Ann Inst Stat Math (2008) 60:883–900
DOI 10.1007/s10463-007-0139-z

Second-order nonlinear least squares estimation

Liqun Wang · Alexandre Leblanc

Received: 16 November 2005 / Revised: 14 November 2006 / Published online: 7 July 2007
© The Institute of Statistical Mathematics, Tokyo 2007

Abstract The ordinary least squares estimation is based on minimization of the squared distance of the response variable to its conditional mean given the predictor variable. We extend this method by including in the criterion function the distance of the squared response variable to its second conditional moment. It is shown that this "second-order" least squares estimator is asymptotically more efficient than the ordinary least squares estimator if the third moment of the random error is nonzero, and both estimators have the same asymptotic covariance matrix if the error distribution is symmetric. Simulation studies show that the variance reduction of the new estimator can be as high as 50% for sample sizes lower than 100. As a by-product, the joint asymptotic covariance matrix of the ordinary least squares estimators for the regression parameter and for the random error variance is also derived, which is only available in the literature for very special cases, e.g. that random error has a normal distribution. The results apply to both linear and nonlinear regression models, where the random error distributions are not necessarily known.

Keywords Nonlinear regression · Asymmetric error distribution · Weighted least squares · Minimum distance estimator · Consistency · Asymptotic normality

1 Introduction

The least squares estimation in (nonlinear) regression models has a long history and its (asymptotic) statistical properties are well-known. See, e.g., Gallant (1987) and Seber and Wild (1989). The ordinary least squares (OLS) estimator minimizes the squared distance of the response variable to its conditional mean given the predic-

L. Wang (✉) · A. Leblanc
Department of Statistics, University of Manitoba, Winnipeg, MB, Canada R3T 2N2
e-mail: liqun_wang@umanitoba.ca

tor variable. This estimator is most efficient, if the random errors in the model are normally distributed. On the other hand, the OLS estimator may not be optimal if the random error distribution is not normal or asymmetric. In order to obtain more efficient estimators in such cases, it is natural to exploit information contained in the higher moments of the data.

In this paper, we study an estimator which minimizes the distances of the response variable and the squared response variable to its first and second conditional moments simultaneously. In particular, we derive the strong consistency and asymptotic normality for this "second-order" least squares (SLS) estimator under general regularity conditions. Moreover, we show that this estimator is asymptotically more efficient than the OLS estimator if the third moment of the random error is nonzero, and both estimators have the same asymptotic covariance matrix if the error distribution is symmetric. Monte Carlo simulation studies show that the variance reduction can be as high as 50% for sample sizes lower than 100. Asymmetric random error distributions in regression problems arise in many applied fields, e.g., in biology, economics, engineering, environmetrics and quantitative finance (e.g., Boos 1987, Hutson 2004, Williams 1997, Theodossiou 1998). They also attract much attention in (Bayesian) robust statistics (e.g., Azzalini and Capitanio 1999, Marazzi and Yohai 2004, Sahu et al. 2003).

The proposed estimation method was first used by Wang (2003, 2004) to deal with the measurement error problems in nonlinear regression models. The theoretical framework used there requires the measurement error variance to be strictly positive. Therefore the results obtained do not apply to the traditional nonlinear least squares setup, which is considered in the present paper. The paper is organized as follows. In Sect. 2 we introduce the second-order least squares estimator and present its consistency and asymptotic normality under some regularity conditions. In Sect. 3 we compare the new estimator with the traditional OLS estimator. We also derive the joint asymptotic covariance matrix of the OLS estimators for θ and σ^2. Section 4 contains Monte Carlo simulations of finite sample performance and comparison of the estimators. Conclusions and discussion are given in Sect. 5, whereas proofs of the theorems are given in Sect. 6.

2 Second-order least squares estimator

Consider the general regression model

$$Y = g(X;\theta) + \varepsilon, \tag{1}$$

where $Y \in \mathbb{R}$ is the response variable, $X \in \mathbb{R}^k$ is the predictor variable, $\theta \in \mathbb{R}^p$ is the unknown regression parameter and ε is the random error satisfying $E(\varepsilon|X) = 0$ and $E(\varepsilon^2|X) = \sigma^2$. The regression function $g(X;\theta)$ can be linear or nonlinear in either X or θ. In addition, we assume that Y and ε have finite fourth moments.

Under the assumption for model (1), the first two conditional moments of Y given X are respectively $E(Y|X) = g(X;\theta)$ and $E(Y^2|X) = g^2(X;\theta) + \sigma^2$. Throughout the paper we denote the parameter vector as $\gamma = (\theta', \sigma^2)'$ and the parameter space as $\Gamma = \Theta \times \Sigma \subset \mathbb{R}^{p+1}$. The true parameter value of model (1) is denoted by $\gamma_0 = (\theta_0', \sigma_0^2)' \in \Gamma$.

Suppose $(Y_i, X_i')'$, $i = 1, 2, \ldots, n$ is an *i.i.d.* random sample. Then the second-order least squares estimator (SLSE) $\hat{\gamma}_{\text{SLS}}$ for γ is defined as the measurable function that minimizes

$$Q_n(\gamma) = \sum_{i=1}^{n} \rho_i'(\gamma) W_i \rho_i(\gamma), \tag{2}$$

where $\rho_i(\gamma) = \left(Y_i - g(X_i; \theta), Y_i^2 - g^2(X_i; \theta) - \sigma^2\right)'$ and $W_i = W(X_i)$ is a 2×2 nonnegative definite matrix which may depend on X_i.

Now we consider the asymptotic properties of the SLSE $\hat{\gamma}_{\text{SLS}}$. For the consistency of $\hat{\gamma}_{\text{SLS}}$ we make the following assumptions, where μ denotes the Lebesgue measure and $\|\cdot\|$ denotes the Euclidean norm in the real space.

Assumption 1 $g(x; \theta)$ is a measurable function of x for every $\theta \in \Theta$, and is continuous in $\theta \in \Theta$ for μ-almost all x.

Assumption 2 $E \|W(X)\| \left(\sup_\Theta g^4(X; \theta) + 1\right) < \infty$.

Assumption 3 The parameter space $\Gamma \subset \mathbb{R}^{p+1}$ is compact.

Assumption 4 For any $\gamma \in \Gamma$, $E[\rho(\gamma) - \rho(\gamma_0)]' W(X)[\rho(\gamma) - \rho(\gamma_0)] = 0$ if and only if $\gamma = \gamma_0$, where $\rho(\gamma) = \left(Y - g(X; \theta), Y^2 - g^2(X; \theta) - \sigma^2\right)'$.

The above regularity conditions are common in the literature of nonlinear regression. In particular, Assumption 1 is usually used to ensure that the objective function $Q_n(\gamma)$ is continuous in γ. Assumption 2 is a moment condition which is sufficient for the uniform convergence of $Q_n(\gamma)$. Similarly, the compactness of the parameter space Γ is often assumed. Finally, Assumption 4 is the usual condition for identifiability of parameters, which guarantees that $Q_n(\gamma)$ has unique minimizer γ_0 in Γ for large n. In practice, it can be a tedious task to check the identifiability condition directly. Rather, it is done in an ad hoc way, since a model is usually identified as long as it is well defined and the parameter space is small enough. In the nonlinear regression literature, Assumption 2 is sometimes expressed through the existence of a function, say $h(x)$, such that $|g(x; \theta)| \leq h(x)$ for all $\theta \in \Theta$ and $E \|W(X)\| h^4(X) < \infty$. In practice, this and other conditions are checked on case by case basis. We demonstrate this through the following examples.

Example 1 Consider the model $Y = \theta_1 e^{\theta_2 X} + \varepsilon$, where $a \leq \theta_1 \leq b$, $c \leq \theta_2 \leq d < 0$ and a, b, c, d are finite. For notational simplicity suppose $\|W\|$ is constant and let $h(x) = \max\{|a|, |b|\}(e^{cx} + e^{dx})$. Then $|g(x; \theta)| \leq h(x)$ for all θ in the parameter space and $Eh^4(X) \leq 8 \max\{|a|^4, |b|^4\}(Ee^{4cX} + Ee^{4dX})$ which is finite as long as X has a finite moment generating function. Therefore Assumption 2 is satisfied.

Example 2 Consider another model $Y = \theta_1/[1 + \exp(\theta_2 + \theta_3 X)] + \varepsilon$, where $a \leq \theta_1 \leq b$ and a, b are finite. Now let $h(x) = \max\{|a|, |b|\}$. Then $|g(x; \theta)| \leq |\theta_1| \leq h(x)$ and $E \|W\| h^4(X)$ will be finite as long as W is properly chosen. Therefore Assumption 2 is easily satisfied.

Theorem 1 *Under Assumptions 1–4, the SLSE* $\hat{\gamma}_{\mathrm{SLS}} \xrightarrow{a.s.} \gamma_0$*, as* $n \to \infty$.

To derive the asymptotic normality for $\hat{\gamma}_{\mathrm{SLS}}$, further regularity conditions are needed.

Assumption 5 θ_0 is an interior point of Θ and, for μ-almost all x, $g(x;\theta)$ is twice continuously differentiable in Θ. Furthermore, the first two derivatives satisfy $E\left\|W(X)\right\| \sup_\Theta \left\|\frac{\partial g(X;\theta)}{\partial\theta}\right\|^4 < \infty$, $E\left\|W(X)\right\| \sup_\Theta \left\|\frac{\partial^2 g(X;\theta)}{\partial\theta\partial\theta'}\right\|^4 < \infty$.

Assumption 6 The matrix $A = E\left[\frac{\partial\rho'(\gamma_0)}{\partial\gamma} W(X) \frac{\partial\rho(\gamma_0)}{\partial\gamma'}\right]$ is nonsingular, where

$$\frac{\partial\rho'(\gamma_0)}{\partial\gamma} = -\begin{pmatrix} \frac{\partial g(X;\theta_0)}{\partial\theta} & 2g(X;\theta_0)\frac{\partial g(X;\theta_0)}{\partial\theta} \\ 0 & 1 \end{pmatrix}$$

and $\partial g(X;\theta_0)/\partial\theta$ is the partial derivative of $g(X;\theta)$ with respect to θ evaluated at θ_0.

Again, Assumptions 5 and 6 are commonly seen regularity conditions which are sufficient for the asymptotic normality of nonlinear estimators. Assumption 5 ensures that the first derivative of $Q_n(\gamma)$ admits a first-order Taylor expansion and the second derivative of $Q_n(\gamma)$ converges uniformly. Assumption 6 implies that the second derivative of $Q_n(\gamma)$ has a nonsingular limiting matrix. In practice these conditions can be checked similarly as in Examples 1 and 2. Throughout this paper, we use **0** to denote the vector of zeros of appropriate dimension.

Theorem 2 *Under Assumptions 1–6, as* $n \to \infty$, $\sqrt{n}(\hat{\gamma}_{\mathrm{SLS}} - \gamma_0) \xrightarrow{L} N(\mathbf{0}, A^{-1}BA^{-1})$*, where*

$$B = E\left[\frac{\partial\rho'(\gamma_0)}{\partial\gamma} W(X)\rho(\gamma_0)\rho'(\gamma_0)W(X)\frac{\partial\rho(\gamma_0)}{\partial\gamma'}\right]. \tag{3}$$

Remark 1 Note that A and B in the above asymptotic covariance matrix depend on the value of γ_0. In practice they can be estimated after $\hat{\gamma}_{\mathrm{SLS}}$ is obtained. From the proof of Theorem 2 in Sect. 6.3 and Lemma 2 in Sect. 6.1, it can be seen that

$$A = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial\rho_i'(\hat{\gamma}_{\mathrm{SLS}})}{\partial\gamma} W_i \frac{\partial\rho_i(\hat{\gamma}_{\mathrm{SLS}})}{\partial\gamma'}\right]$$

and

$$4B = \operatorname*{plim}_{n\to\infty} \frac{1}{n} \frac{\partial Q_n(\hat{\gamma}_{\mathrm{SLS}})}{\partial\gamma} \frac{\partial Q_n(\hat{\gamma}_{\mathrm{SLS}})}{\partial\gamma'},$$

where $\frac{\partial Q_n(\gamma)}{\partial\gamma} = 2\sum_{i=1}^n \frac{\partial\rho_i'(\gamma)}{\partial\gamma} W_i\rho_i(\gamma)$.

The asymptotic covariance $A^{-1}BA^{-1}$ of $\hat{\gamma}_{\mathrm{SLS}}$ depends on the weighting matrix W. A natural question is how to choose W to obtain the most efficient estimator. To answer this question, we first note that, since $\partial\rho'(\gamma_0)/\partial\gamma$ does not depend on Y, matrix B in (3) can be written as $B = E\left[\frac{\partial\rho'(\gamma_0)}{\partial\gamma} WUW \frac{\partial\rho(\gamma_0)}{\partial\gamma'}\right]$, where $U = U(X) = E[\rho(\gamma_0)\rho'(\gamma_0)|X]$. Then, analog to the weighted (nonlinear) least squares estimation, we have

$$A^{-1}BA^{-1} \geq \left(E\left[\frac{\partial\rho'(\gamma_0)}{\partial\gamma} U^{-1} \frac{\partial\rho(\gamma_0)}{\partial\gamma'}\right]\right)^{-1} \tag{4}$$

(in the sense that the difference of the left-hand and right-hand sides is nonnegative definite), and the lower bound is attained for $W = U^{-1}$ in both A and B (e.g., Hansen 1982, Abarin and Wang 2006). By definition, U is nonnegative definite and it will be shown that its determinant is det $U = \sigma_0^2(\mu_4 - \sigma_0^4) - \mu_3^2$, where $\mu_3 = E(\varepsilon^3|X)$ and $\mu_4 = E(\varepsilon^4|X)$. We have the following results.

Corollary 1 *If* $\sigma_0^2(\mu_4 - \sigma_0^4) - \mu_3^2 \neq 0$*, then the optimal weighting matrix is given by*

$$\begin{aligned} U^{-1} = &\frac{1}{\sigma_0^2(\mu_4 - \sigma_0^4) - \mu_3^2} \\ &\times \begin{pmatrix} \mu_4 + 4\mu_3 g(X;\theta_0) + 4\sigma_0^2 g^2(X;\theta_0) - \sigma_0^4 & -\mu_3 - 2\sigma_0^2 g(X;\theta_0) \\ -\mu_3 - 2\sigma_0^2 g(X;\theta_0) & \sigma_0^2 \end{pmatrix} \end{aligned} \tag{5}$$

and the asymptotic covariance matrix of the most efficient SLSE is given by

$$C = \begin{pmatrix} V\left(\hat\theta_{SLS}\right) & \frac{\mu_3}{\mu_4 - \sigma_0^4} V\left(\hat\sigma^2_{SLS}\right) G_2^{-1} G_1 \\ \frac{\mu_3}{\mu_4 - \sigma_0^4} V\left(\hat\sigma^2_{SLS}\right) G_1' G_2^{-1} & V\left(\hat\sigma^2_{SLS}\right) \end{pmatrix}, \tag{6}$$

where

$$V\left(\hat\theta_{SLS}\right) = \left(\sigma_0^2 - \frac{\mu_3^2}{\mu_4 - \sigma_0^4}\right)\left(G_2 - \frac{\mu_3^2}{\sigma_0^2(\mu_4 - \sigma_0^4)} G_1 G_1'\right)^{-1}, \tag{7}$$

$$V\left(\hat\sigma^2_{SLS}\right) = \frac{(\mu_4 - \sigma_0^4)\left(\sigma_0^2(\mu_4 - \sigma_0^4) - \mu_3^2\right)}{\sigma_0^2(\mu_4 - \sigma_0^4) - \mu_3^2 G_1' G_2^{-1} G_1} \tag{8}$$

and

$$G_1 = E\left[\frac{\partial g(X;\theta_0)}{\partial\theta}\right], \quad G_2 = E\left[\frac{\partial g(X;\theta_0)}{\partial\theta}\frac{\partial g(X;\theta_0)}{\partial\theta'}\right]. \tag{9}$$

Note that both the identity matrix I_2 and the optimal weighting matrix U^{-1} satisfy Assumption 6. In practice, however, U^{-1} involves unknown parameters which need to be estimated, before the optimal SLSE is computed. This can be done using the following two-stage procedure. First, minimize $Q_n(\gamma)$ using the identity weight $W = I_2$ to obtain the first-stage estimator $\hat\gamma_{SLS}$. Secondly, estimate the elements of U^{-1} using $\hat\gamma_{SLS}$ and the corresponding moments from the residuals. Finally, minimize $Q_n(\gamma)$ again with $W = \hat U^{-1}$ to obtain the second-stage estimator $\hat{\hat\gamma}_{SLS}$. Since $\hat U^{-1}$ is clearly consistent for U^{-1}, the asymptotic covariance matrix of the two-stage estimator $\hat{\hat\gamma}_{SLS}$

is the same as the right-hand side of (4) and, therefore, $\hat{\hat{\gamma}}_{\rm SLS}$ is asymptotically more efficient than the first-stage estimator $\hat{\gamma}_{\rm SLS}$. The OLS estimator can also be used to estimate U^{-1} if it is consistent and available. More discussion about the so-called feasible weighted least squares estimator can be found in Gallant (1987, Chap. 5).

Remark 2 The calculation of $\hat{\gamma}_{\rm SLS}$ entails numerical minimization of the quadratic form $Q_n(\gamma)$. Like ordinary nonlinear least squares estimation, this can be done using standard procedures such as Newton-Raphson, which is available in most mathematical or statistical computer packages. Given the modern computer power, the extra computational cost of the SLSE over OLSE is minimal.

3 Comparison with OLSE

The OLSE $\hat{\theta}_{\rm OLS}$ for θ is defined as the measurable function that minimizes $S_n(\theta) = \sum_{i=1}^{n}(Y_i - g(X_i;\theta))^2$ and the OLSE for σ^2 is $\hat{\sigma}^2_{\rm OLS} = S_n\left(\hat{\theta}_{\rm OLS}\right)/n$. The consistency and asymptotic normality of $\hat{\theta}_{\rm OLS}$ have been established under various sets of regularity conditions in the literature, e.g., Jennrich (1969) and Wu (1981). Traditionally, asymptotic properties of $\hat{\theta}_{\rm OLS}$ and $\hat{\sigma}^2_{\rm OLS}$ are separately derived and the authors were unable to find a reference giving the joint asymptotic variance-covariance matrix of the two estimators except for the special case where ε has a normal distribution.

In this section, we use a framework similar to the one used in Section 2 to derive the asymptotic covariance matrix of $\hat{\gamma}_{\rm OLS} = \left(\hat{\theta}'_{\rm OLS}, \hat{\sigma}^2_{\rm OLS}\right)'$. Note that $S_n(\theta)$ is a special case of $Q_n(\gamma)$ in (2), where the weighting matrix is taken as

$$W = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

However, this choice of weight violates Assumption 6, because it results in

$$A = \begin{pmatrix} G_2 & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix}$$

which is singular. Therefore, the result of Theorem 2 does not apply to the OLSE. However, the framework of Section 2 is still applicable with minor modification. In fact, the regularity conditions needed for the consistency and asymptotic normality of $\hat{\gamma}_{\rm OLS}$ are similar but weaker than those for the SLSE $\hat{\gamma}_{\rm SLS}$.

Assumption 7 $E\sup_{\Theta} g^2(X;\theta) < \infty$.

Assumption 8 The parameter space $\Theta \subset \mathbb{R}^p$ is compact.

Assumption 9 For any $\theta \in \Theta$, $E[(g(X;\theta) - g(X;\theta_0))^2] = 0$ if and only if $\theta = \theta_0$.

Assumption 10 θ_0 is an interior point of Θ and, for μ-almost all x, $g(x;\theta)$ is twice continuously differentiable in Θ. Furthermore, the first two derivatives satisfy

$$E\sup_{\Theta}\left\|\frac{\partial g(X;\theta)}{\partial\theta}\right\|^2 < \infty,\ E\sup_{\Theta}\left\|\frac{\partial^2 g(X;\theta)}{\partial\theta\partial\theta'}\right\|^2 < \infty.$$

Assumption 11 The matrix $G_2 = E\left[\frac{\partial g(X;\theta_0)}{\partial\theta}\frac{\partial g(X;\theta_0)}{\partial\theta'}\right]$ is nonsingular.

Theorem 3 *1. Under Assumptions 1 and 7–9, as* $n \to \infty$, $\hat{\gamma}_{\mathrm{OLS}} \xrightarrow{a.s.} \gamma_0$.

2. Under Assumptions 1 and 7–11, as $n \to \infty$, $\sqrt{n}(\hat{\gamma}_{\mathrm{OLS}} - \gamma_0) \xrightarrow{L} N(\mathbf{0}, D)$, *where*

$$D = \begin{pmatrix} \sigma_0^2 G_2^{-1} & \mu_3 G_2^{-1} G_1 \\ \mu_3 G_1' G_2^{-1} & \mu_4 - \sigma_0^4 \end{pmatrix}. \tag{10}$$

Now we compare the (optimal) SLSE and OLSE by their asymptotic covariances. First, it is easy to see from (6)–(8) and (10) that the SLSE and OLSE have the same asymptotic variance covariance matrix, if $\mu_3 = 0$. The following theorem shows that, however, the SLSE for θ and σ^2 are respectively more efficient than the corresponding OLSE if $\mu_3 \neq 0$.

Theorem 4 *Suppose* $\mu_3 \neq 0$.

1. $V\left(\hat{\sigma}^2_{\mathrm{OLS}}\right) \geq V\left(\hat{\sigma}^2_{\mathrm{SLS}}\right)$, *with equality holding if and only if* $G_1' G_2^{-1} G_1 = 1$.

2. $V\left(\hat{\theta}_{\mathrm{OLS}}\right) - V\left(\hat{\theta}_{\mathrm{SLS}}\right)$ *is nonnegative definite if* $G_1' G_2^{-1} G_1 = 1$, *and is positive definite if* $G_1' G_2^{-1} G_1 \neq 1$.

Note that the condition $G_1' G_2^{-1} G_1 = 1$ means that G_2^{-1} is a generalized inverse of the matrix $G_1 G_1'$.

4 Monte Carlo simulations

In this section, we investigate the finite sample behavior of the SLSE and compare it with OLSE through some simulation studies. In particular, we consider two models that are commonly used in nonlinear regression literature. The first is an exponential model

$$Y = \theta_1 \exp(\theta_2 X) + \varepsilon,$$

with true parameter values $\theta_1 = 10, \theta_2 = -0.6$ and $\sigma^2 = 2$; and the second is a growth model

$$Y = \frac{\theta_1}{1 + \exp(\theta_2 + \theta_3 X)} + \varepsilon,$$

with $\theta_1 = 10, \theta_2 = 1.5, \theta_3 = -0.8$ and $\sigma^2 = 2$. In both models $X \sim Uniform(0, 20)$ and $\varepsilon = (\chi^2(3) - 3)/\sqrt{3}$ follows a (normalized) $\chi^2(3)$ distribution. All observations are generated independently, and the Monte Carlo means of the OLSE and the SLSE with the estimated optimal weight using OLSE, and their variances (VAR) and mean squared errors (MSE) are computed. For each of the sample sizes $n = 30, 50, 100$ and 200, 1000 Monte Carlo repetitions are carried out. The computation is done using the statistical computing language R for Windows XP on an IBM Workstation with

Table 1 Simulated OLS, SLS, and their variances (VAR) and the mean squared errors (MSE) for exponential model $Y = \theta_1 \exp(\theta_2 X) + \varepsilon$

	OLS	VAR	MSE	SLS	VAR	MSE
$n = 30$						
$\theta_1 = 10$	10.0315	2.0245	2.0255	10.2306	1.6380	1.6895
$\theta_2 = -0.6$	−0.6139	0.0189	0.0190	−0.6282	0.0141	0.0149
$\sigma^2 = 2$	2.0027	0.7656	0.7648	1.7026	0.3093	0.3974
$n = 50$						
$\theta_1 = 10$	10.0238	1.4738	1.4743	10.1880	1.1669	1.2011
$\theta_2 = -0.6$	−0.6109	0.0141	0.0142	−0.6241	0.0100	0.0105
$\sigma^2 = 2$	1.9763	0.5194	0.5194	1.7733	0.2430	0.2941
$n = 100$						
$\theta_1 = 10$	9.9802	0.9863	0.9867	10.1146	0.6428	0.6553
$\theta_2 = -0.6$	−0.6032	0.0074	0.0074	−0.6133	0.0046	0.0048
$\sigma^2 = 2$	2.0061	0.2693	0.2694	1.8891	0.1573	0.1695
$n = 200$						
$\theta_1 = 10$	10.0153	0.5467	0.5469	10.0522	0.3361	0.3384
$\theta_2 = -0.6$	−0.6028	0.0038	0.0038	−0.6054	0.0023	0.0024
$\sigma^2 = 2$	2.0077	0.1129	0.1129	1.9504	0.0774	0.0798

a 2.2 MHz CPU and 4 GB RAM. The CPU time ranges from several to about thirty minutes, depending on the sample sizes.

The simulation results for the exponential model are presented in Table 1, while those for the growth model in Table 2. These results show a clear pattern of significant variance and mean squared error reduction of the SLSE over OLSE. In particular, the variance reduction for σ^2 in both models can be as high as $40-50\%$ in many instances. The only exception is the case of growth model with $n = 200$, where the SLSE has larger variance and MSE than OLSE. This is probably because that in this case the asymptotic variances of the two estimators are very close, so that the difference reflects random sampling or numerical fluctuations only. In addition, finite sample bias can be seen in the SLSE of σ^2. However, this bias is compensated by variance improvement because the corresponding overall MSE is always smaller than that of the OLSE.

5 Discussion

The least squares technique is widely used in regression analysis. If the random errors in the regression model have a normal distribution, then the ordinary least squares estimator is most efficient. However, if the random errors are not normally distributed, then information in the higher moments of the data can be used to construct more efficient estimator.

We have studied a SLSE for a general nonlinear model, where no distributional assumption for the random error is made. It has been shown that the SLSE using the

Table 2 Simulated OLS, SLS, and their variances (VAR) and the mean squared errors (MSE) for growth model $Y = \theta_1/[1 + \exp(\theta_2 + \theta_3 X)] + \varepsilon$

	OLS	VAR	MSE	SLS	VAR	MSE
$n = 30$						
$\theta_1 = 10$	10.0102	0.1064	0.1065	9.9380	0.0740	0.0778
$\theta_2 = 1.5$	1.5601	0.2481	0.2514	1.5221	0.2041	0.2044
$\theta_3 = -0.8$	−0.8374	0.0464	0.0478	−0.8285	0.0370	0.0378
$\sigma^2 = 2$	1.9747	0.8045	0.8052	1.6269	0.2988	0.4377
$n = 50$						
$\theta_1 = 10$	10.0137	0.0605	0.0606	9.9598	0.0437	0.0453
$\theta_2 = 1.5$	1.5516	0.1922	0.1947	1.5271	0.1293	0.1299
$\theta_3 = -0.8$	−0.8282	0.0350	0.0358	−0.8233	0.0242	0.0247
$\sigma^2 = 2$	2.0084	0.5521	0.5522	1.7566	0.2895	0.3485
$n = 100$						
$\theta_1 = 10$	9.9978	0.0348	0.0348	9.9701	0.0291	0.0299
$\theta_2 = 1.5$	1.5318	0.0914	0.0923	1.5132	0.0630	0.0631
$\theta_3 = -0.8$	−0.8132	0.0175	0.0177	−0.8077	0.0115	0.0115
$\sigma^2 = 2$	1.9690	0.2474	0.2481	1.8281	0.1950	0.2244
$n = 200$						
$\theta_1 = 10$	9.9997	0.0161	0.0161	9.9903	0.0167	0.0168
$\theta_2 = 1.5$	1.5123	0.0501	0.0502	1.5110	0.0334	0.0335
$\theta_3 = -0.8$	−0.8095	0.0095	0.0096	−0.8080	0.0064	0.0065
$\sigma^2 = 2$	2.0065	0.1255	0.1255	1.9326	0.1318	0.1363

optimal weight is asymptotically more efficient than the OLSE, if the random error in the model has a nonzero third moment. In the case of a symmetric error distribution, both estimators have the same asymptotic variance covariance matrix. It is also worthwhile to note that in the case of $\mu_3 = 0$ both the SLSE and OLSE for θ and σ^2 are asymptotically orthogonal, even without the normality assumption for the random error. Simulation studies show that the efficiency gain of SLSE over OLSE can be as high as 50% for the variance parameter for sample sizes lower than 100. Given the modern techniques of numerical computation and computer power, the extra computational cost of the SLSE over OLSE is ignorable in practice. Questions that deserve future investigation include finite sample properties, higher order efficiencies and comparisons of the SLSE with other existing estimators.

6 Proofs

6.1 Preliminary

For ease of reading we first restate some existing results which are used in the proofs. For this purpose, let $Z = (Z_1, Z_2, \ldots., Z_n)$ be an *i.i.d.* random sample and $\psi \in \Psi$

a vector of unknown parameters, where the parameter space $\Psi \subset \mathbb{R}^d$ is compact. Further, suppose $Q_n(Z, \psi)$ is a measurable function for each $\psi \in \Psi$ and is continuous in $\psi \in \Psi$ for μ–almost all Z. Then Lemmas 3 and 4 of Amemiya (1973) can be stated as follows.

Lemma 1 *If, as $n \to \infty$, $Q_n(Z, \psi)$ converges $a.s.$ to a nonstochastic function $Q(\psi)$ uniformly for all $\psi \in \Psi$ and $Q(\psi)$ attains a unique minimum at $\psi_0 \in \Psi$, then* $\hat{\psi}_n = \operatorname{argmin}_{\psi \in \Psi} Q_n(Z, \psi) \xrightarrow{a.s.} \psi_0$.

Lemma 2 *If, as $n \to \infty$, $Q_n(Z, \psi)$ converges $a.s.$ to a nonstochastic function $Q(\psi)$ uniformly for all ψ in an open neighborhood of ψ_0, then for any sequence of estimators $\hat{\psi}_n \xrightarrow{a.s.} \psi_0$ it holds $Q_n(Z, \hat{\psi}_n) \xrightarrow{a.s.} Q(\psi_0)$.*

Throughout the proofs, we use the following notations. For any matrix M, its Euclidean norm is denoted as $\|M\| = \sqrt{\operatorname{trace}(M'M)}$, and vec$M$ denotes the column vector consisting of the stacked up columns of M. Further, $\otimes$ denotes the Kronecker product operator.

6.2 Proof of Theorem 1

We show that Assumptions 1–4 are sufficient for all conditions of Lemma 1. First, it is easy to see that Assumption 1 implies that $Q_n(\gamma)$ is measurable and continuous in $\gamma \in \Gamma$ with probability one. Further, by Cauchy-Schwarz inequality and Assumptions 2 and 3 we have

$$E\left[\|W_1\| \sup_{\Theta} (Y_1 - g(X_1;\theta))^2\right] \le 2E\,\|W_1\|\,Y_1^2 + 2E\,\|W_1\| \sup_{\Theta} g^2(X_1;\theta) < \infty$$

and

$$E\left[\|W_1\| \sup_{\Gamma} \left(Y_1^2 - g^2(X_1;\theta) - \sigma^2\right)^2\right] \le 3E\,\|W_1\|\,Y_1^4 + 3E\,\|W_1\| \sup_{\Theta} g^4(X_1;\theta) + 3E\,\|W_1\| \sup_{\Sigma} \sigma^4 < \infty,$$

which imply

$$E \sup_{\Gamma} \rho_1'(\gamma) W_1 \rho_1(\gamma) \le E\,\|W_1\| \sup_{\Gamma} \|\rho_1(\gamma)\|^2 < \infty. \tag{11}$$

It follows from the uniform law of large numbers (ULLN) (Jennrich 1969) that $\frac{1}{n} Q_n(\gamma)$ converges almost surely ($a.s.$) to $Q(\gamma) = E\rho_1'(\gamma) W(Z_1) \rho_1(\gamma)$ uniformly for all $\gamma \in \Gamma$. Since $\rho_1(\gamma) - \rho_1(\gamma_0)$ does not depend on Y_1 we have

$$E[\rho_1'(\gamma_0) W_1 (\rho_1(\gamma) - \rho_1(\gamma_0))] = E[E(\rho_1'(\gamma_0)|X_1) W_1 (\rho_1(\gamma) - \rho_1(\gamma_0))] = 0,$$

which implies $Q(\gamma) = Q(\gamma_0) + E[(\rho_1(\gamma) - \rho_1(\gamma_0))' W_1 (\rho_1(\gamma) - \rho_1(\gamma_0))]$. It follows that $Q(\gamma) \geq Q(\gamma_0)$ and, by Assumption 4, equality holds if and only if $\gamma = \gamma_0$. Thus all conditions of Lemma 1 hold and, therefore, $\hat{\gamma}_{\mathrm{SLS}} \xrightarrow{a.s.} \gamma_0$ follows.

6.3 Proof of Theorem 2

By Assumption 5 the first derivative $\partial Q_n(\gamma)/\partial\gamma$ exists and has a first-order Taylor expansion in Γ. Since $\hat{\gamma}_n \xrightarrow{a.s.} \gamma_0$, for sufficiently large n it holds with probability one

$$\frac{\partial Q_n(\gamma_0)}{\partial\gamma} + \frac{\partial^2 Q_n(\tilde{\gamma}_n)}{\partial\gamma\partial\gamma'}(\hat{\gamma}_{\mathrm{SLS}} - \gamma_0) = \frac{\partial Q_n(\hat{\gamma}_{\mathrm{SLS}})}{\partial\gamma} = \mathbf{0}, \tag{12}$$

where $\|\tilde{\gamma}_n - \gamma_0\| \leq \|\hat{\gamma}_{\mathrm{SLS}} - \gamma_0\|$ and $\mathbf{0}$ is the $(p+1)$-vector of zeros. The first derivative of $Q_n(\gamma)$ in (12) is given by

$$\frac{\partial Q_n(\gamma)}{\partial\gamma} = 2\sum_{i=1}^{n} \frac{\partial\rho_i'(\gamma)}{\partial\gamma} W_i \rho_i(\gamma),$$

where

$$\frac{\partial\rho_i'(\gamma)}{\partial\gamma} = -\begin{pmatrix} \frac{\partial g(X_i;\theta)}{\partial\theta} & 2g(X_i;\theta)\frac{\partial g(X_i;\theta)}{\partial\theta} \\ 0 & 1 \end{pmatrix}.$$

The second derivative of $Q_n(\gamma)$ in (12) is given by

$$\frac{\partial^2 Q_n(\gamma)}{\partial\gamma\partial\gamma'} = 2\sum_{i=1}^{n}\left[\frac{\partial\rho_i'(\gamma)}{\partial\gamma} W_i \frac{\partial\rho_i(\gamma)}{\partial\gamma'} + (\rho_i'(\gamma) W_i \otimes I_{p+1})\frac{\partial \mathrm{vec}(\partial\rho_i'(\gamma)/\partial\gamma)}{\partial\gamma'}\right],$$

where

$$\frac{\partial \mathrm{vec}(\partial\rho_i'(\gamma)/\partial\gamma)}{\partial\gamma'} = -\begin{pmatrix} \frac{\partial^2 g(X_i;\theta)}{\partial\theta\partial\theta'} & \mathbf{0} \\ \mathbf{0} & 0 \\ 2g(X_i;\theta)\frac{\partial^2 g(X_i;\theta)}{\partial\theta\partial\theta'} + 2\frac{\partial g(X_i;\theta)}{\partial\theta}\frac{\partial g(X_i;\theta)}{\partial\theta'} & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix}.$$

By Assumption 5 and Cauchy–Schwarz inequality, we have

$$E\sup_{\Gamma}\left\|\frac{\partial\rho_1'(\gamma)}{\partial\gamma} W_1 \frac{\partial\rho_1(\gamma)}{\partial\gamma'}\right\| \leq E\,\|W_1\|\sup_{\Gamma}\left\|\frac{\partial\rho_1'(\gamma)}{\partial\gamma}\right\|^2$$

$$
\begin{aligned}
&\leq E\,\|W_1\| \sup_{\Gamma}\left(\left\|\frac{\partial g(X_1;\theta)}{\partial\theta}\right\|^2 + 4\left\|g(X_1;\theta)\frac{\partial g(X_1;\theta)}{\partial\theta}\right\|^2 + 1\right)\\
&\leq E\,\|W_1\| \sup_{\Gamma}\left\|\frac{\partial g(X_1;\theta)}{\partial\theta}\right\|^2 + E\,\|W_1\|\\
&\quad +4\left[E\left(\|W_1\| \sup_{\Gamma}\|g(X_1;\theta)\|^4\right)E\left(\|W_1\| \sup_{\Gamma}\left\|\frac{\partial g(X_1;\theta)}{\partial\theta}\right\|^4\right)\right]^{1/2} < \infty.
\end{aligned}
\tag{13}
$$

Similarly, because of (11) and

$$
\begin{aligned}
&E\left(\|W_1\| \sup_{\Gamma}\left\|\frac{\partial \mathrm{vec}(\partial\rho_1'(\gamma)/\partial\gamma)}{\partial\gamma'}\right\|^2\right)\\
&\leq E\,\|W_1\| \sup_{\Gamma}\left(\left\|\frac{\partial^2 g(X_1;\theta)}{\partial\theta\partial\theta'}\right\|^2 +4\left\|\frac{\partial g(X_1;\theta)}{\partial\theta}\right\|^4 +4\left\|g(X_1;\theta)\frac{\partial^2 g(X_1;\theta)}{\partial\theta\partial\theta'}\right\|^2\right)\\
&\leq E\left(\|W_1\| \sup_{\Gamma}\left\|\frac{\partial^2 g(X_1;\theta)}{\partial\theta\partial\theta'}\right\|^2\right) + 4E\left(\|W_1\| \sup_{\Gamma}\left\|\frac{\partial g(X_1;\theta)}{\partial\theta}\right\|^4\right)\\
&\quad +4\left[E\left(\|W_1\| \sup_{\Gamma}\|g(X_1;\theta)\|^4\right)E\left(\|W_1\| \sup_{\Gamma}\left\|\frac{\partial^2 g(X_1;\theta)}{\partial\theta\partial\theta'}\right\|^4\right)\right]^{1/2} < \infty,
\end{aligned}
$$

we have

$$
\begin{aligned}
&E \sup_{\Gamma}\left\|\left(\rho_1'(\gamma)W_1 \otimes I_{p+1}\right)\frac{\partial \mathrm{vec}(\partial\rho_1'(\gamma)/\partial\gamma)}{\partial\gamma'}\right\|\\
&\quad\leq (p+1)E\,\|W_1\| \sup_{\Gamma}\|\rho_1(\gamma)\|\left\|\frac{\partial \mathrm{vec}(\partial\rho_1'(\gamma)/\partial\gamma)}{\partial\gamma'}\right\|\\
&\quad\leq (p+1)\left[E\left(\|W_1\| \sup_{\Gamma}\|\rho_1(\gamma)\|^2\right)E\right.\\
&\qquad\left.\times\left(\|W_1\| \sup_{\Gamma}\left\|\frac{\partial \mathrm{vec}(\partial\rho_1'(\gamma)/\partial\gamma)}{\partial\gamma'}\right\|^2\right)\right]^{1/2} < \infty.
\end{aligned}
\tag{14}
$$

It follows from (13), (14) and the ULLN that

$$
\begin{aligned}
&\frac{1}{n}\frac{\partial^2 Q_n(\gamma)}{\partial\gamma\partial\gamma'} \xrightarrow{a.s.} \frac{\partial^2 Q(\gamma)}{\partial\gamma\partial\gamma'}\\
&\quad = 2E\left[\frac{\partial\rho_1'(\gamma)}{\partial\gamma}W_1\frac{\partial\rho_1(\gamma)}{\partial\gamma'} + (\rho_1'(\gamma)W_1 \otimes I_{p+1})\frac{\partial \mathrm{vec}(\partial\rho_1'(\gamma)/\partial\gamma)}{\partial\gamma'}\right]
\end{aligned}
$$

uniformly for all $\gamma \in \Gamma$. Therefore by Lemma 2 we have

$$\frac{1}{n}\frac{\partial^2 Q_n(\tilde{\gamma}_n)}{\partial\gamma\partial\gamma'} \xrightarrow{a.s.} \frac{\partial^2 Q(\gamma_0)}{\partial\gamma\partial\gamma'} = 2A, \tag{15}$$

where the second equality holds, because

$$\begin{aligned} &E\left[(\rho_1'(\gamma_0)W_1 \otimes I_{p+1})\frac{\partial \text{vec}(\partial\rho_1'(\gamma_0)/\partial\gamma)}{\partial\gamma'}\right] \\ &\quad = E\left[(E(\rho_1'(\gamma_0)|X_1)W_1 \otimes I_{p+1})\frac{\partial \text{vec}(\partial\rho_1'(\gamma_0)/\partial\gamma)}{\partial\gamma'}\right] = 0. \end{aligned}$$

Furthermore, since $\frac{\partial\rho_i'(\gamma)}{\partial\gamma}W_i\rho_i(\gamma)$ are $i.i.d.$ with zero mean, the Central Limit Theorem (CLT) implies that

$$\frac{1}{\sqrt{n}}\frac{\partial Q_n(\gamma_0)}{\partial\gamma} \xrightarrow{L} N(\mathbf{0}, 4B), \tag{16}$$

where B is given in (3). It follows from (12), (15), (16) and Assumption 6, that $\sqrt{n}(\hat{\gamma}_{\text{SLS}} - \gamma_0)$ converges in distribution to $N(\mathbf{0}, A^{-1}BA^{-1})$.

6.4 Proof of Corollary 1

First, by definition the elements of U are $u_{11} = E\left[(Y - g(X;\theta_0))^2\,|X\right] = \sigma_0^2$,

$$\begin{aligned} u_{22} &= E\left[\left(Y^2 - g^2(X;\theta_0) - \sigma_0^2\right)^2 |X\right] \\ &= \mu_4 + 4\mu_3 g(X;\theta_0) + 4\sigma_0^2 g^2(X;\theta_0) - \sigma_0^4 \end{aligned}$$

and

$$\begin{aligned} u_{12} &= E\left[(Y - g(X;\theta_0))\left(Y^2 - g^2(X;\theta_0) - \sigma_0^2\right)|X\right] \\ &= \mu_3 + 2\sigma_0^2 g(X;\theta_0). \end{aligned}$$

It follows that the determinant of U is $\det U = \sigma_0^2\left(\mu_4 - \sigma_0^4\right) - \mu_3^2$ which is nonzero by assumption. Then it is straightforward to calculate the inverse of U which is given by (5). Furthermore, since

$$\frac{\partial\rho'(\gamma_0)}{\partial\gamma} = -\begin{pmatrix} \frac{\partial g(X;\theta_0)}{\partial\theta} & 2g(X;\theta_0)\frac{\partial g(X;\theta_0)}{\partial\theta} \\ 0 & 1 \end{pmatrix},$$

it is also straightforward to show that the lower bound in (4) is given by

$$\left(E\left[\frac{\partial\rho'(\gamma_0)}{\partial\gamma}U^{-1}\frac{\partial\rho(\gamma_0)}{\partial\gamma'}\right]\right)^{-1} = \left(\sigma_0^2(\mu_4 - \sigma_0^4) - \mu_3^2\right)\begin{pmatrix} (\mu_4 - \sigma_0^4)G_2 & -\mu_3 G_1 \\ -\mu_3 G_1' & \sigma_0^2 \end{pmatrix}^{-1}.$$

Finally, the result follows by using the inversion formula for block partitioned matrices (Magnus and Neudecker 1988, p. 11).

6.5 Proof of Theorem 3

The proof of the consistency of $\hat{\gamma}_{\text{OLS}}$ follows the same line as that for Theorem 1. First, Assumption 1 implies that $S_n(\theta)$ is measurable and continuous in θ. Further, Assumption 7 implies

$$E\left[\sup_{\Theta}(Y_1 - g(X_1;\theta))^2\right] \leq 2EY_1^2 + 2E\sup_{\Theta} g^2(X_1;\theta) < \infty.$$

Hence by the ULLN $\frac{1}{n}S_n(\theta)$ converges $a.s.$ to $S(\theta) = E\left[(Y_1 - g(X_1;\theta))^2\right]$ uniformly for all $\theta \in \Theta$. Since

$$\begin{aligned} S(\theta) &= E\left[(Y_1 - g(X_1;\theta_0)^2\right] + E\left[(g(X_1;\theta) - g(X_1;\theta_0))^2\right] \\ &\quad + 2E\left[(Y_1 - g(X_1;\theta_0))(g(X_1;\theta_0) - g(X_1;\theta))\right] \\ &= S(\theta_0) + E\left[(g(X_1;\theta_0) - g(X_1;\theta))^2\right], \end{aligned}$$

it follows from Assumption 9 that $S(\theta) \geq S(\theta_0)$ and equality holds if and only if $\theta = \theta_0$. Therefore, $\hat{\theta}_{\text{OLS}} \xrightarrow{a.s.} \theta_0$ follows from Lemma 1. Moreover, by Lemma 2, $\hat{\sigma}^2_{\text{OLS}} = \frac{1}{n}S_n(\hat{\theta}_{\text{OLS}})$ converges $a.s.$ to $S(\theta_0) = \sigma_0^2$.

The proof of the asymptotic normality of $\hat{\gamma}_{\text{OLS}}$ is similar to that for Theorem 2. First, by Assumption 10, for sufficiently large n, the first derivative $\partial S_n(\theta)/\partial\theta$ admits the first order Taylor expansion

$$\frac{\partial S_n(\theta_0)}{\partial\theta} + \frac{\partial^2 S_n(\tilde{\theta}_n)}{\partial\theta\partial\theta'}(\hat{\theta}_{\text{OLS}} - \theta_0) = \frac{\partial S_n(\hat{\theta}_{\text{OLS}})}{\partial\theta} = \mathbf{0} \tag{17}$$

with probability 1, where $\left\|\tilde{\theta}_n - \theta_0\right\| \leq \left\|\hat{\theta}_{\text{OLS}} - \theta_0\right\|$,

$$\frac{\partial S_n(\theta)}{\partial\theta} = -2\sum_{i=1}^{n}(Y_i - g(X_i;\theta))\frac{\partial g(X_i;\theta)}{\partial\theta}$$

and

$$\frac{\partial^2 S_n(\theta)}{\partial\theta\partial\theta'} = 2\sum_{i=1}^{n}\left[\frac{\partial g(X_i;\theta)}{\partial\theta}\frac{\partial g(X_i;\theta)}{\partial\theta'} - (Y_i - g(X_i;\theta))\frac{\partial^2 g(X_i;\theta)}{\partial\theta\partial\theta'}\right].$$

Similarly, for $\hat{\sigma}^2_{\text{OLS}}$ we have

$$\begin{aligned}\hat{\sigma}^2_{\text{OLS}} &= \frac{1}{n} S_n(\hat{\theta}_{\text{OLS}}) \\ &= \frac{1}{n} S_n(\theta_0) + \frac{1}{n} \frac{\partial S_n(\tilde{\theta}_n)}{\partial \theta'} (\hat{\theta}_{\text{OLS}} - \theta_0),\end{aligned} \tag{18}$$

where $\left\| \tilde{\theta}_n - \theta_0 \right\| \leq \left\| \hat{\theta}_{\text{OLS}} - \theta_0 \right\|$. Here $\tilde{\theta}_n$ may be different from that in (17), but this will not cause confusion subsequently. Combining (17) and (18) we have

$$M_n \left(\hat{\gamma}_{\text{OLS}} - \gamma_0 \right) = \frac{1}{n} \sum_{i=1}^{n} Z_i, \tag{19}$$

where

$$M_n = \begin{pmatrix} \frac{1}{n} \frac{\partial^2 S_n(\tilde{\theta}_n)}{\partial \theta \partial \theta'} & \mathbf{0} \\ -\frac{1}{n} \frac{\partial S_n(\tilde{\theta}_n)}{\partial \theta'} & 1 \end{pmatrix}$$

and

$$Z_i = \begin{pmatrix} 2(Y_i - g(X_i; \theta_0)) \frac{\partial g(X_i; \theta_0)}{\partial \theta} \\ (Y_i - g(X_i; \theta_0))^2 - \sigma_0^2 \end{pmatrix}.$$

By Assumption 10 and Cauchy–Schwarz inequality we have

$$E \sup_{\Theta} \left\| \frac{\partial g(X_1; \theta)}{\partial \theta} \frac{\partial g(X_1; \theta)}{\partial \theta'} \right\| = E \sup_{\Theta} \left\| \frac{\partial g(X_1; \theta)}{\partial \theta} \right\|^2 < \infty$$

and

$$\begin{aligned}&\left(E \sup_{\Theta} \left\| (Y_1 - g(X_1; \theta)) \frac{\partial^2 g(X_1; \theta)}{\partial \theta \partial \theta'} \right\| \right)^2 \\ &\leq E \sup_{\Theta} \| Y_1 - g(X_1; \theta) \|^2 \, E \sup_{\Theta} \left\| \frac{\partial^2 g(X_1; \theta)}{\partial \theta \partial \theta'} \right\|^2 \\ &\leq \left(2E \| Y_1 \|^2 + 2E \sup_{\Theta} g^2(X_1; \theta) \right) E \sup_{\Theta} \left\| \frac{\partial^2 g(X_1; \theta)}{\partial \theta \partial \theta'} \right\|^2 < \infty.\end{aligned}$$

It follow from the ULLN that $\frac{1}{n} \frac{\partial^2 S_n(\theta)}{\partial \theta \partial \theta'} \xrightarrow{a.s.} \frac{\partial^2 S(\theta)}{\partial \theta \partial \theta'}$ uniformly for all $\theta \in \Theta$. Therefore by Lemma 2

$$\frac{1}{n} \frac{\partial^2 S_n(\tilde{\theta}_n)}{\partial \theta \partial \theta'} \xrightarrow{a.s.} \frac{\partial^2 S(\theta_0)}{\partial \theta \partial \theta'} = 2E \left[\frac{\partial g(X_1; \theta_0)}{\partial \theta} \frac{\partial g(X_1; \theta_0)}{\partial \theta'} \right], \tag{20}$$

where the second equality holds because $E\left[(Y_1 - g(X_1;\theta_0))\frac{\partial^2 g(X_1;\theta_0)}{\partial\theta\partial\theta'}\right] = 0$. Similarly, since

$$\begin{aligned}
&\left(E\sup_{\Theta}\left\|(Y_1 - g(X_1;\theta))\frac{\partial g(X_1;\theta)}{\partial\theta}\right\|\right)^2\\
&\le E\sup_{\Theta}\|Y_1 - g(X_1;\theta)\|^2\, E\sup_{\Theta}\left\|\frac{\partial g(X_1;\theta)}{\partial\theta}\right\|^2\\
&\le \left(2E\|Y_1\|^2 + 2E\sup_{\Theta} g^2(X_1;\theta)\right) E\sup_{\Theta}\left\|\frac{\partial g(X_1;\theta)}{\partial\theta}\right\|^2 < \infty,
\end{aligned}$$

we have, by the ULLN and Lemma 2

$$-\frac{1}{n}\frac{\partial S_n(\tilde{\theta}_n)}{\partial\theta'} \xrightarrow{a.s.} 2E\left[(Y_1 - g(X_1;\theta_0))\frac{\partial g(X_1;\theta_0)}{\partial\theta}\right] = 0. \tag{21}$$

Equations (20) and (21) imply that

$$M_n \xrightarrow{a.s.} M = \begin{pmatrix} 2G_2 & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix}. \tag{22}$$

Further, since $\{Z_i\}$ are $i.i.d.$ with zero mean, the CLT implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \xrightarrow{L} N(0, H), \tag{23}$$

where

$$\begin{aligned}
H &= \begin{pmatrix} 4E\left[(Y_1 - g(X_1;\theta_0))^2\frac{\partial g(X_1;\theta_0)}{\partial\theta}\frac{\partial g(X_1;\theta_0)}{\partial\theta'}\right] & 2E\left[(Y_1 - g(X_1;\theta_0))^3\frac{\partial g(X_1;\theta_0)}{\partial\theta}\right] \\ 2E\left[(Y_1 - g(X_1;\theta_0))^3\frac{\partial g(X_1;\theta_0)}{\partial\theta'}\right] & E\left[(Y_1 - g(X_1;\theta_0))^4\right] - \sigma_0^4 \end{pmatrix}\\
&= \begin{pmatrix} 4\sigma_0^2 G_2 & 2\mu_3 G_1 \\ 2\mu_3 G_1' & \mu_4 - \sigma_0^4 \end{pmatrix}.
\end{aligned}$$

Finally by (19), (22) and (23) we have $\sqrt{n}(\hat{\gamma}_{\text{OLS}} - \gamma_0) \xrightarrow{L} N(\mathbf{0}, M^{-1}HM^{-1})$, where $M^{-1}HM^{-1} = D$ is given in (10).

6.6 Proof of Theorem 4

First, since

$$1 - G_1' G_2^{-1} G_1 = E\left[\left(1 - G_1' G_2^{-1}\frac{\partial g(X;\theta_0)}{\partial\theta}\right)^2\right] \ge 0,$$

it is easy to see from (8) that $V\left(\hat{\sigma}^2_{\mathrm{SLS}}\right) \le \mu_4 - \sigma_0^4 = V\left(\hat{\sigma}^2_{\mathrm{OLS}}\right)$, and the equality holds if and only if $G_1' G_2^{-1} G_1 = 1$. To show that $V\left(\hat{\theta}_{\mathrm{OLS}}\right) - V\left(\hat{\theta}_{\mathrm{SLS}}\right)$ is nonnegative definite, we use the inverse formula for block partitioned matrices to rewrite $V\left(\hat{\theta}_{\mathrm{SLS}}\right)$ as

$$\begin{aligned} V\left(\hat{\theta}_{\mathrm{SLS}}\right) &= \left(\sigma_0^2 - \frac{\mu_3^2}{\mu_4 - \sigma_0^4}\right) G_2^{-1} + \frac{\mu_3^2}{(\mu_4 - \sigma_0^4)^2} V\left(\hat{\sigma}^2_{\mathrm{SLS}}\right) G_2^{-1} G_1 G_1' G_2^{-1} \\ &= \sigma_0^2 G_2^{-1} - \frac{\mu_3^2}{\mu_4 - \sigma_0^4}\left(G_2^{-1} - \frac{1}{\mu_4 - \sigma_0^4} V\left(\hat{\sigma}^2_{\mathrm{SLS}}\right) G_2^{-1} G_1 G_1' G_2^{-1}\right) \\ &= V\left(\hat{\theta}_{\mathrm{OLS}}\right) - \frac{\mu_3^2}{\mu_4 - \sigma_0^4} G_2^{-1/2} (I_p - M) G_2^{-1/2}, \end{aligned} \tag{24}$$

where I_p is the p-dimensional identity matrix and $M = \frac{V\left(\hat{\sigma}^2_{\mathrm{SLS}}\right)}{\mu_4 - \sigma_0^4} G_2^{-1/2} G_1 G_1' G_2^{-1/2}$. Because M has the same nonzero eigenvalue as $\frac{V\left(\hat{\sigma}^2_{\mathrm{SLS}}\right)}{\mu_4 - \sigma_0^4} G_1' G_2^{-1} G_1 \le 1$, $I_p - M$ is nonnegative definite. It follows from (24) that $V\left(\hat{\theta}_{\mathrm{OLS}}\right) - V\left(\hat{\theta}_{\mathrm{SLS}}\right)$ is nonnegative definite. Moreover, $I_p - M$, and therefore $V\left(\hat{\theta}_{\mathrm{OLS}}\right) - V\left(\hat{\theta}_{\mathrm{SLS}}\right)$, is positive definite if and only if $G_1' G_2^{-1} G_1 \ne 1$.

Acknowledgements This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors thank two referees and the editor for their constructive comments and suggestions.

References

Abarin, T., Wang, L. (2006). Comparison of GMM with second-order least squares estimator in nonlinear models. *Far East Journal of Theoretical Statistics*, *20*, 179–196.

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica*, *41*, 997–1016.

Azzalini, A., Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society B*, *61*, 579–602.

Boos, D.D. (1987). Detecting skewed errors from regression residuals. *Technometrics*, *29*, 83–90.

Gallant, A.R. (1987). *Nonlinear statistical models*. New York: Wiley.

Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*, 1029–1054.

Hutson, A. (2004). Utilizing the flexibility of the epsilon-skew-normal distribution for common regression problems. *Journal of Applied Statistics*, *31*, 673–683.

Jennrich, R.I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics*, *40*, 633–643.

Magnus, J.R., Neudecker, H. (1988). *Matrix differential calculus with application in statistics and econometrics*. New York: Wiley.

Marazzi, A., Yohai, V.J. (2004). Adaptively truncated maximum likelihood regression with asymmetric errors. *Journal of Statistical Planning and Inference*, *122*, 271–291.

Sahu, S.K., Dey, D.K., Branco, M.D. (2003). A new class of multivariate skew distributions with application to Bayesian regression models. *The Canadian Journal of Statistics*, *31*, 129–150.

Seber, G.A., Wild, C.J. (1989). *Nonlinear regression*. New York: Wiley.

Theodossiou, P. (1998). Fianacial data and the skewed generalized T distribution. *Management Science*, *44*, 1650–1661.

Wang, L. (2003). Estimation of nonlinear Berkson-type measurement error models. *Statistica Sinica*, *13*, 1201–1210.

Wang, L. (2004). Estimation of nonlinear models with Berkson measurement errors. *Annals of Statistics*, *32*, 2559–2579.

Williams, M.S. (1997). A regression technique accounting for heteroscedastic and asymmetric errors. *Journal of Agricultural Biological and Environmental Statistics*, *2*, 108–129.

Wu, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics*, *9*, 501–513.