MULTIPLE LINEAR REGRESSION MODEL UNDER SKEWED ERROR DISTRIBUTIONS: A CASE-STUDY OF FETAL WEIGHT PREDICTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ERKAN KALAFAT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

FEBRUARY 2021

Approval of the thesis:

MULTIPLE LINEAR REGRESSION MODEL UNDER SKEWED ERROR DISTRIBUTIONS: A CASE-STUDY OF FETAL WEIGHT PREDICTION

submitted by **ERKAN KALAFAT** in partial fulfillment of the requirements for the degree of **Master of Science** i**n Statistics, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar

Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ayşen Dener Akkaya

Head of the Department, **Statistics, METU** _____

Prof. Dr. Ayşen Dener Akkaya

Supervisor, **Statistics, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Özlem Türker Bayrak

Inter-Curricular Courses Department, Çankaya University _____

Prof. Dr. Ayşen Dener Akkaya

Statistics, METU _____

Assist. Prof. Dr. Fulya Gökalp Yavuz

Statistics, METU _____

Date: 12.02.2021

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Erkan Kalafat

Signature :

ABSTRACT

MULTIPLE LINEAR REGRESSION MODEL UNDER SKEWED ERROR DISTRIBUTIONS: A CASE-STUDY OF FETAL WEIGHT PREDICTION

Kalafat, Erkan
Master of Science, Statistics
Supervisor: Prof. Dr. Ayşen Dener Akkaya

February 2021, 55 pages

Fetal weight estimation is an essential component of pregnancy care. Current weight estimation methods use parameters obtained from multiple linear regression with least squares or maximum likelihood under normality assumption. However, non-normal errors are very common in real life setting irrespective of the sample size. This work demonstrates the properties of least squares, ridge, maximum likelihood, and modified maximum likelihood estimators of the parameters of multiple linear regression model under skewed error distributions, namely skewed normal and generalized logistic. Performance metrics are obtained from a real-life sample of fetuses with ultrasonographic limb, abdomen and head measurements. Estimation and prediction performance are compared between available and new algorithms using standard error of the estimates, Akaike and Bayesian Information criteria and prediction errors.

Keywords: Multiple linear regression, modified maximum likelihood, ridge, skewed normal, generalized logistic.

ÖZ

ÇARPIK HATA DAĞILIMLARI ALTINDA ÇOKLU DOĞRUSAL REGRESYON MODELİ: FETAL AĞIRLIK TAHMİNİ ÜZERİNE BİR VAKA ÇALIŞMASI

Kalafat, Erkan
Yüksek Lisans, İstatistik
Tez Yöneticisi: Prof. Dr. Ayşen Dener Akkaya

Şubat 2021, 55 sayfa

Fetal ağırlık tahmini, gebelik bakımının önemli bir bileşenidir. Mevcut ağırlık tahmin yöntemleri, normallik varsayımı altında en küçük kareler veya en çok olabilirlik metodlarından elde edilen parametreleri kullanır. Bununla birlikte, normal dağılıma sahip olmayan hata terimi, örnek boyutuna bakılmaksızın gerçek yaşam verilerinde çok yaygındır. Bu çalışma, çarpık hata dağılımları olan çarpık normal ve genelleştirilmiş lojistik altında çoklu doğrusal regresyon modeli için en küçük kareler, ridge, en çok olabilirlik ve uyarlanmış en çok olabilirlik tahmincilerinin özelliklerini göstermektedir. Performans ölçümleri, ultrasonografik uzuv, karın ve baş ölçümleri ile gerçek hayattaki bir fetüs örneğinden elde edilmiştir. Tahmin ve kestirim performansı, tahminlerin standart hatası, Akaike ve Bayesçi bilgi kriteri ve kestirim hataları kullanılarak mevcut ve yeni algoritmalar arasında karşılaştırılmıştır.

Anahtar Kelimeler: Çoklu dğrusal regresyon, uyarlanmış en çok olabilirlik, ridge, çarpık normal, genelleştirilmiş lojistik.

Dedicated to my beloved parents

ACKNOWLEDGEMENTS

The author wishes to express his deepest gratitude to his supervisor Professor Ayşen Dener Akkaya for her limitless enthusiasm, support and patience. This scientific endeavor would have been impossible without her boundless devotion.

The author would also like to thank Professor Asma Khalil for providing the data that was essential to completing this work. Professor Khalil's unending support and guidance over the years are humbly acknowledged.

The intellectual debates with Associate Prof. Zeynep Kalaylıoğlu, which enriched this work significantly are gratefully acknowledged.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AC	Abdominal circumference
AIC	Akaike Information Criterion
BPD	Biparietal diameter
BFGS	Broydon-Fletcher-Goldfarb-Shanno
BIC	Bayesian Information Criterion
BW	Birthweight
FL	Femur length
GL	Generalized logistic
HC	Head circumference
LS	Least squares
ML	Maximum likelihood
MML	Modified maximum likelihood
SE	Standard error
SN	Skewed normal

CHAPTER 1

INTRODUCTION

The ultrasound scans are a vital part of antenatal care, and one of the critical measures of a fetus' wellbeing is fetal weight and growth. The accurate estimation of weight is crucial for detecting at-risk fetuses with growth abnormalities, which would require closer surveillance or intervention. The incorrect classification of a fetus with appropriate weight gain as abnormal could result in unnecessary interventions and maternal anxiety. The failure to detect a fetus with abnormal weight gain until the delivery time has undesired consequences.

The estimation of fetal weight is performed with ultrasound scans measuring the fetal limb length (femur length (FL)), head circumference (HC), and tummy circumference (abdominal circumference (AC)). Hadlock *et al.* (1985) developed a multiple linear regression model in 1985 using 276 pregnant women delivered within a week of the ultrasound scan. Their model consisted of FL, HC, and AC variables, and they reported the coefficient of determination, R^2 value of 97.6% using the least squares (LS) estimators (Hadlock *et al*, 1985). The study population was formed with a balanced sampling of each weight category ranging from less than 1500 grams to more than 4000 grams with 500-gram intervals for each category. Whether the balanced sampling was performed on purpose or unintentionally is not clear. They measured the performance of their model with a metric they called error percentage estimators (Hadlock *et al*, 1985). The formula of error percentage is given in equation 1.1.

$$\left(\frac{Birtweight - Esimated\ Weight}{Birthweight} \times 100\right)$$

(1.1)

They assumed a normal distribution for errors and reported the mean and standard deviation of error percentage as 0 ± 7.5%. However, they did not validate their findings in an external cohort. In the later years, the model (dubbed Hadlock 3) was widely adopted due to its simplicity and superior performance compared to other fetal weight estimation algorithms available at the time.

Many estimated weight algorithms have been proposed over the years, but Hadlock 3 is still the most commonly used fetal weight estimation method (Akhtar 2010, Ben-Haroush 2008, Hammami 2018, Stirnemann 2017). However, Hadlock 3 is far from being precise, and the error percentage of the fetal weight estimation method is mostly dependent on the weight of the baby. The error percentage is more significant for babies on the upper end of the weight spectrum than babies on the lower end. The variance presents a problem for babies with abnormally high weight gain who are suspected to be heavy (macrosomia). Failure to detect such babies can cause complications, and misclassification of normal babies as large may result in unnecessary interventions.

Available fetal weight estimation methods are inadequate to address these problems (Hadlock 1985). Over prediction in the high birth weight ranges and under prediction in the lower birth weight ranges are significant clinical problems. It is clear that more accurate weight estimation methods are needed. Some researchers proposed more advanced imaging methods such as magnetic resonance imaging or three-dimensional ultrasound. However, advanced imaging methods are more expensive, require expertise for their use, and are generally not suited for daily practice. It would be preferable to use more common imaging modalities with measurements most physicians are already familiar with their application. Many published weight algorithms use least squares estimators and do not adequately address the error distribution assumptions. Normality of the errors is not a requirement for the least squares estimation but it has significant implications. Least squares estimators attain the Cramer-Rao bound and are the best linear estimators under normality. A Normal

(Gaussian) distribution is often presumed for errors in large samples due to law of large numbers. However, there is a wealth of literature showing normality assumption for errors often fail in real life problems no matter the sample size (Bailey 2017, Crandall 2015, Tiku 1986). Robust estimation techniques have been developed for non-Gaussian noise and they have shown to be more efficient than available estimators (Tiku 1986, Akkaya 2004). The practical application of such methods to real-life problems such as fetal weight estimation is yet to be conducted. Furthermore, we could not identify any studies exploring the error distribution problems in fetal estimation. This study investigates problems of existing weight estimation algorithms and tries to improve them by using more robust statistical methods employing different error distributions. Maximum likelihood (ML) and modified maximum likelihood (MML) estimators are tried under different error distributions assumption to obtain more efficient estimators. Ridge estimation is tried to limit the effect of collinearity in the design matrix.

1.1 Aim of the study

The aim of this study is developing regression estimates that are robust to non-Gaussian error distributions observed in fetal weight data. New estimates are obtained using various estimation methods such as ridge, ML and MML. Standard errors of the new estimates are compared to existing methods and prediction errors are tested in real-life data.

1.2 Study organization

In chapter 1, we have provided the background for fetal weight estimation, short-comings of current algorithms and proposals for obtaining improved estimates. Also we have outline the aim of the study and organization.

In Chapter 2, theoretical background on existing fetal weight estimation algorithms are briefly summarized. Simulation examples are provided for demonstrating the effect of assumption violations on parameter estimates and their variance.

In Chapter 3, the estimation methodology of various approaches and theoretical properties of skewed error distributions are described in detail.

In Chapter 4, a case study using real-life fetal weight estimation data is outlined. Estimates are developed. Comparisons are made between newly developed and existing algorithms and prediction error figures are provided using a validation set.

Chapter 5 summarized the findings, highlighted the short-comings of the current research and future research proposals.

CHAPTER 2

LITERATURE REVIEW AND THEORETICAL BACKGROUND ON EXISTING FETAL WEIGHT ESTIMATION ALGORITHMS

Ultrasonographic fetal weight estimation is a popular topic, and numerous attempts have been made over the years to improve upon the existing ones (Table 2.1). Some studies focused on special interest groups such as women with diabetes or fetuses suspected to be small or large, while other studies reported unselected populations (Akhtar 2010, Schild 2004). Most studies (Akhtar 2010, Ben-Haroush 2008, Hadlock 1985, Souka 2013) opted to report statistical measures of model fit such as R^2, and more recent ones reported clinical performance metrics, i.e., error percentage (Hammami 2018, Stirnemann 2017). Despite varying sample sizes and performance metrics, ordinary least squares and occasionally maximum likelihood estimators were used by all studies.

The INTERGROWTH-21 research project (Stirnemann, 2017) aimed to create universal fetal weight standards and proposed their version of the fetal weight estimation model. They used a fractional polynomial regression model with two variables and log-transformed the outcome variable. The authors did not report a model fit measure, but according to their report the model underestimated the fetal weight by ~10% on average. Hammami *et al.* (2018) suggested another similar polynomial model. They examined each predictor variable using a combination of linear and fractional polynomial terms and strived to find the model with a significantly better fit. Despite their best efforts, they failed to produce a model with lower prediction errors than Hadlock 3. (Hadlock, 1985) The formula of Hadlock 3, INTERGROWTH-21 and Hamammi 5 are given in equations 2.1, 2.2 and 2.3, respectively.

$$\log_{10}(Weight) = \{1.326 - 0.00326 \times AC \times FL + 0.0107 \times HC + 0.0438 \times AC + 0.158 \times FL\} \quad (2.1)$$

$$\log_e(Weight) = \left\{5.084820 - 54.06633 \times \left(\frac{AC}{100}^3\right) - 95.80076 \times \left(\frac{AC}{100}^3\right) \times \log_e \frac{AC}{100} + 3.136370 \times \frac{HC}{100}\right\} \quad (2.2)$$

$$\log_{10}(Weight) = \{1.42487 - 0.01165 \times HC + 0.03949 \times AC - 0.00028 \times AC^2 + 0.014147 \times FL - 0.00662 \times FL^2\} \quad (2.3)$$

Both methods presumably used ML estimators, unlike Hadlock 3, who used LS estimators. However, the appropriateness of LS or ML estimators in this context has never been investigated. In the next subsection, we will discuss some of the assumptions of LS and ML procedures and whether existing models satisfy them.

Table 2.1 Methodology and performance metrics of contemporary weight algorithms published in the past 20 years.

First author	**Sample size**	**Dependent variable transformation**	**Regression Method**	**Estimator**	**Performance metric (range)**
Ben-Haroush, 2008	5000 – 10.000	Log	Linear	LS	R^2 (0.85 to 0.90)
Akhtar, 2010	1000 – 5000	Log10	Linear	LS	R^2 (0.65 to 0.80)
Kehl, 2012	<1000	Log or Log10	Linear	LS	R^2 (0.65 to 0.75)
Schild, 2004	<1000	None	Polynomial	LS	Percentage error (7.1% ± 5.9)

Halaska, 2006	<1000	Log10	Polynomial	LS	Correlation (0.65 to 0.80)
Souka, 2013	1000 – 10000	None	Polynomial	LS	R^2 (0.85 to 0.90)
Stirnemann, 2017	1000 – 10000	Log	Fractional polynomial	ML	Percentage error (-12.1 to -9.4%)
Hammami, 2018	1000 – 10000	Log10	Fractional polynomial	ML	Percentage error (6.2% ± 4.6)

LS: least squares, ML: maximum likelihood

2.1 Gauss-Markov theorem, normality of errors and conventional estimators

Gauss-Markov theorem states ordinary least squares estimator has the lowest sampling variance among all other linear unbiased estimators given the errors are uncorrelated, has constant variance, and has a mean expected value of zero (2.3),

$$E[\varepsilon_i] = 0, \quad Var(\varepsilon_i) = \sigma^2 < \infty, \forall_i, Cov(\varepsilon_i, \varepsilon_j) = 0, \forall_i \neq \text{j};$$

$$\varepsilon_i = Y_i - \sum_{j=1}^{k} X_{ij} B_j \tag{2.3}$$

However, it is well established that prediction errors are not constant across the weight spectrum (Hadlock 1985, Stirnemann 2017), so constant variance assumption of the Gauss-Markov theorem fails for fetal weight estimation. The reported mean percentage errors are often non-zero in most published algorithms (Table 2.1), so the expected error mean assumption also fails. When the error variance is not constant, the assumption that covariance between all error terms is zero may fail. Neither can we assume that all off-diagonal elements are zero and diagonal elements are constant. Robust estimation of variance could be achieved with 'sandwich estimators' at the expense of useful variable coefficients. However, Huber-White estimators are sensitive to model specification and likely to be erroneous when the model is not correctly specified (Huber, 1967). Empirical evidence shows that LS estimators are not the best linear unbiased estimators of fetal

weight. While some studies used ML to obtain coefficients from fractional polynomial models (Stirnemann, 2017), ML also assumes the normality of errors to obtain variable coefficients. If errors were to be normally distributed, then the ML estimates would be the same as LS estimates. It is clear that LS estimates are not the best linear unbiased estimator of regression variables, and parameter estimation with ML cannot improve upon LS as long as error distribution is assumed to be normal.

2.2 The effect of error distribution on least squares and maximum likelihood estimates

The LS estimators are the best of unbiased linear estimators as long as Gauss-Markov assumptions hold. The LS estimators also converge to maximum likelihood estimates if the error distribution is Gaussian. When one of these assumptions fail, the maximum likelihood procedure is the obvious choice for parameter estimation. However, maximum likelihood is very susceptible to model specification and distribution assumptions of the error. To illustrate, these conditions (2.4) are simulated with the following:

$$y_i = \sum x_{ik}\theta_k + \varepsilon_i$$

where $x_i \sim N(\mu, \sigma), \forall_i$, $\varepsilon_i \sim SN(\xi, \omega, \alpha)$, $\theta_k = 0, \forall_k$ ($i = 1000$ and k= 3).

(2.4)

θ_k (k=1, 2, 3) are assumed to be known and equal to zero. We have chosen θ_k to be zero for easier interpretation of deviations from the true value. Table 2.2 shows estimated parameters using LS, ML under normality, and ML under true error distribution which is skewed normal. Maximum likelihood estimates were obtained using Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm. This quasi-Newton method is suitable for nonlinear approximation and converges faster than expectation-maximization algorithms. Results show LS and ML under normality assumption are the same and they are equally biased with inflated variances when errors are non-Gaussian. Estimates using ML under true error distribution are much closer to the true values regardless of the shape parameter assumptions. Moreover, when the ML function is correctly specified using the true error distribution, the resulting model has a lower Akaike information criterion (AIC) score. The downsides of model misspecification with ML or using LS under non-normality are obtaining biased estimates and inflated variances. R codes for these calculations can be found in Appendix C.

Table 2.2 Parameter estimates using least square, maximum likelihood under-normality and true error distribution.

Parameter	Estimate	SE	Test statistic*	AIC
Data generated with $\boldsymbol{\varepsilon_i \sim SN(\xi, \omega, \alpha);\ \alpha = -5}$				
LS under normality assumption				
θ_1	–0·0154	0·0197	–0·7822	
θ_2	0·0575	0·0195	2·9425	
θ_3	–0·0044	0·0187	–0·2393	
ML under normality assumption				1874·010
θ_1	–0·0154	0·0197	–0·7822	
θ_2	0·0575	0·0195	2·9425	
θ_3	–0·0044	0·0187	–0·2393	
ML under true error distribution				852·083
θ_1	–0·0033	0·0148	–0·2268	
θ_2	0·0333	0·0151	2·2000	
θ_3	0·0034	0·0139	0·2450	
Data generated with $\boldsymbol{\varepsilon_i \sim SN(\xi, \omega, \alpha);\ \alpha = -10}$				
LS under normality assumption				
θ_1	0·0129	0·0197	0·654	
θ_2	0·0217	0·0196	1·408	
θ_3	0·0095	0·0188	0·504	
ML under normality assumption				1886·507
θ_1	0·0129	0·0197	0·654	
θ_2	0·0217	0·0196	1·408	
θ_3	0·0095	0·0188	0·504	
ML under true error distribution				818·231
θ_1	0·0004	0·0010	0·395	
θ_2	0·0005	0·0011	0·461	
θ_3	0·0009	0·0010	0·919	

*Standardized distance from true $\theta_k = 0, \forall_k$

LS: least square, ML: maximum likelihood, N: normal distribution, SN: skewed normal distribution, SE: standard error, AIC: Akaike information criterion

2.3 Significant collinearity of independent variables

Another issue with LS or ML estimators is the collinearity of independent variables. LS estimators will be found as long as the matrix is a full column rank matrix without any exact linear relationship. However, similar to model misspecification and working under non-normality with LS, parameter estimation, variances are affected by the multicollinearity. To illustrate, these conditions (2.5) are simulated with the following:

$$y_i = \sum x_{1i}\theta_1 + x_{2i}\theta_2 + \varepsilon_i$$

$$\text{where } x_1 \sim N(\mu, \sigma), \forall_i,\ x_{2i} = x_{1i} \times \psi_i \sim N(\omega, \tau),\ \varepsilon_i \sim N(\xi, \alpha),$$

$$\theta_1 \textit{ and } \theta_2 \textit{ are known and fixed } (0.5 \textit{ and } 1.2, \textit{respectively})$$

$$(i = 1000 \text{ and } k = 2)$$

(2.5)

The results in table 2.3 show the model matrix becomes increasingly ill-conditioned as the variance of the randomly generated correlation factor ψ decreases. The standard errors of both parameters are inflated with increasing collinearity. Parameter estimation of θ_1 is not affected to an appreciable degree but estimates of θ_2 are grossly inaccurate. Variables used in fetal weight estimation are highly collinear as they belong to the same fetus, and all are anthropomorphic measurements. Some variables such as HC and BPD belong to the same body part and they are essentially a non-linear combinations of each other. The information contained within these variables are likely to be similar but some weight estimation algorithms use them in the same model, despite very high correlation rate. The significant collinearity is likely to bias the parameter estimation and affect the model selection whether we use LS or ML estimates. R codes for these calculations can be found in Appendix D.

Table 2.3 Parameter estimates and standard errors using maximum likelihood under varying collinearity. True values of $\theta_1$and θ_2 are 0.5 and 1.2, respectively.

Parameter	Estimate	SE	$[\boldsymbol{X^T X}]^{-1}$
	$x_{2i} = x_{1i} \times \psi_i \sim N(\omega, \tau)\ \omega = 0.15\ \tau = 0.15$		
θ_1	0.5446	0.0419	$\begin{bmatrix} 0.002 & -0.006 \\ -0.006 & 0.043 \end{bmatrix}$
θ_2	1.0220	0.2030	
	$x_{2i} = x_{1i} \times \psi_i \sim N(\omega, \tau)\ \omega = 0.15\ \tau = 0.05$		
θ_1	0.5550	0.0973	$\begin{bmatrix} 0.010 & -0.060 \\ -0.060 & 0.395 \end{bmatrix}$
θ_2	0.6660	0.6092	
	$x_{2i} = x_{1i} \times \psi_i \sim N(\omega, \tau)\ \omega = 0.15\ \tau = 0.01$		
θ_1	0.6530	0.4590	$\begin{bmatrix} 0.224 & -1.487 \\ -1.487 & 9.895 \end{bmatrix}$
θ_2	0.0100	3.047	
	$x_{2i} = x_{1i} + \psi_i \sim N(\omega, \tau)\ \omega = 0.15\ \tau = 0.001$		
θ_1	0.6383	4.574	$\begin{bmatrix} 22.27 & -148.4 \\ -148.4 & 989.5 \end{bmatrix}$
θ_2	0.1056	30.491	

N: normal distribution, SE: standard error

In summary, both LS and ML estimators encounter significant problems due to assumption violations for this particular real-life problem. Even though many attempts to improve upon the Hadlock 3 model have been made, none were successful (Hamammi 2018). LS and ML estimators' inability to provide better solutions due to model misspecification may explain the issue.

Thus, in Chapter 3, different regression techniques are explained that would work under some or all of these problems.

CHAPTER 3

MULTIPLE LINEAR REGRESSION AND ESTIMATION OF THE PARAMETERS

In this chapter, methods for estimating the parameters of multiple linear regression model are presented using four estimation procedures, namely the LS, ridge, ML with BFGS algorithm and MML.

Regression analysis is a widely used statistical method for modeling and investigating the relationship between variables. More specifically, it aims to describe the relationship between two or more variables of interest. The classical multiple linear regression model formulation is given as follows

$$y_i = \sum_{j=0}^{q} \theta_j x_{ij} + \varepsilon_i \,;\; i = 1, \dots, n \tag{3.1}$$

Traditionally, errors are assumed to have a Gaussian distribution. However, in many real life problems, it is recognized that non-normal distributions are more common (Huber, 1981; Tiku, 1986; Tiku and Akkaya, 2004). In this study, we consider multiple linear regression model assuming two different skewed distributions, namely Generalized logistic (GL) and Skewed Normal (SN) for the error term ε_i.

Some recent work is available on the multiple linear regression model when random error is non-normal (Sazak, 2003; İslam and Tiku, 2004; Sazak, 2006; Tiku, 2008; Akkaya and Tiku, 2010; Bayrak and Akkaya, 2018). It is known that under non-normality, LS estimators are neither efficient nor robust (Islam and Tiku, 2004; Bayrak and Akkaya, 2010) and ML estimators cannot be obtained analytically due to nonlinear terms involved in the likelihood equations. Although numerical solution

can be obtained through recently developed optimization algorithms still they are not very convenient due to convergence problems. Estimates induce bias especially for small samples (Barnett, 1966; Puthenpura and Sinha, 1986; Vaughan, 1992). Moreover it is not possible to identify the properties of the estimators theoretically so that inference about them becomes questionable. The modified maximum likelihood (MML) methodology was introduced by Tiku (1967) and developed by Tiku and Suresh (1992) alleviates the above mentioned problems. The MML estimators have very desirable mathematical and statistical properties: they are;

i) Explicit functions of sample observations and computations become straightforward

ii) More efficient than the LS estimators under non-normality, for all sample sizes but particularly for large n

iii) Asymptotically fully efficient that is unbiased and having minimum variances under very general regularity conditions, and

iv) Robust to plausible deviations from the assumed distributions and to mild data anomalies (e.g. outliers, inliers, misspecification of the distribution).

The generalized logistic distribution for errors is given below;

$$\mathrm{GL}(\varepsilon; \mathrm{b}, \sigma) = \frac{\mathrm{b}}{\sigma}\exp(-\varepsilon/\sigma)/\{1+\exp(-\varepsilon/\sigma)\}^{\mathrm{b}+1}; \quad -\infty < \varepsilon < \infty. \qquad (3.2)$$

where b is the shape parameter, σ is the scale parameter and $\varepsilon_i = y_i - \theta_0 - \sum_{j=1}^{q}\theta_j x_{ij}$ (1≤i ≤n). The expected value and the variance of a GL(b; μ, σ) distribution is

$$\mathrm{E}(.)=\mu+\{\psi(\mathrm{b})-\psi(1)\}\sigma \text{ and } \mathrm{V}(.)=\{\psi'(\mathrm{b})+\psi'(1)\}\sigma^2 \qquad (3.3)$$

where $\psi(.)$is the psi-function and $\psi'(.)$ is its derivative (Abramowitz and Stegun, 1974). The distribution is negatively skewed for b<1 and positively skewed for b>1 and for b=1, it becomes logistic. The Pearson coefficients of skewness and kurtosis of the GL (b; σ) are given below in Table 3.1:

Table 3.1 Skewness and kurtosis values of generalized logistic distribution under different shape parameters, b.

	b	**0.5**	**1**	**2**	**4**	**6**
Skewness	$\mu_3/\mu_2^{3/2}$	-0.855	0	0.577	0.868	0.961
Kurtosis	μ_4/μ_2^2	5.400	4.2	4.333	4.758	4.951

The other distribution we assume for errors is the skewed normal distribution. The SN distribution probability density function is

$$f(\varepsilon;\lambda,\sigma) = \frac{2}{\sigma}\phi(\frac{\varepsilon}{\sigma})\Phi(\lambda\frac{\varepsilon}{\sigma}), \quad -\infty< \varepsilon <\infty, \quad \sigma>0 \tag{3.5}$$

Where λ is the shape parameter, σ is the scale parameter, ϕ(.)= Standard normal pdf, $\Phi(.) = \text{Standard normal cdf}$ and $\varepsilon_i = y_i - \theta_0 - \sum_{j=1}^{q}\theta_j x_{ij}$ (1≤i ≤n, 1<j<q). The expected value and the variance of SN distribution are given by

$$\text{E}(.) = \mu + \sqrt{\frac{2\lambda^2}{\pi(1+\lambda^2)}}\sigma \text{ , V}(.)=(1-\frac{2\lambda^2}{\pi(1+\lambda^2)})\sigma^2 \tag{3.6}$$

Parameter selection for skewed normal distribution was made from a range of values with known skewness and kurtosis, which are outlined in Table 3.2.

Table 3.2 Skewness and kurtosis values of skewed normal distribution under different shape parameters, λ.

	$\boldsymbol{\lambda}$	**0.0**	**1.0**	**2.0**	**3.0**	**4.0**	**5.0**	**10**	**20**	∞
Skewness	$\mu_3/\mu_2^{3/2}$	0	0.14	0.45	0.67	0.78	0.85	0.96	0.99	∞
Kurtosis	μ_4/μ_2^2	3.0	3.06	3.31	3.51	3.63	3.71	3.82	3.86	3.869

The reason for assuming these distributions is due to the fact that in the case study presented in Chapter 4, the residuals ε can be well represented both by the generalized logistic and skewed normal distributions. Different methods can be used to estimate the model parameters.

3.1 Least squares estimation under skewed error distributions

The rationale of least squares is minimizing the sum of squared errors and finding the parameters achieving this condition. LS has some advantages over other estimators. First, it has a very low computational burden and easy to calculate. Second, it does not have any intrinsic distribution assumption.

Least squares estimates are unbiased and best linear estimators as long as Gauss-Markov assumptions hold. Least squares estimates can be calculated with simple matrix operations (3.7)

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{31} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{3n} \end{bmatrix} \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} \quad \boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ \vdots \\ y_n \end{bmatrix}$$

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} \tag{3.7}$$

The variance of parameter estimates are calculated with the following set of matrix operations (3.8):

$$\begin{aligned} E\left[\left(\hat{\theta}-\theta\right)\left(\hat{\theta}-\theta\right)^T\right] &= E[((X^TX)^{-1}X^T\varepsilon)((X^TX)^{-1}X^T\varepsilon)^T] \\ &= E[((X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}] \\ &= [((X^TX)^{-1}X^TE(\varepsilon\varepsilon^T)X(X^TX)^{-1}] \\ &= (X^TX)^{-1}X^T(\sigma^2I)X(X^TX)^{-1} \\ &= (\sigma^2I)(X^TX)^{-1} \end{aligned} \quad (3.8)$$

where, $\sigma^2 = \frac{\varepsilon^T\varepsilon}{n-k}$.

However, the variance estimate of LS and the intercept are biased under non-Gaussian noise. A bias correction procedure is needed before LS estimates can be compared with ML and MML estimates under skewed error distributions. The following equations demonstrate the bias correction procedure for LS under generalized logistic (3.9) and skewed normal distributions (3.10).

i) Generalized logistic distribution

$$\tilde{\theta}_0 = \bar{y} - \{\psi(b) - \psi(1)\}\tilde{\sigma},\ \widetilde{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$\tilde{\sigma} = \frac{s_e}{\sqrt{\psi'(b)+\psi'(1)}};\quad s_e^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y}-\sum_{j=1}^{q}\tilde{\theta}_jx_{ji})^2}{n-q-1};\quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i \quad (3.9)$$

Note that $\tilde{\theta}_0$ and $\tilde{\sigma}$ are bias-corrected estimators.

ii) Skewed normal distribution

$$\tilde{\theta}_0 = \bar{y} - \left\{\sqrt{\frac{2\lambda^2}{\pi(1+\lambda^2)}}\right\}\tilde{\sigma},\ \widetilde{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$\tilde{\sigma} = \frac{s_e}{\sqrt{1-\frac{2\lambda^2}{\pi(1+\lambda^2)}}};\ s_e^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y}-\sum_{j=1}^{q}\tilde{\theta}_j x_{ji})^2}{n-q-1};\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (3.10)$$

Here, $\tilde{\theta}_0$ and $\tilde{\sigma}$ are bias-corrected estimators.

3.2 Ridge estimation

Ridge regression is a shrinkage method aimed at reducing mean squared error at the expense of introducing bias to the estimates. Ridge regression is better for parameter estimation than LS when there is significant collinearity in the data. The aim of ridge regression is introducing bias to the estimates in return for reduced standard errors. The ridge estimator is given by the following

$$\widehat{\boldsymbol{\theta}} = \arg min\,(\boldsymbol{y}-\boldsymbol{X\theta})^{\boldsymbol{T}}(\boldsymbol{y}-\boldsymbol{X\theta}) + \kappa\boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{\theta}$$
$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + \kappa\boldsymbol{I})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \qquad (3.11)$$

κ is the tuning parameter and penalized the parameter estimates proportional to its size. When the tuning parameter is zero the ridge estimator decomposes to least squares and larger values introduce more bias in an attempt to reduce the mean squared error of the estimates. Selection of the tuning parameter is not an exact process. Usually, different values of κ are tried to minimize the mean squared error without causing instability in the parameter estimates. Diagnostic plots of parameter estimates against κ are used to find the inflection point where the introduced bias is too much. However, its use is criticized due to subjective selection process of tuning parameter. The advantages of the ridge estimator is reduced error variance. However, the variance is traded for bias and coefficients shrink towards zero with increasing κ values. They are also extremely susceptible to outliers and share many similarities with LS estimates in this regard. The use of ridge estimation is highly controversial

but it is employed in this work to test whether tackling multicollinearity will improve the prediction process.

3.3 Maximum likelihood estimation under non-normality

Maximum likelihood is a powerful method for estimating unknown parameters. However, parameter estimation and their variance are susceptible to model misspecification. When the true error distribution is Gaussian, the maximum likelihood estimates are the same as LS. However, it is not uncommon for errors to have a non-Gaussian distribution, no matter the sample size. In such cases, correct model specification using the true error distribution will vastly improve the parameter estimation and variance.

Maximization of log-likelihood functions of generalized logistic and skewed normal distribution is complicated. Both distributions include non-linear terms that require special optimization techniques such as expectation-maximization, Nelder-Mead or BFGS. The probability density functions of generalized logistic (skewed) and skewed normal distribution have non-linear terms (see Equations (3.12) and (3.16)) that require special optimization methods We employed the BFGS method in our study to estimate maximum likelihood for these functions. BFGS is a quasi-Newton optimization technique that preconditions the gradient with curvature information. The loss function of the Hessian matrix is approximated step-by-step with information obtained from gradient evaluations. It has a faster convergence rate than expectation maximization algorithms, which is why we preferred it in this study. The maximum likelihood estimates are obtained with maximum likelihood method using the BFGS optimization algorithm. BFGS is a gradient method and the algorithm begins at an initial value of t where;

$$\begin{bmatrix} t_1 \\ \vdots \\ t_k \end{bmatrix} \in \mathbb{R} \text{ and } f(t) \text{ is a differentiable function}$$

BFGS makes an approximation of the Hessian matrix and updates it in each stage to direct the search. Approximation of the $\nabla^2 f(t^*)$ and current values of t_c and $H_c$is performed in several steps. The search direction is given by $d = -H_c^{-1}\nabla f(t_c)$. In the next step the following value of t is approximated with $t_n = t_c + d\lambda$ to allow sufficient decrease. Then the current and approximated next value of t and current Hessian is used to obtain the next Hessian matrix. The gradient norm is checked in each step to test whether convergence is achieved.

3.4 Modified maximum likelihood estimation under non-normality

Numerical optimization procedures are prone to non-convergence, convergence to local maxima in the presence of non-linear functions. Explicit solutions are always preferred due to demonstrated robustness under specific conditions, no computational demand and estimates are guaranteed to be correct. MML is an alternative to ML, offering explicit solutions in place of numerical optimization and more efficient estimates. The MML estimates can be obtained under a variety of distribution assumptions.

3.4.1 MML method under generalized logistic distribution

The likelihood function of generalized logistic distributed error term is

$$\mathrm{L} = \left(\frac{\mathrm{b}}{\sigma}\right)^{\mathrm{n}} \prod_{\mathrm{i}=1}^{\mathrm{n}} \frac{\exp(-z_i)}{[1+\exp(-z_i)]^{\mathrm{b}+1}}; \quad -\infty < \mathrm{z} < \infty, \ \ \sigma > 0, \mathrm{b}>0. \tag{3.12}$$

The corresponding log-likelihood function is

$$\ln\mathrm{L} = \mathrm{n}\ln\mathrm{b} - \mathrm{n}\ln\sigma + \sum_{\mathrm{i}=1}^{\mathrm{n}} \ln\mathrm{e}^{-\mathrm{z}_\mathrm{i}} - (\mathrm{b}+1)\sum_{\mathrm{i}=1}^{\mathrm{n}} \ln[1+\mathrm{e}^{-\mathrm{z}_\mathrm{i}}] \tag{3.13}$$

The maximum likelihood equations for estimating θ_0, θ_j and σ (1<j<q) are as follows:

$$\frac{\partial \ln L}{\partial \theta_0} = \frac{n}{\sigma} - \frac{(b+1)}{\sigma}\sum_{i=1}^{n}\left(\frac{e^{-z_i}}{1+e^{-z_i}}\right) = 0$$

$$\frac{\partial \ln L}{\partial \theta_j} = \frac{1}{\sigma}\sum_{i=1}^{n} x_{ji} - \frac{(b+1)}{\sigma}\sum_{i=1}^{n}\left(\frac{e^{-z_i}}{1+e^{-z_i}}\right) x_{ji} = 0 \quad (j=1,2,3)$$

and

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma}\sum_{i=1}^{n} z_i - \frac{(b+1)}{\sigma}\sum_{i=1}^{n}\left(\frac{e^{-z_i}}{1+e^{-z_i}}\right) z_i = 0 \qquad (3.14)$$

Solutions of the set of equations (3.14) are the ML estimators. These equations do not have explicit solutions due to nonlinear term $g(z_i) = \frac{e^{-z_i}}{1+e^{-z_i}}$ and MML method is used to solve them. The methodology applied to model (3.12) is summarized as follows:

i) To find the MML estimators, first the equations in (3.12) are expressed in terms of the ordered variates $z_{(i)} = (y_{[i]} - \theta_0 - \sum_{j=1}^{q}\theta_j x_{[i]j})/\sigma$, $1 \le i \le n$ (q=3). $(y_{[i]}, x_{[1]j,\ldots,,,}x_{[n]j})$ are the concomitants of $z_{(i)}$, $(1 \le i \le n)$ i.e., the vector of observations $(y_i, x_{i1,\ldots,} x_{iq})$ associated with $z_{(i)}$ obtained by arranging z_i in increasing order of magnitude.

ii) Write $g(z) = \frac{e^{-z}}{(1+e^{-z})}$ and linearize the function $g(z_{(i)})$ by using the first two terms of Taylor series expansions around $E(z_{(i)}) = t_{(i)}$, as follows

$$g(z_{(i)}) \cong g(t_{(i)}) + (z_{(i)} - t_{(i)})\left[\frac{dg(z_{(i)})}{z}\right]_{z_{(i)}=t_{(i)}}$$

$$= \frac{e^{-t_{(i)}}}{\left(1+e^{-t_{(i)}}\right)} + \left(z_{(i)} - t_{(i)}\right)\left[-\frac{e^{-t_{(i)}}}{\left(1+e^{-t_{(i)}}\right)^2}\right]$$

$$= \alpha_i - \beta_i;\ 1 \le i \le n, \tag{3.15}$$

where $\alpha_i = \frac{e^{-t_{(i)}}}{\left(1+e^{-t_{(i)}}\right)^2}(1 + t_{(i)} + e^{-t_{(i)}})$ and $\beta_i = \frac{e^{-t_{1(i)}}}{\left(1+e^{-t_{1(i)}}\right)^2}$

Here $t_{(i)}$ is obtained from the following equation:

$$\int_{-\infty}^{t_{(i)}} f(y)dy = \frac{i}{n+1};\qquad 1 \le i \le n \tag{3.16}$$

where f(.) is the pdf of the generalized logistic distribution and y = ε/σ. The pdf function of generalized logistic distribution is available in a package called `glogis` in R for Windows.

iii) The linear approximations expressed by equation (3.15) is incorporated into the maximum likelihood equations (3.14). Then the resulting set of equations (3.17) have explicit solutions which are the following MML estimators (MMLEs):

$$\widehat{\boldsymbol{\theta}} = \mathbf{K} - \mathbf{D}\widehat{\sigma} \quad \text{and} \quad \widehat{\sigma} = -B + \sqrt{B^2 + 4nC}/2\sqrt{n(n-4)}$$

$$\mathbf{K} = (\mathbf{X}'\boldsymbol{\beta}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\beta}\mathbf{Y}) = (K_j),\ \mathbf{D} = (\mathbf{X}'\boldsymbol{\beta}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\xi}\mathbf{1}) = (D_j)$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{31} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{3n} \end{pmatrix}_{n\times 4},\ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_n \end{pmatrix}_{n\times n};$$

$$\boldsymbol{\xi} = \begin{pmatrix} \Delta_1 & & 0 \\ & \ddots & \\ 0 & & \Delta_n \end{pmatrix}_{n\times n},\ \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1}\ \Delta_i = \alpha_i - \frac{1}{b+1}$$

$$B = (b+1)\sum_{i=1}^{n} \alpha_i [y_{[i]} - \sum_{j=1}^{4} K_j x_{[i]j}],\ C = (b+1)\sum \delta_i \left[y_{[i]} - \sum_{j=1}^{4} K_j x_{[i]j}\right]^2. \quad (3.17)$$

Computations: Using the LS estimators of $\tilde{\theta}_0$ and $\tilde{\theta}_j$ (j=1, 2, 3) and $\tilde{\sigma}$ given in Section 3.1, the estimated residuals

$$\tilde{\varepsilon}_i = y_i - \tilde{\theta}_o - \sum_{j=1}^{3} \theta_j x_{ij} \qquad (1 \le i \le n)$$

are obtained. The next step is to order $\tilde{e}_i$ $(1 \le i \le n)$ and obtain the concomitants $\left(y_{[i]}, x_{[1]j,\ldots,,,} x_{[n]j}\right)$ corresponding to the i^{th} ordered value of $\tilde{\varepsilon}_i$. Using these concomitants the MML estimators are calculated from Equations (3.15). The LS estimators $\tilde{\theta}_0, \tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_3$ and $\tilde{\sigma}$ are then replaced by $\hat{\theta}_0$, $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ and $\hat{\sigma}$, respectively and new concomitants are obtained. The revised MMLEs are computed from these new concomitants. The process is repeated one more time for the estimates to stabilize sufficiently. R codes for these calculations can be found in Appendix B.

3.4.2 MML method under skewed normal distribution

The likelihood function of skewed normal distributed error term becomes

$$L = \left(\frac{2}{\sigma}\right)^{n} \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n} z_i^2} \prod_{i=1}^{n} \Phi\left(\lambda(z_i)\right); \quad z_i = (y_i - \theta_0 - \sum_{j=1}^{q} \theta_j x_{ij})/\sigma;$$
$$-\infty < z < \infty, \ \ \sigma > 0, \lambda \in \mathbb{R}. \tag{3.18}$$

The corresponding log-likelihood function is

$$\ln(L) = n \ln 2 - n \ln \sigma - \frac{n}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} \ln \Phi(\lambda(z_i)).$$

The maximum likelihood equations for estimating θ_0, θ_j and σ (1<j<q) are

$$\frac{\partial \ln L}{\partial \theta_0} = \frac{1}{\sigma} \sum_{i=1}^{n} z_i - \frac{\lambda}{\sigma} \sum_{i=1}^{n} \frac{f(\lambda z_i)}{F(\lambda z_i)} = 0$$

$$\frac{\partial \ln L}{\partial \theta_j} = \frac{1}{\sigma} \sum_{i=1}^{n} z_i x_{ji} - \frac{\lambda}{\sigma} \sum_{i=1}^{n} \frac{f(\lambda z_i)}{F(\lambda z_i)} x_{ji} = 0, \ \ (j = 1, 2, 3)$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma} \sum_{i=1}^{n} z_i^2 - \frac{\lambda}{\sigma} \sum z_i \frac{f(\lambda z_i)}{F(\lambda z_i)} = 0$$

where $h(z) = \frac{f(\lambda z)}{F(\lambda z)}$.

(3.19)

Explicit solution of these five equations require the term $h(z)$ to be linearized using the two terms of Taylor Series Expansion around $E(z_{(i)}) = t_{(i)}$ and the quantiles from skewed normal distribution $t_{(i)}$:

$$h(z_{(i)}) \cong h(t_{(i)}) + (z_{(i)} - t_{(i)}) \left[\frac{dh(z_{(i)})}{z}\right]_{z_{(i)}=t_{(i)}}$$

where $\frac{dh(z_{(i)})}{z} = \frac{\lambda f'(\lambda z_i) \times F(\lambda z_i) - \lambda f(\lambda z_i)^2}{F(\lambda z_i)^2}$, $f'(\lambda t_{(i)}) = -\lambda t_{(i)} f(\lambda t_{(i)})$;

f(.) and F(.) are standard normal pdf and cdf, respectively.

$$= \alpha_i - \beta_i;\ 1 \le i \le n,$$

(3.20)

Here $\alpha_i = \frac{f(\lambda t_{(i)})}{F(\lambda t_{(i)})} + t_{(i)}\beta_i$ and $\beta_i = \left[\frac{f(\lambda t_{(i)}) \times (\lambda^2 t_{(i)} F(\lambda t_{(i)}) + \lambda f(\lambda t_{(i)}))}{F(\lambda t_{(i)})^2}\right]$.

The quantiles of skewed normal distribution $t_{(i)}$ are obtained from the following equation:

$$\int_{-\infty}^{t_{(i)}} f(t)dt = \frac{i}{n+1};\qquad 1 \le i \le n$$

(3.21)

where f(.) is the pdf of the skewed normal distribution and $y = \varepsilon/\sigma$. The pdf function of skewed normal distribution is available in a package called `SN` in R for Windows. To obtain the MMLEs, the steps given in Section 3.4.1 are repeated and the following MMLEs are obtained:

$$\widehat{\boldsymbol{\theta}} = \mathbf{K} - \lambda \mathbf{D}\hat{\sigma} \quad \text{and} \quad \hat{\sigma} = -B + \sqrt{B^2 + 4nC}/2\sqrt{n(n-4)}$$

$$\mathbf{K} = (\mathbf{X}'\boldsymbol{\delta}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\delta}\mathbf{Y}) = (K_j),\ \mathbf{D} = (\mathbf{X}'\boldsymbol{\delta}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\alpha}\mathbf{1}) = (D_j)$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{31} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{3n} \end{pmatrix}_{n\times 4},\ \boldsymbol{\delta} = \begin{pmatrix} \delta_1 & & 0 \\ & \ddots & \\ 0 & & \delta_n \end{pmatrix}_{n\times n};\ \delta_i = 1 + \lambda\beta_i$$

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_n \end{pmatrix}_{n\times n},\ \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1}$$

$$B = \sum_{i=1}^{n} \alpha_i [y_{[i]} - \sum_{j=1}^{4} K_j x_{[i]j}],\ C = \sum \delta_i \left[y_{[i]} - \sum_{j=1}^{4} K_j x_{[i]j}\right]^2.$$

(3.22)

R codes for these calculations can be found in Appendix A.

CHAPTER 4

CASE STUDY: FETAL WEIGHT PREDICTION

In this chapter we use the aforementioned methods to estimate fetal weight through multiple regression model and compare the efficiencies of the model parameters. The data include real-life recording of 10932 anonymzed pregnancies who were scanned within 10 days of delivery and had records of gestational age at scan, AC, HC, FL and birth weight. Hadlock 3, INTERGROWTH-21 aand Hammami 5 estimates are applied to the data to obtain the residuals. The formula of Hadlock 3, INTERGROWTH-21 and Hamammi 5 are given in equations 4.1, 4.2 and 4.3, respectively.

$$\log_{10}(Weight) = \{1.326 - 0.00326 \times AC \times FL + 0.0107 \times HC + 0.0438 \times AC + 0.158 \times FL\} \quad (4.1)$$

$$\log_e(Weight) = \left\{5.084820 - 54.06633 \times \left(\frac{AC}{100}^3\right) - 95.80076 \times \left(\frac{AC}{100}^3\right) \times \log_e \frac{AC}{100} + 3.136370 \times \frac{HC}{100}\right\} \quad (4.2)$$

$$\log_{10}(Weight) = \{1.42487 - 0.01165 \times HC + 0.03949 \times AC - 0.00028 \times AC^2 + 0.014147 \times FL - 0.00662 \times FL^2\} \quad (4.3)$$

Descriptive statistics and variable distributions of residuals from existing models were investigated using the whole sample. Then, the sample was divided randomly 1:1 for training and validation. LS, ridge, maximum likelihood, modified maximum likelihood estimates are obtained usign the training set. Variance of the estimates were derived from either asymptotic variance-covariance matrices or bootstrapping. Prediction errors were obtained from the validation set using the estimates derived from the training set. Results were tabulated or plotted for visual inspection.

4.1 Descriptive statistics and distribution of residuals obtained with existing regression models

In a sample of 10932 pregnancies, we applied the Hadlock 3 estimates to AC, FL and HC measurements and obtained the residuals. The residuals have the following characteristics (Table 4.1). Hadlock 3 estimates does not have an expected mean of zero and the variance is not constant for all data subsets. Residuals were not homoscedastic and the expected mean was dependent on the estimated value (Figure 4.1). Although, the available algorithm used least square estimators, they apparently do not satisfy the assumptions of Gauss-Markov theorem. Hence, the estimators of Hadlock 3 do not have the smallest variance among all other unbiased estimators. Similar problems also exists with INTERGROWTH-21 and Hammami 5 estimates. The residuals of Hadlock 3, INTERGROWTH-21 and Hammami 5 all have skewed distributions (Figure 4.2).

Table 4.1 Characteristics of Hadlock 3 estimates and errors in a population of 10932 pregnancies

N	$E[Y_i]$	ε_i
10932	Any $E[Y_i]$	-28.1 ± 302.7
1	$E[Y_i]$ < 1500 grams	-
74	$1500 \leq E[Y_i]$ < 2000 grams	-103.1 ± 193.2
534	$2000 \leq E[Y_i]$ < 2500 grams	-67.6 ± 251
1456	$2500 \leq E[Y_i]$ < 3000 grams	-27.3 ± 272
3300	$3000 \leq E[Y_i]$ < 3500 grams	-8.6 ± 280.4
3830	$3500 \leq E[Y_i]$ < 4000 grams	46.1 ± 303.8
1518	$4000 \leq E[Y_i]$ < 4500 grams	123.5 ± 337.0
218	$4500 \leq E[Y_i]$ < 6000 grams	246.1 ± 375.5
1	$6000 \leq E[Y_i]$	-

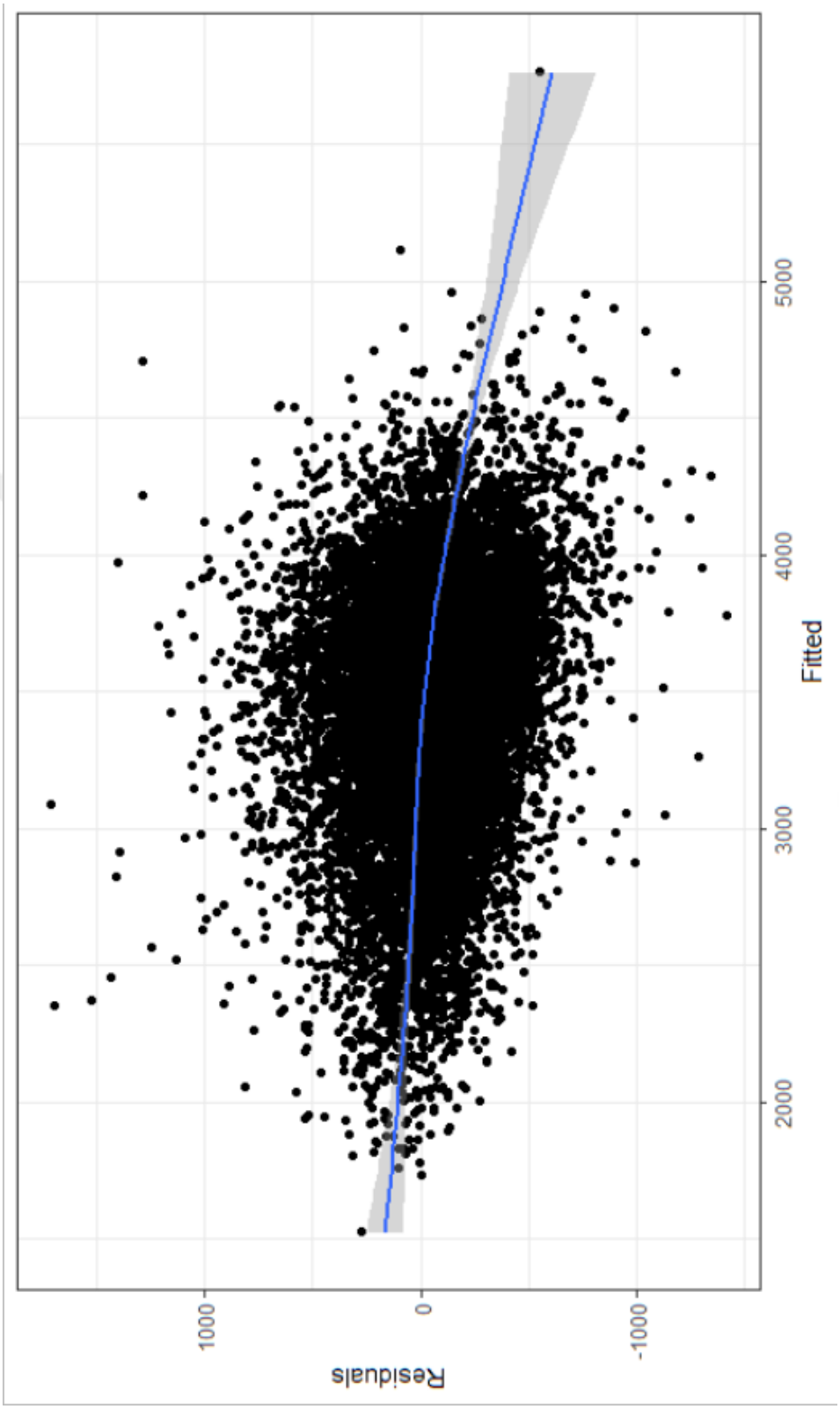
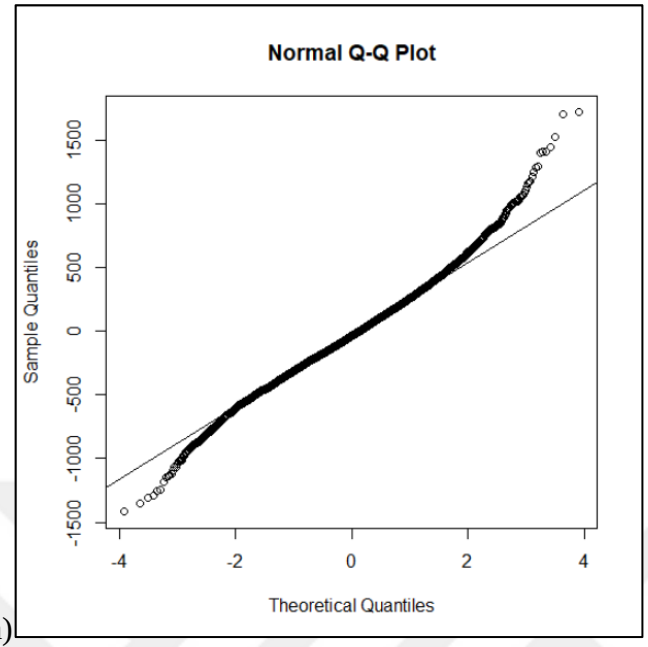
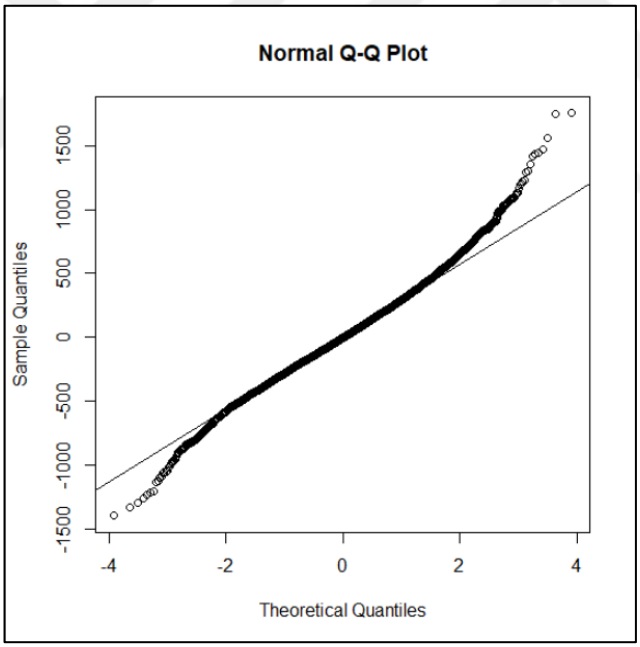

Residuals
1000
0
-1000
2000
3000
4000
5000
Fitted

Figure 4.1. Residuals versus fitted values for Hadlock 3 model.

(a)



(b)



(c)

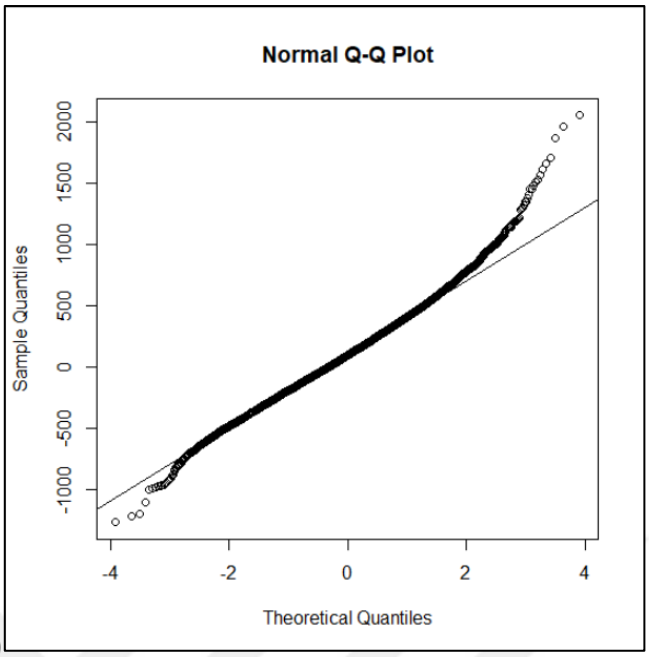


Figure 4.2. Quantile-quantile plots of residuals obtained from (a) Hadlock 3, (b) INTERGROWTH-21 and (c) Hammami 5 estimates.

We investigated possible distribution that may better fit the residuals. Residuals from Hadlock 3, INTERGROWTH-21 and Hamammi 5 estimates all have long tails and slightly right skewed distributions (Table 4.2). We bootstrapped samples sized 1000 with replacement over 10.000 iterations and the variance of skewness and kurtosis values were very low, indicating the distribution is consistently long-tailed and right skewed.

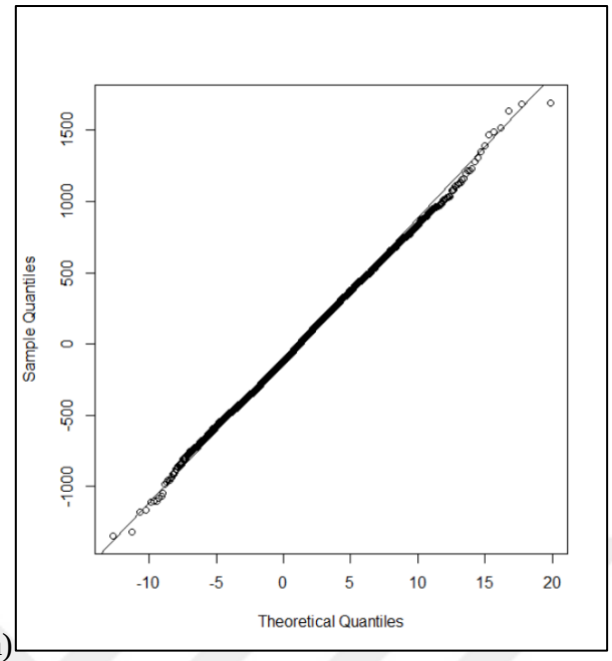
Table 4.2 The skewness and kurtosis values of estimation errors according to Hadlock 3, INTERGROWTH-21 and Hammami 5 models.

Models	Skewness	Kurtosis	Skewness, subsets	Kurtosis, subsets
Hadlock 3	-0.231	4.045	-0.222 ± 0.018	3.939 ± 0.217
INTERGROWTH − 21	-0.316	4.002	-0.302 ± 0.019	3.841 ± 0.311
Hamammi 5	-0.215	4.043	-0.207 ± 0.018	3.936 ± 0.213

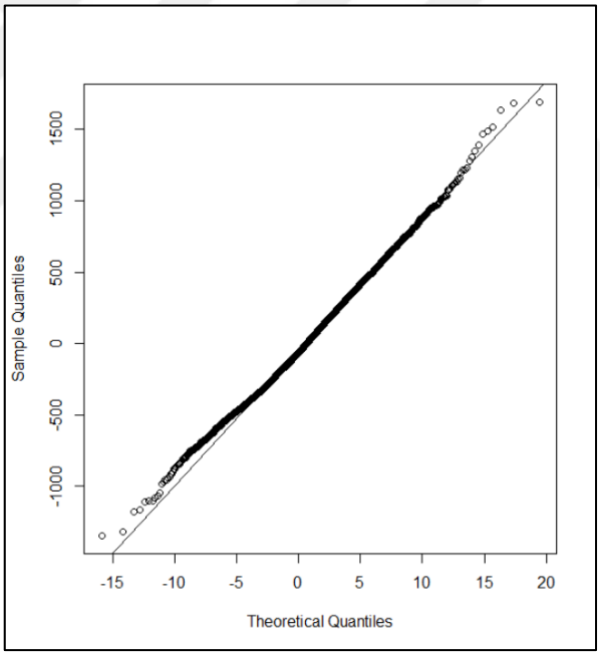
Values are presented as mean and standard deviation

We hypothesized generalized logistic distribution (skewed) and skewed normal distribution may provide a better fit to the data. We have tried GL distribution with b values of 1.5 and 1.2. Also, we have tried SN distribution with a λ value of 1.95. The assumptions were derived from theoretical distribution parameter outline in the previous section. The best fit to the residual was provided by generalized logistic distribution with b of 1.5. The coefficient of determination was over 0.999 as opposed to 0.995 for skew-normal and 0.989 for normal distribution. Quantile-quantile plots visually confirmed GL with b=1.5 has the best fit, followed by GL with b= 1.2 and SN with $\lambda = 1.95$ (Figure 4.3)

(a)


Sample Quantiles
1500
1000
500
0
-500
-1000
-10
-5
0
5
10
15
20
Theoretical Quantiles

(b)


Sample Quantiles
1500
1000
500
0
-500
-1000
-15
-10
-5
0
5
10
15
20
Theoretical Quantiles

(c)

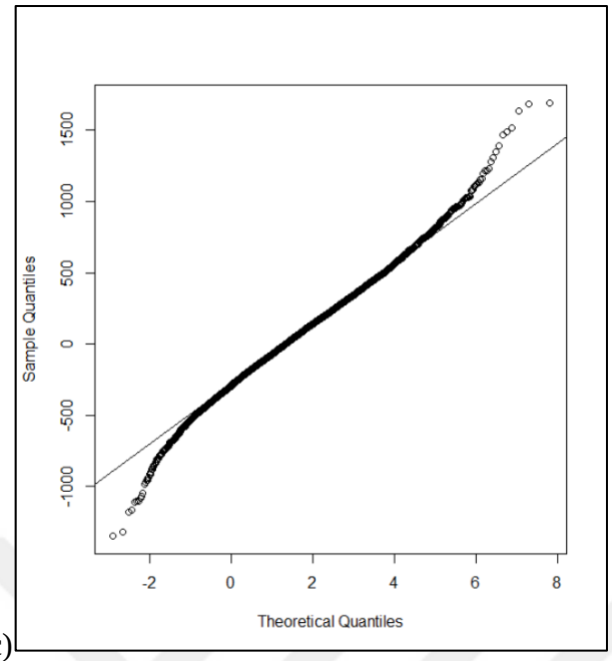


Figure 4.3. Quantile-quantile plots of residuals obtained from least squares estimated fitted to (a) GL (b=1.5), (b) GL (b=1.2) and (c) SN (λ=1.95).

Collinearity is also a significant problem for regression models. Significant correlation between the independent variables causes Determinant of X^TX to be ill-conditioned. Collinearity causes the inverse of X^TX to be highly sensitive to changes in the model matrix and as a results, the coefficients have large variance. We observed significant correlation between independent variables in our sample. There was moderate correlation between measurements of AC, FL and HC (Table 4.3, Figure 4.4).

Table 4.3 Correlation matrix of predictor and outcome variables. Data derived from measurements of 10932 babies.

Variables	$\boldsymbol{X_{AC}}$	$\boldsymbol{X_{HC}}$	$\boldsymbol{X_{FL}}$	$\boldsymbol{Y_{BW}}$
$\boldsymbol{X_{AC}}$	1.00	0.618	0.590	0.827
$\boldsymbol{X_{HC}}$	0.618	1.00	0.498	0.645
$\boldsymbol{X_{FL}}$	0.590	0.498	1.00	0.617
$\boldsymbol{Y_{BW}}$	0.827	0.645	0.617	1.00

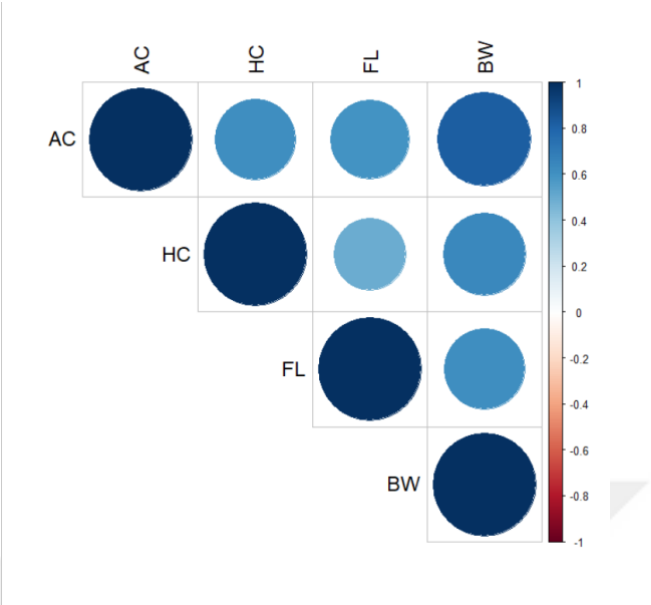


Figure 4.4. Correlation plot of the variables.

Several methods exist that can detect whether correlation is severe enough to affect regression results. Some examples include Farrar chi-square, condition number and sub-setting the data. Farrar Glauber test and condition number analyses both suggested there is problematic collinearity in Hadlock 3, Hammami 5, INTERGROWTH-21 models (Table 4.4). Even in the simple linear model using variables AC, HC and FL had problematic collinearity according to the tests.

Table 4.4 Collinearity diagnostics of fetal weight estimation models

Models	Farrar Chi-square	Condition Number
Hadlock 3	Yes	Yes
Hammami 5	Yes	Yes
INTERGROWTH − 21	Yes	Yes
Simple linear*	Yes	Yes

$$*Y_{birth\ weight} = \theta_0 + \theta_{AC}X_{AC} + \theta_{HC}X_{HC} + \theta_{FL}X_{FL} + \varepsilon$$

The hallmark of an ill-conditioned matrix due to collinearity is the sensitivity of regression coefficients to change in data. This property could be demonstrated either perturbation of the dataset or sub-setting the data. We bootstrapped samples sized 1000 over 100.000 iterations and looked at the change in regression coefficients. The mean regression coefficients of AC, FL and HC were 14.1, 24.9 and 7.5. The regression coefficients of AC, FL and HC ranged from 11.8 to 16.2, 10.6 to 40.9 and 3.9 to 11.2, respectively. There were substantial changes in the regression coefficients within the bootstrapped subsets indicating problematic collinearity even in the most simple regression model (Table 4.5).

Table 4.5 Model coefficients obtained from bootstrapped samples sized 1000 over 100.000 iterations.

Model coefficients*	Mean value	95% confidence intervals	Minimum-maximum
θ_{AC}	14.1	13.1 – 15.1	11.8 – 16.2
θ_{FL}	24.9	18.7 – 31.1	10.6 – 40.9
θ_{HC}	7.5	5.8 – 9.1	3.9 – 11.2

$$Y_{birth\ weight} = X_{AC} + X_{HC} + X_{FL}$$

4.2 Regression parameters with ridge, maximum likelihood and modified maximum likelihood estimators and their variance

We used the training set to estimate parameters with methods described in Chapter 3 (Table 4.6). Standard errors (SE) of the model parameters under SN distribution are obtained with a simple non-parametric bootstrapping procedure. The data is resampled with a sample size equal to the original with replacement of 300 iterations. Parameter estimates are stored in each iteration and standard errors are calculated from the bootstrapped samples. The LS estimate of σ is 300.19 with a SE of 2.87. Bias-corrected LS estimate of σ for GL vastly improve to 175.91 with a SE of 2.14. Ridge estimate of σ is 299.19 with a SE of 2.86. ML and MML estimates of σ under GL (b=1.2) distribution have the lowest value. While the ML and MML estimates for σ under GL(b=1.5) are near to each other. MML estimates have much higher efficiencies with a reduced standard error of 1.19 compared to 2.06. The findings are similar when compared to bias corrected LS estimates with a SE of 2.14 compared to 1.19 for MMLEs. The efficiency gains with MML estimates are 44.39% when compared to bias-corrected LS estimators and 42.2% compared to ML estimates.

Table 4.6 Parameter estimates using, LS, bias-corrected LS, ML with BFGS optimization and MMLE for skewed-normal and generalized logistic distributions.

Estimators (Training data)	$\theta_0\ (SE)$	$\theta_{AC}\ (SE)$	$\theta_{HC}(SE)$	$\theta_{FL}\ (SE)$	$\sigma\ (SE)$
LS	–5892·74	13·84	7·74	26·12	300·19
	(108·12)	(0·22)	(0·38)	(1·39)	(2·87)
LS (bias–corrected)	–5943·43	13·84	7·74	26·12	175·91
	(–)	(0·22)	(0·38)	(1·39)	(2·14)
Ridge ($\kappa = 47$)	–5742·02	12·73	8·31	26·72	299·19
	(108·02)	(0·22)	(0·38)	(1·41)	(2·86)
ML					
– SN ($\lambda = 1 \cdot 95$)	–5696·63	13·91	7·63	23·72	297·19
	(106·99)	(0·22)	(0·37)	(1·37)	(2·95)
– GL (b = 1·2)	–5895·79	14·05	7·79	24·19	175·37
	(103·3)	(0·21)	(0·36)	(1·36)	(1·96)
– GL (b = 1·5)	–5896·28	13·97	7·73	24·08	186·32
	(103·4)	(0·22)	(0·36)	(1·36)	(2·06)
MML					
– SN ($\lambda = 1 \cdot 95$)	–5702·13	14·15	7·65	23·48	349·82
	(64·43)	(0·14)	(0·22)	(0·83)	(2·11)
– GL (b = 1·2)	–5924·58	13·86	7·69	25·88	178·61
	(9·94)	(0·19)	(0·33)	(1·24)	(1·11)
– GL (b = 1·5)	–5926·01	13·78	7·70	25·44	189·64
	(3·93)	(0·22)	(0·38)	(1·40)	(1·19)

4.3 Fetal weight estimation and comparison with existing algorithms

In this section, we use the validation dataset in order to obtain the prediction errors. Predicted birthweights are subtracted from actual birthweight to calculate the residuals. Root mean squared errors are calculated for each estimators and absolute prediction error percentages are presented.

The best estimators are ridge and MML under GL (b=1.2) with root mean squared errors of 2387.32 and 2422.09 versus 2443.13 for Hadlock 3. Both ridge and MML under GL (b=1.2) had significantly reduced absolute prediction error above 20% and between 10 and 20% (Table 4.7). The significance was tested with McNemar's chi-squared test and there was a significant reduction for prediction errors between 10 to 20% (21.6 vs 20.7%) and more than 20% (2.9 vs 2.4%) when MML GL estimators were used instead of Hadlock 3. Overall efficiency gain for ridge and MML under GL (b=1.2) were 1.2% and 1.4% for prediction errors above 10%.

Table 4.7 Absolute prediction error percentages and root mean squared errors of the models for the validation set.

Prediction error %	Hadlock 3	LS (corrected)	Ridge ($\kappa = 47$)	MML SN (λ=1.95)	MML GL (b =1·5)	MML GL (b =1·2)
RMSE	2443·13	2539·16	2387·32	2451·28	2556·04	2422·09
< 1%	486 (8·9)	530 (9·7)	521 (9·5)	506 (9·2)	526 (9·6)	519 (9·5)
1 to 5%	1952 (35·7)	1862 (34·0)	1951 (35·7)	1846 (33·8)	1848† (33·8)	1940 (35·5)
5 to 10%	1684 (30·8)	1711 (31·3)	1721 (31·5)	1710 (31·3)	1735 (31·7)	1744 (31·9)
10 to 20%	1183 (21·6)	1227 (22·4)	1133* (20·7)	1204 (22·0)	1224 (22·4)	1130* (20·7)
> 20%	161 (2·9)	136* (2·5)	140* (2·6)	200* (3·6)	133* (2·4)	133* (2·4)

* P< 0.05 McNemar's chi-squared test when compared to Hadlock 3

Values are given as count and percentage of the total

We also investigate the model fit with Akaike and Bayesian information criterion (Table 4.8). The model estimates obtained from the training set using Ridge, LS and MML under different error distribution are used to calculate the AIC and BIC scores. Hadlock 3 estimates are also used for calculation. The highest AIC and BIC scores are obtained from MML under SN distribution with an AIC score of 86246.3 vs 77918.24 of Hadlock 3; a BIC score of 86266.12 vs 77935.05 of Hadlock 3. The remainder of estimators including Ridge **(**$\kappa = 47$**)**, MML GL (b = 1·2) and MML GL (b = 1·5) have lower AIC and BIC scores compared to Hadlock 3. The lowest AIC and BIC scores were provided by MML under GL (b = 1·5) with an AIC score of 77553.92 vs 77918.24 of Hadlock 3; a BIC score of 77810.62 vs 77935.05 of Hadlock 3. The AIC and BIC results agree with our previous findings that correct model specification with MML improves the model fit compared to Hadlock 3 and other estimation methods such as Ridge.

Table 4.8 Model fit according to Akaike and Bayesian information criterion in the validation set.

Estimator	*AIC*	*BIC*
Hadlock 3	77918·24	77935·05
Ridge ($\kappa = 47$)	77722·65	77742·47
LS	77691·77	77711·58
MML		
– SN ($\lambda = 1 \cdot 95$)	86246·30	86266·12
– GL (b = 1·2)	77810·62	77830·44
– GL (b = 1·5)	77553·92	77573·74

AIC: Akaike Information Criterion, BIC: Bayesian Information Criterion

CHAPTER 5

DISCUSSION

The current work presents estimator properties of LS and ML under skewed error distribution, develops MML estimators for GL and SN error distributions. Furthermore, MML estimator for fetal weight prediction are presented along with real-life performance metrics. MML under GL error distribution has significantly improved RMSE and absolute prediction error percentage compared to Hadlock 3, which is the current standard of fetal weight estimation. Although, the overall margin of improvement is minimal in our work, recent studies which tried to improve upon Hadlock 3 could not come up with a better estimator.

5.1 Contribution of findings to the available literature

Our work adds to the body of literature on fetal weight estimation. The observed performance improvement is significant since no other study to date had better performance compared to Hadlock 3, including contemporary ones that used very large datasets and complex polynomial equations. Furthermore, our study presents a new parameter estimation technique using MML under SN error distributions.

5.2 Limitations of the current study and future research

Our work outlines some of the theoretical problems with fetal weight estimation but not all of them. Fetal weight estimation using ultrasonographic assessment is subject to measurement errors. The role of measurement errors on prediction and parameter estimation performance should be investigated in future studies. Moreover,

regression techniques used in this study assume dependent variables are non-stochastic. In fact, ultrasonographic measurements of AC, FL and HC are stochastic and regression methods using stochastic design elements is highly likely to improve the prediction more than MML under GL error distribution.

5.3 Conclusions

We demonstrate LS and ML estimators are significantly affected by model mispecification and multicollinearity. The investigated solutions include ML and MML estimator under skewed normal and generalized logistic error distributions for model misspecification and ridge estimation for multicollinearity. The real life data of 10932 fetuses is divided 1:1 as training and validation datasets. Parameter are estimated from the training set and tested on the validation set. All hypothesized problems (i.e., error misspecification, collinearity) are shown to exist using real data. Generalized logistic distribution and skewed normal are tried as potential candidates for error terms and generalized logistic provides the best fit. MML estimators under GL error distribution are more efficient compared to LS and ML estimators under normality. Ridge estimators are not more efficient compared to LS or ML estimates but offers lower mean squared error. Ridge and MML estimator under GL offer better prediction performance compared to current standard. This is inline with our hypothesis that fetal weight prediction errrors are strictly non-normal with long tails and slightly skewed. Correct model specification with ML and MML offer more efficient estimates with MML being the most efficient. Furthermore, prediction accuracy is improved with MML estimates. The overall efficiency gain in prediction is between 1 to 2% range and stricly involve the tails , i.e., gross absolute prediction error more than 10%. However, no algorithm up to date was able to outperform Hadlock 3 and these findings are significant with this context.

REFERENCES

Abramowitz, M. (1974). Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables, Dover Publications, Inc.

Akhtar W, Ali A, Aslam M, Saeed F, Salman, Ahmad N. Birth weight estimation--a sonographic model for Pakistani population. JPMA The Journal of the Pakistan Medical Association. 2010;60(7):517-20.

Akkaya, A. D., & Tiku, M. L. (2008). Robust estimation in multiple linear regression model with non-Gaussian noise. Automatica, 44(2), 407-417. doi:https://doi.org/10.1016/j.automatica.2007.06.029

Bailey, D. C. (2017). Not Normal: the uncertainties of scientific measurements. Royal Society open science, 4(1), 160600.

Barnett, V.D., (1966). Evaluation of the maximum likelihood estimator where the likelihood equation has multiple roots. Biometrika 53, 151-165.

Bayrak, Ö. T. and A. D. Akkaya (2011). Autoregressive models with stochastic design variables and nonnormal innovations. Proceedings of the 5th international conference on Applied mathematics, simulation, modelling. Corfu Island, Greece, World Scientific and Engineering Academy and Society (WSEAS): 197–201.

Ben-Haroush A, Melamed N, Mashiach R, Meizner I, Yogev Y. New regression formulas for sonographic weight estimation within 10, 7, and 3 days of delivery. J Ultrasound Med. 2008;27(11):1553-8.

Beta, J., Khan, N., Khalil, A., Fiolna, M., Ramadan, G., & Akolekar, R. (2019). Maternal and neonatal complications of fetal macrosomia: systematic review and meta-analysis. Ultrasound Obstet Gynecol, 54, 308-318.

Crandall, S. and B. Ratra (2015). NON-GAUSSIAN AB ERROR DISTRIBUTIONS OF LMC DISTANCE MODULI MEASUREMENTS. The Astrophysical Journal 815(2): 87.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. The Review of Economics and Statistics, 49(1), 92-107. doi:10.2307/1937887

Hadlock, F. P., Harrist, R. B., & Martinez-Poyer, J. (1991). In utero analysis of fetal growth: a sonographic weight standard. *Radiology, 181*(1), 129-133. doi:10.1148/radiology.181.1.1887021

Hammami, A., Mazer Zumaeta, A., Syngelaki, A., Akolekar, R., & Nicolaides, K. H. (2018). Ultrasonographic estimation of fetal weight: development of new model and assessment of performance of previous models. *Ultrasound Obstet Gynecol, 52*, 35-43.

Halaska MG, Vlk R, Feldmar P, Hrehorcak M, Krcmar M, Mlcochova H, et al. Predicting term birth weight using ultrasound and maternal characteristics. Eur J Obstet Gynecol Reprod Biol. 2006;128(1-2):231-5.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, Berkeley, Calif., University of California Press.

Huber, P. J. (1981). Regression. Robust Statistics**:** 153-198.

Islam, M. Q. and M. L. Tiku (2005). "Multiple Linear Regression Model Under Nonnormality." Communications in Statistics - Theory and Methods **33**(10): 2443-2467.

Kacem, Y., Cannie, M. M., Kadji, C., Dobrescu, O., Lo Zito, L., Ziane, S., Jani, J. C. (2013). Fetal weight estimation: comparison of two-dimensional US and MR imaging assessments. *Radiology, 267*, 902-910.

Kehl S, Schmidt U, Spaich S, Schild RL, Sütterlin M, Siemer J. What are the limits of accuracy in fetal weight estimation with conventional biometry in two-dimensional ultrasound? A novel postpartum study. Ultrasound in Obstetrics & Gynecology. 2012;39(5):543-8.

Mazzone, E., Dall'Asta, A., Kiener, A. J. O., Carpano, M. G., Suprani, A., Ghi, T., & Frusca, T. (2019). Prediction of fetal macrosomia using two-dimensional and three-dimensional ultrasound. *Eur J Obstet Gynecol Reprod Biol, 243*, 26-31.

Puthenpura, S. and N. K. Sinha (1986). "Modified maximum likelihood method for the robust estimation of system parameters from very noisy data." Automatica **22**(2): 231-235.

Read, C. B., & Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*.

Sazak, H. S. (2003). " ESTIMATION AND HYPOTHESIS TESTING IN STOCHASTIC REGRESSION." Available at: https://etd.lib.metu.edu.tr/upload/3/724294/index.pdf

Sazak, H. S., et al. (2006). "Regression Analysis with a Stochastic Design Variable." International Statistical Review **74**(1): 77-88.

Schild RL, Fell K, Fimmers R, Gembruch U, Hansmann M. A new formula for calculating weight in the fetus of ≤ 1600 g. Ultrasound in Obstetrics & Gynecology. 2004;24(7):775-80.

Sovio, U., White, I. R., Dacey, A., Pasupathy, D., & Smith, G. C. S. (2015). Screening for fetal growth restriction with universal third trimester ultrasonography in nulliparous women in the Pregnancy Outcome Prediction (POP) study: a prospective cohort study. *The Lancet, 386*, 2089-2097.

Stirnemann, J., Villar, J., Salomon, L. J., Ohuma, E., Ruyan, P., Altman, D. G., . . . Papageorghiou, A. T. (2017). International estimated fetal weight standards of the INTERGROWTH-21(st) Project. *Ultrasound Obstet Gynecol, 49*, 478-486.

Tiku, M. L. (1967). "Estimating the mean and standard deviation from a censored normal sample." Biometrika **54**(1-2): 155-165.

Tiku, M. L., Tan, W. Y., and Balakrishnan, N. (1986). Robust Inference. Marcel Dekker, New York.

Tiku, M. L. and R. P. Suresh (1992). "A new method of estimation for location and scale parameters." Journal of Statistical Planning and Inference **30**(2): 281-292.

Tiku, M.L. and Akkaya, A.D. (2004) Robust Estimation and Hypothesis Testing. New Age International (P) Publishers, New Delhi, 337.

Vaughan, D. C. (1992). On the Tiku-Suresh method of estimation. Communications in Statistics. Theory and Methods, 21:451–469

APPENDICES

A. Parameter estimation using least square, maximum likelihood under-normality and true error distribution.

```
 set.seed(123)
library(sn)
library(glogis)
library(e1071)
library(bbmle)
x1=rnorm(1000,0,1)
x2=rnorm(1000,0,1)
x3=rnorm(1000,0,1)
b0=1
b1=0
b2=0
b3=0
alpha=-5
alpha=-10
error=rsn(1000,2,1, alpha)
y1sn=b0+x1*b1+x2*b2+x3*b3+error2
mod1=lm(y1sn~x1+x2+x3)
summary(mod1)
mll <- lm_mll(y1sn ~ x1+x2+x3)
summary(mle(mll))
mod3=selm(y1sn~x1+x2+x3)
summary(mod3)
```

B. Parameter estimation and standard errors using maximum likelihood under varying collinearity

```
 LIBRARY(STATS4)
LM_MLL <- FUNCTION(FORMULA, DATA = ENVIRONMENT(FORMULA))
{
  Y <- MODEL.RESPONSE(MODEL.FRAME(FORMULA, DATA))
  X <- MODEL.MATRIX(FORMULA, DATA)
  B0 <- NUMERIC(NCOL(X))
  NAMES(B0) <- COLNAMES(X)
  FUNCTION(B=B0, SIGMA=1)
    -SUM(DNORM(Y, X %*% B, SIGMA, LOG=TRUE))
}
SET.SEED(123)
X1=RNORM(1000,0,1)
X2=X1*RNORM(1000,0.15,0.15)
X=MATRIX(C(X1,X2),NCOL = 2)
SOLVE(T(X)%*%X)
B0=1
B1=0.5
B2=1.2
(B1-0.50168954 )/0.04399872
(B2-1.02201618  )/0.20306875
Y1NORM=X1*B1+X2*B2+ERROR1
MLL <- LM_MLL(Y1NORM ~ X1+X2)
SUMMARY(MLE(MLL, LOWER=C(-INF,-INF, 0.01)))
S
```

```
ET.SEED(123)
X1=RNORM(1000,0,1)
X2=X1*RNORM(1000,0.15,0.05)
X=MATRIX(C(X1,X2),NCOL = 2)
SOLVE(T(X)%*%X)
B0=1
B1=0.5
B2=1.2
(B1-0.55507692 )/0.09735583
(B2-0.66608896)/0.60920645
Y1NORM=X1*B1+X2*B2+ERROR1
MLL <- LM_MLL(Y1NORM ~ X1+X2)
SUMMARY(MLE(MLL, LOWER=C(-INF,-INF, 0.01)))
SET.SEED(123)
X1=RNORM(1000,0,1)
X2=X1*RNORM(1000,0.15,0.01)
X=MATRIX(C(X1,X2),NCOL = 2)
SOLVE(T(X)%*%X)
B0=1
B1=0.5
B2=1.2
(B1-0.65300875 )/0.45908187
(B2-0.01000000)/3.04711089
Y1NORM=X1*B1+X2*B2+ERROR1
MLL <- LM_MLL(Y1NORM ~ X1+X2)
SUMMARY(MLE(MLL, LOWER=C(-INF,-INF, 0.01)))
SET.SEED(123)
X1=RNORM(1000,0,1)
X2=X1*RNORM(1000,0.15,0.001)
X=MATRIX(C(X1,X2),NCOL = 2)
SOLVE(T(X)%*%X)
Y1NORM=X1*B1+X2*B2+ERROR1
MLL <- LM_MLL(Y1NORM ~ X1+X2)
SUMMARY(MLE(MLL, LOWER=C(-INF,-INF, 0.01)))
```

(B1-0.63832176)/4.57477067

(b2-0.10561094)/30.49122867

C. Modified maximum likelihood estimation for skewed normal distribution

```
  n1=10932
locationerror=0
scaleerror=1
shapeerror=1.95
empirical=ppoints(n1)
ti=qsn(p=empirical,
     xi=locationerror,
     omega=scaleerror,
     alpha=shapeerror)
betai=((shapeerror^2*ti*dsn(shapeerror*ti,0,1,shapeerror)*psn(shapeerror*ti,0,1,shapeerror))+
(shapeerror*dsn(shapeerror*ti,0,1,shapeerror)^2))/(psn(shapeerror*ti,0,1,shapeerror)^2)
hist(betai)
plot(betai,type="l")
alfai=(dsn(ti*shapeerror,0,1,shapeerror)/psn(ti*shapeerror,0,1,shapeerror))+betai*ti
plot(betai,type="l")
xi=matrix(c(rep(1,n1),x1,x2,x3),ncol = 4)
yi=matrix(c(y),ncol = 1)
lseestimates=solve(t(xi)%*%xi)%*%t(xi)%*%yi
sqrt(sum((yi-xi%*%lseestimates)^2)/(n1-4))
wi=yi-xi%*%lseestimates
j=matrix(c(w,x1,x2,x3,y),ncol = 5)
orderj=j[order(j[,1]),]
deltai=1+shapeerror*betai
al1=alfai*shapeerror
diagdelta=(diag(deltai))
n1matrix=matrix(rep(1,n1),nrow = 1)
diagalfa=diag(al1)
xi=cbind(rep(1,n1),orderj[,2:4])
yi=orderj[,5]
k=solve(t(xi)%*%diagdelta%*%xi)%*%(t(xi)%*%diagdelta%*%yi)
```

```
d=solve(t(xi)%*%diagdelta%*%xi)%*%(t(xi)%*%diagalfa%*%t(n1matrix))
b=sum(alfai*(yi-k[1,1]-k[2,1]*(xi[,2])-
            k[3,1]*(xi[,3])-
            k[4,1]*(xi[,4])))
c=sum(deltai*(yi-k[1,1]-k[2,1]*(xi[,2])-
             k[3,1]*(xi[,3])-
             k[4,1]*(xi[,4]))^2)
sigmahat=(-b+sqrt(b^2+4*c*n1))/
  (2*sqrt(n1*(n1-4)))
thetahat=k-d*sigmahat

#iteration2


w=yi-xi%*%thetahat
j=matrix(c(w,x1,x2,x3,y),ncol = 5)
orderj=j[order(j[,1]),]
deltai=1+shapeerror*betai
al1=alfai*shapeerror
diagdelta=(diag(deltai))
n1matrix=matrix(rep(1,n1),nrow = 1)
diagalfa=diag(al1)
xi=cbind(rep(1,n1),orderj[,2:4])
yi=orderj[,5]
k=solve(t(xi)%*%diagdelta%*%xi)%*%(t(xi)%*%diagdelta%*%yi)
d=solve(t(xi)%*%diagdelta%*%xi)%*%(t(xi)%*%diagalfa%*%t(n1matrix))

b=sum(alfai*(yi-k[1,1]-k[2,1]*(xi[,2])-
            k[3,1]*(xi[,3])-
            k[4,1]*(xi[,4])))
c=sum(deltai*(yi-k[1,1]-k[2,1]*(xi[,2])-
             k[3,1]*(xi[,3])-
             k[4,1]*(xi[,4]))^2)
sigmahat=(-b+sqrt(b^2+4*c*n1))/
  (2*sqrt(n1*(n1-4)))
thetahat=k-d*sigmahat
```

```
#iteration3
w=yi-xi%*%thetahat
j=matrix(c(w,x1,x2,x3,y),ncol = 5)
orderj=j[order(j[,1]),]
deltai=1+shapeerror*betai
al1=alfai*shapeerror
diagdelta=(diag(deltai))
n1matrix=matrix(rep(1,n1),nrow = 1)
diagalfa=diag(al1)
xi=cbind(rep(1,n1),orderj[,2:4])
yi=orderj[,5]
k=solve(t(xi)%*%diagdelta%*%xi)%*%(t(xi)%*%diagdelta%*%yi)
d=solve(t(xi)%*%diagdelta%*%xi)%*%(t(xi)%*%diagalfa%*%t(n1matrix))
b=sum(alfai*(yi-k[1,1]-k[2,1]*(xi[,2])-
            k[3,1]*(xi[,3])-
            k[4,1]*(xi[,4])))
c=sum(deltai*(yi-k[1,1]-k[2,1]*(xi[,2])-
             k[3,1]*(xi[,3])-
             k[4,1]*(xi[,4]))^2)
sigmahat=(-b+sqrt(b^2+4*c*n1))/
  (2*sqrt(n1*(n1-4)))
thetahat=k-d*sigmahat
```

D. Modified maximum likelihood estimation for generalized logistic distribution

```
n1=length(x1)
ii=ppoints(n1)
shapegl=1.5
tt=rep(na,n1)
for (i in 1:n1) {
  tt[i]=-log((ii[i]^(-1/shapegl))-1)
}
plot(tt,type="l")
var1=exp(tt)
betai=var1/(1+var1)^2
alfai=(1/(1+var1))+betai*tt
deltai=alfai-(1/(shapegl+1))
mod1=lm(y~x1+x2+x3)
w=mod1$residuals
j=matrix(c(w,x1,x2,x3,y),ncol = 5)
orderj=j[order(j[,1]),]
diagbeta=(diag(betai))
diagdelta=(diag(deltai))
n1matrix=matrix(rep(1,n1),nrow = 1)
xi=cbind(rep(1,n1),orderj[,2:4])
yi=orderj[,5]
k=solve(t(xi)%*%diagbeta%*%xi)%*%(t(xi)%*%diagbeta%*%yi)
d=solve(t(xi)%*%diagbeta%*%xi)%*%(t(xi)%*%diagdelta%*%t(n1matrix))
b=sum(deltai*(yi-k[1,1]-k[2,1]*(xi[,2])-
              k[3,1]*(xi[,3])-
              k[4,1]*(xi[,4])))
c=sum(betai*(yi-k[1,1]-k[2,1]*(xi[,2])-
               k[3,1]*(xi[,3])-
               k[4,1]*(xi[,4]))^2)
b1=b*(shapegl+1)
c1=c*(shapegl+1)
```

```
sigmahat=(-b1+sqrt(b1^2+4*c1*n1))/
  (2*sqrt(n1*(n1-4)))

thetahat=k-d*sigmahat
#iter2
w=yi-xi%*%thetahat
j=matrix(c(w,x1,x2,x3,y),ncol = 5)
orderj=j[order(j[,1]),]
diagbeta=(diag(betai))
diagdelta=(diag(deltai))
n1matrix=matrix(rep(1,n1),nrow = 1)
xi=cbind(rep(1,n1),orderj[,2:4])
yi=orderj[,5]

k=solve(t(xi)%*%diagbeta%*%xi)%*%(t(xi)%*%diagbeta%*%yi)
d=solve(t(xi)%*%diagbeta%*%xi)%*%(t(xi)%*%diagdelta%*%t(n1matrix))
b=sum(deltai*(yi-k[1,1]-k[2,1]*(xi[,2])-
              k[3,1]*(xi[,3])-
              k[4,1]*(xi[,4])))
c=sum(betai*(yi-k[1,1]-k[2,1]*(xi[,2])-
             k[3,1]*(xi[,3])-
             k[4,1]*(xi[,4]))^2)
b1=b*(shapegl+1)
c1=c*(shapegl+1)
sigmahat=(-b1+sqrt(b1^2+4*c1*n1))/
  (2*sqrt(n1*(n1-4)))
thetahat=k-d*sigmahat
#iter3
w=yi-xi%*%thetahat
j=matrix(c(w,x1,x2,x3,y),ncol = 5)
orderj=j[order(j[,1]),]
diagbeta=(diag(betai))
diagdelta=(diag(deltai))
n1matrix=matrix(rep(1,n1),nrow = 1)
xi=cbind(rep(1,n1),orderj[,2:4])
yi=orderj[,5]
```

```
k=solve(t(xi)%*%diagbeta%*%xi)%*%(t(xi)%*%diagbeta%*%yi)
d=solve(t(xi)%*%diagbeta%*%xi)%*%(t(xi)%*%diagdelta%*%t(n1matrix))

b=sum(deltai*(yi-k[1,1]-k[2,1]*(xi[,2])-
              k[3,1]*(xi[,3])-
              k[4,1]*(xi[,4])))
c=sum(betai*(yi-k[1,1]-k[2,1]*(xi[,2])-
             k[3,1]*(xi[,3])-
             k[4,1]*(xi[,4]))^2)
b1=b*(shapegl+1)
c1=c*(shapegl+1)
sigmahat=(-b1+sqrt(b1^2+4*c1*n1))/
  (2*sqrt(n1*(n1-4)))
thetahat=k-d*sigmahat
```