SCIENCE CHINA
Mathematics

• ARTICLES •

July 2015 Vol. 58 No. 7: 1523–1536
doi: 10.1007/s11425-015-4976-6

Corrected empirical likelihood for a class of generalized linear measurement error models

YANG YiPing[1], LI GaoRong[2,3,*] & TONG TieJun[4]

[1]*College of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing* 400067*, China;*
[2]*Beijing Center for Scientific and Engineering Computing, Beijing University of Technology, Beijing* 100124*, China;*
[3]*College of Applied Sciences, Beijing University of Technology, Beijing* 100124*, China;*
[4]*Department of Mathematics, Hong Kong Baptist University, Hong Kong, China*

Email: yeepingyang@gmail.com, ligaorong@bjut.edu.cn, tongt@hkbu.edu.hk

Received March 19, 2014; accepted July 22, 2014; published online February 28, 2015

Abstract Generalized linear measurement error models, such as Gaussian regression, Poisson regression and logistic regression, are considered. To eliminate the effects of measurement error on parameter estimation, a corrected empirical likelihood method is proposed to make statistical inference for a class of generalized linear measurement error models based on the moment identities of the corrected score function. The asymptotic distribution of the empirical log-likelihood ratio for the regression parameter is proved to be a Chi-squared distribution under some regularity conditions. The corresponding maximum empirical likelihood estimator of the regression parameter π is derived, and the asymptotic normality is shown. Furthermore, we consider the construction of the confidence intervals for one component of the regression parameter by using the partial profile empirical likelihood. Simulation studies are conducted to assess the finite sample performance. A real data set from the ACTG 175 study is used for illustrating the proposed method.

Keywords generalized linear model, empirical likelihood, measurement error, corrected score

MSC(2010) Primary 62G05, 62J12; Secondary 62G20

Citation: Yang Y P, Li G R, Tong T J. Corrected empirical likelihood for a class of generalized linear measurement error models. Sci China Math, 2015, 58: 1523–1536, doi: 10.1007/s11425-015-4976-6

1 Introduction

Generalized linear models arise frequently in practice and have attracted considerable research interest, such as Gamma regression, inverse Gaussian regression, Poisson regression and logistic regression. Let $\mu = E(Y|X)$, $\mathrm{Var}(Y|X) = V(\mu, \phi)$, where ϕ is a known dispersion parameter and $V(\cdot)$ is a known variance function. The generalized linear model of Y given X is

$$g(\mu) = X^{\mathrm{T}}\beta, \tag{1.1}$$

where $g(\cdot)$ is a known link function, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a $p \times 1$ vector of unknown regression parameter. Wedderburn [25] considered the quasi-likelihood estimator. Firth [4] studied the efficiency of quasi-likelihood. Chen and Cui [1] improved the efficiency of parameter estimation of the quasi-likelihood by employing the empirical likelihood and incorporating extra constraints. See [16] for more details about generalized linear models.

*Corresponding author

However, the situation occurs frequently in medical research, where the covariates may not be exactly observable. For example, the CD4 cell counts in AIDS dataset (see [11,14]), and the systolic blood pressure and the high density lipoprotein in the Framingham Heart Study dataset (see [15,29]) are measured with errors. If one ignores these measurement errors, the estimators and inference may be biased. We need to correct the resulting bias. Let W be the observed value of the covariate X. We assume an additive measurement error model,

$$W = X + U, \tag{1.2}$$

where $U \sim N(0, \Sigma_{uu})$ is independent of (X, Y). If some elements of X are measured without errors, the corresponding elements of U and related variance components in Σ_{uu} are set to zero. The measurement error problem has been widely studied. Stefanski and Carroll [22] derived the efficient scores in a structural generalized linear measurement-error model. Stefanski [21] and Nakamura [17] obtained the corrected score functions of some generalized linear models, such as linear regression, Gamma regression, inverse Gamma regression and Poisson regression. Stefanski [21] showed that the corrected score for logistic regression does not exist. Huang and Wang [6] proposed consistent functional methods for logistic regression in which some covariates were not accurately ascertainable. Liang et al. [12], Li and Xue [9] and Liang and Li [13] studied partially linear measurement error models. Zhou and Liang [33], and Zhang et al. [31] developed semiparametric profile least-squares method based estimation procedures for semiparametric varying-coefficient partially linear models with error-prone linear covariates. Yi et al. [29,30] discussed the simultaneous inference and bias analysis for longitudinal data with covariate measurement errors and missing responses. Recently, a class of variable selection procedures for measurement error models have been developed, see for example, [13,15,32]. The purpose of this paper is to construct the confidence region of β for a class of generalized linear measurement error models.

To construct the confidence region for β, the direct way is to use the asymptotic normal distribution of the estimator of β by a plug-in estimator of the asymptotic covariance matrix. But this may cause larger errors for the confidence region since the estimator of the asymptotic covariance matrix is very complicated. The asymptotic confidence intervals for β can also be constructed by using a bootstrap method. The main advantage of the bootstrap method is that it does not reply on the asymptotic distribution of β. However, statistical properties of the bootstrap method with measurement error data need to be investigated. This is beyond the scope of this paper. Taking these issues into account, we recommend using the empirical likelihood method to construct the confidence region for β. The empirical likelihood was introduced for linear regression models by Owen [18]. Kolaczyk [7] extended the model framework to generalized linear models. Cui and Chen [2], and Cui and Kong [3] applied the empirical likelihood to linear and semi-linear errors-in-covariables models. Stute et al. [23] discussed the empirical likelihood inference in nonlinear errors-in-covariables models with validation data. Other related works include [8,10,24,26–28,34] and among others.

The corrected empirical likelihood method has many important features. First, to eliminate the effects of measurement errors on parameter estimation, we consider a correction for score functions based on the moment identities in [6,17]. Second, the empirical likelihood does not involve the plug-in estimate for the limiting variance. Third, the shape and orientation of the empirical likelihood-based confidence region are determined entirely by the data.

The remainder of the paper is organized as follows. In Section 2, the corrected empirical log-likelihood ratio for some generalized linear measurement error models is proposed, and its asymptotic distribution is shown to be the Chi-squared distribution with p degrees of freedom. We further obtain the maximum empirical likelihood estimator of the regression coefficient, and investigate its asymptotic properties. Section 3 reports the results of simulation studies and an application to a real data set. Finally, we conclude the paper in Section 4 and present the proofs in Appendix.

2 Methodology and results

2.1 Corrected score function

Let $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a random sample from model (1.1). If X_i is observed, an unbiased score

function of β is

$$\eta_i(\beta; X_i, Y_i) = \frac{\partial \mu_i}{\partial \beta} V_i^{-1}(\mu, \phi)(Y_i - \mu_i). \tag{2.1}$$

When X_i is subject to error and W_i is the observable value of X_i, $\eta_i(\beta; W_i, Y_i)$ by a direct replacement of X_i by W_i will lead to a biased score function, i.e., $E[\eta_i(\beta; W_i, Y_i)] = 0$ will not always hold. Motivated by the idea of Nakamura [17], we construct an unbiased score function $\eta_i^*(\Sigma_{uu}, \beta; W_i, Y_i)$ for β such that

$$E[\eta_i^*(\Sigma_{uu}, \beta; W_i, Y_i)] = 0.$$

To find the unbiased score function, by the moment identities associated with the error model (1.2), we have

$$E(W|X) = X, \qquad E(WW^{\mathrm{T}}|X) = XX^{\mathrm{T}} + \Sigma_{uu}, \tag{2.2}$$

$$E[W \exp(W^{\mathrm{T}}\beta)|X] = (X + \Sigma_{uu}\beta) \exp(X^{\mathrm{T}}\beta + \beta^{\mathrm{T}}\Sigma_{uu}\beta/2), \tag{2.3}$$

$$E[W \exp(-W^{\mathrm{T}}\beta)|X] = (X - \Sigma_{uu}\beta) \exp[-X^{\mathrm{T}}\beta + \beta^{\mathrm{T}}\Sigma_{uu}\beta/2], \tag{2.4}$$

$$E[W \exp(-2W^{\mathrm{T}}\beta)|X] = (X - 2\Sigma_{uu}\beta) \exp[-2X^{\mathrm{T}}\beta + 2\beta^{\mathrm{T}}\Sigma_{uu}\beta]. \tag{2.5}$$

In what follows, we construct the corrected score function for various measurement error regression models that are widely used in practice.

(1) Gamma measurement error regression models. Let Y follow the Gamma distribution with probability density

$$f(y) = \frac{1}{\Gamma(\phi)\theta^{\phi}} y^{\phi-1} \mathrm{e}^{-y/\theta},$$

where ϕ is known, θ is a canonical parameter, and $\Gamma(\cdot)$ is the Gamma function. The mean and variance of Y given X are $\mu = \phi\theta$ and $\mathrm{Var}(Y|X) = \mu^2/\phi$, respectively. Consider the log linear measurement error model

$$\begin{cases} \log(\mu_i) = X_i^{\mathrm{T}}\beta, \\ W_i = X_i + U_i. \end{cases}$$

By (2.1), the score function is

$$\eta_i(\beta; X_i, Y_i) = \phi X_i[Y_i \exp(-X_i^{\mathrm{T}}\beta) - 1].$$

By the moment identities (2.2) and (2.4), we have the following corrected score function,

$$\eta_i^*(\Sigma_{uu}, \beta; W_i, Y_i) = \phi(W_i + \Sigma_{uu}\beta) \exp(-W_i^{\mathrm{T}}\beta - \beta^{\mathrm{T}}\Sigma_{uu}\beta/2)Y_i - \phi W_i.$$

(2) Inverse Gaussian or Wald measurement error regression models. Let Y follow the inverse Gaussian distribution with mean μ and variance $\mathrm{Var}(Y|X) = \phi\mu^3$. Consider the log linear measurement error model

$$\begin{cases} \log(\mu_i) = X_i^{\mathrm{T}}\beta, \\ W_i = X_i + U_i. \end{cases}$$

By the moment identities (2.4) and (2.5), the corrected score function is given by

$$\begin{aligned} \eta_i^*(\Sigma_{uu}, \beta; W_i, Y_i) = {} & \phi^{-1}(W_i + 2\Sigma_{uu}\beta) \exp(-2W_i^{\mathrm{T}}\beta - 2\beta^{\mathrm{T}}\Sigma_{uu}\beta)Y_i \\ & - \phi^{-1}(W_i + \Sigma_{uu}\beta) \exp(-W_i^{\mathrm{T}}\beta - \beta^{\mathrm{T}}\Sigma_{uu}\beta/2). \end{aligned}$$

(3) Poisson measurement error regression models. Let Y follow the Poisson distribution with mean μ. Then $\mathrm{Var}(Y|X) = \mu$. Consider the log linear measurement error model

$$\begin{cases} \log(\mu_i) = X_i^{\mathrm{T}}\beta, \\ W_i = X_i + U_i. \end{cases}$$

We have the following corrected score function,

$$\eta_i^*(\Sigma_{uu},\beta;W_i,Y_i)=W_iY_i-(W_i-\Sigma_{uu}\beta)\exp(W_i^{\mathrm{T}}\beta-\beta^{\mathrm{T}}\Sigma_{uu}\beta/2).$$

(4) Binary logistic measurement error regression models. We consider the logistic measurement error regression model

$$\begin{cases}P(Y=1|X)=\dfrac{1}{1+\exp(-X^{\mathrm{T}}\beta)},\\ W=X+U,\end{cases}$$

with mean $\mu=[1+\exp(-X^{\mathrm{T}}\beta)]^{-1}$ and variance $\mathrm{Var}(Y|X)=\mu(1-\mu)$. Then the score function is

$$\eta_i(\beta;X_i,Y_i)=X_i\{Y_i-[1+\exp(-X_i^{\mathrm{T}}\beta)]^{-1}\}.$$

Note that the corrected score function does not exist for this model because the term $X_i[1+\exp(-X_i^{\mathrm{T}}\beta)]^{-1}$ cannot be corrected by Nakamura [17]. Instead, we follow Huang and Wang [6] and introduce the proper weight for the score function $\eta_i(\beta;X_i,Y_i)$ so that $\eta_{i,\omega}(\beta;W_i,Y_i)$ is corrected. Define

$$\eta_{i,\omega}(\beta;X_i,Y_i)=\omega(\beta;X_i,Y_i)\eta_i(\beta;X_i,Y_i),$$

where

$$\omega(\beta;X_i,Y_i)=1+\exp(X_i^{\mathrm{T}}\beta).$$

If there is a function $\eta_i^*(\Sigma_{uu},\beta;W_i,Y_i)$ such that

$$E\left[\eta_i^*(\Sigma_{uu},\beta;W_i,Y_i)|X_i,Y_i\right]=\eta_{i,\omega}(\beta;X_i,Y_i),$$

then $\eta_i^*(\Sigma_{uu},\beta;W_i,Y_i)$ is an unbiased score function. By the moment identities (2.2) and (2.3), the corrected score function is

$$\eta_i^*(\Sigma_{uu},\beta;W_i,Y_i)=W_iY_i+(W_i+\Sigma_{uu}\beta)\exp(-W_i^{\mathrm{T}}\beta-\beta^{\mathrm{T}}\Sigma_{uu}\beta/2)Y_i-W_i.$$

2.2 Empirical likelihood

Note that $\{\eta_i^*(\Sigma_{uu},\beta;W_i,Y_i);1\leqslant i\leqslant n\}$ are independent of each other with

$$E\{\eta_i^*(\Sigma_{uu},\beta;W_i,Y_i)\}=0.$$

The empirical log-likelihood ratio function for β is defined by

$$l(\Sigma_{uu},\beta)=-2\max\left\{\sum_{i=1}^n\log(np_i)\,\Big|\,p_i\geqslant 0,\sum_{i=1}^n p_i=1,\sum_{i=1}^n p_i\eta_i^*(\Sigma_{uu},\beta;W_i,Y_i)=0\right\},$$

where $p_i\,(i=1,\ldots,n)$ are nonnegative real numbers. By the Lagrange multiplier method, it can be shown that

$$l(\Sigma_{uu},\beta)=2\sum_{i=1}^n\log[1+\lambda^{\mathrm{T}}\eta_i^*(\Sigma_{uu},\beta;W_i,Y_i)],\tag{2.6}$$

where λ is a $p\times 1$ vector that is the solution to

$$\sum_{i=1}^n\frac{\eta_i^*(\Sigma_{uu},\beta;W_i,Y_i)}{1+\lambda^{\mathrm{T}}\eta_i^*(\Sigma_{uu},\beta;W_i,Y_i)}=0.\tag{2.7}$$

Denote $\Omega=E\{\eta^*(\Sigma_{uu},\beta;W,Y)\eta^{*\mathrm{T}}(\Sigma_{uu},\beta;W,Y)\}$. Suppose that the parameter space Θ is compact. The asymptotic distribution of the empirical log-likelihood ratio statistic $l(\Sigma_{uu},\beta)$ is established in Theorem 2.1.

Theorem 2.1. *Suppose that the matrix Ω is not singular, $E(XX^{\mathrm{T}}) < \infty$ and $E(UU^{\mathrm{T}}) < \infty$. If β is the true parameter and $n \to \infty$, then*

$$l(\Sigma_{uu}, \beta) \xrightarrow{\mathcal{L}} \chi_p^2,$$

where $\xrightarrow{\mathcal{L}}$ denotes the convergence in distribution and χ_p^2 is the Chi-squared distribution with p degrees of freedom.

Let $\chi_p^2(\alpha)$ be the α quantile of χ_p^2 for $0 < \alpha < 1$. Theorem 2.1 implies that an approximate $1-\alpha$ confidence region for β is given by

$$R_\alpha(\check{\beta}) = \{\check{\beta} \mid l(\check{\beta}) \leqslant \chi_p^2(1-\alpha)\}.$$

Remark 2.1. The naive empirical log-likelihood ratio (NELR) that is neglecting the measurement errors is not a Chi-squared distribution, because $\eta_i(\beta; W_i, Y_i)$, by a direct replacement of X_i by W_i, is a biased score function.

We can also define the maximizer of $-l(\beta)$, say $\hat{\beta}$, as the maximum empirical likelihood estimator of β. As the number of constraints equals the number of parameters, it may be shown that the optimal $p_i = 1/n$ (see Owen [18], Qin and Lawless [20]). Then, the maximum empirical likelihood estimator for β is the solution of

$$\sum_{i=1}^{n} \eta_i^*(\Sigma_{uu}, \beta; W_i, Y_i) = 0.$$

We state the asymptotic normality of the maximum empirical likelihood estimator in the following theorem.

Theorem 2.2. *Under the conditions of Theorem* 2.1*, if Γ is a positive definite matrix and $n \to \infty$, then*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \Gamma^{-1}\Omega(\Gamma^{-1})^{\mathrm{T}}),$$

where

$$\Gamma = E\{\partial \eta^*(\Sigma_{uu}, \beta; W, Y)/\partial\beta\}.$$

To construct the confidence region for β based on Theorem 2.2, we need to use the plug-in estimator for the covariance of β. By the moment method, the consistent estimators of Γ and Ω are

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial \eta_i^*(\Sigma_{uu}, \hat{\beta}; W_i, Y_i)}{\partial\beta}\right\},$$

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\{\eta_i^*(\Sigma_{uu}, \hat{\beta}; W_i, Y_i)\eta_i^{*\mathrm{T}}(\Sigma_{uu}, \hat{\beta}; W_i, Y_i)\}.$$

But, Σ_{uu} is usually unknown in practice. When Σ_{uu} is unknown, the measurement error covariance matrix Σ_{uu} can be estimated by partial replication (see Liang et al. [12]). We observe

$$W_{ij} = X_i + U_{ij}, \quad j = 1, \ldots, m_i.$$

Let $\overline{W}_i$ be the sample mean of the replicates, then, a consistent, unbiased method of moment estimate for Σ_{uu} is

$$\hat{\Sigma}_{uu} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(W_{ij} - \overline{W}_i)(W_{ij} - \overline{W}_i)^{\mathrm{T}}}{\sum\limits_{j=1}^{n}(m_i - 1)}.$$

Note that $\mathrm{Cov}(\overline{W}_i) = m_i^{-1}\Sigma_{uu}$, we replace W_i and Σ_{uu} by $\overline{W}_i$ and $m_i^{-1}\hat{\Sigma}_{uu}$, respectively. Hence, the unbiased score function is $\eta_i^*(m_i^{-1}\hat{\Sigma}_{uu}, \beta; \overline{W}_i, Y_i)$ and the corresponding log-likelihood ratio denotes $l(\hat{\Sigma}_{uu}, \beta)$. Throughout this section, we assume that $\frac{1}{n}\sum_{i=1}^{n} m_i^{-1}$ converges to a finite constant as $n \to \infty$.

Theorem 2.3. *Under the conditions of Theorem* 2.1*, as* $n \to \infty$*, we still have the following conclusion,*

$$l(\hat{\Sigma}_{uu}, \beta) \xrightarrow{\mathcal{L}} \chi_p^2.$$

We may maximize $\{-l(\hat{\Sigma}_{uu}, \beta)\}$ to obtain an estimator for the parameter β, say $\tilde{\beta}$. We give the following asymptotic distribution of $\tilde{\beta}$.

Theorem 2.4. *Under the conditions of Theorem* 2.1*, if* Γ^* *is a positive definite matrix and* $n \to \infty$*, then*

$$\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, \Gamma^{*-1}\Omega^*(\Gamma^{*-1})^{\mathrm{T}}),$$

where

$$\Gamma^* = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E\{\partial \eta^*(m_i^{-1}\Sigma_{uu}, \beta; \overline{W}_i, Y_i)/\partial\beta\}$$

and

$$\Omega^* = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E\{\eta^*(m_i^{-1}\Sigma_{uu}, \beta; \overline{W}_i, Y_i)\eta^{*\mathrm{T}}(m_i^{-1}\Sigma_{uu}, \beta; \overline{W}_i, Y_i)\}.$$

The consistent estimators of Γ^* and Ω^* are easily defined as

$$\hat{\Gamma}^* = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial \eta_i^*(\frac{1}{m_i}\hat{\Sigma}_{uu}, \tilde{\beta}; \overline{W}_i, Y_i)}{\partial \beta} \right\},$$

$$\hat{\Omega}^* = \frac{1}{n} \sum_{i=1}^{n} \left\{ \eta_i^* \left(\frac{1}{m_i}\hat{\Sigma}_{uu}, \tilde{\beta}; \overline{W}_i, Y_i \right) \eta_i^{*\mathrm{T}} \left(\frac{1}{m_i}\hat{\Sigma}_{uu}, \tilde{\beta}; \overline{W}_i, Y_i \right) \right\}.$$

2.3 Partial profile empirical likelihood

When β is a more than two-dimensional vector, we cannot graph the joint confidence region for β. However, we can calculate the confidence intervals of one component of β by constructing a partial profile empirical likelihood ratio. Let e_r be the unit vector of length p with 1 at position r for $r = 1, \ldots, p$. With an argument similar to Xue and Zhu [28], the estimators of the r-th component β_r are

$$\hat{\beta}_r = e_r^{\mathrm{T}}\hat{\beta} \quad \text{and} \quad \tilde{\beta}_r = e_r^{\mathrm{T}}\tilde{\beta}$$

when Σ_{uu} is known and unknown, respectively. Write

$$\eta_i^*(\beta) \equiv \eta_i^*(\Sigma_{uu}, \beta; W_i, Y_i)$$

and

$$\hat{\eta}_i^*(\beta) \equiv \eta_i^* \left(\frac{1}{m_i}\hat{\Sigma}_{uu}, \beta; \overline{W}_i, Y_i \right).$$

Let

$$\eta_{ir}^*(\beta_r) = e_r^{\mathrm{T}}\hat{\Gamma}^{-1}\eta_i^*(\hat{\beta}_1, \ldots, \hat{\beta}_{r-1}, \beta_r, \hat{\beta}_{r+1}, \ldots, \hat{\beta}_p),$$

$$\hat{\eta}_{ir}^*(\beta_r) = e_r^{\mathrm{T}}\hat{\Gamma}^{*-1}\hat{\eta}_i^*(\tilde{\beta}_1, \ldots, \tilde{\beta}_{r-1}, \beta_r, \tilde{\beta}_{r+1}, \ldots, \tilde{\beta}_p).$$

Then, the partial profile empirical log-likelihood ratios for β_r with known and unknown Σ_{uu} are defined by

$$l_r(\Sigma_{uu}, \beta_r) = -2\max\left\{ \sum_{i=1}^{n} \log(np_i) \,\middle|\, p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i\eta_{ir}^*(\beta_r) = 0 \right\},$$

and

$$l_r(\hat{\Sigma}_{uu}, \beta_r) = -2\max\left\{ \sum_{i=1}^{n} \log(np_i) \,\middle|\, p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i\hat{\eta}_{ir}^*(\beta_r) = 0 \right\}.$$

Under the assumptions of Theorems 2.1 and 2.3, we will show that the asymptotic distributions of $l_r(\Sigma_{uu}, \beta_r)$ and $l_r(\hat{\Sigma}_{uu}, \beta_r)$ are standard Chi-squared with 1 degree of freedom.

Theorem 2.5. *Under the assumptions of Theorem* 2.1*, as* $n \to \infty$*, we have*

$$l_r(\Sigma_{uu}, \beta_r) \xrightarrow{\mathcal{L}} \chi_1^2.$$

Under the assumptions of Theorem 2.3*, we have*

$$l_r(\hat{\Sigma}_{uu}, \beta_r) \xrightarrow{\mathcal{L}} \chi_1^2.$$

Applying Theorem 2.5, we can construct the approximate $1-\alpha$ confidence intervals for β_r.

Remark 2.2. The above method can be applied to construct the confidence regions for any two different components (β_r, β_s). The method is as follows: First, $\eta_{irs}^*(\beta_r, \beta_s)$ or $\hat{\eta}_{irs}^*(\beta_r, \beta_s)$ is obtained by replacing the other components of β except for (β_r, β_s) in $\eta_i^*(\beta)$ or $\hat{\eta}_i^*(\beta)$ with their estimators and using (e_r, e_s) to replace e_r. Second, we construct the empirical likelihood confidence region for (β_r, β_s) by using $l_{rs}(\Sigma_{uu}, \beta_r, \beta_s)$ or $l_{rs}(\hat{\Sigma}_{uu}, \beta_r, \beta_s)$.

3 Numerical studies

In this section, we investigate the finite sample performance of the proposed method via simulation studies and real data analysis. We will report the results for the Poisson measurement error regression model and the logistic measurement error regression model. Other models have the similar results and are therefore omitted. We only give the one-dimensional and two-dimensional cases in the simulation studies. When β is a three-dimensional vector, say $\beta = (\beta_1, \beta_2, \beta_3)^{\mathrm{T}}$, we can calculate the confidence intervals of β_1, β_2 and β_3 by Theorem 2.5, respectively. Because the simulation results are similar to those of the one-dimensional and two-dimensional cases, therefore we omit this case. But we consider the three-dimensional case in real data analysis. The following questions are considered in the simulation studies.

(1) Compare the empirical distribution of the corrected empirical log-likelihood ratio (CELR) and the naive empirical log-likelihood ratio (NELR) that is neglecting the measurement errors with a direct replacement of X by W with the theory distribution in Theorem 2.3.

(2) Compare the confidence intervals of β obtained by the three methods: the naive empirical log-likelihood (NEL) that is neglecting the measurement errors, the corrected empirical likelihood method (CEL) based on Theorem 2.3 and the normal approximation method (NA) based on Theorem 2.4.

3.1 One-dimensional case

We consider the Poisson measurement error regression model. The response variable Y is generated from Poisson(μ) with

$$\log(\mu) = X^{\mathrm{T}}\beta, \quad W = X + U,$$

where $\beta = 1$, X is generated from $N(1,1)$, and U is generated from $N(0, \sigma_u^2)$. We take $\sigma_u = 0.2, 0.4$ and 0.6 to represent different levels of measurement errors. σ_u is unknown, to estimate σ_u, two replicates of W are generated. The sizes of the samples are $n = 100, 150$, and 200. For each setting, we simulate 2000 times to assess the performance.

First, we compare the finite sample distribution of $l(\beta)$ with the asymptotic distribution of χ_1^2 when $n = 150$. The Q-Q plots of CELR and NELR are given in Figure 1. It is clear to see that the distribution of CELR can be approximated by χ_1^2 for the different levels of measurement errors, but the distribution of NELR performs not well because the naive score function is biased.

To evaluate the performance of the confidence intervals of β, the averages of 95% confidence intervals and the corresponding coverage probabilities are computed. The results are reported in Table 1. From Table 1, we can see the following results. Firstly, the confidence intervals based on NEL are biased and the coverage probabilities are low when the level of the measurement error is high. Secondly, the confidence

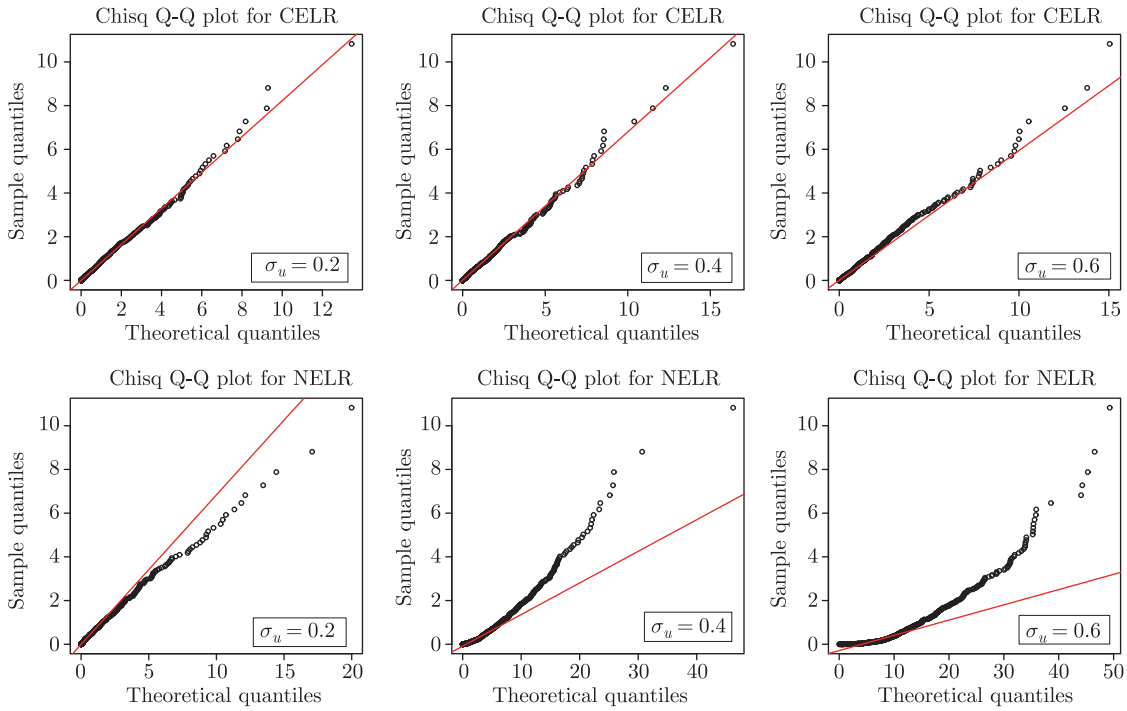

Figure 1 The Q-Q plots of CELR and NELR with $\sigma_u = 0.2, 0.4$ and 0.6 when $n = 150$ for the Poisson measurement error regression model

Table 1 The average lengths and coverage probabilities of β based on NEL, CEL and NA for the Poisson measurement error regression model when the nominal level is 0.95

		Average lengths			Coverage probabilities		
n	σ_u	NEL	CEL	NA	NEL	CEL	NA
100	0.2	0.0895	0.0907	0.0925	0.8966	0.9142	0.9048
	0.4	0.1102	0.1169	0.1211	0.7850	0.9108	0.8760
	0.6	0.1466	0.1537	0.1582	0.5610	0.8834	0.8506
150	0.2	0.0740	0.0755	0.0774	0.9075	0.9305	0.9110
	0.4	0.0954	0.1013	0.1024	0.7370	0.9115	0.9030
	0.6	0.1284	0.1326	0.1342	0.4320	0.9040	0.8730
200	0.2	0.0661	0.0666	0.0669	0.8970	0.9310	0.9190
	0.4	0.0857	0.0897	0.0902	0.6890	0.9135	0.9040
	0.6	0.1203	0.1178	0.1202	0.3455	0.9065	0.8980

intervals and coverage probabilities based on CEL and NA depend on the measurement error and the sample size. The confidence intervals increase and the coverage probabilities decrease as the measurement error increases for a fixed sample size. The average interval lengths decrease as the sample size increases while the corresponding coverage probabilities increase for a fixed level of measurement error. Thirdly, CEL gives shorter interval lengths and higher coverage probabilities than NA.

3.2 Two-dimensional case

We generate data from the logistic measurement error regression model

$$\text{logit}(Y = 1|X) = X^{\mathrm{T}}\beta, \qquad W = X + U,$$

where $\beta = (0.4, 0.8)^{\mathrm{T}}$, X_1 and X_2 are generated from $N(0,1)$, and U is generated from a two-dimensional normal distribution with mean 0 and covariates matrix $\Sigma_{uu} = 0.4^2 I_2$. We generate two replicates of W to estimate covariates matrix because Σ_{uu} is unknown. We compute the finite sample distributions of CELR and NELR with $n = 100, 150$, and 200 and report the results in Figure 2. From Figure 2, we can see that the distribution of CELR provides a better approximation to χ^2_2 than that of NELR.

The average confidence regions for β based on 2000 simulation runs and its coverage probabilities are computed with $n = 150$ and $\sigma_u = 0.4$. Figure 3 shows that CEL gives smaller confidence region than NA. The empirical coverage probabilities of NEL, CEL and NA are 0.902, 0.918 and 0.908, respectively.

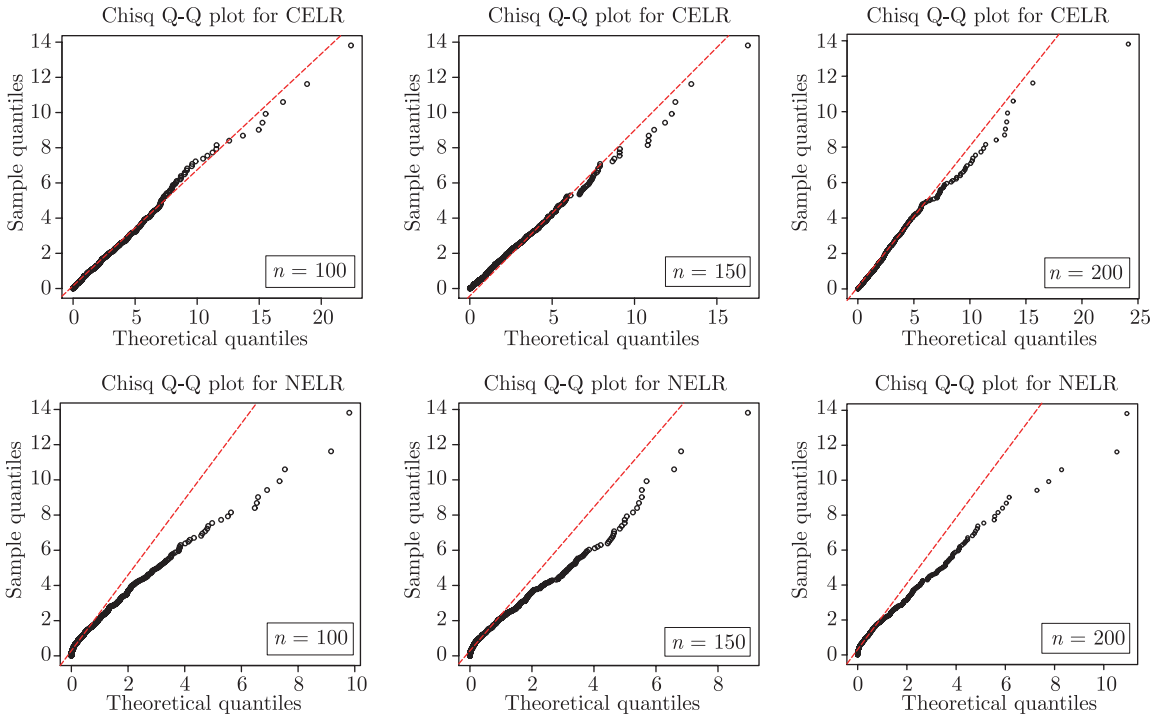

Figure 2 The Q-Q plots of CELR and NELR with $n = 100, 150$ and 200 when $\sigma_u = 0.4$ for the logistic measurement error regression model

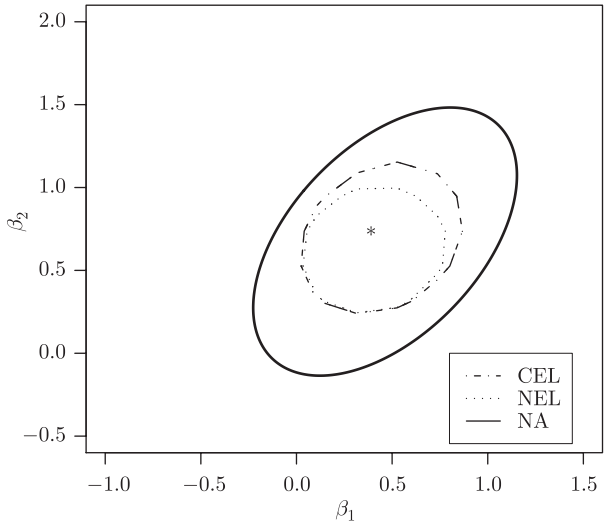

Figure 3 The 95% average confidence regions for (β_1, β_2), based on NEL (dotted curve), CEL (dotted-dashed curve), and NA (solid curve) when $n = 150$ and $\sigma_u = 0.4$

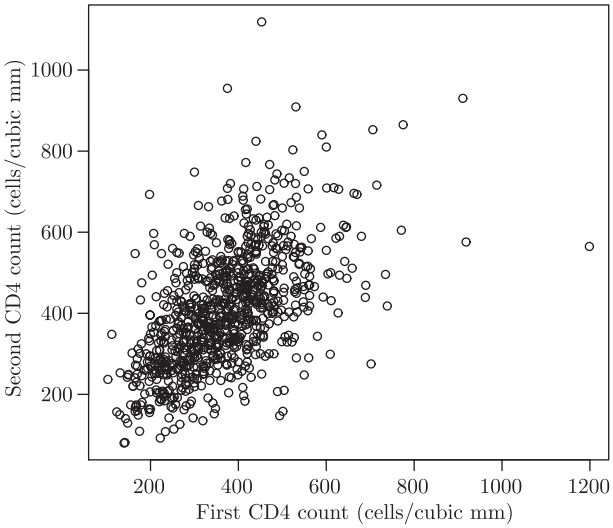

Figure 4 Duplicated baseline CD4 count measurements from 885 antiretroviral-naive patients in the ACTG 175 study

Table 2 The estimators and confidence intervals of β based on NEL, CEL and NA in ACTG 175 study

	Estimators		Confidence intervals		
β	NEL	CEL/NA	NEL	CEL	NA
β_1	4.8929	8.1066	(4.1929, 5.5929)	(7.5808, 8.5074)	(7.6182, 8.5949)
β_2	−1.1173	−1.6656	(−1.6693, −0.4173)	(−1.8041, −1.5797)	(−2.1173, −1.2138)
β_3	−0.0469	−0.0038	(−0.6568, 0.5257)	(−0.6401, 0.5921)	(−0.6212, 0.6136)

3.3 A real example

In this subsection, we analyze an AIDS clinical study conducted by the AIDS Clinical Trials Group (ACTG 175, see [6]). The data set is available in the R package "speff2trial". This randomized trial was designed to evaluate treatment with either one or two nucleosides in HIV-infected adults having a screening CD4 count between 200 and 500 and no history of AIDS-defining illness (Hammer et al. [5]). We are interested in the relationship between the true baseline CD4 count, the history of intravenous drug use and the symptomatic HIV infection. We analyze 885 antiretroviral naive patients who have duplicated baseline CD4 count measurements. Figure 4 shows the duplicated baseline CD4 count measurements from 885 patients. We adopt the logistic regression model with three covariates. Y is the symptomatic HIV infection (0 = asymptomatic, 1 = symptomatic). We take $X_1 = 1$ for the intercept term. X_2 is the true baseline log(CD4). The CD4 count is subject to the measurement error, σ_u^2 can be estimated to be 0.0565 by replication experiments. X_3 is the history of intravenous drug use (0 = no, 1 = yes). The estimators and confidence intervals of β based on NEL, CEL and NA are reported in Table 2. From Table 2, we see that the history of intravenous drug use has no significant effect on the symptomatic HIV infection, and CEL gives shorter confidence intervals than NA. As expected, when the measurement error is taken into account, we find a somewhat stronger negative association between the CD4 count and the symptomatic HIV infection.

4 Conclusion

In this paper, a corrected technique for constructing the empirical log-likelihood ratio is proposed for a class of generalized linear measurement error models. The idea of "correction" comes from Stefanski [21], Nakamura [17], and Huang and Wang [6]. The corrected empirical log-likelihood ratio has an asymptotic

Chi-squared distribution, whereas the naive empirical log-likelihood is not asymptotically Chi-square distributed because the naive score function is biased. The advantages of our proposed method are demonstrated in simulation studies and in a real data example. Finally, we note that the methodology in this paper is general and can be readily extended to some other models, including the commonly used generalized partially linear measurement error models.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 11301569, 11471029 and 11101014), the Beijing Natural Science Foundation (Grant No. 1142002), the Science and Technology Project of Beijing Municipal Education Commission (Grant No. KM201410005010), Hong Kong Research Grant (Grant No. HKBU202711) and Hong Kong Baptist University FRG Grants (Grant Nos. FRG2/11-12/110 and FRG1/13-14/018). The authors thank the editor, associate editor and the referees for insightful comments that led to an improvement of an earlier manuscript.

References

1 Chen S X, Cui H J. An extended empirical likelihood for generalized linear models. Statist Sinica, 2003, 13: 69–81
2 Cui H J, Chen S X. Empirical likelihood confidence region for parameter in the errors-in-variables models. J Multivariate Anal, 2003, 84: 101–115
3 Cui H J, Kong E F. Empirical likelihood confidence region for parameters in semi-linear errors-in-variables models. Scand J Statist, 2006, 33: 153–168
4 Firth D. On the efficiency of quasi-likelihood estimation. Biometrika, 1987, 74: 233–245
5 Hammer S M, Katzenstein D A, Hughes M D, et al. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. New England J Medicine, 1996, 335: 1081–1090
6 Huang Y, Wang C Y. Consistent functional methods for logistic regression with errors in covariates. J Amer Statist Assoc, 2001, 96: 1469–1482
7 Kolaczyk E D. Empirical likelihood for generalized linear models. Statist Sinica, 1994, 4: 199–218
8 Li G R, Lin L, Zhu L X. Empirical likelihood for varying coefficient partially linear model with diverging number of parameters. J Multivariate Anal, 2012, 105: 85–111
9 Li G R, Xue L G. Empirical likelihood confidence region for the parameter in a partially linear errors-in-variables model. Comm Statist Theory Methods, 2008, 37: 1552–1564
10 Li G R, Zhu L X, Xue L G, et al. Empirical likelihood inference in partially linear single-index models for longitudinal data. J Multivariate Anal, 2010, 101: 718–732
11 Liang H. Generalized partially linear mixed-effects models incorporating mismeasured covariates. Ann Inst Statist Math, 2009, 61: 27–46
12 Liang H, Härdle W, Carroll R J. Estimation in a semiparametric partially linear errors-in-variables model. Ann Statist, 1999, 27: 1519–1535
13 Liang H, Li R Z. Variable selection for partially linear models with measurement errors. J Amer Statist Assoc, 2009, 104: 234–248
14 Lin X H, Carroll R J. Nonparametric function estimation for clustered data when the predictor is measured without/with error. J Amer Statist Assoc, 2000, 95: 520–534
15 Ma Y Y, Li R Z. Variable selection in measurement error models. Bernoulli, 2010, 16: 274–300
16 McCullagh P, Nelder J A. Generalized Linear Models. London: Chapman and Hall, 1989
17 Nakamura T. Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. Biometrika, 1990, 77: 127–137
18 Owen A B. Empirical likelihood ratio confidence intervals for a single functional. Biometrika, 1988, 75: 237–249
19 Owen A B. Empirical likelihood ratio confidence regions. Ann Statist, 1990, 18: 90–120
20 Qin J, Lawless J F. Empirical likelihood and general estimating equations. Ann Statist, 1994, 22: 300–325
21 Stefanski L. Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. Comm Statist Theory Methods, 1989, 18: 4335–4358
22 Stefanski L, Carroll R. Conditional scores and optimal scores for generalized linear measurement-error models. Biometrika, 1987, 74: 703–716
23 Stute W, Xue L G, Zhu L X. Empirical likelihood inference in nonlinear errors-in-covariables models with validation data. J Amer Statist Assoc, 2007, 102: 332–346
24 Wang Q H, Rao J N K. Empirical likelihood-based inference in linear errors-in-covariables models with validation data. Biometrika, 2002, 89: 345–358

25 Wedderburn R W M. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. Biometrika, 1974, 61: 439–447
26 Xue L G. Empirical likelihood confidence intervals for response mean with data missing at random. Scand J Statist, 2009, 36: 671–685
27 Xue L G. Empirical likelihood for linear models with missing responses. J Multivariate Anal, 2009, 100: 1353–1366
28 Xue L G, Zhu L X. Empirical likelihood for a varying coefficient model with longitudinal data. J Amer Statist Assoc, 2007, 102: 642–654
29 Yi G Y, Liu W, Wu L. Simultaneous inference and bias analysis for longitudinal data with covariate measurement error and missing responses. Biometrics, 2011, 67: 67–75
30 Yi G Y, Ma Y Y, Carroll R J. A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. Biometrika, 2012, 99: 151–165
31 Zhang W W, Li G R, Xue L G. Profile inference on partially linear varying-coefficient errors-in-variables models under restricted condition. Comput Statist Data Anal, 2011, 55: 3027–3040
32 Zhao P X, Xue L G. Variable selection for semiparametric varying coefficient partially linear errors-in-variables models. J Multivariate Anal, 2010, 101: 1872–1883
33 Zhou Y, Liang H. Statistical inference for semiparametric varying-coefficient partially linear models with error-prone linear covariates. Ann Statist, 2009, 37: 427–458
34 Zhu L X, Xue L G. Empirical likelihood confidence regions in a partially linear single-index model. J R Stat Soc Ser B Stat Methodol, 2006, 68: 549–570

Appendix: Proof of theorems

Proof of Theorem 2.1. First, we will show that

$$\max_{1\leqslant i\leqslant n}\|\eta_i^*(\beta)\| = o_p(n^{1/2}). \tag{A.1}$$

It is well known that for any sequence of i.i.d. random variables $\{\xi_i, 1 \leqslant i \leqslant n\}$ with $E(\xi_i^2) < \infty$, $\max_{1\leqslant i\leqslant n}|\xi_i| \xrightarrow{\text{a.s.}} 0$, which implies that

$$\max_{1\leqslant i\leqslant n}|\eta_{i,s}^*(\beta)| = o_p(n^{1/2}).$$

Since $E|\eta_{i,s}^*(\beta)|^2 < \infty$, where $\eta_{i,s}^*$ is the s-th component of η_i^*, this proves (A.1).

By using the same arguments that are used in (2.14) in Owen [19], we can prove that

$$\|\lambda\| = O_p(n^{-1/2}). \tag{A.2}$$

From (2.7), we have

$$n^{-1}\sum_{i=1}^n \eta_i^*(\beta)(1-\lambda^{\mathrm{T}}\eta_i^*(\beta)) + n^{-1}\sum_{i=1}^n \eta_i^*(\beta)\frac{\lambda^{\mathrm{T}}\eta_i^*(\beta)\eta_i^{*\mathrm{T}}(\beta)\lambda}{1+\lambda^{\mathrm{T}}\eta_i^*(\beta)} = 0. \tag{A.3}$$

The last term on the left-hand side in (A.3) has a norm bounded by

$$n^{-1}\sum_{i=1}^n \|\eta_i^*(\beta)\|^3\|\lambda\|^2|1+\lambda^{\mathrm{T}}\eta_i^*(\beta)|^{-1} = o(n^{1/2})O_p(n^{-1})O_p(1) = o_p(n^{-1/2}).$$

Therefore,

$$\lambda = \left[\frac{1}{n}\sum_{i=1}^n \eta_i^*(\beta)\eta_i^{*\mathrm{T}}(\beta)\right]^{-1}\frac{1}{n}\sum_{i=1}^n \eta_i^*(\beta) + o_p(n^{-1/2}).$$

By (A.1) and (A.2), we have

$$\max_{1\leqslant i\leqslant n}|\lambda^{\mathrm{T}}\eta_i^*(\beta)| = O_p(n^{-1/2})o_p(n^{1/2}) = o_p(1).$$

Hence, applying the Taylor expansion of $\log(\cdot)$ around 1 in (2.6), we have for some r_i between 1 and $1+\lambda^{\mathrm{T}}\eta_i^*(\beta)$ $(i=1,\ldots,n)$,

$$\begin{aligned} l(\Sigma_{uu},\beta) &= 2\sum_{i=1}^{n}\left\{\lambda^{\mathrm{T}}\eta_i^*(\beta)-\frac{1}{2}(\lambda^{\mathrm{T}}\eta_i^*(\beta))^2+\frac{1}{3}\frac{(\lambda^{\mathrm{T}}\eta_i^*(\beta))^3}{(1+r_i)^3}\right\} \\ &= 2\lambda^{\mathrm{T}}\sum_{i=1}^{n}\eta_i^*(\beta)-\lambda^{\mathrm{T}}\sum_{i=1}^{n}\eta_i^*(\beta)\eta_i^{*\mathrm{T}}(\beta)\lambda+o_p(1) \\ &= \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i^*(\beta)\right]^{\mathrm{T}}\left[\frac{1}{n}\sum_{i=1}^{n}\eta_i^*(\beta)\eta_i^{*\mathrm{T}}(\beta)\right]^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i^*(\beta)\right]+o_p(1). \end{aligned} \tag{A.4}$$

It is easy to see that

$$E\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i^*(\beta)\right)=0,$$

and

$$\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i^*(\beta)\right)=\Omega.$$

Using the central limit theorem, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i^*(\beta)\xrightarrow{\mathcal{L}}N(0,\Omega). \tag{A.5}$$

By the law of large numbers, we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\eta_i^*(\beta)\eta_i^{*\mathrm{T}}(\beta)\xrightarrow{P}\Omega.$$

Using this together with (A.4), we obtain

$$l(\Sigma_{uu},\beta)=\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i^*(\beta)\right]^{\mathrm{T}}\Omega^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i^*(\beta)\right]+o_p(1). \tag{A.6}$$

By (A.5), we have

$$\Omega^{-\frac{1}{2}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i^*(\beta)\xrightarrow{\mathcal{L}}N(0,I_p), \tag{A.7}$$

where I_p is the $p\times p$ identity matrix. Results (A.6) and (A.7) together lead to Theorem 2.1. □

Proof of Theorem 2.2. Applying the Taylor expansion, we have

$$\sqrt{n}(\hat{\beta}-\beta)=\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\eta_i^*(\beta)}{\partial\beta}\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i^*(\beta)\right\}+o_p(1).$$

By the central limit theorem, together with (A.5), we can prove Theorem 2.2. □

Proof of Theorem 2.3. Because $\hat{\Sigma}_{uu}$ is a consistent, unbiased moment estimator of Σ_{uu}, we have

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_i^*(\beta)\hat{\eta}_i^{*\mathrm{T}}(\beta)\xrightarrow{P}\Omega^*, \tag{A.8}$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\eta}_i^*(\beta)\xrightarrow{\mathcal{L}}N(0,\Omega^*). \tag{A.9}$$

In a similar way to Theorem 2.1, we can prove that $l(\hat{\Sigma}_{uu}, \beta)$ is asymptotically equivalent to

$$\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\eta}_i^*(\beta)\right]^{\mathrm{T}}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_i^*(\beta)\hat{\eta}_i^{*\mathrm{T}}(\beta)\right]^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\eta}_i^*(\beta)\right].$$

This together with (A.8) and (A.9) proves Theorem 2.3. □

The proofs of Theorems 2.4 and 2.5 are similar to those of Theorems 2.2 and 2.3, we hence omit their proofs.