

Second-Order Least Squares Method for Statistical Inference in  
Linear Models with Asymmetric Errors

by

Sanha Kil

A thesis submitted to the Faculty of Graduate and Postdoctoral Studies of  
The University of Manitoba  
in partial fulfillment of the requirements of the degree of

Master of Science

Department of Statistics  
University of Manitoba  
Winnipeg

Copyright © 2026 by Sanha Kil

## **Abstract**

The ordinary least squares estimator (OLSE) is the best linear unbiased estimator (BLUE) when the Gauss–Markov assumptions are satisfied. However, if the random error term has an asymmetric distribution, more efficient estimators can be constructed by considering higher moments of the data. Wang (2003, 2004) showed that the second-order least squares estimator (SLSE) can be constructed using the first two conditional moments of the response variable. Additionally, Wang and Leblanc (2008) showed that the SLSE is asymptotically more efficient than the OLSE in nonlinear models when the third moment of the error is nonzero. We apply this approach to linear models and develop narrower confidence intervals and more powerful Wald-type hypothesis tests, as demonstrated through simulation studies. Furthermore, we develop a numerical algorithm with improved stability for computing the SLSE. The thesis also presents two real-data applications.

# Acknowledgements

First, I express my gratitude to my primary supervisor, Dr. Liquan Wang, and my co-supervisor, Dr. Yuliya Martsynyuk, whose guidance and expertise have been instrumental throughout my research on the numerical optimization of the SLSE. Without their continuous support, this thesis would not have been possible.

I also sincerely thank my advisory committee members, Drs. Xiaoting Li and Po Yang, for their valuable suggestions and constructive feedback, which greatly improved the quality of this work.

Special thanks go to the Department of Statistics and the Faculty of Graduate Studies at the University of Manitoba, as well as the University of Manitoba Graduate Fellowship (UMGF), for the financial support provided throughout my study.

Finally, I am grateful to my parents for their constant encouragement and belief in me. I would also like to express my appreciation to Corrie and Mike for their great support and help, and to Maggie for her unwavering love and support throughout my graduate program.

# Dedication

*To Maggie, for her genuine support and love*

# Contents

|                                                               |           |
|---------------------------------------------------------------|-----------|
| <b>Abstract</b>                                               | <b>i</b>  |
| <b>Acknowledgements</b>                                       | <b>i</b>  |
| <b>List of Tables</b>                                         | <b>v</b>  |
| <b>1 Introduction</b>                                         | <b>1</b>  |
| 1.1 Motivation . . . . .                                      | 1         |
| 1.2 Linear Regression and OLSE . . . . .                      | 2         |
| 1.3 Second-Order Least Squares Estimation . . . . .           | 4         |
| 1.4 Objectives and Contributions of the Thesis . . . . .      | 5         |
| <b>2 SLSE theory</b>                                          | <b>7</b>  |
| 2.1 SLSE Objective Function . . . . .                         | 7         |
| 2.2 Asymptotic Properties of the SLSE . . . . .               | 9         |
| 2.3 Efficiency of the SLSE Compared to the OLSE . . . . .     | 11        |
| <b>3 Numerical Optimization for the SLSE</b>                  | <b>13</b> |
| 3.1 Iteratively Reweighted Newton Method . . . . .            | 13        |
| 3.1.1 Newton-Raphson Algorithm . . . . .                      | 13        |
| 3.1.2 Iteratively Reweighted Newton Method for SLSE . . . . . | 15        |
| 3.2 Stabilization of the Newton Method . . . . .              | 17        |

|          |                                                             |           |
|----------|-------------------------------------------------------------|-----------|
| 3.2.1    | Ridge Regularization . . . . .                              | 18        |
| 3.2.2    | Armijo-Type Line Search . . . . .                           | 19        |
| 3.2.3    | Reweighted Ridge–Newton Method with Line Search . . . . .   | 21        |
| 3.3      | Numerical Stability Comparison . . . . .                    | 24        |
| <b>4</b> | <b>Simulation</b>                                           | <b>25</b> |
| 4.1      | Simulation Design . . . . .                                 | 25        |
| 4.2      | Numerical Stability of the Optimization Algorithm . . . . . | 27        |
| 4.3      | Finite-Sample Performance of the SLSE . . . . .             | 28        |
| 4.4      | Hypothesis Testing Performance . . . . .                    | 35        |
| <b>5</b> | <b>Real Data Application</b>                                | <b>40</b> |
| 5.1      | Australian Institute of Sport (AIS) Dataset . . . . .       | 40        |
| 5.2      | Health Insurance Dataset . . . . .                          | 43        |
| <b>6</b> | <b>Conclusion</b>                                           | <b>46</b> |
| <b>7</b> | <b>Appendix</b>                                             | <b>50</b> |
| 7.1      | Preliminary . . . . .                                       | 50        |
| 7.2      | Derivation of gradient and Hessian matrix . . . . .         | 51        |
| 7.3      | Proof of Theorem 1 . . . . .                                | 53        |
| 7.4      | Proof of Theorem 2 . . . . .                                | 55        |
| 7.5      | R simulation code . . . . .                                 | 56        |

# List of Tables

|     |                                                                                                                                                                                                                              |    |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1 | Bad sample rates across sample sizes for the standard reweighted Newton method and the stabilized Reweighted Ridge–Newton method with line search.                                                                           | 24 |
| 4.1 | Average ASE/SE ratios for OLSE and SLSE across simulation settings. . . .                                                                                                                                                    | 30 |
| 4.2 | OLSE vs. SLSE, and OLS CI vs. SLS CI for Model 1 . . . . .                                                                                                                                                                   | 33 |
| 4.3 | OLSE vs. SLSE, and OLS CI vs. SLS CI for Model 2 . . . . .                                                                                                                                                                   | 34 |
| 4.4 | Power comparison under different error distributions. $X_1 = 1$ , $X_2 \sim \text{Unif}(1, 20)$ , $X_3 \sim \mathcal{N}(5, 4)$ . . . . .                                                                                     | 37 |
| 4.5 | Power comparison under different error distributions. $X_1 \sim \text{Beta}(2, 5)$ , $(X_2, X_3) \sim \mathcal{N}\left(\begin{bmatrix} 20 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0.50.5 & 1 \end{bmatrix}\right)$ . . . . . | 37 |
| 5.1 | Comparison of OLSE and SLSE for the AIS dataset . . . . .                                                                                                                                                                    | 42 |
| 5.2 | Comparison of OLSE and SLSE for the standardized health insurance dataset                                                                                                                                                    | 45 |

# 1

## Introduction

### 1.1 Motivation

Linear models are commonly used in regression analysis across various scientific fields. Under the Gauss–Markov assumptions, the ordinary least squares estimator (OLSE) is considered the best linear unbiased estimator (BLUE). However, in many real-world situations, such as in environmental research and economics, data can be asymmetrically distributed. When the errors are skewed, the OLSE may not be the most efficient choice. In these cases, accounting for higher-order moments can help achieve more efficient estimates.

One example is the second-order least squares (SLS) method introduced by Wang and Leblanc (2008), which uses the first two conditional moments of the response variable given the predictors. The SLS estimator (SLSE) is shown to be asymptotically more efficient than the OLSE when the model error has a nonzero third moment, while the two estimators are asymptotically equivalent otherwise. Moreover, the SLSE attains the optimal semiparametric efficiency bound (Kim and Ma, 2012). This framework is important because asymmetric data is common in real-world studies. Such uneven data distributions often appear in practical data analysis. Beguería and Vicente-Serrano (2006), Green (1984a), Wright et al. (1982), Yousof et al. (2018), Castro-Camilo et al. (2022), and Nandy et al. (2025).

Since its introduction, the SLS method has been applied to a variety of regression settings, including measurement error models (Wang, 2003, 2004), mixed effects models (Wang, 2007), censored linear models (Abarin and Wang, 2009), and dynamic models (Salamh and Wang, 2021a,b). More recently, it has also been studied in the context of optimal design for linear models (Gao and Zhou, 2014; Bose and Mukerjee, 2015; Yin and Zhou, 2017; Yeh and Zhou, 2021; He and Yue, 2022; He et al., 2024). Despite the increasing volume of research on SLS, most of the literature has focused on point estimation and design issues.

Although the asymptotic theory of the SLSE has been worked out for general nonlinear regression models, its properties in linear models haven't been studied much. Specifically, methods for constructing confidence intervals and conducting hypothesis tests based on the SLSE remain underdeveloped. Also, unlike the OLSE, the SLSE doesn't have a closed-form formula, which makes numerical computation tricky and sometimes unstable, requiring more robust optimization algorithms.

The main issue is not whether the SLSE significantly improves point estimates, but whether the decrease in asymptotic variance results in narrower confidence intervals and increased power in finite samples while controlling type I error. This thesis addresses these questions by deriving additional properties of the SLSE in linear models and examining its performance in confidence interval construction and hypothesis testing. Wald-type statistics are built for inference, demonstrating that the SLSE provides more precise confidence intervals and more powerful tests than the OLSE when errors are asymmetric. For practical implementation, we also introduce a robust numerical algorithm based on an iteratively reweighted Ridge-Newton method with Armijo backtracking line search to compute the SLSE.

## 1.2 Linear Regression and OLSE

Consider the linear regression model:

$$Y_i = X_i' \beta + \epsilon_i, \quad i = 1, 2, \dots, n, \quad (1.1)$$

where  $Y_i$  is the ordinary response variable,  $X_i$  is a vector of covariates,  $\beta$  is the vector of unknown regression coefficients, and  $\epsilon_i$  is the random error. We assume that the error term has a mean of zero,  $E(\epsilon|X) = 0$ , and a constant variance,  $E(\epsilon^2|X) = \sigma^2$ . Given the independent and identically distributed (*i.i.d.*) data  $(y_i, x_i)$  for  $i = 1, 2, \dots, n$ , the ordinary least squares estimator (OLSE), denoted as  $\hat{\beta}_{OLS}$ , is obtained by minimizing the sum of squared residuals.

$$RSS(\beta) = \sum_{i=1}^n (y_i - x_i' \beta)^2$$

and RSS can be written in matrix form:

$$RSS(\beta) = (Y - X\beta)'(Y - X\beta)$$

where  $Y = (y_1, \dots, y_n)'$  and  $X = (x_1, \dots, x_n)'$ . Then OLSE can be obtained as such:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

The OLSE is unbiased,  $E(\hat{\beta}_{OLS}) = \beta$ , and the variance of estimators  $Var(\hat{\beta}_{OLS})$  (conditional on  $X$ ) is given by  $Var(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}$ , and  $\sigma^2$  is typically estimated after obtaining  $\hat{\beta}_{OLS}$  as

$$\hat{\sigma}_{OLS}^2 = \frac{1}{n-p} RSS(\hat{\beta}_{OLS}).$$

In addition to its finite-sample properties under Gauss–Markov conditions, the ordinary least squares estimator also has desirable properties in large samples, such as consistency.

Under standard regularity conditions, it holds that

$$\hat{\beta}_{OLS} \xrightarrow{p} \beta, \quad n \rightarrow \infty.$$

This result is derived from the law of large numbers and does not require any distributional assumption.

Moreover, when the normality assumption  $\epsilon \sim N(0, \sigma^2 I_n)$  is held, the conditional distribution of OLSE can be stated as follows:

$$\hat{\beta}_{OLS} \sim N(\beta, \sigma^2 (X'X)^{-1})$$

Thus, supposing that  $n^{-1}X'X \xrightarrow{p} M_{xx}$ , it holds that

$$\sqrt{n}(\hat{\beta}_{OLS} - \beta) \xrightarrow{d} N(0, \sigma^2 M_{xx}^{-1}),$$

given that standard regularity conditions hold.

As discussed in Section 1.1, under the Gauss-Markov assumptions, the OLSE is the best linear unbiased estimator (BLUE). The normality assumption is not required for the Gauss-Markov result, even though it leads to the exact finite-sample normal distribution of the OLSE. However, if the model error follows an asymmetric distribution, taking into account only the first moment (the mean) fails to capture the information from the skewness of the distribution. Therefore, the higher moments of the data can be utilized to build a more efficient estimator. This provides a reason for the use of the second-order least squares (SLS) method, which accounts for the first two conditional moments of  $Y$  given  $X$  ( $E(Y|X)$  and  $E(Y^2|X)$ ) to improve estimation efficiency and statistical power in the presence of asymmetric errors. The SLS method is introduced in the next section.

### 1.3 Second-Order Least Squares Estimation

The second-order least squares (SLS) method builds upon the ordinary least squares approach by including information from both the first and second conditional moments of the response variable. For the linear regression model described in (1.1), the first two conditional moments

of  $Y$  given  $X$  are

$$E(Y_i|X_i) = X_i'\beta, \quad E(Y_i^2|X_i) = (X_i'\beta)^2 + \sigma^2, \quad i = 1, 2, \dots, n,$$

where  $\sigma^2 = \text{Var}(\epsilon|X)$ . While the ordinary least squares estimator is obtained from only the first conditional moment, the second-order least squares (SLS) method uses both the first and second moments to utilize information related to the skewness of the error distribution.

Wang and Leblanc (2008) introduced the idea by defining an estimator that minimizes the quadratic formula based on these two conditional moments. Given an *i.i.d.* sample  $(y_i, x_i)$ , the SLSE is to minimize

$$Q_n(\gamma) = \sum_{i=1}^n \rho_i(\gamma)' W_i \rho_i(\gamma), \quad (1.9)$$

where

$$\rho_i(\gamma) = \begin{pmatrix} y_i - x_i'\beta \\ y_i^2 - (x_i'\beta)^2 - \sigma^2 \end{pmatrix}, \quad (1.10)$$

and  $W_i$  is a nonnegative definite weighting matrix that may depend on the covariates  $x_i$ .

Under suitable regularity conditions, Wang and Leblanc (2008) showed that the SLSE is strongly consistent and asymptotically normal. More importantly, they demonstrated that the SLSE is asymptotically more efficient than the OLSE when the third moment of the error distribution is nonzero in general nonlinear models, while the two estimators are asymptotically equivalent when the error distribution is symmetric.

## 1.4 Objectives and Contributions of the Thesis

The primary objective of this thesis is to investigate the second-order least squares estimator in the context of statistical inference for linear regression models with asymmetric error distributions. While the asymptotic efficiency of the SLSE has been established in nonlinear models by Wang and Leblanc (2008), most existing work has focused on point estimation,

leaving inference-related problems underexplored.

This thesis makes several contributions. First, we focus the SLS framework on linear regression models and derive the asymptotic variance forms of the regression coefficients and the error variance for the SLSE. Based on these results, we construct Wald-type confidence intervals and hypothesis tests for the regression parameters and investigate their theoretical properties.

Second, we address the numerical challenges associated with computing the SLSE. We propose a stabilized estimation algorithm based on an iteratively reweighted Ridge–Newton method combined with Armijo backtracking line search. This approach improves numerical stability by regularizing the Hessian matrix and ensuring reliable convergence under skewed error distributions, while keeping the computation time relatively low.

Third, we conduct Monte Carlo simulation studies to evaluate the finite-sample performance of the SLSE compared to the OLSE. The comparison focuses on confidence interval length, coverage probability, and power testing. The simulation results show that the SLSE produces more precise confidence intervals and more powerful hypothesis tests in the presence of asymmetric errors as sample size increases, while showing similar performance under symmetric error distributions.

Finally, we evaluate the proposed methodology through applications to real datasets, including data from the Australian Institute of Sports (AIS) and a health insurance dataset. These examples show the practical benefits of including second-order moment information.

## 2

# SLSE theory

## 2.1 SLSE Objective Function

The second-order least squares estimator is defined through a quadratic objective function constructed from the first two conditional moments of the response variable. For the linear regression model introduced in (1.1), the first two conditional moments are given by

$$\mathbb{E}(Y_i | X_i) = X_i' \beta, \quad \mathbb{E}(Y_i^2 | X_i) = (X_i' \beta)^2 + \sigma^2, \quad i = 1, 2, \dots, n. \quad (2.1.1)$$

Let  $\gamma = (\beta', \sigma^2)' \in \mathbb{R}^{p+1}$  denote the vector of unknown parameters. To incorporate both moment conditions into estimation, define the moment-based residual vector

$$\rho_i(\gamma) = \begin{pmatrix} Y_i - X_i' \beta \\ Y_i^2 - (X_i' \beta)^2 - \sigma^2 \end{pmatrix}, \quad i = 1, 2, \dots, n. \quad (2.1.2)$$

The SLSE is obtained by minimizing the quadratic form of these residuals:

$$Q_n(\gamma) = \sum_{i=1}^n \rho_i(\gamma)' W_i \rho_i(\gamma), \quad (2.1.3)$$

where  $W_i = W(X_i)$  is a  $2 \times 2$  nonnegative definite weight matrix that may depend on the

covariates  $X_i$ .

The estimator is then defined as

$$\hat{\gamma}_{SLS} = \arg \min_{\gamma \in \Gamma} Q_n(\gamma), \quad (2.1.4)$$

where  $\Gamma \subset \mathbb{R}^{p+1}$  is the parameter space.

The choice of the weight matrix  $W_i$  plays an important role in determining the efficiency of the estimator. In particular, it has been shown by Wang and Leblanc (2008) that the most efficient SLSE is obtained by taking

$$W_i = \mathbb{E}[\rho_i(\gamma_0)\rho_i(\gamma_0)' \mid X_i]^{-1}, \quad (2.1.5)$$

where  $\gamma_0 = (\beta_0', \sigma_0^2)'$  denotes the true parameter value.

For linear regression models, the optimal weight matrix holds the form:

$$W_i = \eta \begin{pmatrix} \mu_4 - \sigma_0^4 + 4\mu_3(X_i'\beta_0) + 4\sigma_0^2(X_i'\beta_0)^2 & -\mu_3 - 2\sigma_0^2(X_i'\beta_0) \\ -\mu_3 - 2\sigma_0^2(X_i'\beta_0) & \sigma_0^2 \end{pmatrix}, \quad (2.1.6)$$

where  $\mu_j = \mathbb{E}(\epsilon^j \mid X)$  for  $j = 3, 4$ , and  $\eta = [\sigma_0^2(\mu_4 - \sigma_0^4) - \mu_3^2]^{-1}$ .

It is important to note that the objective function  $Q_n(\gamma)$  cannot be optimized as a closed-form minimizer, in contrast to the ordinary least squares estimator. As a result, numerical optimization methods are required to compute the SLSE in practice.

Finally, we note that the classical least squares objective function is a special case of (2.1.3) when the weight matrix is chosen as

$$W_i = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

in which case the objective function reduces to

$$Q_n(\gamma) = \sum_{i=1}^n (Y_i - X_i' \beta)^2,$$

which depends only on the first moment condition. This shows that the SLSE includes the ordinary least squares framework and utilizes the additional information from the second conditional moment on top of that.

## 2.2 Asymptotic Properties of the SLSE

In this section, we establish the large-sample properties of the second-order least squares estimator (SLSE). Under the assumptions below, the SLSE is consistent and asymptotically normal.

**Assumption 1.** *The error  $\varepsilon$  satisfies  $E(\varepsilon \mid X) = 0$  and  $E(\varepsilon^2 \mid X) = \sigma_0^2 > 0$  and has finite moments  $\mu_j = E(\varepsilon^j \mid X)$  for  $j = 3, 4$ , and  $\sigma_0^2(\mu_4 - \sigma_0^4) - \mu_3^2 > 0$ .*

**Assumption 2.** *The predictor  $X$  satisfies  $E(\|X\|^4) < \infty$  and  $M_{xx} = E(XX')$  is positive definite.*

**Assumption 3.** *The true value  $\gamma_0 \in \text{int}(\Gamma)$  and the parameter space  $\Gamma \subset \mathbb{R}^{p+1}$  is compact.*

Under Assumption 1, the weight matrix  $W$  is nonsingular since  $\sigma_0^2(\mu_4 - \sigma_0^4) - \mu_3^2 > 0$ . Thus, the determinant of  $W$  is nonzero.

Let  $\mu_x = \mathbb{E}(X)$  and  $M_{xx} = \mathbb{E}(XX')$ . Let  $\gamma = (\beta', \sigma^2)' \in \mathbb{R}^{p+1}$  denote the parameter vector, and let  $\gamma_0 = (\beta_0', \sigma_0^2)'$  denote the true value.

**Theorem 1.**

Suppose that Assumptions 1-3 hold. Then, as  $n \rightarrow \infty$ , the SLSE satisfies:

1. Consistency:

$$\hat{\gamma}_{SLS} \xrightarrow{a.s.} \gamma_0.$$

2. Asymptotic Normality:

$$\sqrt{n}(\hat{\gamma}_{SLS} - \gamma_0) \xrightarrow{d} N(0, C),$$

where the asymptotic covariance matrix  $C$  is given by

$$C = \begin{pmatrix} V(\hat{\beta}_{SLS}) & \xi V(\hat{\sigma}_{SLS}^2) M_{xx}^{-1} \mu_x \\ \xi V(\hat{\sigma}_{SLS}^2) \mu_x' M_{xx}^{-1} & V(\hat{\sigma}_{SLS}^2) \end{pmatrix}, \quad (2.2.1)$$

with

$$V(\hat{\beta}_{SLS}) = \sigma_0^2 (1 - \xi) (M_{xx} - \xi \mu_x \mu_x')^{-1}, \quad (2.2.2)$$

$$V(\hat{\sigma}_{SLS}^2) = \frac{(\mu_4 - \sigma_0^4)(1 - \xi)}{1 - \xi \mu_x' M_{xx}^{-1} \mu_x}, \quad (2.2.3)$$

and

$$\xi = \frac{\mu_3^2}{\sigma_0^2(\mu_4 - \sigma_0^4)}. \quad (2.2.4)$$

This result shows the feasibility of replacing the unknown asymptotic covariance matrix with a plug-in estimator and constructing Wald-type confidence intervals for individual regression coefficients for the SLS.

The proof follows from a Taylor expansion of the estimating equations together with the uniform law of large numbers and the central limit theorem. Full details are provided in the Appendix.

From Theorem 1, we see that when  $\mu_3 = 0$  and thus  $\xi = 0$ , the SLSE  $\hat{\beta}_{SLS}$  and  $\hat{\sigma}_{SLS}^2$  have the same asymptotic variance as the  $V(\hat{\beta}_{OLS}) = \sigma_0^2 M_{xx}^{-1}$  and  $V(\hat{\sigma}_{OLS}^2) = \mu_4 - \sigma_0^4$ , respectively. However, the following theorem shows that in general the SLSE is asymptotically more efficient than the OLSE when  $\mu_3 \neq 0$ .

A special case arises when the linear model includes an intercept term. In this case, the first covariate is  $X_1 = 1$  with probability 1. Let  $e_1 = (1, 0, \dots, 0)'$  be the standard unit vector. It follows that  $M_{xx}e_1 = \mu_x$ , that is,  $\mu'_x M_{xx}^{-1} \mu_x = \mu'_x e_1 = 1$ .

After applying the Sherman-Morrison formula, the inverse matrix expands to:

$$(M_{xx} - \xi \mu_x \mu'_x)^{-1} = M_{xx}^{-1} + \frac{\xi}{1 - \xi} M_{xx}^{-1} \mu_x \mu'_x M_{xx}^{-1} = M_{xx}^{-1} + \frac{\xi}{1 - \xi} e_1 e'_1.$$

Therefore, the asymptotic variance of the SLSE can be written as:

$$V(\hat{\beta}_{SLS}) = \sigma_0^2 [(1 - \xi) M_{xx}^{-1} + \xi e_1 e'_1].$$

The variance difference then becomes  $V(\hat{\beta}_{OLS}) - V(\hat{\beta}_{SLS}) = \sigma_0^2 \xi (M_{xx}^{-1} - e_1 e'_1)$ .

In contrast, when the model does not include an intercept and the covariates are centered such that  $\mu_x = 0$ ,

$$V(\hat{\beta}_{SLS}) = \sigma_0^2 (1 - \xi) M_{xx}^{-1}.$$

## 2.3 Efficiency of the SLSE Compared to the OLSE

In this section, we dive into the asymptotic efficiency of the SLSE relative to the OLSE. Although Theorem 1 establishes the strong consistency and asymptotic normality of the SLSE, it does not evaluate the efficiency gain obtained by accounting for higher-order moment information. The following theorem proves the efficiency of the SLSE.

### Theorem 2.

Suppose that **Assumptions 1-3** hold. If  $\mu_3 \neq 0$ , then  $0 < \xi < 1$ , and:

1.

$$V(\hat{\beta}_{OLS}) - V(\hat{\beta}_{SLS})$$

is a non-negative definite matrix. Furthermore, it is strictly positive definite if and only if  $\mu'_x M_{xx}^{-1} \mu_x \neq 1$ .

2.

$$V(\hat{\sigma}_{OLS}^2) - V(\hat{\sigma}_{SLS}^2) \geq 0,$$

with equality holding if and only if  $\mu'_x M_{xx}^{-1} \mu_x = 1$ .

Theorem 2 shows that, when the error distribution is asymmetric, the SLSE has an asymptotic variance no larger than that of the OLSE. The non-negative definiteness of  $V(\hat{\beta}_{OLS}) - V(\hat{\beta}_{SLS})$  guarantees that using second-order moment information does not increase asymptotic uncertainty. Instead, the SLSE consistently reduces the variance of the regression coefficient estimates. The same efficiency gain appears to the estimator of the error variance, since  $V(\hat{\sigma}_{OLS}^2) \geq V(\hat{\sigma}_{SLS}^2)$ . The improvement for all components of  $\hat{\beta}$  only happens when  $\mu'_x M_{xx}^{-1} \mu_x \neq 1$ .

# 3

## Numerical Optimization for the SLSE

### 3.1 Iteratively Reweighted Newton Method

#### 3.1.1 Newton-Raphson Algorithm

The second-order least squares estimator (SLSE) does not have a closed-form solution. Unlike the ordinary least squares estimator, which can be obtained through matrix inversion  $(X'X)^{-1}$ , the SLSE is defined as the minimizer of a nonlinear quadratic objective function involving both the first and second conditional moments of the response variable. Therefore, numerical optimization algorithms are required to compute the estimator in practice. Among the available optimization algorithms, the Newton-Raphson (NR) algorithm is one of the most widely used iterative methods due to its fast local convergence and effectiveness in nonlinear estimation problems.

As shown by Wang and Leblanc (2008), the SLSE can be computed using standard nonlinear optimization procedures such as the Newton-Raphson method. The Newton-Raphson algorithm is a second-order optimization method that updates parameter estimates using both the gradient vector and the Hessian matrix of the objective function. In nonlinear optimization, second-order methods are particularly efficient because the local curvature information is thoroughly considered and often achieve quadratic convergence when the initial

estimate is sufficiently close to the minimizer Nocedal and Wright (2006); Dennis and Schnabel (1983); Gill et al. (1981).

Let  $Q_n(\gamma)$  denote the SLSE objective function defined in Chapter 2, where

$$\gamma = (\beta', \sigma^2)'$$

Suppose  $\gamma^{(t)}$  is the estimate at iteration  $t$ . Then the Newton–Raphson update is given by

$$\gamma^{(t+1)} = \gamma^{(t)} - [\nabla^2 Q_n(\gamma^{(t)})]^{-1} \nabla Q_n(\gamma^{(t)}),$$

where  $\nabla Q_n(\gamma^{(t)})$  is the gradient vector and  $\nabla^2 Q_n(\gamma^{(t)})$  is the Hessian matrix at the current iterate.

The Newton-Raphson method is derived from a second-order Taylor expansion of the objective function around the current estimate. Specifically, for a twice continuously differentiable function  $Q_n(\gamma)$ ,

$$Q_n(\gamma) \approx Q_n(\gamma^{(t)}) + \nabla Q_n(\gamma^{(t)})'(\gamma - \gamma^{(t)}) + \frac{1}{2}(\gamma - \gamma^{(t)})'\nabla^2 Q_n(\gamma^{(t)})(\gamma - \gamma^{(t)}).$$

Minimizing this quadratic approximation with respect to  $\gamma$  leads to the Newton-Raphson update formula above.

One important advantage of the NR algorithm is its fast convergence. When the Hessian matrix is positive definite and the initial value is sufficiently close to the minimizer, the algorithm works as a quadratic convergence, meaning the estimation error decreases quickly every iteration Nocedal and Wright (2006). This property makes Newton-Raphson computationally effective for many nonlinear estimation problems.

Despite its strong theoretical convergence properties, the Newton-Raphson method frequently becomes unstable in practice when the Hessian matrix is nearly singular or not positive definite. Such numerical instability is common in optimization problems on highly

skewed error distributions or extreme observations. In the SLSE framework, the objective function depends on second-order moment information, which can generate unstable Hessian matrices under asymmetric errors. Consequently, the standard Newton-Raphson algorithm may not converge, or produce unreliable estimates in finite samples such as  $Var(\hat{\sigma}_{SLSE}^2) \leq 0$ .

Moreover, the SLSE objective function can be understood through the lens of iteratively reweighted optimization methods. Similar weighted updating ideas can be found in iteratively reweighted least squares (IRLS) algorithms commonly used in generalized linear models and robust estimation Green (1984b). In each iteration, the optimization algorithm updates parameter estimates based on local curvature information obtained from the current iterate.

These numerical difficulties resulted in the development of stabilized optimization procedures for the SLSE. In the following sections, we derive the gradient and Hessian matrices of the SLSE objective function and investigate the instability of the standard Newton-Raphson algorithm through simulation studies. We then introduce stabilization techniques based on ridge regularization and Armijo-type line search to improve convergence and numerical robustness.

### 3.1.2 Iteratively Reweighted Newton Method for SLSE

To apply the Newton-Raphson algorithm to the second-order least squares estimator (SLSE), clear forms of the gradient vector and Hessian matrix of the objective function are required. Remember that the SLSE objective function is defined as

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

where

$$\rho_i(\gamma) = \begin{pmatrix} y_i - x_i' \beta \\ y_i^2 - (x_i' \beta)^2 - \sigma^2 \end{pmatrix},$$

and  $W_i$  is a non-negative definite weight matrix, and the optimal weight matrix in the context of the SLS in linear models is shown in 2.1.6.

In practical implementation, the optimal weight matrix  $W_i$  depends on unknown error moments and parameter values. Therefore, the optimization is carried out using an iteratively reweighted procedure. At iteration  $t$ , the third and fourth error moments are estimated from the residuals obtained using the current parameter estimate  $\gamma^{(t)}$ , and the corresponding weight matrices  $W_i^{(t)}$  are calculated. These weight matrices are then fixed while the Newton-type search direction is calculated for that iteration. After updating the parameter vector, the residual moments and weight matrices are recalculated for the next iteration.

Thus, the derivatives used in each Newton step are taken with respect to  $\rho_i(\gamma)$ , while  $W_i^{(t)}$  is treated as fixed. The procedure should therefore be interpreted as an iteratively reweighted Newton-type method rather than the ordinary Newton-Raphson method applied directly to an objective where the estimated weight matrices change with respect to  $\gamma$ .

Using the Newton-Raphson algorithm introduced in the previous section, the parameter vector is updated iteratively according to

$$\gamma^{(t+1)} = \gamma^{(t)} - [H_n(\gamma^{(t)})]^{-1} G_n(\gamma^{(t)}),$$

where

$$G_n(\gamma) = \nabla Q_n(\gamma)$$

is the gradient vector and

$$H_n(\gamma) = \nabla^2 Q_n(\gamma)$$

is the Hessian matrix.

For the SLSE objective function, the gradient takes the form

$$G_n(\gamma) = \sum_{i=1}^n \left( \frac{\partial \rho_i(\gamma)}{\partial \gamma} \right)' W_i \rho_i(\gamma),$$

while the Hessian matrix takes the following form

$$H_n(\gamma) = \sum_{i=1}^n \left[ \left( \frac{\partial \rho_i(\gamma)}{\partial \gamma} \right)' W_i \left( \frac{\partial \rho_i(\gamma)}{\partial \gamma} \right) + (\rho_i(\gamma)' W_i \otimes I_{p+1}) \frac{\partial \text{vec} \left( \frac{\partial \rho_i(\gamma)}{\partial \gamma} \right)}{\partial \gamma'} \right].$$

These expressions establish the starting point of the iteratively reweighted Newton optimization procedure used throughout this thesis. Since the objective function involves non-linear second-order moment terms, the gradient and Hessian matrices are more complicated than those that appear in ordinary least squares estimation.

In the appendix, detailed derivations of the gradient vector and Hessian matrix are presented.

## 3.2 Stabilization of the Newton Method

Wang et al. (2023) studied numerical optimization for SLS estimation and proposed an Alternate Updating (AU) method to address the non-convexity of the SLS objective function. The approach developed in this thesis follows a different direction, using an iteratively reweighted Newton-type procedure combined with Hessian regularization and an Armijo-type line search to improve numerical stability.

While the iteratively reweighted Newton method offers an efficient approach for the SLSE, the optimization often encounters numerical stability issues in finite samples. The Hessian matrix of the SLSE objective can become nearly singular or indefinite, especially under highly asymmetric error distributions or when extreme observations are present. These instabilities can lead to excessively large Newton directions, causing the algorithm to diverge or fail during matrix inversion.

These numerical challenges are particularly evident in the SLSE, as its objective function incorporates second-order moment information, leading to curvature structures that are more complex than those found in standard least squares estimation. Therefore, stabilization

methods are necessary to ensure the optimization algorithm converges more reliably.

In this section, two stabilization methods are introduced. First, ridge regularization is used to improve the condition and positive definiteness of the Hessian matrix. Second, an Armijo-type line search procedure is introduced to control the step size of the Newton update and guarantee sufficient decrease of the objective function at each iteration.

### 3.2.1 Ridge Regularization

One of the most common sources of numerical instability in the Newton-Raphson algorithm is from the inversion of the Hessian matrix. In practice, the Hessian matrix may become nearly singular or indefinite, especially when the objective function has strong local curvature or when extreme observations are present in the data Dennis and Schnabel (1983); Nocedal and Wright (2006). In such situations, the Newton direction may become excessively large or fail to provide a descent direction, resulting in unstable parameter updates or divergence of the optimization procedure.

To stabilize the optimization process, ridge regularization is implemented on the Hessian matrix. The basic idea is to stabilize the Hessian matrix by adding a positive identity matrix each check time until achieving positive definiteness. Specifically, instead of using the standard Newton update

$$\gamma^{(t+1)} = \gamma^{(t)} - H_n(\gamma^{(t)})^{-1} G_n(\gamma^{(t)}),$$

we consider the enhanced update

$$\gamma^{(t+1)} = \gamma^{(t)} - [H_n(\gamma^{(t)}) + \lambda I]^{-1} G_n(\gamma^{(t)}),$$

where  $\lambda > 0$  is the ridge parameter and  $I$  denotes the identity matrix with proper rank.

The addition of  $\lambda I$  increases the eigenvalues of the Hessian matrix and improves its numerical condition. In particular, when the Hessian matrix is nearly singular or indefinite,

the ridge parameter helps guarantee positive definiteness and stabilizes the matrix inversion required in the Newton update. Consequently, the optimization procedure with the ridge regularization becomes more robust in finite samples.

This type of regularization is closely related to Levenberg-Marquardt type algorithms commonly used in nonlinear optimization (Nocedal and Wright (2006); Gill et al. (1981)). In terms of the SLSE, ridge regularization is essential because the second-order moment component of the objective function can generate highly unstable curvature under strongly skewed error distributions.

In practice, the ridge parameter will be selected in a flexible manner during the optimization procedure. When the Hessian matrix is well-conditioned, a small ridge parameter is chosen, and the algorithm behaves similarly to the standard Newton method. On the other hand, when instability is detected, larger ridge values can be introduced to stabilize the update direction.

### 3.2.2 Armijo-Type Line Search

In Newton-type optimization methods, the parameter is usually updated by following a descent direction that is based on the local curvature of the objective function. However, directly taking the full Newton step may lead to instability when the objective function shows significant curvature or the current iterate is distant from the target. In such situations, overly large parameter updates may increase the objective function value or produce an unstable Hessian during the optimization process.

To improve convergence stability, an Armijo-type backtracking line search procedure is implemented to control the step size of the iterative update. Let

$$d^{(t)} = - [\nabla^2 f(\theta^{(t)})]^{-1} \nabla f(\theta^{(t)})$$

denote the Newton search direction at iteration  $t$ . Instead of taking the full Newton step,

the parameter vector is updated according to

$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t d^{(t)},$$

where  $\alpha_t \in (0, 1]$  is a step-size parameter selected through a backtracking line search process.

The Armijo line search determines how big a step to take by ensuring that the objective function decreases sufficiently. Specifically, the step size  $\alpha_t$  is chosen such that

$$f(\theta^{(t)} + \alpha_t d^{(t)}) \leq f(\theta^{(t)}) + c\alpha_t \nabla f(\theta^{(t)})' d^{(t)},$$

where  $c \in (0, 1)$  is a small constant. This condition ensures that changing the parameter leads to a noticeable decrease in the value of the objective function, compared to the usual way of moving in the descent direction.

In practice, the algorithm begins with an initial step size, typically  $\alpha_t = 1$ , corresponding to the full Newton step. If the Armijo condition is not satisfied, the step size is repeatedly reduced as such:

$$\alpha_t \leftarrow \rho \alpha_t,$$

where  $\rho \in (0, 1)$  is a shrinkage factor. This backtracking procedure continues until the sufficient decrease condition is met.

The Armijo-type line search improves the robustness of the Newton-type optimization algorithm by preventing excessive parameter updates and stabilizing convergence in the face of local curvature fluctuations (Armijo (1966); Nocedal and Wright (2006)). In the following section, the line search procedure is combined with ridge regularization to build a stabilized optimization algorithm to compute the SLSE.

### 3.2.3 Reweighted Ridge–Newton Method with Line Search

The ridge regularization and Armijo-type line search procedures introduced in the previous sections can be combined to construct a stabilized optimization algorithm for computing the SLSE. This method integrates the iteratively reweighted Newton update with Hessian regularization and flexible control of step-size to improve numerical stability and convergence.

At each iteration, the gradient vector and Hessian matrix of the SLSE objective function are first evaluated using the current parameter estimates. To stabilize the curvature matrix, a ridge parameter is added to the Hessian matrix before inversion. Then, the search direction  $d_k$  is used in an Armijo-type backtracking line search procedure to determine the step size satisfying the decrease condition.

To give more details, let

$$G_n(\gamma^{(t)}) = \nabla Q_n(\gamma^{(t)})$$

and

$$H_n(\gamma^{(t)}) = \nabla^2 Q_n(\gamma^{(t)})$$

denote the gradient vector and Hessian matrix of the SLS objective function measured at iteration  $t$ . The stabilized search direction is obtained by

$$d^{(t)} = -[H_n(\gamma^{(t)}) + \lambda_t I]^{-1} G_n(\gamma^{(t)}),$$

where  $\lambda_t > 0$  is the ridge parameter at iteration  $t$ . The parameter vector is then updated according to

$$\gamma^{(t+1)} = \gamma^{(t)} + \alpha_t d^{(t)},$$

where  $\alpha_t$  is determined through an Armijo-type backtracking line search applied to the ridge–Newton direction  $d^{(t)}$ . Starting with the full step  $\alpha_t = 1$ , the candidate update

$$\gamma_{\text{cand}} = \gamma^{(t)} + \alpha_t d^{(t)}$$

is accepted if the variance component remains positive and the sufficient-decrease condition

$$Q_n(\gamma^{(t)} + \alpha_t d^{(t)}) \leq Q_n(\gamma^{(t)}) + c_1 \alpha_t G_n(\gamma^{(t)})^\top d^{(t)}$$

is satisfied, where  $c_1 = 10^{-4}$ . Otherwise, the step size is reduced according to

$$\alpha_t \leftarrow \frac{1}{2} \alpha_t$$

until the conditions are satisfied. Thus, ridge regularization is first used to obtain a stable Newton search direction, while the line search subsequently controls the distance moved along that direction.

This algorithm combines the fast local convergence properties of Newton-type algorithms with some additional stabilization tools that bring robustness under heavily skewed asymmetric error distributions and unstable Hessian matrices. Notably, ridge regularization improves the condition of the Hessian matrix, while the line search procedure prevents large parameter updates and stabilizes convergence of the optimization process.

The stabilized optimization algorithm introduced here significantly enhances the numerical stability of SLSE computations when using finite samples. In the next section, we compare the performance of the standard reweighted Newton method with this stabilized approach.

**Algorithm: Reweighted Ridge–Newton Method**

- 1: **Input:** Data  $(Y, X)$ , initial value  $\gamma^{(0)}$ , tolerance  $\delta$ , maximum iteration  $K_{\max}$ .
- 2: Set  $t = 0$  and  $\Delta = \infty$ .
- 3: **while**  $\Delta > \delta$  and  $t < K_{\max}$  **do**
- 4:   Compute the residuals

$$e_i^{(t)} = y_i - x_i' \beta^{(t)}, \quad i = 1, \dots, n,$$

and the corresponding residual moments

$$\hat{\mu}_j^{(t)} = \frac{1}{n} \sum_{i=1}^n \left( e_i^{(t)} \right)^j, \quad j = 3, 4.$$

- 5:   Recompute the weight matrices  $W_i^{(t)}$  using the current parameter estimates and residual moments  $\hat{\mu}_3^{(t)}$  and  $\hat{\mu}_4^{(t)}$ .
- 6:   With  $W_i^{(t)}$  held fixed, evaluate the Gradient  $G_t = G(\gamma^{(t)})$  and Hessian  $H_t = H(\gamma^{(t)})$ .
- 7:   Choose the smallest  $\lambda \in \{10^{-8}, 10^{-7}, 10^{-6}, \dots\}$  such that

$$H_t + \lambda I_{p+1} \text{ is positive definite.}$$

- 8:   Set the ridge–Newton direction

$$d_t = -(H_t + \lambda I_{p+1})^{-1} G_t.$$

- 9:   Choose  $\alpha \in \{1, 1/2, 1/4, \dots\}$  such that

$$\gamma_{p+1}^{(t)} + \alpha d_{t,p+1} > 0$$

and

$$Q_n(\gamma^{(t)} + \alpha d_t) \leq Q_n(\gamma^{(t)}) + c_1 \alpha G_t' d_t.$$

- 10:   Update

$$\gamma^{(t+1)} = \gamma^{(t)} + \alpha d_t, \quad \Delta = \|\gamma^{(t+1)} - \gamma^{(t)}\|.$$

- 11:   Set  $t \leftarrow t + 1$ .

- 12: **end while**

- 13: **Output:**  $\hat{\gamma} = \gamma^{(t)}$ .

### 3.3 Numerical Stability Comparison

To check whether the numerical stability of the proposed optimization methods is achieved, a simulation study was conducted under several asymmetric error distributions and regression models. The simulation setup is identical to that later used in the Monte Carlo studies in Chapter 4. In particular, both the intercept and non-intercept regression models were examined under different skewnesses in the error distribution.

Table 3.1 shows the proportion of bad samples generated by the standard reweighted Newton method and the proposed Reweighted Ridge–Newton method with line search. A bad sample is a simulation replicate in which the optimization procedure failed to converge or produced numerically unstable parameter estimates.

Table 3.1: Bad sample rates across sample sizes for the standard reweighted Newton method and the stabilized Reweighted Ridge–Newton method with line search.

| Model   | Error distribution | Reweighted Newton |           |           | Ridge–Newton + Line Search |           |           |
|---------|--------------------|-------------------|-----------|-----------|----------------------------|-----------|-----------|
|         |                    | $n = 50$          | $n = 100$ | $n = 300$ | $n = 50$                   | $n = 100$ | $n = 300$ |
| Model 1 | Normal             | 13.76%            | 3.47%     | 0.02%     | 0%                         | 0%        | 0%        |
| Model 1 | Exponential        | 8.22%             | 0.73%     | 0.02%     | 0%                         | 0%        | 0%        |
| Model 1 | Weibull(0.6, 1)    | 2.81%             | 0.04%     | 0.02%     | 0.02%                      | 0%        | 0%        |
| Model 2 | Normal             | 16.54%            | 7.15%     | 0.02%     | 0%                         | 0%        | 0%        |
| Model 2 | Exponential        | 28.41%            | 13.22%    | 2.64%     | 0%                         | 0%        | 0%        |
| Model 2 | Weibull(0.6, 1)    | 37.14%            | 30.37%    | 20.43%    | 0.02%                      | 0.02%     | 0%        |

# 4

## Simulation

### 4.1 Simulation Design

In this chapter, Monte Carlo simulation studies were conducted to evaluate the finite-sample performance of the SLSE and compare it with the OLSE under both symmetric and asymmetric error distributions. The simulation settings were chosen to investigate the efficiency gain of the SLSE, the accuracy of asymptotic variance estimation, the coverage probabilities of confidence intervals, and the power performance of Wald-type hypothesis tests.

Two linear regression models were considered throughout the simulation studies. Model 1 is an intercept model defined as

$$Y_i = \beta_0 + X_{i2}\beta_1 + X_{i3}\beta_2 + \varepsilon_i,$$

where

$$X_{i2} \sim \text{Unif}(1, 20), \quad X_{i3} \sim N(5, 4).$$

The model 2 is a non-intercept model defined as

$$Y_i = X_{i1}\beta_0 + X_{i2}\beta_1 + X_{i3}\beta_2 + \varepsilon_i,$$

where

$$X_{i1} \sim \text{Beta}(2, 5),$$

and

$$(X_{i2}, X_{i3})^\top \sim N \left( \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right).$$

The regression coefficients were fixed as

$$\beta = (\beta_0, \beta_1, \beta_2)' = (0.5, 1, -1.5)',$$

for the entire simulation, and all observations were generated independently.

To study the effect of asymmetry in the error distribution, three different error distributions were considered. First, the normal distribution  $N(0, 1)$  was used as a symmetric benchmark with zero skewness. Second, the exponential distribution  $\text{Exp}(1)$  was considered to represent moderately asymmetric errors with skewness equal to 2. Finally, the Weibull distribution  $\text{Weibull}(0.6, 1)$  was used to represent strongly asymmetric and heavy-tailed errors, having skewness approximately equal to 4.59. The Exponential and Weibull distributions were standardized so that the generated errors had mean zero and variance of 1.

Most of the simulation studies were conducted for sample sizes  $n = 50, 100$ , and  $300$ . For each simulation setting, 5000 Monte Carlo repetitions were generated. In Figure 4.1 on the large-sample behavior of the ratio of the average lengths of the OLS- and SLS-based CIs, we used  $n = 50, 100, 300, 500, 1000, 5000, 10000, 20000, 50000$

The SLSE was computed using the stabilized Reweighted Ridge-Newton method combined with the Armijo-type backtracking line search developed in Chapter 3, throughout the simulation studies. For each Monte Carlo sample, the Newton iterations were initialized using the ordinary least squares estimates,  $\gamma^{(0)} = \left( \hat{\beta}'_{\text{OLS}}, \hat{\sigma}_{\text{OLS}}^2 \right)'$ . The OLSE could provide a reasonable initial value to avoid a random initialization, which is important because Newton-type optimization is local and can be sensitive to the starting value. The maximum number of

Newton iterations was set to 100, and convergence was declared when  $\|\gamma^{(t+1)} - \gamma^{(t)}\| < 10^{-6}$ . Averaging over the sample sizes  $n = 50, 100$ , and 300, the mean numbers of Newton iterations for Model 1 were approximately 1.60, 2.15, and 2.27 under the normal, standardized exponential, and standardized Weibull error distributions, respectively. The corresponding averages for Model 2 were approximately 2.05, 7.66, and 8.95. Thus, the algorithm generally converged within only a few iterations for Model 1 and for Model 2 under normal errors, while more iterations were required for Model 2 under asymmetric errors.

For each simulation setting, the OLSE and SLSE were compared in terms of empirical bias, asymptotic standard error (ASE), empirical standard error (SE), coverage probability (CP), hypothesis testing power, and confidence interval (CI) length ratio (R), which is  $R_n = \frac{\text{Average OLS CI Length}}{\text{Average SLS CI Length}}$  for each sample size  $n$ .

Based on the asymptotic normality, confidence intervals for each component  $\beta_j$  can be constructed as

$$\hat{\beta}_j \pm z_{\alpha/2} \sqrt{\hat{V}_{jj}/n},$$

where  $\hat{V}_{jj}$  is the  $j$ -th diagonal element of  $\hat{V} = \text{Var}(\hat{\beta}_{SLSE})$ .

Hypothesis testing can be performed using the Wald statistic

$$Z_j = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{\hat{V}_{jj}/n}},$$

for the two-sided hypotheses  $H_0 : \beta_j = \beta_{j0}$  vs  $H_1 : \beta_j \neq \beta_{j0}$ , where  $\beta_{j0}$  is known.

## 4.2 Numerical Stability of the Optimization Algorithm

As discussed in Chapter 3, numerical instability may occur when computing the SLSE using the standard iteratively reweighted Newton method. In particular, under strongly asymmetric error distributions, the Hessian matrix of the SLSE objective function may become nearly singular or indefinite, leading to unstable parameter updates, failure of matrix inversion, or

divergence of the optimization algorithm.

During the Monte Carlo simulation studies, such numerical instability created bad samples, where a bad sample was calculated as  $\frac{failures}{failures+MCrepetition}$ . The bad samples occurred most commonly under the heavily skewed Weibull error distribution with skewness of 4.59, especially for smaller sample sizes and the non-intercept model. These numerical difficulties were mainly caused by ill-conditioned Hessian matrices generated by extreme observations.

To address these difficulties, all simulation studies presented in this chapter were conducted using the stabilized Reweighted Ridge-Newton method with Armijo-type backtracking line search developed in Chapter 3. The ridge regularization improved the conditioning of the Hessian matrix and stabilized the matrix inversion process, while the Armijo line search controlled large Newton updates.

The stabilized optimization algorithm significantly reduced the number of numerical failures across all simulation settings and provided a stable computation of the SLSE, particularly under extremely asymmetric error distributions. The numerical stability comparisons between the standard reweighted Newton method and the stabilized optimization procedure are presented in Table 3.1

### 4.3 Finite-Sample Performance of the SLSE

The normal error distribution is not included in Figure 4.1 because it serves as a symmetric benchmark. When the error distribution is normal,  $\mu_3 = 0$ , and the OLSE and SLSE are asymptotically equivalent, as established in Chapter 2. The results reported in Table 4.1- 4.3 are consistent with this property, with the confidence interval length ratio approaching one as the sample size increases. Therefore, Figure 4.1 focuses on the exponential and Weibull error distributions where  $\mu_3 \neq 0$  and the efficiency gain of the SLSE becomes more obvious.

The finite-sample results presented in Table 4.2 and 4.3 show a number of important properties of the second-order least squares estimator (SLSE) under both symmetric and

asymmetric error distributions. Overall, the simulation results show that the SLSE maintains a good finite-sample point estimation relative to the OLSE while achieving huge efficiency gains when the asymmetric errors are present.

First, both the OLSE and the SLSE have small empirical bias across all simulation settings. For both Model 1 and Model 2, the Monte Carlo means of the estimators remain close to the true parameter values even under heavily skewed error distributions. This pattern is consistent with the asymptotic consistency results established in Chapter 2. Furthermore, the bias decreases as the sample size increases, indicating convergence of both estimators toward the true parameter values.

The asymptotic standard errors (ASE) and empirical standard errors (SE) are also shown to be close to each other across most simulation settings. As the sample size increases, the ASE values become increasingly close to the empirical SE values for both the OLSE and SLSE. This result provides empirical evidence for the asymptotic variance expressions derived in Theorem 1. In particular, the convergence between ASE and empirical SE becomes more visible at  $n = 300$ , indicating that the asymptotic approximation of the ASE becomes reliable for moderate sample sizes.

Table 4.1 presents the ASE/SE ratios for both the OLSE and SLSE across different simulation settings. Values close to 1 indicate good alignment between the asymptotic standard error and the empirical standard error. Overall, the ASE/SE ratios move closer to 1 as the sample size increases, providing empirical evidence for the asymptotic normality results established in Chapter 2. Under the symmetric normal error distribution, the convergence is relatively fast for both estimators. In contrast, under the heavily skewed Weibull distribution, the SLSE shows slower convergence for smaller sample sizes, where the ASE tends to underestimate the empirical SE. However, the ratios improve gradually as  $n$  increases and become approximately close to 1 by  $n = 300$ , indicating that the asymptotic approximation becomes reliable for moderate sample sizes even under strongly asymmetric error distributions.

The coverage probability (CP) results further support the asymptotic properties of the

Table 4.1: Average ASE/SE ratios for OLSE and SLSE across simulation settings.

| Model   | Error Distribution | $n$ | OLS ASE/SE | SLS ASE/SE |
|---------|--------------------|-----|------------|------------|
| Model 1 | Normal             | 50  | 0.991      | 0.918      |
| Model 1 | Normal             | 100 | 0.995      | 0.956      |
| Model 1 | Normal             | 300 | 0.999      | 0.985      |
| Model 1 | Exp(1)             | 50  | 0.978      | 0.902      |
| Model 1 | Exp(1)             | 100 | 0.991      | 0.956      |
| Model 1 | Exp(1)             | 300 | 0.997      | 0.986      |
| Model 1 | Weibull(0.6, 1)    | 50  | 0.965      | 0.842      |
| Model 1 | Weibull(0.6, 1)    | 100 | 0.982      | 0.914      |
| Model 1 | Weibull(0.6, 1)    | 300 | 0.994      | 0.973      |
| Model 2 | Normal             | 50  | 0.985      | 0.920      |
| Model 2 | Normal             | 100 | 0.992      | 0.949      |
| Model 2 | Normal             | 300 | 0.989      | 0.973      |
| Model 2 | Exp(1)             | 50  | 0.975      | 0.901      |
| Model 2 | Exp(1)             | 100 | 0.991      | 0.960      |
| Model 2 | Exp(1)             | 300 | 0.997      | 0.987      |
| Model 2 | Weibull(0.6, 1)    | 50  | 0.952      | 0.816      |
| Model 2 | Weibull(0.6, 1)    | 100 | 0.974      | 0.891      |
| Model 2 | Weibull(0.6, 1)    | 300 | 0.991      | 0.964      |

SLSE. Under the symmetric normal error distribution, at small sample sizes, the SLSE intervals are undercovered because the asymptotic variance approximation is not fast enough. The approximation improves as  $n$  increases. This result is consistent with Theorem 1, which states that the OLSE and SLSE are asymptotically equivalent when the third moment of the error distribution is zero.

Under asymmetric error distributions, however, clear differences between the two estimators begin to appear. For the standardized exponential and Weibull error distributions, the SLSE consistently produces smaller ASE and SE values than the OLSE, demonstrating improved estimation efficiency. The improvement becomes more substantial under the heavily skewed Weibull distribution with skewness of approximately 4.59. These results highlight that the second conditional moment utilized by the SLSE carries additional information when the error distribution is asymmetric.

The confidence interval length ratio  $R_n$  further shows the efficiency gain of the SLSE.

Recall that

$$R_n = \frac{\text{Average OLS CI Length}}{\text{Average SLS CI Length}}.$$

Values of  $R_n > 1$  indicate that the SLSE produces shorter confidence intervals than the OLSE. Under the normal distribution, the ratio remains close to 1, reflecting the asymptotic equivalence between the two estimators. In contrast, under asymmetric error distributions, the ratio increases significantly. Especially under the Weibull error distribution, the ratio exceeds 1.5 and nears 2 in several simulation settings, implying that the SLSE produces considerably more precise confidence intervals.

Figure 4.1 illustrates the asymptotic convergence pattern of the confidence interval length ratio  $R_n$  across increasing sample sizes. Under both the exponential and Weibull error distributions, the ratio initially fluctuates for smaller sample sizes but gradually stabilizes as  $n$  increases, suggesting convergence toward a limiting efficiency constant. This pattern indicates that the efficiency gain of the SLSE is achieved asymptotically.

Overall, the finite-sample results demonstrate that the SLSE maintains performance comparable to the OLSE under symmetric errors while providing substantial improvements in estimation efficiency and confidence intervals under asymmetric error distributions. The simulation results thus support the theoretical properties established in Chapter 2 and show the advantages of accounting for the second-order moment in the estimation.

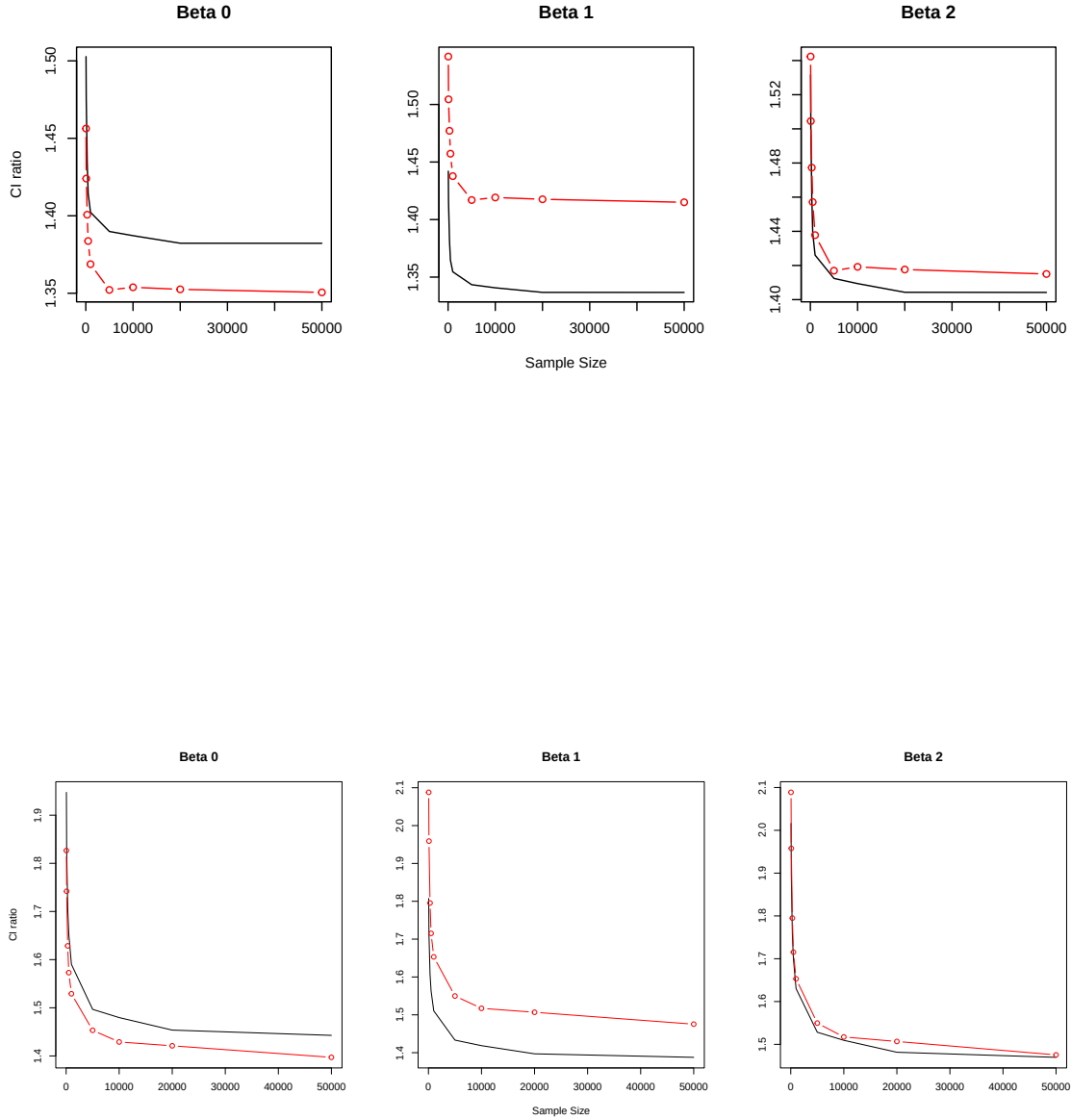


Figure 4.1: OLS/SLS CI length ratio convergence of model 1 (above) and model 2 (below) with  $\varepsilon \sim \text{Exp}(1)$  (black) and  $\varepsilon \sim \text{Weibull}(0.6, 1)$  (red) and  $n = 50, 100, 300, 500, 1000, 5000, 10000, 20000, 50000$

Table 4.2: OLSE vs. SLSE, and OLS CI vs. SLS CI for Model 1

|                           | OLS     |        |        |        | SLS     |        |        |        | $R_n$  |
|---------------------------|---------|--------|--------|--------|---------|--------|--------|--------|--------|
| Parameter                 | OLSE    | ASE    | SE     | CP     | SLSE    | ASE    | SE     | CP     |        |
| <b>N(0,1)</b>             |         |        |        |        |         |        |        |        |        |
| $n = 50$                  |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.4993  | 0.4649 | 0.4695 | 0.9436 | 0.5009  | 0.4392 | 0.4800 | 0.9224 | 1.0584 |
| $\beta_1$                 | 1.0000  | 0.0251 | 0.0256 | 0.9390 | 0.9999  | 0.0236 | 0.0262 | 0.9136 | 1.0615 |
| $\beta_2$                 | -1.5000 | 0.0729 | 0.0737 | 0.9428 | -1.5003 | 0.0687 | 0.0751 | 0.9194 | 1.0615 |
| $n = 100$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.5041  | 0.3238 | 0.3236 | 0.9432 | 0.5044  | 0.3147 | 0.3287 | 0.9320 | 1.0292 |
| $\beta_1$                 | 1.0001  | 0.0175 | 0.0178 | 0.9400 | 1.0001  | 0.0170 | 0.0182 | 0.9294 | 1.0307 |
| $\beta_2$                 | -1.5006 | 0.0508 | 0.0505 | 0.9430 | -1.5007 | 0.0493 | 0.0513 | 0.9318 | 1.0307 |
| $n = 300$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.4998  | 0.1855 | 0.1886 | 0.9466 | 0.4998  | 0.1838 | 0.1891 | 0.9424 | 1.0095 |
| $\beta_1$                 | 1.0001  | 0.0100 | 0.0100 | 0.9514 | 1.0001  | 0.0099 | 0.0100 | 0.9496 | 1.0100 |
| $\beta_2$                 | -1.4999 | 0.0290 | 0.0297 | 0.9450 | -1.4999 | 0.0287 | 0.0298 | 0.9402 | 1.0100 |
| <b>std-Exp(1)</b>         |         |        |        |        |         |        |        |        |        |
| $n = 50$                  |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.4977  | 0.4582 | 0.4635 | 0.9453 | 0.4999  | 0.3130 | 0.3329 | 0.9312 | 1.4638 |
| $\beta_1$                 | 1.0001  | 0.0247 | 0.0251 | 0.9459 | 0.9999  | 0.0159 | 0.0165 | 0.9440 | 1.5524 |
| $\beta_2$                 | -1.4997 | 0.0720 | 0.0736 | 0.9395 | -1.4997 | 0.0464 | 0.0509 | 0.9257 | 1.5520 |
| $n = 100$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.4957  | 0.3216 | 0.3278 | 0.9483 | 0.4972  | 0.2245 | 0.2282 | 0.9448 | 1.4329 |
| $\beta_1$                 | 1.0000  | 0.0174 | 0.0175 | 0.9497 | 0.9999  | 0.0115 | 0.0117 | 0.9480 | 1.5165 |
| $\beta_2$                 | -1.4994 | 0.0504 | 0.0518 | 0.9432 | -1.4995 | 0.0332 | 0.0342 | 0.9430 | 1.5165 |
| $n = 300$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.5012  | 0.1851 | 0.1858 | 0.9467 | 0.4994  | 0.1327 | 0.1336 | 0.9495 | 1.3945 |
| $\beta_1$                 | 0.9999  | 0.0100 | 0.0100 | 0.9486 | 1.0000  | 0.0068 | 0.0069 | 0.9489 | 1.4702 |
| $\beta_2$                 | -1.5000 | 0.0290 | 0.0292 | 0.9466 | -1.4999 | 0.0197 | 0.0199 | 0.9504 | 1.4701 |
| <b>std-Weibull(0.6,1)</b> |         |        |        |        |         |        |        |        |        |
| $n = 50$                  |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.4939  | 0.4375 | 0.4691 | 0.9375 | 0.4986  | 0.2392 | 0.2598 | 0.9234 | 1.8290 |
| $\beta_1$                 | 1.0006  | 0.0236 | 0.0253 | 0.9498 | 1.0002  | 0.0113 | 0.0120 | 0.9447 | 2.0930 |
| $\beta_2$                 | -1.4999 | 0.0686 | 0.0731 | 0.9457 | -1.5000 | 0.0328 | 0.0364 | 0.9381 | 2.0930 |
| $n = 100$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.5014  | 0.3130 | 0.3245 | 0.9435 | 0.4994  | 0.1801 | 0.1869 | 0.9357 | 1.7374 |
| $\beta_1$                 | 0.9998  | 0.0169 | 0.0177 | 0.9454 | 1.0000  | 0.0086 | 0.0089 | 0.9499 | 1.9546 |
| $\beta_2$                 | -1.4999 | 0.0490 | 0.0510 | 0.9478 | -1.4999 | 0.0251 | 0.0264 | 0.9425 | 1.9548 |
| $n = 300$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.4965  | 0.1827 | 0.1871 | 0.9448 | 0.4979  | 0.1126 | 0.1149 | 0.9429 | 1.6217 |
| $\beta_1$                 | 1.0002  | 0.0099 | 0.0099 | 0.9505 | 1.0001  | 0.0055 | 0.0055 | 0.9548 | 1.7843 |
| $\beta_2$                 | -1.4999 | 0.0285 | 0.0292 | 0.9478 | -1.5000 | 0.0160 | 0.0163 | 0.9496 | 1.7843 |

Table 4.3: OLSE vs. SLSE, and OLS CI vs. SLS CI for Model 2

|                           | OLS     |        |        |        | SLS     |        |        |        | $R_n$  |
|---------------------------|---------|--------|--------|--------|---------|--------|--------|--------|--------|
| Parameter                 | OLSE    | ASE    | SE     | CP     | SLSE    | ASE    | SE     | CP     |        |
| <b>N(0,1)</b>             |         |        |        |        |         |        |        |        |        |
| $n = 50$                  |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.5067  | 0.7463 | 0.7554 | 0.9432 | 0.5063  | 0.6997 | 0.7638 | 0.9204 | 1.0666 |
| $\beta_1$                 | 0.9987  | 0.1117 | 0.1119 | 0.9456 | 0.9987  | 0.1050 | 0.1133 | 0.9276 | 1.0637 |
| $\beta_2$                 | -1.5013 | 0.1565 | 0.1569 | 0.9442 | -1.5015 | 0.1465 | 0.1598 | 0.9246 | 1.0682 |
| $n = 100$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.4916  | 0.5185 | 0.5271 | 0.9442 | 0.4902  | 0.5023 | 0.5329 | 0.9328 | 1.0321 |
| $\beta_1$                 | 1.0008  | 0.0778 | 0.0785 | 0.9464 | 1.0010  | 0.0755 | 0.0794 | 0.9356 | 1.0307 |
| $\beta_2$                 | -1.4972 | 0.1088 | 0.1107 | 0.9402 | -1.4974 | 0.1053 | 0.1120 | 0.9304 | 1.0328 |
| $n = 300$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.4996  | 0.2961 | 0.2960 | 0.9462 | 0.4991  | 0.2928 | 0.2980 | 0.9424 | 1.0112 |
| $\beta_1$                 | 0.9999  | 0.0444 | 0.0444 | 0.9489 | 1.0000  | 0.0439 | 0.0446 | 0.9445 | 1.0107 |
| $\beta_2$                 | -1.4999 | 0.0621 | 0.0628 | 0.9477 | -1.4999 | 0.0614 | 0.0631 | 0.9444 | 1.0114 |
| <b>std-Exp(1)</b>         |         |        |        |        |         |        |        |        |        |
| $n = 50$                  |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.5006  | 0.7346 | 0.7544 | 0.9423 | 0.5032  | 0.4862 | 0.5430 | 0.9248 | 1.5110 |
| $\beta_1$                 | 0.9996  | 0.1100 | 0.1127 | 0.9402 | 0.9987  | 0.0758 | 0.0833 | 0.9243 | 1.4517 |
| $\beta_2$                 | -1.5002 | 0.1538 | 0.1583 | 0.9459 | -1.5020 | 0.0999 | 0.1098 | 0.9290 | 1.5400 |
| $n = 100$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.4938  | 0.5149 | 0.5180 | 0.9479 | 0.4989  | 0.3489 | 0.3591 | 0.9424 | 1.4760 |
| $\beta_1$                 | 1.0001  | 0.0771 | 0.0779 | 0.9438 | 0.9991  | 0.0544 | 0.0560 | 0.9399 | 1.4180 |
| $\beta_2$                 | -1.4996 | 0.1078 | 0.1113 | 0.9436 | -1.4997 | 0.0717 | 0.0751 | 0.9391 | 1.5040 |
| $n = 300$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.4980  | 0.2940 | 0.2966 | 0.9456 | 0.5007  | 0.2044 | 0.2095 | 0.9416 | 1.4390 |
| $\beta_1$                 | 0.9997  | 0.0441 | 0.0445 | 0.9449 | 0.9989  | 0.0319 | 0.0328 | 0.9428 | 1.3860 |
| $\beta_2$                 | -1.5006 | 0.0617 | 0.0622 | 0.9517 | -1.4997 | 0.0422 | 0.0426 | 0.9472 | 1.4650 |
| <b>std-Weibull(0.6,1)</b> |         |        |        |        |         |        |        |        |        |
| $n = 50$                  |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.5041  | 0.7005 | 0.7609 | 0.9360 | 0.5203  | 0.3632 | 0.4841 | 0.9030 | 1.9290 |
| $\beta_1$                 | 0.9995  | 0.1049 | 0.1125 | 0.9340 | 0.9996  | 0.0584 | 0.0737 | 0.8997 | 1.7940 |
| $\beta_2$                 | -1.4980 | 0.1466 | 0.1584 | 0.9390 | -1.4997 | 0.0734 | 0.0970 | 0.9078 | 1.9980 |
| $n = 100$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.5047  | 0.5016 | 0.5191 | 0.9491 | 0.5112  | 0.2728 | 0.3222 | 0.9251 | 1.8390 |
| $\beta_1$                 | 0.9992  | 0.0752 | 0.0776 | 0.9408 | 1.0000  | 0.0438 | 0.0506 | 0.9228 | 1.7160 |
| $\beta_2$                 | -1.5004 | 0.1051 | 0.1104 | 0.9435 | -1.5001 | 0.0553 | 0.0655 | 0.9260 | 1.9020 |
| $n = 300$                 |         |        |        |        |         |        |        |        |        |
| $\beta_0$                 | 0.4946  | 0.2918 | 0.2973 | 0.9456 | 0.4981  | 0.1715 | 0.1790 | 0.9438 | 1.7010 |
| $\beta_1$                 | 1.0004  | 0.0438 | 0.0442 | 0.9496 | 1.0001  | 0.0274 | 0.0283 | 0.9412 | 1.6020 |
| $\beta_2$                 | -1.4993 | 0.0612 | 0.0610 | 0.9517 | -1.5001 | 0.0350 | 0.0361 | 0.9479 | 1.7520 |

## 4.4 Hypothesis Testing Performance

In this section, Wald-type hypothesis tests were constructed for both the OLSE and SLSE using their corresponding asymptotic variance estimators. The empirical power was evaluated across a sequence of parameter values near the true regression coefficients for sample sizes  $n = 50, 100$ , and  $300$ .

Since the normal error case represents the symmetric benchmark where  $\mu_3 = 0$  and no asymptotic efficiency gain is expected from the SLSE, the power comparisons are presented for the exponential and Weibull error distributions, where the effect of asymmetry can be shown directly.

Tables 4.4 and 4.5 show a direct comparison of the empirical rejection probabilities for the OLSE and SLSE under the exponential and Weibull error distributions. At the true parameter values, the rejection probabilities are close to the nominal 5% level, especially as the sample size increases. This suggests that the Wald-type tests based on both estimators control the Type I error rate reasonably well in moderate sample sizes. However, some inflation is observed for the SLSE when the sample size is small, and the error distribution is strongly skewed. This indicates that the asymptotic variance approximation may be less accurate in small samples when the error distribution has severe asymmetry.

For parameter values away from the null hypothesis, the SLSE typically produces larger empirical power than the OLSE across both regression models. The difference is especially clear under the Weibull error distribution, where the SLSE reaches high power with a smaller sample size than the OLSE. For example, in Model 1, the gain is large for both slope parameters, showing that the reduction in variance directly improves the ability to detect deviation from the null hypothesis. Similar patterns are observed in Model 2, although the increase in power is smaller than in Model 1 because the non-intercept makes estimation more difficult. The table results show that the SLSE improves testing power under asymmetric errors, but the small-sample results should be interpreted with some caution when the skewness is strong.

Figures 4.2 and 4.3 present the estimated power curves for Model 1 and Model 2, respectively. In all simulation settings, the SLSE consistently shows a higher power than the OLSE, reflecting the variance reduction achieved by incorporating second-order moment information. The difference between the two methods becomes prominent under the heavily skewed Weibull error distribution, where the efficiency gain of the SLSE is strongest.

The power curves shown in Figures 4.2 and 4.3 demonstrate that the efficiency gain of the SLSE translates directly into stronger hypothesis testing performance. Across both regression models and all regression coefficients, the SLSE produces steeper power curves than the OLSE, indicating faster convergence toward power 1 as the parameter values go further away from the null hypothesis.

For smaller sample sizes, the SLSE shows more inflated power behavior due to finite-sample underestimation of the asymptotic variance under strong skewness. However, as the sample size increases, the power curves stabilize and continue to demonstrate the asymptotic efficiency advantage of the SLSE. To sum up, the results indicate that the variance reduction achieved by the SLSE leads to shorter confidence intervals, as well as stronger inferential performance under asymmetric error distributions.

Table 4.4: Power comparison under different error distributions.  $X_1 = 1$ ,  $X_2 \sim \text{Unif}(1, 20)$ ,  $X_3 \sim \mathcal{N}(5, 4)$ .

|           |       | $\varepsilon \sim \text{Exp}(1)$ |        |           |        |           |        | $\varepsilon \sim \text{Weibull}(0.6, 1)$ |        |           |        |           |        |
|-----------|-------|----------------------------------|--------|-----------|--------|-----------|--------|-------------------------------------------|--------|-----------|--------|-----------|--------|
|           |       | $n = 50$                         |        | $n = 100$ |        | $n = 300$ |        | $n = 50$                                  |        | $n = 100$ |        | $n = 300$ |        |
|           |       | OLS                              | SLS    | OLS       | SLS    | OLS       | SLS    | OLS                                       | SLS    | OLS       | SLS    | OLS       | SLS    |
| $\beta_0$ | 0.50  | 0.0568                           | 0.0694 | 0.0516    | 0.0548 | 0.0496    | 0.0536 | 0.0626                                    | 0.0744 | 0.0590    | 0.0654 | 0.0530    | 0.0546 |
|           | 0.72  | 0.0998                           | 0.1578 | 0.1348    | 0.2094 | 0.2396    | 0.4000 | 0.1378                                    | 0.2616 | 0.1554    | 0.3150 | 0.2792    | 0.5538 |
|           | 0.94  | 0.2014                           | 0.3570 | 0.3350    | 0.5402 | 0.6744    | 0.8960 | 0.2828                                    | 0.5332 | 0.3592    | 0.6828 | 0.6948    | 0.9418 |
| $\beta_1$ | 1.00  | 0.0550                           | 0.0580 | 0.0492    | 0.0500 | 0.0468    | 0.0454 | 0.0512                                    | 0.0544 | 0.0578    | 0.0496 | 0.0472    | 0.0460 |
|           | 1.055 | 0.6234                           | 0.9168 | 0.8848    | 0.9954 | 1.0000    | 1.0000 | 0.6742                                    | 0.9808 | 0.8752    | 0.9986 | 0.9988    | 1.0000 |
|           | 1.11  | 0.9810                           | 1.0000 | 1.0000    | 1.0000 | 1.0000    | 1.0000 | 0.9622                                    | 1.0000 | 0.9974    | 1.0000 | 1.0000    | 1.0000 |
| $\beta_2$ | -1.50 | 0.0570                           | 0.0722 | 0.0598    | 0.0596 | 0.0508    | 0.0574 | 0.0510                                    | 0.0588 | 0.0518    | 0.0520 | 0.0498    | 0.0504 |
|           | -1.39 | 0.3646                           | 0.6602 | 0.5938    | 0.8934 | 0.9608    | 0.9998 | 0.4360                                    | 0.8824 | 0.6500    | 0.9754 | 0.9600    | 0.9998 |
|           | -1.28 | 0.8532                           | 0.9882 | 0.9834    | 0.9998 | 1.0000    | 1.0000 | 0.8634                                    | 0.9966 | 0.9712    | 1.0000 | 1.0000    | 1.0000 |

Table 4.5: Power comparison under different error distributions.  $X_1 \sim \text{Beta}(2, 5)$ ,  $(X_2, X_3) \sim \mathcal{N}\left(\begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$ .

|           |       | $\varepsilon \sim \text{Exp}(1)$ |        |           |        |           |        | $\varepsilon \sim \text{Weibull}(0.6, 1)$ |        |           |        |           |        |
|-----------|-------|----------------------------------|--------|-----------|--------|-----------|--------|-------------------------------------------|--------|-----------|--------|-----------|--------|
|           |       | $n = 50$                         |        | $n = 100$ |        | $n = 300$ |        | $n = 50$                                  |        | $n = 100$ |        | $n = 300$ |        |
|           |       | OLS                              | SLS    | OLS       | SLS    | OLS       | SLS    | OLS                                       | SLS    | OLS       | SLS    | OLS       | SLS    |
| $\beta_0$ | 0.50  | 0.0548                           | 0.0730 | 0.0518    | 0.0528 | 0.0496    | 0.0550 | 0.0628                                    | 0.0988 | 0.0486    | 0.0766 | 0.0484    | 0.0544 |
|           | 0.72  | 0.0748                           | 0.1074 | 0.0762    | 0.1112 | 0.1184    | 0.2100 | 0.0918                                    | 0.1818 | 0.0874    | 0.1852 | 0.1354    | 0.2968 |
|           | 0.94  | 0.1194                           | 0.1968 | 0.1594    | 0.2806 | 0.3404    | 0.5900 | 0.1560                                    | 0.3398 | 0.1918    | 0.4438 | 0.3676    | 0.7288 |
| $\beta_1$ | 1.00  | 0.0602                           | 0.0708 | 0.0538    | 0.0598 | 0.0514    | 0.0598 | 0.0642                                    | 0.0968 | 0.0572    | 0.0756 | 0.0528    | 0.0548 |
|           | 1.055 | 0.0996                           | 0.1642 | 0.1330    | 0.2098 | 0.2532    | 0.4282 | 0.1384                                    | 0.2772 | 0.1690    | 0.3392 | 0.2862    | 0.5608 |
|           | 1.11  | 0.2126                           | 0.3554 | 0.3364    | 0.5582 | 0.7030    | 0.9174 | 0.2924                                    | 0.5496 | 0.3986    | 0.7020 | 0.7150    | 0.9528 |
| $\beta_2$ | -1.50 | 0.0534                           | 0.0682 | 0.0572    | 0.0632 | 0.0490    | 0.0522 | 0.0618                                    | 0.0952 | 0.0554    | 0.0744 | 0.0482    | 0.0534 |
|           | -1.39 | 0.1204                           | 0.2122 | 0.1846    | 0.3452 | 0.4368    | 0.7510 | 0.1338                                    | 0.3588 | 0.2012    | 0.5442 | 0.4410    | 0.8942 |
|           | -1.28 | 0.3186                           | 0.6162 | 0.5500    | 0.8630 | 0.9508    | 0.9992 | 0.3834                                    | 0.8424 | 0.5944    | 0.9598 | 0.9482    | 0.9996 |

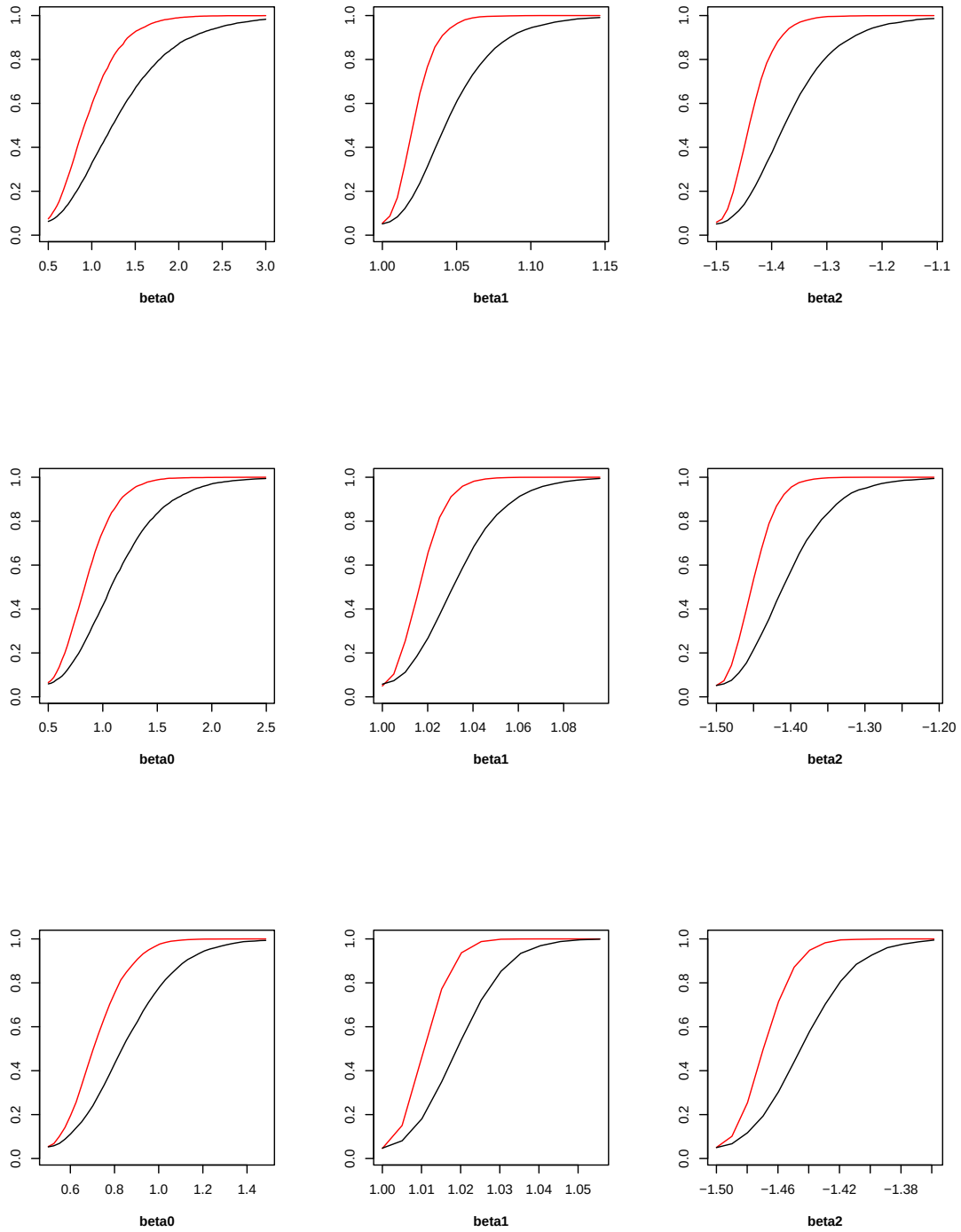


Figure 4.2: Power of OLS (black) and SLS (red) for Model 1 when  $\varepsilon \sim \text{Weibull}(0.6, 1)$  and  $n = 50$  (top), 100, and 300 (bottom)

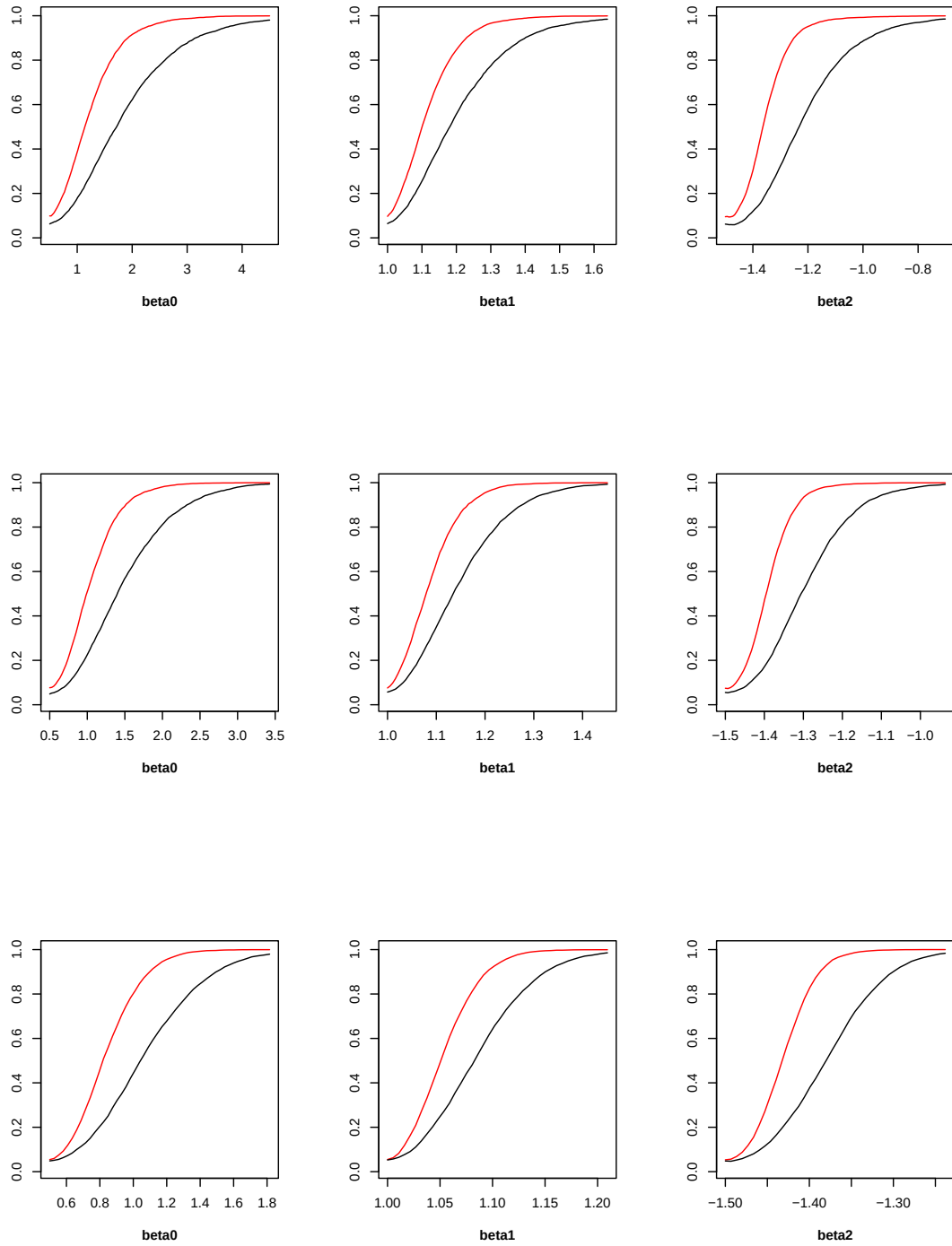


Figure 4.3: Power of OLS (black) and SLS (red) for Model 2 when  $\varepsilon \sim \text{Weibull}(0.6, 1)$  and  $n = 50$  (top), 100, and 300 (bottom)

# 5

## Real Data Application

### 5.1 Australian Institute of Sport (AIS) Dataset

To study the practical performance of the SLSE, the algorithm developed in Chapter 3 was applied to the Australian Institute of Sport (AIS) dataset, which was originally collected in 1991 and made available in 1994 (Telford and Cunningham, 1991; Cook and Weisberg, 1994). The AIS dataset contains biomedical measurements collected from athletes and has been widely used in regression analysis studies with continuous responses.

In this application, plasma ferritin concentration (Fe) was used as the response variable, while body mass index (BMI) and lean body mass (LBM) were included as explanatory variables. The linear regression model considered for this dataset is given by

$$Fe_i = \beta_0 + \beta_1 \text{BMI}_i + \beta_2 \text{LBM}_i + \varepsilon_i, \quad i = 1, 2, \dots, 202$$

where  $Fe_i$  denotes the plasma ferritin concentration for the  $i$ -th individual and  $\varepsilon_i$  is the random error term satisfying  $E(\varepsilon_i | X_i) = 0$ .

The AIS dataset is useful for evaluating the SLSE because the response variable shows an asymmetric pattern. Since the SLSE uses information from both the first and second moments of the response variable, it should yield more reliable statistical inference, especially when

the error distribution is asymmetric.

The OLSE and the SLSE were fitted using the same regression model. The SLSE was computed using the stabilized Reweighted Ridge-Newton optimization algorithm with Armijo-type line search developed in Chapter 3. Statistical inference was conducted using asymptotic standard errors and Wald-type confidence intervals.

Figure 5.1 presents the histogram of the residuals along with the normal Q-Q plot for the fitted AIS regression model. The histogram shows moderate asymmetry in the residual distribution, while the Q-Q plot shows that the distribution noticeably deviates from the normal distribution. These figures suggest that the normality assumption is not fully satisfied, giving room for the use of the SLSE framework.

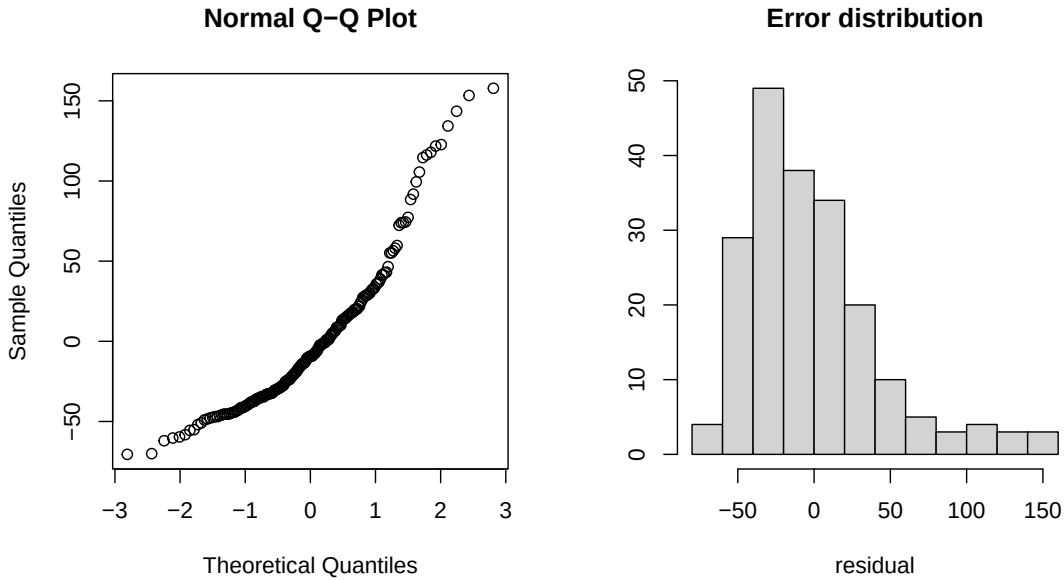


Figure 5.1: OLS residual histogram and normal Q-Q plot for the AIS dataset.

The estimation and inference results for the AIS dataset are summarized in Table 5.1. Overall, the OLSE and SLSE produced similar point estimates for the regression coefficients, especially for covariates, indicating that both estimators capture similar linear relationships between the response variable and covariates. However, the estimated standard errors and

confidence interval lengths was observed to be largely different.

In particular, the SLSE consistently produced smaller standard errors and narrower confidence intervals compared with the OLSE. The reduction in confidence interval length shows the gain in statistical efficiency by including second-order moment information into the estimation process. The results aligned with the asymptotic theory developed in Chapter 2 and the Monte Carlo simulation results presented in Chapter 4.

Table 5.1: Comparison of OLSE and SLSE for the AIS dataset

| Variable  | OLSE     |         |                     |          | SLSE     |         |                     |         |
|-----------|----------|---------|---------------------|----------|----------|---------|---------------------|---------|
|           | Estimate | SE      | 95% CI              | Length   | Estimate | SE      | 95% CI              | Length  |
| Intercept | -30.8571 | 25.8511 | [-81.5253, 19.8110] | 101.3363 | -12.6490 | 18.9672 | [-49.8247, 24.5267] | 74.3514 |
| BMI       | 2.5653   | 1.5816  | [-0.5347, 5.6652]   | 6.1999   | 2.3743   | 1.1530  | [0.1144, 4.6342]    | 4.5198  |
| LBM       | 0.7529   | 0.3466  | [0.0737, 1.4322]    | 1.3585   | 0.5398   | 0.2526  | [0.0447, 1.0350]    | 0.9904  |

The reduction in standard errors also had impacts on the inference for BMI. Under the OLSE, the 95% confidence interval for the BMI coefficient was

$$[-0.5347, 5.6652],$$

which contains zero. Thus, the BMI coefficient was not statistically significant at the 5% significance level based on the OLSE. In contrast, the SLSE produced the narrower 95% confidence interval

$$[0.1144, 4.6342],$$

which does not include zero. Therefore, the BMI coefficient was statistically significant at the 5% level under the SLSE. This example shows how the reduction in estimation uncertainty gained from the SLSE can affect the inference.

Furthermore, the stabilized optimization algorithm converged successfully without numerical instability, demonstrating the practical feasibility of the proposed computational procedure in real data applications.

## 5.2 Health Insurance Dataset

To further evaluate the performance of the SLSE on the real dataset, the Reweighted Ridge–Newton with line search was used again on a health insurance dataset containing demographic and medical information on individuals (Choi, 2018). The dataset includes variables such as age, sex, body mass index (BMI), number of children, smoking status, and residential region together with individual medical charges.

Sex was coded as 0 for female and 1 for male, while smoking status was coded as 0 for smokers and 1 for non-smokers. Since residential region is a nominal categorical variable with four levels, it was represented using three indicator variables for northeast, southwest, and southeast, with northwest used as the reference category. To improve the numerical stability of the SLSE optimization procedure, the response variable and covariates, including the dummy variables, were centered and standardized to have sample mean zero and sample standard deviation one. As a result, the reported regression coefficients correspond to standardized variables rather than the original measurement units.

Let  $Y_i^*$  denote the standardized medical charges and let the superscript  $*$  denote the corresponding standardized covariates. The linear regression model for this dataset is

$$\begin{aligned} Y_i^* = & \beta_1 Age_i^* + \beta_2 Sex_i^* + \beta_3 BMI_i^* + \beta_4 Children_i^* + \beta_5 Smoker_i^* \\ & + \beta_6 NE_i^* + \beta_7 SW_i^* + \beta_8 SE_i^* + \varepsilon_i, \quad i = 1, \dots, 1338, \end{aligned}$$

where  $NE_i$ ,  $SW_i$ , and  $SE_i$  are indicator variables for the northeast, southwest, and southeast regions, respectively, with northwest as a reference category. The random error satisfies  $E(\varepsilon_i | X_i) = 0$ .

Health insurance data are known to have large asymmetry because a relatively small proportion of individuals commonly account for extremely large healthcare costs. Thus, the health insurance dataset provides a setting for evaluating the efficiency gain of the SLSE under asymmetric error distributions.

Figure 5.2 shows a histogram of the residuals alongside a normal Q-Q plot for the health

insurance regression model. The residuals histogram appears right-skewed, and the Q-Q plot shows clear deviations from the normal line. These observations suggest that the assumption of normality, which is important for the accuracy of the OLSE relative to the SLSE, might not hold.

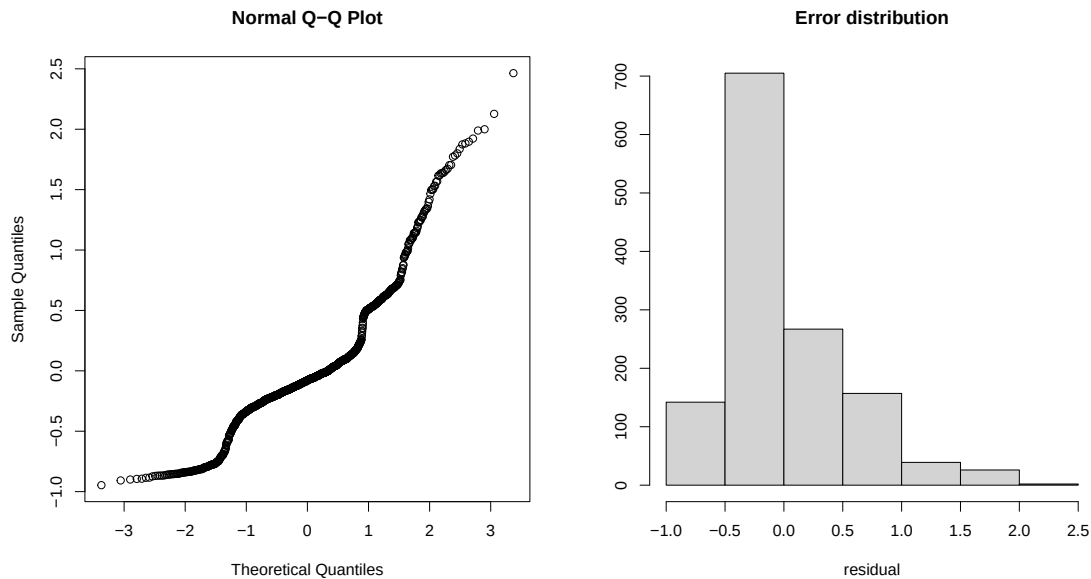


Figure 5.2: OLS residual histogram and normal Q-Q plot for the health insurance dataset.

The estimation and inference results are summarized in Table 5.2. The OLSE and SLSE produced the same qualitative inferential conclusions for the model, even though some of the point estimates were substantially different. Age, BMI, number of children, and smoking status were statistically significant at the 5% level under both estimators, while sex was not statistically significant under either estimator.

The three regional indicator variables were also not statistically significant under either method. For example, using northwest as the reference category, the SLSE 95% confidence intervals were  $[-0.0050, 0.0332]$  for northeast,  $[-0.0272, 0.0111]$  for southwest, and  $[-0.0365, 0.0034]$  for southeast.

Despite producing the same significance conclusions, the SLSE yielded substantially smaller standard errors and narrower confidence intervals than the OLSE for every regression coefficient. For example, the confidence interval length for age decreased from 0.0541 under the OLSE to 0.0313 under the SLSE, while that for BMI decreased from 0.0564 to 0.0326. Overall, the SLSE confidence intervals were approximately 42% shorter than the OLSE intervals. This result is consistent with the efficiency improvement supported by the asymptotic theory when the third error moment is nonzero.

Table 5.2: Comparison of OLSE and SLSE for the standardized health insurance dataset

| Variable           | OLSE     |        |                    |        | SLSE     |        |                    |        |
|--------------------|----------|--------|--------------------|--------|----------|--------|--------------------|--------|
|                    | Estimate | SE     | 95% CI             | Length | Estimate | SE     | 95% CI             | Length |
| age                | 0.2980   | 0.0138 | [0.2710, 0.3251]   | 0.0541 | 0.3028   | 0.0080 | [0.2872, 0.3184]   | 0.0313 |
| sex                | -0.0054  | 0.0137 | [-0.0324, 0.0215]  | 0.0539 | -0.0042  | 0.0079 | [-0.0198, 0.0114]  | 0.0312 |
| bmi                | 0.1708   | 0.0144 | [0.1426, 0.1990]   | 0.0564 | 0.0750   | 0.0083 | [0.0587, 0.0913]   | 0.0326 |
| children           | 0.0473   | 0.0137 | [0.0205, 0.0742]   | 0.0538 | 0.0430   | 0.0079 | [0.0275, 0.0585]   | 0.0311 |
| smoker             | -0.7950  | 0.0138 | [-0.8220, -0.7680] | 0.0540 | -0.5958  | 0.0080 | [-0.6114, -0.5802] | 0.0312 |
| region (northeast) | 0.0125   | 0.0168 | [-0.0205, 0.0455]  | 0.0660 | 0.0141   | 0.0097 | [-0.0050, 0.0332]  | 0.0382 |
| region (southwest) | -0.0215  | 0.0169 | [-0.0546, 0.0116]  | 0.0662 | -0.0080  | 0.0098 | [-0.0272, 0.0111]  | 0.0383 |
| region (southeast) | -0.0251  | 0.0176 | [-0.0596, 0.0094]  | 0.0690 | -0.0166  | 0.0102 | [-0.0365, 0.0034]  | 0.0399 |

In addition, the stabilized optimization algorithm converged successfully throughout the estimation procedure without numerical instability. Overall, the health insurance dataset analysis demonstrates that the SLSE can provide substantially more efficient statistical inference than the OLSE in regression settings involving asymmetric response distributions.

# 6

## Conclusion

This thesis examined the properties and performance of the second-order least squares estimator (SLSE) for linear regression models when the error distribution is asymmetric. While the ordinary least squares estimator (OLSE) is considered the best linear unbiased estimator under Gauss-Markov assumptions, its performance can decrease under skewed error distributions. Inspired by the work of Wang and Leblanc (2008), the study explored the SLSE by using both the first and second conditional moments of the response variable to utilize additional information from the higher-order moments of the data.

The theoretical properties of the SLSE were first established for linear regression models. In particular, the asymptotic variance formulas of the regression coefficients and the error variance estimator were derived, and the asymptotic efficiency of the SLSE relative to the OLSE was studied. It was shown that when the third moment of the error distribution is zero, which includes symmetric error distributions, the SLSE and OLSE are asymptotically equivalent. However, when the error distribution is asymmetric and the third moment is nonzero, the SLSE reduces asymptotic variance by utilizing second-order moment information. Consequently, the SLSE provides more statistically efficient estimation and inference under asymmetric error distributions.

Since the SLSE objective function  $Q_n(\gamma)$  does not have a closed-form solution, numer-

ical optimization algorithms were required for practical implementation. To compute the estimator reliably, an iteratively reweighted Newton method was developed together with stabilization methods such as ridge regularization and Armijo-type backtracking line search. The stabilized Reweighted Ridge-Newton algorithm with line search significantly improved numerical robustness and convergence under heavily skewed error distributions where the standard Newton method frequently produced unstable Hessian matrices and convergence failures. The simulation studies demonstrated that the proposed stabilization procedure almost completely eliminated bad samples while maintaining computational efficiency.

Monte Carlo simulation studies were conducted to evaluate the finite-sample performance of the SLSE and compare it with the OLSE under both symmetric and asymmetric error distributions. The simulation results showed that both estimators remained approximately unbiased in all settings. Under symmetric normal errors, the SLSE and OLSE showed nearly identical performance, which is consistent with the asymptotic theory. In contrast, under asymmetric error distributions such as the Exponential and Weibull distributions, the SLSE consistently produced smaller standard errors and shorter confidence intervals while maintaining coverage probabilities close to the nominal level for moderately large sample sizes. The simulation results also demonstrated that the asymptotic standard errors of the SLSE converged to the empirical standard errors as the sample size increased, supporting the asymptotic normality results derived in Chapter 2. Furthermore, the SLSE generally achieved more powerful Wald-type hypothesis tests under asymmetric error distributions due to its improved estimation efficiency.

Two real data applications were presented to demonstrate the practical usefulness of the SLSE in linear models with improved numerical algorithms. The AIS dataset and the health insurance dataset both presented asymmetric error distributions, making them suitable settings for the SLSE. In both applications, the SLSE consistently produced narrower confidence intervals and smaller standard errors than the OLSE while keeping the regression relationship. These results indicated that incorporating second-order conditional moment

information can improve statistical inference in practical regression problems for asymmetric data.

Overall, findings in the thesis demonstrate that the SLSE can be an efficient alternative to the OLSE when the error distribution is asymmetric. By incorporating additional moment information, the SLSE achieves improved estimation efficiency and more precise statistical inference while having comparable performance under symmetric errors. In addition, the stabilized optimization algorithm developed in this thesis enables practical computation of the estimator even under highly skewed distributions.

Although the SLSE demonstrated substantial efficiency gains under asymmetric error distributions, several limitations occurred. First, even though the enhanced numerical algorithm substantially improves computational stability, it does not completely remove bad samples. The simulation results show that bad sample rates still occur, especially when sample sizes are small or asymmetric error distributions are heavily skewed. A clear pattern emerges where the bad sample rate increases with greater skewness in the error distribution. This suggests that the numerical difficulty of computing the SLSE is closely related to the degree of asymmetry in the error distribution. Therefore, while the improved algorithm reduces the frequency of unstable samples, one needs to be cautious when using the SLSE under highly skewed error distributions.

Second, the asymptotic approximation of the SLSE may be less accurate in finite samples when the error distribution has severe skewness. This behavior was observed in the simulation studies, where the asymptotic standard errors underestimated the empirical standard errors for smaller sample sizes under the Weibull error distribution. Although the approximation improved as the sample size increased, finite-sample inference still struggles with highly asymmetric data.

A few directions for future research could be considered. First, the SLSE framework may be extended to causal inference problems, since asymmetric outcomes and treatment-response distributions commonly occur in observational studies. Including second-order moment in-

formation into causal effect estimation procedures may improve efficiency and robustness of estimators under non-normal errors. Second, the SLSE may be further developed for measurement error models, where higher-order moment information may help recover additional information from mismeasured covariates. Since second-order least squares methods have already been studied for certain measurement error, further study into statistical inference and computational methods for these models would be an option. Third, extending the SLSE framework to high-dimensional regression problems can be another important direction. Combining SLSE methodology with regularization techniques such as Lasso or penalized estimation may provide efficient estimation procedures for asymmetric high-dimensional data.

# 7

## Appendix

### 7.1 Preliminary

Wang and Leblanc (2008) established the consistency and asymptotic normality of the SLSE for general nonlinear regression models. In the linear regression framework considered in this thesis, several technical conditions and derivations become considerably simpler. The following lemmas are adapted from Amemiya (1985) and Wang and Leblanc (2008)). Let  $Q_n(Z, \psi)$  denote a measurable function of  $Z$  which is continuous in  $\psi$ , where  $\psi \in \Psi$  and  $\Psi$  is compact.

**Lemma 1.** If, as  $n \rightarrow \infty$ ,  $\sup_{\Psi} |Q_n(Z, \psi) - Q(\psi)| \rightarrow 0$ , where  $Q(\psi)$  is a non-stochastic function and has 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  $n \rightarrow \infty$ ,  $\sup_{\Psi} |Q_n(Z, \psi) - Q(\psi)| \rightarrow 0$  and  $\psi_0$  is an interior point of  $\Psi$ , then for any sequence of estimators  $\hat{\psi}_n \xrightarrow{a.s.} \psi_0$  it holds  $Q_n(Z, \hat{\psi}_n) \rightarrow Q_n(\psi_0)$ .

## 7.2 Derivation of gradient and Hessian matrix

We derive the gradient and Hessian for the given data  $(y_i, x_i), i = 1, 2, \dots, n$ . To evaluate the gradient, we have

$$\nabla_{\gamma} \rho'_i(\gamma) = - \begin{pmatrix} x_i & 2(x'_i \beta) x_i \\ 0 & 1 \end{pmatrix}$$

$$\nabla_{\gamma} \rho'_i(\gamma) W_i = -\eta \begin{pmatrix} [\mu_4 - \sigma^4 + 2\mu_3(x'_i \beta)] x_i & -\mu_3 x_i \\ -2\sigma^2(x'_i \beta) - \mu_3 & \sigma^2 \end{pmatrix}$$

and therefore

$$\nabla_{\gamma} \rho'_i(\gamma) W_i \rho_i(\gamma) = -\eta \begin{pmatrix} [(\mu_4 - \sigma^4)(y_i - x'_i \beta) - \mu_3(y_i - x'_i \beta)^2 + \mu_3 \sigma^2] x_i \\ \sigma^2(y_i - x'_i \beta)^2 - \sigma^4 - \mu_3(y_i - x'_i \beta) \end{pmatrix}$$

Further, to evaluate the Hessian, we have

$$\nabla_{\gamma} \rho'_i(\gamma) W_i \nabla_{\gamma'} \rho_i(\gamma) = \eta \begin{pmatrix} (\mu_4 - \sigma^4) x_i x'_i & -\mu_3 x_i \\ -\mu_3 x'_i & \sigma^2 \end{pmatrix}.$$

Since

$$\nabla_{\gamma'} \text{vec}(\nabla_{\gamma} \rho'_i(\gamma)) = - \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 2x_i x'_i & 0 \\ 0 & 0 \end{pmatrix},$$

we have

$$\begin{aligned} & (\rho'_i(\gamma) W_i \otimes I_{p+1}) \nabla_{\gamma'} \text{vec}(\nabla_{\gamma} \rho'_i(\gamma)) \\ &= -2\eta \begin{pmatrix} [\sigma^2(y_i - x'_i \beta)^2 - \sigma^4 - \mu_3(y_i - x'_i \beta)] x_i x'_i & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

and therefore

$$H_i = \eta \begin{pmatrix} [\mu_4 + \sigma^4 - 2\sigma^2(y_i - x'_i\beta)^2 + 2\mu_3(y_i - x'_i\beta)]x_i x'_i & -\mu_3 x_i \\ -\mu_3 x'_i & \sigma^2 \end{pmatrix}.$$

Therefore, the gradient is

$$\begin{aligned} G(\gamma) &= \sum_{i=1}^n \nabla_{\gamma} \rho'_i(\gamma) W_i \rho_i(\gamma) \\ &= -\eta \sum_{i=1}^n \begin{pmatrix} [(\mu_4 - \sigma^4)(y_i - x'_i\beta) - \mu_3(y_i - x'_i\beta)^2 + \mu_3\sigma^2]x_i \\ \sigma^2(y_i - x'_i\beta)^2 - \sigma^4 - \mu_3(y_i - x'_i\beta) \end{pmatrix} \end{aligned}$$

and Hessian is

$$\begin{aligned} H(\gamma) &= \sum_{i=1}^n \nabla_{\gamma} \rho'_i(\gamma) W_i \nabla_{\gamma'} \rho_i(\gamma) + (\rho'_i(\gamma) W_i \otimes I_{p+1}) \nabla_{\gamma'} \text{vec}(\nabla_{\gamma} \rho'_i(\gamma)) \\ &= \eta \sum_{i=1}^n \begin{pmatrix} [\mu_4 + \sigma^4 - 2\sigma^2(y_i - x'_i\beta)^2 + 2\mu_3(y_i - x'_i\beta)]x_i x'_i & -\mu_3 x_i \\ -\mu_3 x'_i & \sigma^2 \end{pmatrix}. \end{aligned}$$

### 7.3 Proof of Theorem 1

(1) First, denote  $q_i(\gamma) = \rho_i(\gamma)'W_i\rho_i(\gamma)$  and

$$W_i = \begin{pmatrix} w_{i,11} & w_{i,12} \\ w_{i,21} & w_{i,22} \end{pmatrix},$$

where  $w_{i,11} = \mu_4 - \sigma_0^4 + 4\mu_3(X_i'\beta_0) + 4\sigma_0^2(X_i'\beta_0)^2$ ,  $w_{i,12} = w_{i,21} = -\mu_3 - 2\sigma_0^2(X_i'\beta_0)$ ,  $w_{i,22} = w_{22} = \sigma_0^2$ . Then it is easy to see that the quadratic form

$$q_i(\gamma) = w_{i,11}(Y_i - X_i'\beta)^2 + 2w_{i,12}(Y_i - X_i'\beta)(Y_i^2 - (X_i'\beta)^2 - \sigma^2) + w_{22}(Y_i^2 - (X_i'\beta)^2 - \sigma^2)^2.$$

can be expanded in terms of the powers of  $Y_i$  and  $\|X\|_i$  and the highest orders in this expansion are  $Y_i^4$ ,  $Y_i^3\|X\|_i$ ,  $Y_i^2\|X\|_i^2$  and  $\|X\|_i^4$ . Further, since  $\gamma$  appears only in the coefficients of these terms and the parameter space  $\Gamma$  is compact, by assumption 2 it holds  $E[\max_{\Gamma} q_i(\gamma)] < \infty$ . It follows from the uniform law of large numbers that  $\sup_{\Gamma} |Q_n(\gamma)/n - Q(\gamma)| \rightarrow 0$ , where  $Q(\gamma) = E[q_i(\gamma)]$  (Amemiya (1985)). In order to show that  $Q(\gamma)$  has unique minimizer at the true value  $\gamma_0$ , note that, for any  $\gamma \in \Gamma$ ,

$$Q(\gamma) = Q(\gamma_0) + E[(\rho_i(\gamma) - \rho_i(\gamma_0))'W_i(\rho_i(\gamma) - \rho_i(\gamma_0))].$$

It follows that  $Q(\gamma) \geq Q(\gamma_0)$  with equality holding if and only if

$$E[(\rho_i(\gamma) - \rho_i(\gamma_0))'W_i(\rho_i(\gamma) - \rho_i(\gamma_0))] = 0.$$

Since  $W_i$  is positive definite with probability one, the above equality holds if and only if

$$\rho_i(\gamma) - \rho_i(\gamma_0) = [X_i'(\beta - \beta_0), (\beta - \beta_0)'X_iX_i'(\beta + \beta_0) + \sigma^2 - \sigma_0^2]' = 0$$

with probability one, implying  $\beta = \beta_0$  and  $\sigma^2 = \sigma_0^2$ . Finally, it follows from Lemma 1 that  $\hat{\gamma} \xrightarrow{p} \gamma_0$ .

(2) First, by Taylor expansion of the score function, we have

$$G_n(\hat{\gamma}) = G_n(\gamma_0) + H_n(\bar{\gamma})(\hat{\gamma} - \gamma_0), \quad (1)$$

where  $\|\bar{\gamma} - \gamma_0\| \leq \|\hat{\gamma} - \gamma_0\|$ ,

$$\begin{aligned} G_n(\gamma) &= \nabla_{\gamma} Q_n(\gamma) \\ &= \sum_{i=1}^n \nabla_{\gamma} \rho_i(\gamma)' W_i \rho_i(\gamma), \end{aligned}$$

and

$$\begin{aligned} H_n(\gamma) &= \nabla_{\gamma\gamma'}^2 Q_n(\gamma) \\ &= \sum_{i=1}^n \nabla_{\gamma} \rho_i(\gamma)' W_i \nabla_{\gamma'} \rho_i(\gamma) + (\rho_i(\gamma)' W_i \otimes I_{p+1}) \nabla_{\gamma'} \text{vec}(\nabla_{\gamma} \rho_i(\gamma)'). \end{aligned}$$

Similar to the proof (1), we can show that  $\sup_{\gamma} |H_n(\gamma)/n - H(\gamma)| \xrightarrow{p} 0$ , where  $H(\gamma) = E[\nabla_{\gamma\gamma'}^2 Q(\gamma)]$ , and hence, by the ULLN and Lemma 2, it holds

$$\begin{aligned} H_n(\bar{\gamma}) &\xrightarrow{p} H(\gamma_0) \\ &= E[\nabla_{\gamma} \rho_i(\gamma_0)' W_i \nabla_{\gamma'} \rho_i(\gamma_0) + (\rho_i(\gamma_0)' W_i \otimes I_{p+1}) \nabla_{\gamma'} \text{vec}(\nabla_{\gamma} \rho_i(\gamma_0)')] \\ &= A, \end{aligned}$$

where the last equality holds because the second order derivative  $\nabla_{\gamma'} \text{vec}(\nabla_{\gamma} \rho_i(\gamma)')$  does not depend on  $Y_i$  and hence the second term has expectation zero. It follows from (1) that

$$\sqrt{n}(\hat{\gamma} - \gamma_0) = A^{-1} \frac{1}{\sqrt{n}} G_n(\gamma_0) + o_p(1), \quad (2)$$

Further, since  $E[\nabla_{\gamma}\rho'_i(\gamma_0)W_i\rho_i(\gamma_0)] = 0$ , by the Central Limit Theorem,  $\frac{1}{\sqrt{n}}G_n(\gamma_0) \xrightarrow{d} N(0, B)$ , where

$$\begin{aligned} B &= E[\nabla_{\gamma}\rho_i(\gamma_0)'W_i\rho_i(\gamma_0)\rho_i(\gamma_0)'W_i\nabla_{\gamma'}\rho_i(\gamma_0)] \\ &= E[\nabla_{\gamma}\rho_i(\gamma_0)'W_iE[\rho_i(\gamma_0)\rho_i(\gamma_0)'|X_i]W_i\nabla_{\gamma'}\rho_i(\gamma_0)] \\ &= E[\nabla_{\gamma}\rho_i(\gamma_0)'W_iW_i^{-1}W_i\nabla_{\gamma'}\rho_i(\gamma_0)] \\ &= A. \end{aligned}$$

Thus, it follows from (2) that  $\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{d} N(0, A^{-1})$ . Finally, using the expression

$$\nabla\rho'_i(\gamma_0) = -\begin{pmatrix} X_i & 2(X'_i\beta_0)X_i \\ 0 & 1 \end{pmatrix},$$

we have

$$A = \eta \begin{pmatrix} (\mu_4 - \sigma_0^4)M_{xx} & -\mu_3\mu_x \\ -\mu_3\mu'_x & \sigma_0^2 \end{pmatrix}.$$

Therefore the asymptotic covariance matrix  $C = A^{-1}$  can be obtained by using the block matrix inversion formula.

## 7.4 Proof of Theorem 2

(1). First, since  $\mu_3 \neq 0$ , by A1 we have  $0 < \xi < 1$ . Further, since

$$V(\hat{\beta}_{SLS})^{-1} = \frac{1}{\sigma_0^2(1 - \xi)}(M_{xx} - \xi\mu_x\mu'_x)$$

and  $V(\hat{\beta}_{OLS})^{-1} = \frac{1}{\sigma_0^2}M_{xx}$ , we have

$$\sigma_0^2[V(\hat{\beta}_{SLS})^{-1} - V(\hat{\beta}_{OLS})^{-1}] = \frac{\xi}{1 - \xi}(M_{xx} - \mu_x\mu'_x)$$

$$= \frac{\xi}{1 - \xi} \Sigma_{xx}$$

which is nonnegative definite, where  $\Sigma_{xx}$  is the covariance matrix of  $X$ . It follows that  $V(\hat{\beta}_{OLS}) - V(\hat{\beta}_{SLS})$  is nonnegative definite. Further, since

$$\Sigma_{xx} = M_{xx}^{1/2}(I - M_{xx}^{-1/2}\mu_x\mu_x'M_{xx}^{-1/2})M_{xx}^{1/2}$$

and  $M_{xx}^{-1/2}\mu_x\mu_x'M_{xx}^{-1/2}$  has the same nonzero eigenvalue as  $\mu_x'M_{xx}^{-1}\mu_x$ ,  $\Sigma_{xx}$ , and therefore  $V(\hat{\beta}_{OLS}) - V(\hat{\beta}_{SLS})$ , is positive definite if and only if  $\mu_x'M_{xx}^{-1}\mu_x \neq 1$ .

(2). First, since

$$1 - \mu_x'M_{xx}^{-1}\mu_x = E[(1 - \mu_x'M_{xx}^{-1}X)^2] \geq 0,$$

we have  $\mu_x'M_{xx}^{-1}\mu_x \leq 1$ . It follows that

$$V(\hat{\sigma}_{OLS}^2) - V(\hat{\sigma}_{SLS}^2) = (\mu_4 - \sigma_0^4)\left(1 - \frac{1 - \xi}{1 - \xi\mu_x'M_{xx}^{-1}\mu_x}\right) \geq 0$$

with equality holding if and only if  $\mu_x'M_{xx}^{-1}\mu_x = 1$ .

## 7.5 R simulation code

The following R code was used to conduct the Monte Carlo simulation study in Chapter 4.

```
library(MASS) # for multivariate normal generation

set.seed(7934599)

#####
# Simulation setup
#####

mc = 5000
n = c(50, 100, 300)

beta0 = c(0.5, 1, -1.5)
```

```

sigma2 = 1
gamma0 = c(beta0, sigma2)

p = length(gamma0)      # number of parameters including sigma^2
q = length(beta0)       # number of regression coefficients
z_stat = qnorm(0.975)

# Choose simulation setting
model_type = "nonintercept" # "intercept" or "nonintercept"
error_type = "weibull"      # "normal", "exponential", or "weibull"

# Range of null values for power calculation
range = 20
b0_range = seq(beta0[1], beta0[1] + 4.0, length.out = range)
b1_range = seq(beta0[2], beta0[2] + 0.8, length.out = range)
b2_range = seq(beta0[3], beta0[3] + 1.0, length.out = range)

# Values used in the thesis power tables
b1_range[2] = 1.02
b1_range[3] = 1.04
b2_range[2] = -1.45
b2_range[3] = -1.40

#####
# Storage matrices
#####

olse = matrix(0, p, length(n))
slse = matrix(0, p, length(n))

SE_OLS = matrix(0, p, length(n))
SE_SLS = matrix(0, p, length(n))

ASE_OLS = matrix(0, p, length(n))
ASE_SLS = matrix(0, p, length(n))

CP_OLS = matrix(0, p, length(n))
CP_SLS = matrix(0, p, length(n))

LEN_OLS = matrix(0, p, length(n))
LEN_SLS = matrix(0, p, length(n))

```

```

LEN_RATIO = matrix(0, p, length(n))

power_OLS_b0 = matrix(0, range, length(n))
power_SLS_b0 = matrix(0, range, length(n))
power_OLS_b1 = matrix(0, range, length(n))
power_SLS_b1 = matrix(0, range, length(n))
power_OLS_b2 = matrix(0, range, length(n))
power_SLS_b2 = matrix(0, range, length(n))

bad_sample = matrix(0, 4, length(n))
rownames(bad_sample) = c("nonconvergence", "negative_variance",
                          "invalid_variance", "line_search_failure")

#####
# Standardized Weibull(0.6, 1) error
#####

std_weibull_error = function(n, shape = 0.6, scale = 1) {
  m1 = scale * gamma(1 + 1 / shape)
  m2 = scale^2 * gamma(1 + 2 / shape)

  (rweibull(n, shape = shape, scale = scale) - m1) / sqrt(m2 - m1^2)
}

#####
# Data generation
#####

generate_data = function(n, model_type, error_type) {

  if (model_type == "intercept") {
    x1 = rep(1, n)
    x2 = runif(n, 1, 20)
    x3 = rnorm(n, 5, 2)
    x = cbind(x1, x2, x3)
  }

  if (model_type == "nonintercept") {
    x1 = rbeta(n, 2, 5)

    mu_BN = c(2, 0)

```

```

sigma_BN = matrix(c(1, 0.5, 0.5, 1), nrow = 2)
data = mvrnorm(n = n, mu = mu_BN, Sigma = sigma_BN)

x2 = data[, 1]
x3 = data[, 2]
x = cbind(x1, x2, x3)
}

if (error_type == "normal") {
  eps = rnorm(n)
}

if (error_type == "exponential") {
  eps = rexp(n, rate = 1) - 1
}

if (error_type == "weibull") {
  eps = std_weibull_error(n, shape = 0.6, scale = 1)
}

y = as.numeric(x %*% beta0 + sqrt(sigma2) * eps)

list(x = x, y = y)
}

#####
# SLSE objective function Q_n(gamma)
#####

Q_at = function(gamma, x, y) {
  beta = gamma[1:q]
  s2 = gamma[p]

  if (!is.finite(s2) || s2 <= 0) {
    return(Inf)
  }

  xb = as.numeric(x %*% beta)
  e = as.numeric(y - xb)

  mu3 = mean(e^3)

```

```

mu4 = mean(e^4)

den = s2 * (mu4 - s2^2) - mu3^2

if (!is.finite(den) || den <= 0) {
  return(Inf)
}

eta = 1 / den

r1 = e
r2 = y^2 - xb^2 - s2

w11 = eta * (mu4 - s2^2 + 4 * mu3 * xb + 4 * s2 * xb^2)
w12 = eta * (-mu3 - 2 * s2 * xb)
w22 = eta * s2

sum(w11 * r1^2 + 2 * w12 * r1 * r2 + w22 * r2^2)
}

#####
# Monte Carlo simulation
#####

for (k in 1:length(n)) {

  ols = matrix(0, p, mc)
  sls = matrix(0, p, mc)

  ose = matrix(0, p, mc)
  ase = matrix(0, p, mc)

  cp_o = matrix(0, p, mc)
  cp_s = matrix(0, p, mc)

  len_ols = matrix(0, p, mc)
  len_sls = matrix(0, p, mc)

  rej_ols_b0 = matrix(0, range, mc)
  rej_sls_b0 = matrix(0, range, mc)
  rej_ols_b1 = matrix(0, range, mc)

```

```

rej_sls_b1 = matrix(0, range, mc)
rej_ols_b2 = matrix(0, range, mc)
rej_sls_b2 = matrix(0, range, mc)

cat("MC_repetition=", mc, ", Sample_size=", n[k], "\n")

for (j in 1:mc) {

  repeat {

    error_happened = FALSE

    tryCatch({

      #####
      # Generate data
      #####

      dat = generate_data(n[k], model_type, error_type)
      x = dat$x
      y = dat$y

      #####
      # Ordinary least squares estimator
      #####

      lmols = lm.fit(x = x, y = y)

      beta_ols = as.numeric(lmols$coefficients)
      res_ols = as.numeric(y - x %>% beta_ols)

      sig2_ols = sum(res_ols^2) / (n[k] - q)
      mu4_ols = mean(res_ols^4)

      ols[, j] = c(beta_ols, sig2_ols)

      beta_ols_se = sqrt(sig2_ols * diag(solve(t(x) %>% x)))
      sig2_ols_se = sqrt((mu4_ols - sig2_ols^2) / n[k])

      ose[, j] = c(beta_ols_se, sig2_ols_se)

```

```
#####
# SLSE using reweighted Ridge-Newton method with line search
#####

gam1 = ols[, j]
gam0 = rep(0, p)

nr = 0
max_iter = 100
tol = 1e-6

line_search_failed = FALSE
invalid_variance = FALSE

while (sqrt(sum((gam1 - gam0)^2)) > tol && nr < max_iter) {

  gam0 = gam1

  beta = gam0[1:q]
  s2 = gam0[p]

  if (!is.finite(s2) || s2 <= 0) {
    invalid_variance = TRUE
    break
  }

  res = as.numeric(y - x %*% beta)

  mu3 = mean(res^3)
  mu4 = mean(res^4)

  den = s2 * (mu4 - s2^2) - mu3^2

  if (!is.finite(den) || den <= 0) {
    invalid_variance = TRUE
    break
  }

  eta = 1 / den

#####
```

```

# Gradient vector
#####

G1 = t(x) %*% ((mu4 - s2^2) * res -
               mu3 * res^2 +
               mu3 * s2)

G2 = sum(s2 * res^2 - s2^2 - mu3 * res)

G = -eta * rbind(G1, G2)

#####
# Pure Hessian matrix
# This matches the Hessian formula derived in Chapter 3.
#####

H11 = t(x) %*% ((mu4 + s2^2 +
                2 * mu3 * res -
                2 * s2 * res^2) * x)

H12 = -mu3 * colSums(x)
H22 = n[k] * s2

H = eta * cbind(rbind(H11, H12), c(H12, H22))
H = (H + t(H)) / 2

#####
# Adaptive ridge regularization
#####

lambda = 1e-8
lambda_max = 1e8
Ipp = diag(p)
chol_ok = FALSE

repeat {
  H_r = H + lambda * Ipp
  ch = try(chol(H_r), silent = TRUE)

  if (!inherits(ch, "try-error")) {
    chol_ok = TRUE

```

```

        break
    }

    lambda = 10 * lambda

    if (lambda > lambda_max) {
        break
    }
}

if (!chol_ok) {
    ev = eigen(H, symmetric = TRUE)
    ev$values = pmax(ev$values, 1e-10)
    H_r = ev$vectors %*% diag(ev$values) %*% t(ev$vectors)
}

#####
# Newton direction
#####

step_dir = try(solve(H_r, G), silent = TRUE)

if (inherits(step_dir, "try-error")) {
    step_dir = qr.solve(H_r, G)
}

d = -as.numeric(step_dir)

# Safety check: ensure descent direction for Armijo search
directional_derivative = as.numeric(t(G) %*% d)

if (!is.finite(directional_derivative) || directional_derivative >= 0) {
    d = -as.numeric(G)
    directional_derivative = -sum(G^2)
}

#####
# Armijo backtracking line search based on  $Q_n(\gamma)$ 
#####

Q0 = Q_at(gam0, x, y)

```

```

alpha = 1
c1 = 1e-4
shrink = 0.5
min_alpha = 1e-8

accepted = FALSE

repeat {

  cand = gam0 + alpha * d

  if (cand[p] <= 0) {
    alpha = shrink * alpha

    if (alpha < min_alpha) {
      break
    }

    next
  }

  Q_new = Q_at(cand, x, y)

  if (is.finite(Q_new) &&
      Q_new <= Q0 + c1 * alpha * directional_derivative) {
    gam1 = cand
    accepted = TRUE
    break
  }

  alpha = shrink * alpha

  if (alpha < min_alpha) {
    break
  }
}

if (!accepted) {
  line_search_failed = TRUE
  break
}

```

```

    }

    nr = nr + 1
}

#####
# Bad sample checks
#####

if (nr == max_iter) {
  bad_sample["nonconvergence", k] =
    bad_sample["nonconvergence", k] + 1
  next
}

if (gam1[p] <= 0) {
  bad_sample["negative_variance", k] =
    bad_sample["negative_variance", k] + 1
  next
}

if (invalid_variance) {
  bad_sample["invalid_variance", k] =
    bad_sample["invalid_variance", k] + 1
  next
}

if (line_search_failed) {
  bad_sample["line_search_failure", k] =
    bad_sample["line_search_failure", k] + 1
  next
}

sls[, j] = gam1

#####
# Asymptotic standard errors for SLSE
#####

res_sls = as.numeric(y - x %*% sls[1:q, j])

```

```

s2_sls = sls[p, j]
mu3_sls = mean(res_sls^3)
mu4_sls = mean(res_sls^4)

xi = mu3_sls^2 / (s2_sls * (mu4_sls - s2_sls^2))

xbar = colMeans(x)
Mxx = t(x) %*% x / n[k]

if (model_type == "intercept") {

  E11 = matrix(0, nrow = q, ncol = q)
  E11[1, 1] = 1

  avb = s2_sls * ((1 - xi) * solve(Mxx) + xi * E11) / n[k]

} else {

  avb = s2_sls * (1 - xi) *
    solve(Mxx - xi * xbar %*% t(xbar)) / n[k]
}

avs = (mu4_sls - s2_sls^2) * (1 - xi) /
  as.numeric(1 - xi * t(xbar) %*% solve(Mxx) %*% xbar) / n[k]

sdb = sqrt(diag(avb))
sds = sqrt(avs)

if (sum(is.nan(sdb)) + is.nan(sds) == 0) {

  ase[, j] = c(sdb, sds)

  #####
  # Confidence interval length and coverage probability
  #####

  len_ols[, j] = 2 * z_stat * ose[, j]
  len_sls[, j] = 2 * z_stat * ase[, j]

  cp_o[, j] = (ols[, j] - z_stat * ose[, j]) <= gamma0 &
    gamma0 <= (ols[, j] + z_stat * ose[, j])

```

```

cp_s[, j] = (sls[, j] - z_stat * ase[, j]) <= gamma0 &
  gamma0 <= (sls[, j] + z_stat * ase[, j])

#####
# Wald test rejection rates for power calculation
#####

for (i in 1:range) {

  # beta_0
  rej_ols_b0[i, j] =
    b0_range[i] > ols[1, j] + z_stat * ose[1, j] |
    b0_range[i] < ols[1, j] - z_stat * ose[1, j]

  rej_sls_b0[i, j] =
    b0_range[i] > sls[1, j] + z_stat * ase[1, j] |
    b0_range[i] < sls[1, j] - z_stat * ase[1, j]

  # beta_1
  rej_ols_b1[i, j] =
    b1_range[i] > ols[2, j] + z_stat * ose[2, j] |
    b1_range[i] < ols[2, j] - z_stat * ose[2, j]

  rej_sls_b1[i, j] =
    b1_range[i] > sls[2, j] + z_stat * ase[2, j] |
    b1_range[i] < sls[2, j] - z_stat * ase[2, j]

  # beta_2
  rej_ols_b2[i, j] =
    b2_range[i] > ols[3, j] + z_stat * ose[3, j] |
    b2_range[i] < ols[3, j] - z_stat * ose[3, j]

  rej_sls_b2[i, j] =
    b2_range[i] > sls[3, j] + z_stat * ase[3, j] |
    b2_range[i] < sls[3, j] - z_stat * ase[3, j]
}

break
}

```

```

    }, error = function(e) {
      error_happened <- TRUE
    })

    if (error_happened) {
      bad_sample["invalid_variance", k] =
        bad_sample["invalid_variance", k] + 1
      next
    }
  }
}

#####
# Store Monte Carlo summaries
#####

olse[, k] = rowMeans(ols)
slse[, k] = rowMeans(sls)

SE_OLS[, k] = apply(ols, 1, sd)
SE_SLS[, k] = apply(sls, 1, sd)

ASE_OLS[, k] = rowMeans(ose)
ASE_SLS[, k] = rowMeans(ase)

CP_OLS[, k] = rowMeans(cp_o)
CP_SLS[, k] = rowMeans(cp_s)

LEN_OLS[, k] = rowMeans(len_ols)
LEN_SLS[, k] = rowMeans(len_sls)
LEN_RATIO[, k] = LEN_OLS[, k] / LEN_SLS[, k]

power_OLS_b0[, k] = rowMeans(rej_ols_b0)
power_SLS_b0[, k] = rowMeans(rej_sls_b0)

power_OLS_b1[, k] = rowMeans(rej_ols_b1)
power_SLS_b1[, k] = rowMeans(rej_sls_b1)

power_OLS_b2[, k] = rowMeans(rej_ols_b2)
power_SLS_b2[, k] = rowMeans(rej_sls_b2)
}

```

# Bibliography

- Abarin, T. and Wang, L. (2009). Second-order least squares estimation of censored regression models. *Journal of Statistical Planning and Inference*, 139(2):125–135.
- Amemiya, T. (1985). *Advanced econometrics*. Harvard University Press.
- Armijo, L. (1966). Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3.
- Beguiría, S. and Vicente-Serrano, S. M. (2006). Mapping the hazard of extreme rainfall. *Journal of Applied Meteorology and Climatology*, 45(1):108–124.
- Bose, M. and Mukerjee, R. (2015). Optimal design measures under asymmetric errors. *Journal of Statistical Planning and Inference*, 159:28–36.
- Castro-Camilo, D., Huser, R., and Rue, H. (2022). Practical strategies for generalized extreme value-based regression models. *Environmetrics*, 33(6):e2742.
- Choi, M. (2018). Medical cost personal datasets. Kaggle dataset. Accessed 2026-06-16.
- Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. John Wiley & Sons, New York.
- Dennis, J. E. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, New Jersey.

- Gao, L. L. and Zhou, J. (2014). New optimal design criteria. *Journal of Statistical Planning and Inference*, 149:140–151.
- Gill, P. E., Murray, W., and Wright, M. H. (1981). *Practical Optimization*. Academic Press, London.
- Green, P. J. (1984a). Iteratively reweighted least squares for maximum likelihood estimation. *Journal of the Royal Statistical Society: Series B*, 46(2):149–170.
- Green, P. J. (1984b). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society: Series B*, 46(2):149–192.
- He, L. and Yue, R. X. (2022). Il-optimal designs under slse. *Metrika*, 85(1):53–66.
- He, L., Yue, R. X., and Du, A. (2024). Optimal designs for comparing curves in regression models with asymmetric errors. *Journal of Statistical Planning and Inference*, 228:46–58.
- Kim, M. and Ma, Y. (2012). Efficiency of slse. *Annals of the Institute of Statistical Mathematics*, 64(4):751–764.
- Nandy, A., Basu, A., and Ghosh, A. (2025). Robust inference for linear regression models with possibly skewed error distribution. *Journal of Computational and Applied Mathematics*, 463:116502.
- Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, New York, 2 edition.
- Salamh, M. and Wang, L. (2021a). Second-order least squares estimation in nonlinear time series models. *Econometrics*, 9(4):41.
- Salamh, M. and Wang, L. (2021b). Second-order least squares method for dynamic panel data models. *Journal of Risk and Financial Management*, 14(9):410.

- Telford, R. D. and Cunningham, R. B. (1991). Sex, sport, and body-size dependency of hematology in highly trained athletes. *Medicine and Science in Sports and Exercise*, 23(7):788–794.
- Wang, L. (2003). Estimation of nonlinear berkson-type measurement error models. *Statistica Sinica*, 13(4):1201–1210.
- Wang, L. (2004). Estimation of nonlinear models with berkson measurement errors. *The Annals of Statistics*, 32(6):2559–2579.
- Wang, L. (2007). A unified approach to estimation of nonlinear mixed effects. *Canadian Journal of Statistics*, 35(2):233–248.
- Wang, L. and Leblanc, A. (2008). Second-order nonlinear least squares estimation. *Annals of the Institute of Statistical Mathematics*, 60(4):883–900.
- Wang, X., Kong, L., and Wang, L. (2023). Numerical optimization and computation for second-order least squares estimation. *Pacific Journal of Optimization*, 19:315–334.
- Wright, B. D., Green, P. J., and Braiden, P. M. (1982). Quantitative analysis of delayed fracture observed in stress rate tests on brittle materials. *Journal of Materials Science*, 17(11):3227–3234.
- Yeh, C. K. and Zhou, J. (2021). Properties of optimal regression designs under slse. *Statistical Papers*, 62(1):75–92.
- Yin, Y. and Zhou, J. (2017). Optimal designs using slse. *Statistica Sinica*, pages 1841–1856.
- Yousof, H. M., Jahanshahi, S. M. A., Ramires, T. G., Aryal, G. R., and Hamedani, G. G. (2018). A new distribution for extreme values: regression model, characterizations and applications. *Journal of Data Science*, 16(4):677–706.