top(z-Z_i)}{h_{2,n}}\right) Y_i}{\sum_{i=1}^{n} K_2\left(\frac{\alpha_n^\top(z-Z_i)}{h_{2,n}}\right)} \tag{2.1}$$

where $h_{2,n}$ is a bandwidth and $K_2(\cdot)$ is a kernel function. To avoid technical difficulties due to small values in the denominator of $\widehat{u}_n(\cdot)$, we define a truncation version of $\widehat{u}_n(\cdot)$.

Let $\widehat{f}_{n,\mathcal{Z}}(z) = (n_2 h_{2,n})^{-1}\sum_{i=1}^{n} K_2\left(\frac{\widehat{\alpha}_n(z-Z_i)}{h_{2,n}}\right)$ and $\widehat{f}_{\tau_n,\mathcal{Z}}(z) = \max\{\widehat{f}_{n,\mathcal{Z}}(z), \tau_n\}$ for some positive constant sequence τ_n tending to zero. The truncated version of $\widehat{u}_n(z)$ is then defined by

$$\widehat{u}_{\tau_n}(z) = \frac{\widehat{u}_n(z)\widehat{f}_{n,\mathcal{Z}}(z)}{\widehat{f}_{\tau_n,\mathcal{Z}}(z)}.$$

We then can define a final estimator of β, $\widehat{\beta}_{nN}$ say, by replacing $u(\cdot)$ in $\widetilde{\beta}_N$ with $\widehat{u}_{\tau_n}(x)$. That is,

$$\widehat{\beta}_{n,N} = \left(\sum_{j=n+1}^{n+N}\mathbf{X}_j\mathbf{X}_j^\top\right)^{-1}\sum_{j=n+1}^{n+N}\mathbf{X}_j\widehat{u}_{\tau_n}(Z_j).$$

Let $\Sigma = E\mathbf{X}\mathbf{X}^\top$ and $V_1(\beta) = E[(u(Z) - \mathbf{X}^\top\beta)^2\mathbf{X}\mathbf{X}^\top] + \lambda E[(Y - u(Z))^2\mathbf{X}\mathbf{X}^\top]$, where $\lambda = \frac{N}{n}$.

THEOREM 2.1 Under all the assumptions listed in the Appendix, we have

$$\sqrt{N}(\widehat{\beta}_{n,N} - \beta) \xrightarrow{\mathcal{L}} N(0, V(\beta)),$$

where $V(\beta) = \Sigma^{-1} V_1(\beta) \Sigma^{-1}$.

REMARK 2.1 The first term in the asymptotic covariance of $\widehat{\beta}_{n,N}$ is the contribution of the primary data, the Fisher information for β in the primary sample by the regression relationship between $u(z)$ and X. The second term represents the extra cost due to estimation of unknown $u(Z)$.

REMARL 2.2 The asymptotic covariance of $\widehat{\beta}_{n,N}$ can be estimated consistently by

$$V_{n,N} = \Sigma_{n,N}^{-1} \widehat{V}_1(\widehat{\beta}_{n,N}) \Sigma_{n,N}^{-1}$$

where

$$\Sigma_{n,N} = \frac{1}{N} \sum_{j=n+1}^{n+N} \mathbf{X}_j \mathbf{X}_j^\top$$

and

$$\widehat{V}_1(\widehat{\beta}_{n,N}) = \frac{1}{N} \sum_{j=n+1}^{n+N} [(\widehat{u}_{\gamma_n}(Z_j) - \mathbf{X}_j^\top \widehat{\beta}_{n,N})^2 \mathbf{X}_j \mathbf{X}_j^\top] + \frac{N}{n^2} \sum_{i=1}^{n} [(Y_i - \widehat{u}_{\gamma_n}(Z_i))^2 \mathbf{X}_i \mathbf{X}_i^\top].$$

REMARK 2.3 To use information sufficiently, one may define the estimator of β to be

$$\widetilde{\beta}_{n,N} = \widetilde{\Sigma}_{n,N}^{-1} \widetilde{A}_{n,N},$$

where $\widetilde{\Sigma}_{n,N} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top + \frac{1}{N} \sum_{k=n+1}^{n+N} \mathbf{X}_k \mathbf{X}_k^\top$ and $\widetilde{A}_{n,N} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i Y_i + \frac{1}{N} \sum_{k=n+1}^{n+N} \mathbf{X}_k u_{\tau_n}(Z_k)$. In most applications, however, the primary data set is large relative to the validation data, i.e., λ is large. For example, in the nurses health study described by Rosner *et al* (1989), $\lambda = 517.6$. In such cases, there is little information about β in the validation data set, and there will be little difference between $\widetilde{\beta}_{n,N}$ and $\widehat{\beta}_{n,N}$. On the other hand, It is much simpler to consider $\widehat{\beta}_{n,N}$. For these reasons, we consider $\widehat{\beta}_{n,N}$ only.

3 Estimated and adjusted empirical likelihood for β

We first give some motivations for defining an estimated empirical likelihood. Let $A_j(\beta) = \mathbf{X}_j(u(Z_j) - \mathbf{X}_j^\top \beta)$. Then, we have $EA_j(\beta) = 0$. Let F_p be the distribution function which assigns probability p_j at point $(\mathbf{X}_j, \widetilde{Y}_j)$, respectively, for $j = n+1, \cdots, n+N$. Then, we may define the empirical loglikelihood, evaluated at β, as

$$l_N(\beta) = -2 \max \sum_{j=n+1}^{n+N} \log(Np_j),$$

where the maximum is over $\sum_{j=n+1}^{n+N} p_j A_j(\beta) = 0$ and $\sum_{j=n+1}^{n+N} p_j = 1$. If β is the true parameter, then $l_N(\beta)$ can be shown to be asymptotically distributed as a standard χ^2 with p degrees of freedom. However, this result cannot be used to make inference about β because $l_N(\beta)$ contains the unknown terms $u(Z_j)$, and hence β is not identifiable. Naturally, we replace $u(\cdot)$ in $l_N(\beta)$ by an estimator of it and define an estimated empirical log-likelihood, $\widehat{l}_N(\beta)$. Here, we replace $u(\cdot)$ in $l_N(\beta)$ with $\widehat{u}_{\tau_n}(\cdot)$ and define an estimated empirical log-likelihood by

$$\widehat{l}_N(\beta) = -2 \max_{\sum_{j=n+1}^{n+N} p_j \widehat{A}_j(\beta) = 0} \sum_{j=n+1}^{n+N} (NP_j), \tag{3.1}$$

where $\widehat{A}_j(\beta)$ is $A_j(\beta)$ with $u(\cdot)$ replaced with $\widehat{u}_{\tau_n}(\cdot)$.

By using the Lagrange multiplier method, the optimal values of $p_j's$ satisfying (3.1) can be shown to be

$$p_j = \frac{1}{N} \frac{1}{1 + \lambda^\top \widehat{A}_j(\beta)},$$

where λ is the solution of the equation

$$\frac{1}{N} \sum_{j=n+1}^{n+N} \frac{\widehat{A}_j(\beta)}{1 + \lambda^\top \widehat{A}_j(\beta)} = 0. \tag{3.2}$$

This yields

$$\widehat{l}_{n,N}(\beta) = 2 \sum_{j=n+1}^{n+N} \log\{1 + \lambda^\top (\widehat{u}_{\tau_n}(Z_j) - \mathbf{X}_j^\top \beta)\}. \tag{3.3}$$

Let $V_0(\beta) = E[\mathbf{X}\mathbf{X}^\top(u(Z) - \mathbf{X}^\top\beta)^2]$.

THEOREM 3.1 Under the assumptions given in the Appendix, if β is the true parameter, we have

$$\widehat{l}_{n,N}(\beta) \xrightarrow{\mathcal{L}} \sum_{i=1}^{p} w_i \chi^2_{1,i},$$

where $\chi^2_{1,i}$ for $1 \le i \le p$ are independent χ^2_1 variables and $w_1, w_2, \cdots, w_p$ are the eigenvalues of $D(\beta) = V_0^{-1}(\beta)V_1(\beta)$ with $V_1(\beta)$ defined in Theorem 2.1.

To apply Theorem 3.1 to construct a confidence region or inteval for β, we must estimate the unkown weights w_i consistently. By the "plug in" method, $V_1(\beta)$ and $V_0(\beta)$ can be estimated consistently by $\widehat{V}_1(\widehat{\beta}_{n,N})$, which is defined in Section 2, and

$$\widehat{V}_0(\widehat{\beta}_{n,N}) = N^{-1} \sum_{j=n+1}^{n+N} [\mathbf{X}_j\mathbf{X}_j^\top(\widehat{u}_{\tau_n}(Z_j) - \mathbf{X}_j^\top\widehat{\beta}_{n,N})^2]$$

respectively. This implies that the eigenvalues of $\widehat{D}(\widehat{\beta}_{n,N}) = \widehat{V}_0^{-1}(\widehat{\beta}_{n,N})\widehat{V}_1(\widehat{\beta}_{n,N})$, $\widehat{w}_i$ say, estimate w_i consistently for $i = 1, 2, \cdots, p$. Let $\widehat{c}_\alpha$ be the $1-\alpha$ quantile of the conditional distribution of the weighted sum $\widehat{S} = \widehat{w}_1\chi^2_{1,1} + \cdots + \widehat{w}_p\chi^2_{1,p}$ given the data. Then the confidence region for β with asymptotically correct coverage probability $1-\alpha$ can be defined to be

$$\widehat{I}_\alpha(\widetilde{\beta}) = \{\widetilde{\beta} : \widehat{l}_{n,N}(\widetilde{\beta}) \le \widehat{c}_\alpha\}.$$

In practice, the conditional distribution of the weighted sum $\widehat{S}$ given the data $\{(\mathbf{X}_i, Y_i, \widetilde{Y}_i)_{i=1}^n\}$ and $\{(\mathbf{X}_j, \widetilde{Y}_j)_{j=n+1}^{n+N}\}$ can be obtained using Monte Carlo simulations by repeatedly generating independent samples $\chi^2_{1,1}, \cdots, \chi^2_{1,p}$ from the χ^2_1 distribution.

In the absence of measurement error, $D(\beta)$ reduces to an identity matrix so that $w_i = 1$ for $1 \le i \le p$ and Theorem 3.1 reduces to Owen's (1991) result that the empirical loglikelihood is asymptotically χ^2_p. Next, we define an adjusted empirical log-likelihood whose asymptotic distribution is a standard chi-square.

Let

$$\widehat{H}(\beta) = \{\sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\}\{\sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\}^\top.$$

By examing the asymptotic expansion of $\widehat{l}_{n,N}(\beta)$, we define an adjusted empirical loglikelihood by

$$\widehat{l}_{ad}(\beta) = \widehat{r}(\beta)\widehat{l}_{n,N}(\beta), \tag{3.4}$$

which can be proved to be asymptotically χ^2_p, where

$$\widehat{r}(\beta) = \frac{\mathrm{tr}(\widehat{V}^{-1}(\beta)\widehat{H}(\beta))}{\mathrm{tr}(\widehat{V}_0^{-1}(\beta)\widehat{H}(\beta))}.$$

THEOREM 3.2. Under the regularity conditions given in the appendix, if β is the true value of the parameter, we have

(a) as $n \to \infty$

$$\widehat{l}_{ad}(\beta) \xrightarrow{\mathcal{L}} \chi^2_p,$$

where χ^2_p is a standard chi-square random variable with p degrees of freedom.

(b)

$$\lim_{n\to\infty} P(\beta \in \widehat{I}_{ad,\alpha}(\widetilde{\beta})) = 1 - \alpha,$$

where $\widehat{I}_{ad,\alpha}(\widetilde{\beta}) = \{\widetilde{\beta} : \widehat{l}_{ad}(\widetilde{\beta}) \le \chi^2_{p,\alpha}\}$ with $\chi^2_{p,\alpha}$ the $1-\alpha$ quantile of the χ^2_p distribution for $0 < \alpha < 1$.

4 Estimated and adjusted empirical likelihoods for linear combinations of β

In practice, statisticians are often confronted with the problem of constructing confidence intervals or regions for a particular regression coefficient, a subvector of β or linear combinations of β. To address this problem, we develop empirical likelihood method to make inference for a vector of linear combinations $\theta = C\beta$ of β, where $C = (C_1, C_2)$, C_1 is a $k \times k$ matrix and C_2 is a $k \times (p-k)$ matrix. For example, θ is the subvector of the first k regression coefficients if $C_1 = I_k$ and $C_2 = 0$. If $k = 1$, then θ reduces to a single linear combination, which includes an individual regression coefficients and the mean response at a given X level as special cases. Without loss of generality, we assume that C_1^{-1} exists.

Let $\gamma = (\theta^\top, \beta_{0(k)}^\top)^\top$, where $\beta_{0(k)}$ denotes the column subvector of the last $p-k$ elements of β. Let $\mathbf{X}_i = (\mathbf{X}_{i1}^\top, \mathbf{X}_{i2}^\top)^\top$, where $\mathbf{X}_{i1}$ and $\mathbf{X}_{i2}$ are $k \times 1$ and $(p-k) \times 1$ subvectors. Let $\widetilde{\mathbf{X}}_i = (\widetilde{X}_{i1}^\top, \widetilde{X}_{i2}^\top)^\top = (\mathbf{X}_{i1}^{-1} C_1^{-1}, \mathbf{X}_{i2}^\top - \mathbf{X}_{i1}^\top C_1^{-1} C_2)^\top$. Then, model (1.2) reduces to

$$u(Z_j) = \widetilde{\mathbf{X}}_j^\top \gamma + \epsilon, j = n+1, \cdots, n+N.$$

If $u(\cdot)$ was known, the LSE of γ can be defined to be $\widetilde{\gamma}_{n,N} = (\sum_{j=n+1}^{n+N} \widetilde{\mathbf{X}}_j \widetilde{\mathbf{X}}_j^\top)^{-1}$ $(\sum_{j=n+1}^{n+N} \widetilde{\mathbf{X}}_j u(Z_j))$. Let $\widetilde{\beta}_{n(k)}$ denote the subvector of the last $p-k$ elements of $\widetilde{\gamma}_{n,N}$. We have

$$E\{\widetilde{X}_{j1}(u(Z_j) - \widetilde{X}_{j1}^\top \theta - \widetilde{X}_{j2}^\top \widetilde{\beta}_{n(k)}\} = 0, j = n+1, \cdots, n+N.$$

Let $\widehat{\gamma}_{n,N}$ be $\widetilde{\gamma}_{n,N}$ with $u(\cdot)$ replaced by $\widehat{u}_{\tau_n}(\cdot)$. Let $\widehat{\beta}_{n(k)}$ denote the subvector of the last $p-k$ elements of $\widehat{\gamma}_{n,N}$.

Similar to the previous section, for a given θ, we introduce the following auxiliary variables

$$\widehat{W}_j(\theta) = \widetilde{X}_{j1}(\widehat{u}_{\tau_n}(Z_j) - \widetilde{X}_{j1}^\top \theta - \widetilde{X}_{j2}^\top \widehat{\beta}_{n(k)})$$

and define an estimated empirical log-likelihood function

$$\widetilde{l}_{n,N}(\theta) = 2 \sum_{j=n+1}^{n+N} \log(1 + \zeta^\top \widehat{W}_j(\theta)),$$

where ζ is the solution of the following equation

$$\sum_{j=n+1}^{n+N} \frac{\widehat{W}_j(\theta)}{1 + \zeta^\top \widehat{W}_j(\theta)} = 0.$$

Let $\widetilde{\mathbf{X}} = (\widetilde{X}_1^\top, \widetilde{X}_2^\top)^\top$ and

$$\begin{aligned}
K &= E(\widetilde{X}_1 \widetilde{X}_2^\top) \\
P &= E(\widetilde{X}_2 \widetilde{X}_2^\top) \\
\eta &= \widetilde{X}_1 - E(\widetilde{X}_1 \widetilde{\mathbf{X}}^\top)\{E[\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top]\}^{-1}\widetilde{\mathbf{X}} \\
&\quad + E[\widetilde{X}_1 \widetilde{X}_1^\top]\{E(\widetilde{X}_1 \widetilde{X}_1^\top) - K^\top P^{-1} K\}^{-1}(\widetilde{X}_1 - K^\top P^{-1} \widetilde{X}_2), \\
V_0^*(\theta) &= E[\widetilde{X}_1 \widetilde{X}_1^\top (u(Z) - \widetilde{X}_1^\top \theta - \widetilde{X}_2^\top \beta_{(k)})^2], \\
V_1^*(\theta) &= E[(u(Z) - \widetilde{X}_1^\top \theta - \widetilde{X}_2^\top \beta_{(k)})^2 \eta\eta^\top]] + \lambda E[(Y - u(Z))^2 \eta\eta^\top]
\end{aligned}$$

THEOREM 4.1 Under the assumptions listed in the Appendix, we have

$$\widetilde{l}_{n,N}(\theta) \xrightarrow{\mathcal{L}} \sum_{i=1}^{k} \widetilde{w}_{1,i}\chi_{1,i},$$

where $\widetilde{w}_{1,i}, 1 \le i \le k$ are the eigenvalues of $V_0^{*-1}(\theta)\widetilde{V}_1^*(\theta)$ and $\chi_{1,i}^2$ is independent standard χ_1^2 variables for $i = 1, 2, \cdots, k$.

Let

$$\begin{aligned}
K_N &= \frac{1}{N}\sum_{j=n+1}^{n+N} \widetilde{X}_{j1}\widetilde{X}_{j2}^\top, \\
P_N &= \frac{1}{N}\sum_{j=n+1}^{n+N} \widetilde{X}_{j2}\widetilde{X}_{j2}^\top, \\
\eta_j &= \widetilde{X}_{j1} - \left(\frac{1}{N}\sum_{j=n+1}^{n+N} \widetilde{X}_{j1}\widetilde{\mathbf{X}}_j^\top\right)\left(\frac{1}{N}\sum_{j=n+1}^{n+N} \widetilde{\mathbf{X}}_j\widetilde{\mathbf{X}}_j^\top\right)^{-1}\widetilde{\mathbf{X}}_j, \\
&\quad + \left(\frac{1}{N}\sum_{j=n+1}^{n+N} \widetilde{X}_{j1}\widetilde{X}_{j1}^\top\right)\left(\frac{1}{N}\sum_{j=n+1}^{n+N} \widetilde{X}_{j1}\widetilde{X}_{j1}^\top - K_N^\top P_N^{-1}K_N\right)^{-1}\left(\widetilde{X}_{j1} - K_N^\top P_N^{-1}\widetilde{X}_{j2}\right), \\
\widehat{V}_0^*(\theta) &= \frac{1}{N}\sum_{j=n+1}^{n+N} \widetilde{\mathbf{X}}_j\widetilde{\mathbf{X}}_j^\top(\widehat{u}_{\tau_n}(Z_j) - \widetilde{X}_{j1}^\top\theta - \widetilde{X}_{j2}^\top\widehat{\beta}_{(k)})^2, \\
\widehat{V}_1^*(\theta) &= \frac{1}{N}\sum_{j=N+1}^{n+N} [(\widehat{u}_{\tau_n}(Z_j) - \widetilde{X}_{j1}^\top\theta - \widetilde{X}_{j2}^\top\widehat{\beta}_{(k)})^2\widehat{\eta}_j\widehat{\eta}_j^\top] \\
&\quad + \frac{N}{n^2}\sum_{i=1}^{n}[(Y_i - \widehat{u}_{\tau_n}(Z_i))^2\widehat{\eta}_i\widehat{\eta}_i^\top]
\end{aligned}$$

and

$$\widetilde{H}(\theta) = \left(\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N} W_j(\theta)\right)\left(\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N} W_j(\theta)\right)^\top.$$

An adjusted empirical loglikelihood is then defined by

$$\widetilde{l}_{ad}(\theta) = \widetilde{r}(\theta)\widetilde{l}_{n,N}(\theta),$$

where

$$\widetilde{r}_n(\theta) = \frac{\mathrm{tr}(\widehat{V}_1^{*-1}(\theta)\widetilde{H}(\theta))}{\mathrm{tr}(\widehat{V}_0^{*-1}(\theta)\widetilde{H}(\theta))}.$$

THEOREM 4.2 Assume the assumptions listed in the Appendix, if θ is the true value of the parameter,we have

(a) as $n \to \infty$

$$\widetilde{l}_{ad}(\theta) \xrightarrow{\mathcal{L}} \chi^2_k,$$

where χ^2_k is a standard chi-square random variable with k degrees of freedom.

(b)

$$\lim_n P(\theta \in \widetilde{I}_{ad,a}(\widetilde{\theta})) = 1 - \alpha,$$

where $\widetilde{I}_{ad,\alpha}(\widetilde{\theta}) = \{\widetilde{\theta} : \widetilde{l}_{ad}(\widetilde{\theta}) \leq \chi_{k,\alpha}\}$ with $\chi^2_{k,\alpha}$ the $1-\alpha$ quantile of the χ^2_k distribution for $0 < \alpha < 1$.

5 Simulation Studies

We conducted simulation to better understand the finite-sample performances of the proposed inferential procedures.

In our simulation studies, we consider the two cases of $p = 1$ and $p = 2$. For the case of $p = 1$, The surrogates $\widetilde{Y}$ were generated as the standard normal random variables. The linear model considered was $Y = \mathbf{X}^\top \beta + e$, where $\beta = 2.30$ and $\mathbf{X}$ was generated from a standard exponential distribution, while e given $Z = (\mathbf{X}, \widetilde{Y})$ was normally distributed with mean $(\alpha^\top Z)^2 - 2.30\mathbf{X} + 0.69$ and variance $\sigma^2 = 1$, where $\alpha = (1.23, 0.32)^\top$. We estimate α using α_n given in Section 2. The simulation were run with validation data and primary data sizes of $(n, N) = (10, 30), (30, 90), (60, 180), (10, 50), (30, 150)$ and $(60, 300)$. The bandwidths $h_{1,n} = n^{-\frac{15}{96}}$ and $h_{2,n} = n^{-\frac{2}{5}}$, and the kernel functions $K_1(\cdot)$ and $K_2(\cdot)$ are taken to be

$$K_1(u) = \begin{cases} -\frac{15}{8}u^2 + \frac{9}{8}, & -1 \leq u \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

and

$$K_2(u) = \begin{cases} \frac{15}{16}(1 - 2u^2 + u^4), & -1 \leq u \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

b_n and τ_n were taken to be $n^{-\frac{1}{24}}$ respectively. We calculated the coverage probabilities and the average lengths of the confidence intervals, with nominal level 0.90

and 0.95, respectively, by using 5000 simulation runs. The simulation results are presented in Tables 1 and 2.

Insert Tables 1 and 2 here

From Tables 1 and 2, the estimated and adjusted empirical log-likelihood confidence intervals have higher coverage accuracies and shorter average lengths than the normal approximation based confidence intervals. It is easily observed that the normal approximation based confidence intervals are consistently over-covering, but they do this by using long intervals. The adjusted empirical log-likelihood performs slightly better than the estimated empirical log-likelihood in terms of coverage accuracies and average lengths.

For the case of $p = 2$. The surrogates $\widetilde{Y}$ were generated as the standard normal random variables. Consider the linear model (1.1) with $\beta = (-1.24, 3.10)^\top$, where $\mathbf{X}$ was generated from a standard bivariate normal distribution with correlation coefficient $\rho = 0.62$; while e given $Z = (X^\top, \widetilde{Y})^\top$ was normally distributed with mean $(\alpha^\top Z)^2 - \mathbf{X}^\top\beta - 4.79$ and variance $\sigma^2 = 1$, where $\alpha = (0.25, -1.31, 1.85)^\top$. We estimate α using α_n given in Section 2. The simulation were run with the same validation data and primary data sizes as in the case of $p = 1$. Also, $h_{1,n}$, $h_{2,n}$, $K_1(\cdot)$, $K_2(\cdot)$, b_n and τ_n were taken to be the same as in the case of $p = 1$. We calculated the coverage probabilities of the confidence intervals, with nominal level 0.90 and 0.95, respectively, by using 5000 simulation runs. The simulation results are reported in Tables 3.

Insert Table 3 here

From Table 3, the normal approximation method leads to significantly lower coverage probabilities than the norminal levels when sample sizes are not large. The estimated and adjusted empirical log-likelihood methods outperform the normal approximation based methods in terms of coverage accuracy when sample sizes are small or moderate. Generally, the adjusted empirical log-likelihood behaves better slightly than the estimated empirical log-likelihood.

From Tables 1, 2 and 3, all the methods perform better in terms of coverage accuracies and average lengths as N increases with n constants. However, this improvement is small compared to increasing both n and N.

6 Appendix

(C.u): $u(\cdot)$ has bounded partial derivatives of order two

(C.X): $E[X_{ir}^4] < \infty, r = 1, 2, \cdots, p$

(C.$\widetilde{Y}$): $E|\widetilde{Y}|^4 < \infty$

(C.Y): $\sup_{z\in\mathcal{Z}} E[Y^2|Z = z] < \infty$

(C.e)i: $E[e|Z] = 0$

ii: $\sup_{z\in\mathcal{Z}} E[e^2|Z = z] < \infty$

(C.Σ): $E[\mathbf{X}\mathbf{X}^\tau]$ is a positive definite matrix

(C.K_1)i: $K_1(\cdot)$ is symmetric about 0 with bounded support $[-1, 1]$

ii: $\int_{-1}^1 K_1(u)\,du = 1$ and $\int_{-1}^1 u^i K_1(u)\,du = 0, i = 1, 2, 3$

(C.$h_{1,n}$): As $n \to \infty, h_{1,n} \sim n^{-c_1}$, $b \sim n^{-c_2}$ with positive numbers c_1 and c_2 satisfying that $\frac{1}{8} + \frac{c_2}{4} < c_1 < \frac{1}{4} - c_2$, and the notation "$\sim$" means that two quatities have the same coverage order.

(C.K_2): $K_2(\cdot)$ is a bounded nonnegative kernel function of order one with bounded support.

(C.$h_{2,n}$)i: $nh_{2,n}^{\frac{3}{2}}\tau_n \to \infty$

ii: $\frac{nh_{2,n}^3}{\tau_n} \to 0$

(C.f)i: $\sqrt{N}E\{\|\mathbf{X}u(Z)\|I[f_{\mathcal{Z}}(Z) \leq \tau_n]\} \longrightarrow 0$

ii: $f_{\mathcal{Z}}(z)$ has bounded partial derivatives of order one.

(C.Nn): $\frac{N}{n} \to \lambda$, where λ is a nonnegative constant

(C.R^*): $R_i^*(y)$ for $i = 1, 2, \cdots, p+2$, and $f_Y(y)$ are 3-times differentable and their third derivatives satisfy the following conditions: there exists a neighborhood of the orign, say U_1, and a constant $c > 0$ such that for any $u \in U$

$$\begin{aligned} |f_Y^{(3)}(y+u) - f_Y^{(3)}(u)| &\leq c|u|, \\ |R_i^{*(3)}(y+u) - R_i^{*(3)}(u)| &\leq c|u|, i = 1, 2, \cdots, p \end{aligned}$$

(C.R)i: For pair $1 \leq i, l \leq p+2$ and for any $u \in U$

$$|R_i(y+u)R_l(y+u) - R_i(y)R_l(y)| \leq c|u|$$

ii: $\sqrt{n}ER_i(Y)R_l(Y)I[f_Y(Y) < b] = o(1)$ as $n \to \infty$, for $1 \leq i, l \leq d$, where $I[\cdot]$ is the indicator function and b satisfies ($C.h_{1,n}$)

Remark: (C.X), (C.Y), (C.K_1), (C.$h_{1,n}$), (C.R^*), (C.R) are used in Zhu and Fang (1997) to obtain the result $\widehat{\alpha}_n - \alpha = O_p(n^{-\frac{1}{2}})$. The remaining conditions are standard.

Clearly

$$\widehat{\beta}_{n,N} - \beta = \Sigma_{n,N}^{-1} A_{n,N} \tag{A.1}$$

and

$$\Sigma_{n,N} \xrightarrow{p} \Sigma, \tag{A.2}$$

where $\Sigma_{n,N}$ and Σ are defined in Section 2 and

$$A_{n,N} = \frac{1}{N} \sum_{k=n+1}^{n+N} \mathbf{X}_k(\widehat{u}_{\tau_n}(Z_k) - \mathbf{X}_k^\top \beta)$$

To prove Theorem 2.1, we need prove the following lemmas

Lemma A.1 Under the assumptions of Theorem 2.1, we have

$$\begin{aligned} A_{n,N} &= \tfrac{1}{N}\textstyle\sum_{k=n+1}^{n+N} \mathbf{X}_k(u(Z_k) - \mathbf{X}_k^\top \beta) \\ &\quad + \tfrac{1}{nN}\textstyle\sum_{k=n+1}^{n+N}\sum_{i=1}^{n} \frac{\mathbf{X}_k(Y_i - u(Z_k))K_2\left(\frac{\alpha^\top(Z_i - Z_k)}{h_{2,n}}\right)}{h_{2,n} f_{\mathcal{Z}}(Z_k)} + o_p(N^{-\frac{1}{2}}) \end{aligned} \tag{A.3}$$

Let $\widetilde{u}_{\tau_n}(\cdot)$ be $\widehat{u}_{\tau_n}(\cdot)$ with α_n replaced By α. Let $f_{\tau_n}(\cdot) = \max\{f_{\mathcal{Z}}(\cdot), \tau_n\}$ and $u_{\tau_n}(\cdot) = \frac{u(\cdot)f_{\mathcal{Z}}(\cdot)}{f_{\tau_n}(\cdot)}$. Then, we have

$$\begin{aligned} A_{n,N} &= \tfrac{1}{N}\sum_{k=n+1}^{n+N}\mathbf{X}_k(u(Z_k)-\mathbf{X}_k^\top\beta) \\ &\quad +\tfrac{1}{N}\sum_{k=n+1}^{n+N}\mathbf{X}_k(u_{\tau_n}(Z_k)-u(Z_k)) \\ &\quad +\tfrac{1}{N}\sum_{k=n+1}^{n+N}\mathbf{X}_k(\widetilde{u}_{\tau_n}(Z_k)-u_{\tau_n}(Z_k)) \\ &\quad +\tfrac{1}{N}\sum_{k=n+1}^{n+N}\mathbf{X}_k(\widehat{u}_{\tau_n}(Z_k)-\widetilde{u}_{\tau_n}(Z_k)) \end{aligned} \tag{A.4}$$

Let R, S and T be the second, third and fourth terms on the right hand side of (A.4). For any $\epsilon > 0$, we have

$$\begin{aligned} & P(\sqrt{N}\|R\| > \epsilon) \\ \leq\ & P(\tfrac{1}{\sqrt{N}}\sum_{k=n+1}^{n+N}\|\mathbf{X}_k u(Z_k)\|[f_{\mathcal{Z}}(Z_k) < \tau_n] > \epsilon) \\ \leq\ & \epsilon^{-1}\sqrt{N}E\{\|\mathbf{X}u(Z)\|I[f_{\mathcal{Z}}(Z) < \tau_n]\} \longrightarrow 0 \end{aligned} \tag{A.5}$$

by condition (C.f). This proves

$$R = o_p(n^{-\frac{1}{2}}). \tag{A.6}$$

Let

$$\triangle_n(z) = \widehat{f}_{n,\mathcal{Z}}(z) - f(z)$$

and

$$\triangle_{\tau_n}(z) = \widetilde{f}_{\tau_n,\mathcal{Z}}(z) - f(z).$$

By some algebra, we have

$$\begin{aligned} S &= \tfrac{1}{nN}\sum_{k=n+1}^{n+N}\sum_{i=1}^{n}\frac{\mathbf{X}_k(Y_i-u(Z_k))K_2\left(\frac{\alpha^\top(Z_i-Z_k)}{h_{2,n}}\right)}{h_{2,n}f_{\mathcal{Z}}(Z_k)} \\ &\quad +r_{nN,1}+r_{nN,2}+r_{nN,3}+r_{nN,4}, \end{aligned} \tag{A.7}$$

where

$$\begin{aligned} r_{n,N1} &= -\tfrac{1}{N}\frac{\mathbf{X}_k(\widehat{u}_n(Z_k)-u(Z_k))\widetilde{f}_{n,\mathcal{Z}}(Z_k)(f_{\tau_n,\mathcal{Z}}(Z_k)-f(Z_k))}{f_{\tau_n,\mathcal{Z}}(Z_k)f_{\mathcal{Z}}(Z_k)} \\ r_{n,N2} &= \tfrac{1}{N}\sum_{k=n+1}^{n+N}\frac{\mathbf{X}_k u(Z_k)[f_{\tau_n,\mathcal{Z}}(Z_k)\widehat{f}_{n,\mathcal{Z}}(Z_k)-f_{\mathcal{Z}}(Z_k)\widehat{f}_{b_n}(Z_k)]}{f^2_{\tau_n,\mathcal{Z}}(Z_k)} \\ r_{n,N3} &= -\tfrac{1}{N}\sum_{k=n+1}^{n+N}\frac{\mathbf{X}_k(\widetilde{u}_n(Z_k)\widetilde{f}_{n,\mathcal{Z}}(Z_k)-u(Z_k)f_{\mathcal{Z}}(Z_k))\triangle_{\tau_n}(Z_k)}{f^2_{\tau_n,\mathcal{Z}}(Z_k)} \\ r_{n,N4} &= \tfrac{1}{N}\sum_{k=n+1}^{n+N}\frac{\mathbf{X}_k\widetilde{u}_n(Z_k)\widetilde{f}_n(Z_k)\triangle^2_{b_n}(Z_k)}{f^2_{\tau_n}(Z_k)\widetilde{f}_{\tau_n}(Z_k)} \end{aligned}$$

Note that

$$r_{n,N1} = \frac{1}{nN} \sum_{k=n+1}^{n+N} \sum_{i=n}^{n} \frac{\mathbf{X}_k (Y_i - u(Z_k)) K_2 \left(\frac{\alpha^\top (Z_i - Z_k)}{h_{2,n}} \right)}{h_{2,n} f_{\tau_n, \mathcal{Z}}(Z_k) f_{\mathcal{Z}}(Z_k)} (f_{\tau_n}(Z_k) - f(Z_k)).$$

Hence, by (C.f) and (C.K_2) and some standard arguments we have

$$\begin{aligned} \sqrt{N} E \|r_{n,N1}\| &\leq \frac{\sqrt{N}}{n} \sum_{i=1}^{n} E\{\|\mathbf{X}_i (Y_i - u(Z_i))\| I[f(Z_i) < \tau_n]\} \\ &\leq \sqrt{N} E\{\|\mathbf{X}(Y - u(Z))\| I[f_{\mathcal{Z}}(Z) < \tau_n]\} \longrightarrow 0 \end{aligned} \tag{A.8}$$

Let

$$C_n(Z_k) = \frac{\mathbf{X}_k u(Z_k) (f_{\tau_n}(Z_k) \widehat{f}_{n,\mathcal{Z}}(Z_k) - f_{\mathcal{Z}}(Z_k) \widehat{f}_{\tau_n,\mathcal{Z}}(Z_k))}{f^2_{\tau_n}(Z_k)}$$

Then, we have

$$\begin{aligned} r_{n,N2} &= \frac{1}{N} \sum_{k=n+1}^{n+N} C_n(Z_k) I[f_{\mathcal{Z}}(Z_k) < \tau_n, \widehat{f}_{n,\mathcal{Z}}(Z_k) \leq -\tau_n] \\ &\quad + \frac{1}{N} \sum_{k=n+1}^{n+N} C_n(Z_k) I[f_{\mathcal{Z}}(Z_k) < \tau_n, -\tau_n < \widehat{f}_{n,\mathcal{Z}}(Z_k) < \tau_n] \\ &\quad + \frac{1}{N} \sum_{k=n+1}^{n+N} C_n(Z_k) I[f_{\mathcal{Z}}(Z_k) \geq 2\tau_n, \widehat{f}_{n,\mathcal{Z}}(Z_n) < \tau_n] \\ &\quad + \frac{1}{N} \sum_{k=n+1}^{n+N} C_n(Z_k) I[\tau_n \leq f_{\mathcal{Z}}(Z_k) < 2\tau_n, \widehat{f}_{n,\mathcal{Z}}(Z_k) < \tau_n] \\ &\quad + \frac{1}{N} \sum_{k=n+1}^{n+N} \sum_{k=n+1}^{n+N} C_n(Z_k) I[f_{\mathcal{Z}}(Z_k) < \tau_n, \widehat{f}_{n,\mathcal{Z}}(Z_k) \geq 2\tau_n] \\ &\quad + \frac{1}{N} \sum_{k=n+1}^{n+N} C_n(Z_k) I[f_{\mathcal{Z}}(Z_k) < \tau_n, \tau_n \leq \widehat{f}_{n,\mathcal{Z}}(Z_k) < 2\tau_n) \\ &:= \sum_{i=1}^{6} J_{n,Ni} \end{aligned} \tag{A.9}$$

For any $\epsilon > 0$, we have

$$P(\sqrt{N} |J_{n,N1}| > \epsilon) \leq P(\sup_z |\widehat{f}_{n,\mathcal{Z}}(z) - f_{\mathcal{Z}}(z)| > \tau_n). \tag{A.10}$$

By some standard arguments, it can be shown that

$$P(\sup_z |\widehat{f}_{n,\mathcal{Z}}(z) - f_{\mathcal{Z}}(z)| > \tau_n) \longrightarrow 0 \tag{A.11}$$

by (C.K_2) and (C.$h_{2,n}$). This together with (A.10) proves

$$|J_{n,N1}| = o_p(N^{-\frac{1}{2}}). \tag{A.12}$$

It is easy to see that $|f_{\tau_n}(Z_k)\widehat{f}_n(Z_k) - f_{\mathcal{Z}}(Z_k)\widehat{f}_{\tau_n}(Z_k)| \leq 2\tau_n^2$ as $f(Z_k) < \tau_n$ and $-\tau_n \leq \widehat{f}_n(Z_k) \leq \tau_n$. Hence, we have

$$|J_{n,N2}| \leq \frac{2}{N}\sum_{k=n+1}^{n+N} \|\mathbf{X}_k u(Z_k)\| I[f(Z_k) < \tau_n]$$

By Markov inequality and condition (C.f), we get

$$J_{n,N2} = o_p(N^{-\frac{1}{2}}). \tag{A.13}$$

If $f(Z_i) > \tau_n$ and $\widehat{f}_n(Z_i) < \tau_n$, we have

$$|f_{\tau_n}(Z_i)\widehat{f}_n(Z_i) - f(Z_i)\widehat{f}_{\tau_n}(Z_i)| \leq \tau_n f(Z_i). \tag{A.14}$$

This together with the fact $f_{\tau_n}(Z_i) \geq \tau_n$ and $f_{\tau_n}(Z_i) \geq f(Z_i)$ proves

$$|J_{n,N3}| \leq \frac{1}{N}\sum_{k=n+1}^{n+N} \|\mathbf{X}_k u(Z_k)\| I[f_{\mathcal{Z}}(Z_k) \geq 2\tau_n, \widehat{f}_{n,\mathcal{Z}}(Z_k) < \tau_n] \tag{A.15}$$

and

$$\begin{aligned} |J_{n,N4}| &\leq \tfrac{1}{N}\sum_{k=n+1}^{n+N} \|\mathbf{X}_k u(Z_k)\| I[\tau_n < f(Z_k) < 2\tau_n, \widehat{f}_{n,\mathcal{Z}}(Z_I) < \tau_n] \\ &\leq \tfrac{1}{N}\sum_{k=n+1}^{n+N} \|\mathbf{X}_k u(Z_k)\| I[f_{\mathcal{Z}}(Z_k) \leq 2\tau_n]. \end{aligned} \tag{A.16}$$

By (A.15) and (A.11), for any $\epsilon > 0$ we have

$$P(\sqrt{N}|J_{n,N3}| > \epsilon) \leq P[\sup_z |\widehat{f}_{n,\mathcal{Z}}(z) - f_{\mathcal{Z}}(z)| > \tau_n) \to 0.$$

This proves

$$J_{n,N3} = o_p(N^{-\frac{1}{2}}). \tag{A.17}$$

Similarly, we have

$$J_{n,N5} = o_p(N^{-\frac{1}{2}}). \tag{A.18}$$

By (A.16) and (C.f), similar to (A.13) we have

$$J_{n,N4} = o_p(N^{-\frac{1}{2}}). \tag{A.19}$$

Similarly, we have

$$|J_{n,N6}| = o_p(N^{-\frac{1}{2}})$$

This together with (A.9), (A.12), (A.13), (A.16)-(A.20) together prove

$$r_{n,N2} = o_p(N^{-\frac{1}{2}}). \tag{A.20}$$

Let

$$\begin{aligned} A_n(z) &= (nh_{2,n})^{-1} \sum_{i=1}^{n} (Y_i - u(Z_i)) K_2\left(\tfrac{\alpha^\top (z - Z_i)}{h_{2,n}}\right) \\ B_n(z) &= (nh_{n,2})^{-1} \sum_{i=1}^{n} (u(Z_i) - u(z)) K_2\left(\tfrac{\alpha^\top (z - Z_i)}{h_{2,n}}\right) \end{aligned}$$

Then, we have

$$\begin{aligned} r_{n,N3} &= -\tfrac{1}{N} \sum_{k=n+1}^{n+N} \tfrac{\mathbf{X}_k A_n(Z_k) \triangle_{\tau_n}(Z_k)}{f^2_{\tau_n,\mathcal{Z}}(Z_k)} - \tfrac{1}{N} \sum_{k=n+1}^{n+N} \tfrac{\mathbf{X}_k B_n(Z_k) \triangle_{\tau_n}(Z_k)}{f^2_{\tau_n,\mathcal{Z}}(Z_k)} \\ &\quad -\tfrac{1}{N} \sum_{k=n+1}^{n+N} \tfrac{\mathbf{X}_k u(Z_k) \triangle_n(Z_k) \triangle_{\tau_n}(Z_k)}{f^2_{\tau_n,\mathcal{Z}}(Z_k)} := r_{n,N31} + r_{n,N32} + r_{n,N33} \end{aligned} \tag{A.21}$$

Clearly,

$$|r_{n,N31}| \leq \left(\frac{1}{N} \sum_{k=n+1}^{n+N} \Big| \frac{A_n(Z_k)\mathbf{X}_k}{f^2_{\tau_n,\mathcal{Z}}(Z_k)} \Big| \right) \sup_z |\triangle_{\tau_n}(z)|.$$

Standard arguments can be used to prove that

$$\frac{1}{N} \sum_{k=n+1}^{n+N} \Big| \frac{A_n(Z_k)\mathbf{X}_k}{f^2_{\tau_n,\mathcal{Z}}(Z_k)} \Big| = O_p((nh_{n,2})^{-\frac{1}{2}} \tau_n^{-2}) \tag{A.22}$$

and

$$\sup_z |\triangle_{\tau_n}(z)| \leq \sup_z |\triangle_n(z)| = O_p((nh_{2,n})^{-\frac{1}{2}}) + O_p(h^2_{2,n}) \tag{A.23}$$

Hence, (A.22) and (A.23) together with (C.$h_{2,n}$) prove

$$r_{n,N31} = o_p(N^{-\frac{1}{2}}). \tag{A.24}$$

Clearly,

$$\|r_{n,N32}\| \leq \frac{1}{N} \left\| \sum_{k=n+1}^{n+N} \frac{\mathbf{X}_k B_n(Z_k)}{f_{\tau_n}(Z_k)} \right\| \frac{\sup_z |\triangle_{\tau_n}(z)|}{\tau_n} \tag{A.25}$$

By (C.u), (C.X) and (C.K_2), standard arguments can be used to prove

$$\frac{1}{N} \left\| \sum_{k=n+1}^{n+N} \frac{\mathbf{X}_k B_n(Z_k)}{f_{\tau_n,\mathcal{Z}}(Z_k)} \right\| = O_p((nh_{2,n})^{-\frac{1}{2}} \tau_n^{-2}). \tag{A.26}$$

By (A.23), (A.25), (A.26) and (C.$h_{2,n}$), it follows that

$$\|r_{n,N32}\| = o_p(N^{-\frac{1}{2}}). \tag{A.27}$$

For $r_{n,N33}$, we have

$$\|r_{n,N33}\| \leq \tau_n^{-2}(\sup_z |\triangle_n(z)\|)^2 \left(\frac{1}{N}\sum_{k=n+1}^{n+N} \|\mathbf{X}_k u(Z_k)\|\right).$$

By (A.23) and conditions (C.$h_{2,n}$), it follows that

$$\|r_{n,N33}\| = o_p(N^{-\frac{1}{2}}). \tag{A.28}$$

(A.21), (A.24), (A.27) and (A.28) together prove

$$r_{n,N3} = o_p(N^{-\frac{1}{2}}). \tag{A.29}$$

Note that $\widetilde{f}_n(Z_k)/\widetilde{f}_{\tau_n}(Z_k) \leq 1$ and hence we have

$$\|r_{n,N4}\| \leq \frac{1}{N}\sum_{k=n+1}^{n+N} \|\frac{\mathbf{X}_k \widetilde{u}_n(Z_k)}{f^2_{\tau_n}(Z_k)}\triangle_n^2(Z_k)\|.$$

Similar to (A.28), we get

$$r_{n,N4} = o_p(N^{-\frac{1}{2}}). \tag{A.30}$$

By Theorem 2 of Zhu and Fang (1996), we have

$$\alpha_n - \alpha = O_p(n^{-\frac{1}{2}}). \tag{A.31}$$

Hence, the same arguments as those used in the proof of Härdle and Stoke (1981) can be applied to the proof of the following

$$P(\sup_z |\widehat{f}_{\tau_n,\mathcal{Z}}(z) - \widetilde{f}_{\tau_n,\mathcal{Z}}(z)| > \tau_n) \longrightarrow 0 \tag{A.32}$$

as $\tau_n N^{\frac{2}{5}} \to \infty$, which is implied by (C.$h_{2,n}$). Using (A.31), (A.32) and the inequality of Theorem 3.3 of Härdle and Stoke (1989), we have

$$T = o_p(N^{-\frac{1}{2}}). \tag{A.33}$$

By (A.4), (A.6), (A.7), (A.8), (A.20), (A.29), (A.30) and (A.33), we prove Lemma A.1.

Lemma A.2. Under the assumptions of Theorem 2.1, we have

$$\sqrt{N}A_{n,N} \xrightarrow{\mathcal{L}} N(0, V_1(\beta)).$$

Proof. Let $V_i = (Y_i, \widetilde{Y}_i, \mathbf{X}_i)$ and $W_k = (\mathbf{X}_k, \widetilde{Y}_k)$. Let

$$\Psi_n(V_i, W_k; h_{2,n}) = \mathbf{X}_k(u(Z_k) - \mathbf{X}_k^\top \beta) + \frac{\mathbf{X}_k(Y_i - u(Z_k))K_2\left(\frac{\alpha^\top(Z_i - Z_k)}{h_{2,n}}\right)}{h_{2,n} f_{\mathcal{Z}}(Z_k)}$$

and

$$U_{n,N} = \frac{1}{n\sqrt{N}} \sum_{i=1}^{n} \sum_{k=n+1}^{n+N} \Psi(V_i, W_k; h_{2,n}).$$

Clearly, $U_{n,N}$ is a two sample statistic. By (C.u), we have

$$E[\Psi_n(V, W; h_{2,n})|V] \longrightarrow X(Y - u(Z)) \tag{A.34}$$

By (C.u) and (C.K_2), we have

$$\begin{array}{rcl} & & E[\Psi_n(V, W; h_{2,n})|W] \\ & = & X(u(Z) - X^\top \beta) \\ & & + \frac{X \int (u(z) - u(Z))K_2\left(\frac{\alpha^\top(z-Z)}{h_{2,n}}\right) f(z)\, dz}{h_{2,n} f_{\mathcal{Z}}(Z)} \\ & \longrightarrow & X(u(Z) - X^\top \beta) \end{array} \tag{A.35}$$

Clearly,

$$\begin{array}{rcl} E[\Psi_n(V, W; h_{2,n})] & = & E\{E[\Psi_n(V, W; h_{2,n})|W]\} \\ & = & E\frac{\mathbf{X} \int (u(z) - u(Z))K_2\left(\frac{\alpha^\top(z-Z)}{h_{2,n}}\right) f_{\mathcal{Z}}(z)\, dz}{h_{2,n} f_{\mathcal{Z}}(Z)}. \end{array} \tag{A.36}$$

By derivative mean theorem and (C.u), we have

$$\Big| \int (u(z) - u(Z))K_2\left(\frac{\alpha^\top(z - Z)}{h_{2,n}}\right) f_{\mathcal{Z}}(z)\, dz \Big| \le c h_{2,n}^2 f_{\mathcal{Z}}(Z). \tag{A.37}$$

By (A.36) and (A.37), we get

$$\|E\Psi_n(V, W; h_{2,n})\| \le C h_{2,n}. \tag{A.38}$$

Condition (C.$h_{2,n}$) implies $nh_{2,n}^2 \to 0$. This together with (A.38) proves

$$\sqrt{N}E\Psi_n(V, W; h_{2,n}) \longrightarrow 0. \tag{A.39}$$

Lemma B.1 of Sepanski and Lee (1995) together with (A.34), (A.35) and (A.39) proves Lemma A.2.

Proof of Theorem 2.1. Theorem 2.1 is a direct result of (A.1), (A.2), Lemma A.1 and A.2.

Proof of Theoerm 3.1 By Wang and Rao (2002), standard arguments can be used to prove

$$\begin{cases} \max_{n+1\le j\le n+N} \widehat{A}_j(\beta) &= o_p(N^{\frac{1}{2}}), \\ \frac{1}{N}\sum\limits_{j=n+1}^{n+N} \widehat{A}_j^\top(\beta)\widehat{A}_j(\beta) &= O_p(1) \\ \lambda &= O_p(N^{-\frac{1}{2}}). \end{cases} \tag{A.40}$$

Applying Taylor's expansion to (3.3), and using (A.40), we get

$$\widehat{l}_{n,N}(\beta) = 2\sum_{j=n+1}^{n+N}\{\lambda_N^\top \widehat{A}_j(\beta) - \frac{1}{2}(\lambda_N^\top \widehat{A}_j(\beta))^2\} + o_p(1). \tag{A.41}$$

Applying Taylor's expansion to (3.2), and using (A.40), we have

$$\sum_{j=n+1}^{n+N} \lambda_N \widehat{A}_j(\beta) = \sum_{j=n+1}^{n+N} (\lambda_N^\top \widehat{A}_j(\beta))^2 + o_p(1) \tag{A.42}$$

and

$$\lambda_N = \left(\sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\widehat{A}_j^\top(\beta)\right)^{-1} \sum_{j=n+1}^{n+N} \widehat{A}_j(\beta) + o_p(n^{-\frac{1}{2}}). \tag{A.43}$$

(A.41), (A.42) and (A.43) together yield

$$\widehat{l}_{n,N}(\beta) = \{\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\}^\top \widehat{V}_0^{-1}(\beta)\{\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\} + o_p(1). \tag{A.44}$$

where $\widehat{V}_0(\cdot)$ is defined in Section 3.

It can be shown that $\widehat{V}_0(\beta) \xrightarrow{p} V_0(\beta)$ by the fact $\widehat{\alpha}_n - \alpha = O_p(n^{-\frac{1}{2}})$ and some standard arguments. This together with (A.44) proves

$$\widehat{l}_{n,N}(\beta) = \{\frac{1}{\sqrt{N}}V^{-\frac{1}{2}}(\beta)\sum_{j=n+1}^{n+N} \widehat{A}_j(\beta)\}^\top D(\beta)\{\frac{1}{\sqrt{N}}V^{-\frac{1}{2}}(\beta)\sum_{j=n+1}^{n+N} \widehat{A}(\beta)\} + o_p(1) \tag{A.45}$$

where $D(\beta) = V^{\frac{1}{2}}(\beta)V_0^{-1}(\beta)V^{\frac{1}{2}}(\beta)$.

Using arguments similar to Wang and Rao (2002), Theorem 3.1 can be proved by Lemma A.2 and (A.45).

Proof of Theorem 3.2

By Lemma A.2 and the facts $\widehat{V}_1(\beta) \xrightarrow{p} V_1(\beta)$ and $\widehat{V}_0(\beta) \xrightarrow{p} V_0(\beta)$, it can be shown

$$\widehat{r}(\beta) = O_p(1). \tag{A.46}$$

This together with (A.44) proves

$$\widehat{l}_{ad}(\beta) = \{\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widehat{A}_j(\beta)\}\widehat{V}_1^{-1}(\beta)\{\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widehat{A}_j(\beta)\} + o_p(1).$$

By Lemma A.2 and the fact $\widehat{V}_1(\beta) \xrightarrow{p} V_1(\beta)$, Theorem 3.2 (i) is then proved. Theorem 3.2 (ii) is a direct result of (i).

Proof of Theorem 4.1. Similar to (A.44), we have

$$\widetilde{l}_{n,N}(\theta) = \{\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widehat{W}_j(\theta)\}\widehat{V}_0^{*^{-1}}(\theta)\{\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widehat{W}_j(\theta)\} + o_p(1), \tag{A.47}$$

where $\widehat{V}_0^*(\theta)$ is defined in Section 4.

Observe that

$$\begin{aligned}\tfrac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widehat{W}_j(\theta) &= \tfrac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widetilde{X}_{j1}(\widehat{u}_{\tau_n}(Z_j) - \widetilde{X}_j^\top\gamma) \\ &\quad + \tfrac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widetilde{X}_{j1}\widetilde{\mathbf{X}}_j^\top(\gamma - \widehat{\gamma}_n) \\ &\quad + \tfrac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widetilde{X}_{j1}\widetilde{X}_{j1}^\top(\widehat{\theta}_n - \theta) \\ &:= T_{N1} + T_{N2} + T_{N3}\end{aligned} \tag{A.48}$$

Standard arguments can be used to get

$$T_{N2} = -\{E[\widetilde{X}_{j1}\widetilde{\mathbf{X}}_j^\top]\}\{E\widetilde{\mathbf{X}}_j\widetilde{\mathbf{X}}_j^\top\}\{\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widetilde{\mathbf{X}}_j(u_{\tau_n}(Z_j) - \widetilde{\mathbf{X}}_j^\top\gamma)\} + o_p(1) \tag{A.49}$$

and

$$\begin{aligned}T_{N3} &= E\{\widetilde{X}_{j1}\widetilde{X}_{j1}^\top\}\{E(\widetilde{X}_{j1}\widetilde{X}_{j1}^\top) - K^\top P^{-1}K\}^{-1} \\ &\quad \times\big[\tfrac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}(\widetilde{X}_{j1} - K^\top P^{-1}\widetilde{X}_{j2})(\widehat{u}_{\tau_n}(Z_j) - \widetilde{\mathbf{X}}_j^\top\gamma)] + o_p(1).\end{aligned} \tag{A.50}$$

It follows from (A.48)—(A.50)

$$\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\widehat{W}_j(\theta)=\frac{1}{\sqrt{N}}\sum_{j=n+1}^{n+N}\eta_j(\widehat{u}_{\tau_n}(Z_j)-\widetilde{\mathbf{X}}_j^{\top}\gamma)+o_p(1)\xrightarrow{\mathcal{L}}N(0,V_1^*(\theta)),\quad(A.51)$$

where $V_1^*(\theta)$ is as defined in Section 4. Note that $\widehat{V}_0^*(\theta)\xrightarrow{p}V_0^*(\theta)$, where $V_0^*(\theta)$ is defined in Section 4. This together with (A.47) proves Theorem 4.1 by arguments similar to Lemmas A.1 and A.2.

Proof of Theorem 4.2 The arguments are similar to that of Theorem 3.2 from Theorem 4.1.

Acknowledgements. The research was supported by the National Natural Science Foundation of China (key grant 10231030) and Humboldt-Universität Berlin–Sonderforschungsbereich 373.

REFERENCES

Buonaccorsi, J.P. (1996). Measurement error in the response in the general linear model. J. Amer. Statist. Assoc., 91(434), 633-642.

Carroll R.J. and Stefanski, L.A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. J. Amer. Statist. Assoc., 85,652-663.

Carroll, R.J., Knickerbocker, R. K. and Wang, C. Y.(1995) Dimension reduction in a semiparametric regression model with errors in covariates. The Annals of Statistics 23 161-181.

Carroll,R.J. and Wand, M.P.(1991). Semiparametric estimation in logistic measure error models. J. Roy. Statist.soc., Ser B **53** 652-663.

Carroll, R.J., Ruppert,D. and Stefanski, L.W. (1995). Measurement Error in Nonlinear Models. Chapman and Hall, New York.

Duan,N and Li, K.C.(1991). Slicing regression: a link-free regression method. Ann. Statist. 19, 505-530.

Duncan, G. and Hill, D.(1985). An investigations of the extent and consequences of measurement error in labor-economics survey data. Journal of Labor Economics 3 508-532.

Fuller, W.A.(1987). Measurement error models. New York, John Wiley & Sons, Inc.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). J. Amer. Statist. Assoc. 86 337-342.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for single functional. *Biometrika* **75**, 237-249.

Owen, A.(1991). Empirical likelihood for linear models. *Ann. Statist.* **19**, 1725-1747.

Pepe, M.S.(1992). Inference using surrogate outcome data and a validation sample. Biometrika 79 355-365.

Pepe, M.S. and Fleming, T.R.(1991). A general nonparametric method for dealing with errors in missing or surrogate covariate data. J. Amer. Statist. Assoc. 86 108-113.

Pepe, M. S., Reilly, M. and Fleming, T. R. (1994). Auxiliary outcome data and the mean score method. J. Statist. Plan. Inference 42 137-160.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. Biometrika 82, 299-314.

Rosner, B., Willett, W.C. and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Statist. Med, 8, 1075-1093

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

Sepanski, J.H. and Lee, L.F.(1995). Semiparametric Estimation of nonlinear error-in-variables models with validation study. J. Nonparametric Statistics 4 365-394.

Wang, Q.H.(1999). Estimation of partial linear error-in-variables model. Journal of Multivariate Analysis 69 30-64.

Wang, Q.H. (2000). Estimation of linear error-in-covariables models with validation data under random censorship. Journal of Multivariate Analysis 74, 245-266.

Wang, Q.H. and Rao,J.N.K.(2002). Empirical likelihood-based in linear errors-in-covariables models with validation data. Biometrika 89, 345-358.

Wittes, J., Lakatos, E. and Probstfied, J.(1989). Surrogate endpoints in clinical trails: Cardiovascular diseases. Statist. Med. 8, 415-425.

Zhu, L.X. and Fang, K.T.(1996). Asymptotics for kernel estimator of sliced inverse regression. Ann. Statist., 24, 1053-1068.

Table 1. Simulated coverage probabilities of the $1-\alpha$ normal approximation (NA),estimated empirical likelihood (EEL) and adjusted empirical likelihood (ADEL) confidence intervals for β

	$1-\alpha = 90\%$			$1-\alpha = 95\%$		
Sample size	NA	EEL	ADEL	Normal	EEL	ADEL
(10,30)	0.975	0.841	0.847	0.992	0.917	0.925
(30,90)	0.942	0.878	0.889	0.981	0.934	0.937
(60,180)	0.910	0.894	0.897	0.961	0.945	0.947
(10,50)	0.970	0.844	0.833	0.984	0.921	0.928
(30,150)	0.934	0.881	0.894	0.969	0.939	0.941
(60,300)	0.912	0.894	0.899	0.958	0.946	0.945

Table 2. Simulated average lengths of the $1-\alpha$ normal approximation (NA),estimated empirical likelihood (EEL) and adjusted empirical likelihood (ADEL) confidence intervals for β

	$1-\alpha=90\%$			$1-\alpha=95\%$		
Sample size	NA	EEL	ADEL	Normal	EEL	ADEL
(10,30)	2.202	1.685	1.512	2.487	1.714	1.658
(30,90)	1.751	1.450	1.386	2.125	1.543	1.486
(60,180)	1.327	0.971	0.874	1.641	1.072	0.985
(10,50)	2.171	1.600	1.496	2.478	1.682	1.646
(30,150)	1.747	1.446	1.274	2.086	1.539	1.473
(60,300)	1.256	0.968	0.802	1.607	0.998	0.980

Table 3. Simulated coverage probabilities of the $1-\alpha$ normal approximation (NA),estimated empirical likelihood (EEL) and adjusted empirical likelihood (ADEL) confidence intervals for β

	$1-\alpha=90\%$			$1-\alpha=95\%$		
Sample size	NA	EEL	ADEL	Normal	EEL	ADEL
(10,30)	0.847	0.855	0.862	0.894	0.909	0.920
(30,90)	0.859	0.864	0.882	0.918	0.926	0.931
(60,180)	0.892	0.892	0.895	0.939	0.943	0.942
(10,50)	0.842	0.849	0.861	0.903	0.914	0.928
(30,150)	0.863	0.868	0.889	0.922	0.928	0.936
(60,300)	0.894	0.895	0.893	0.942	0.947	0.944