NEW BIASED ESTIMATORS TO HANDLE THE PROBLEM OF MULTICOLLINEARITY

by

NG SET FOONG

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

March 2008

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my deep gratitude to those who have contributed throughout the duration of my research work.

First and foremost, I would like to extend my gratitude to my supervisors, Associate Professor Dr. Low Heng Chin and Professor Quah Soon Hoe for their guidance, invaluable advice and continuous concern throughout this thesis.

Sincere thanks to the Dean of the School of Mathematical Sciences and the staff in Universiti Sains Malaysia especially the staff in the School of Mathematical Sciences for their assistance and cooperation.

Finally, I would like to thank all my family members, my colleagues and friends for their support and sharing. Last but not least, my sincere thanks and appreciation to my mother and my husband for their encouragement, understanding and support throughout the period spent in completing this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

		Page
y^*	dependent variable	1
$w_1, w_2, ..., w_p$	independent variables	1
ϕ_j	parameter of a linear regression model with variables y^* and $w_1, w_2, ..., w_p$, $j = 0, 1, 2, ..., p$	1
ε^*	error term of linear regression model with variables y^* and $w_1, w_2, ..., w_p$	1
$\mathbf{Y}^*$	$n \times 1$ vector of the observed random variables	2
$\mathbf{W}$	$n \times (p+1)$ matrix of the known independent variables	2
$\boldsymbol{\varphi}$	$(p+1) \times 1$ vector of parameters of a linear regression model with variables $\mathbf{Y}^*$ and $\mathbf{W}$	2
$\boldsymbol{\varepsilon}^*$	$n \times 1$ vector of errors of a linear regression model with variables $\mathbf{Y}^*$ and $\mathbf{W}$, such that $\boldsymbol{\varepsilon}^* \sim \mathrm{N}(\mathbf{0}, \sigma_*^2 \mathbf{I}_n)$	2
σ_*^2	variance of error term ε^*	2
$\mathbf{I}_n$	identity matrix of dimension $n \times n$	2
$\hat{\boldsymbol{\varphi}}$	the least squares estimator of the parameter $\boldsymbol{\varphi}$	4
$\mathbf{Y}$	$n \times 1$ vector of standardized dependent variables	9
$\mathbf{Z}$	$n \times p$ matrix of standardized independent variables	9
y_i	element of the vector of standardized dependent variable, $\mathbf{Y}$, $i = 1, 2, ..., n$	10
y_i^*	element of the vector $\mathbf{Y}^*$, $i = 1, 2, ..., n$	10
$\overline{y}^*$	mean of the dependent variable y^*	10
z_{ij}	element of the matrix of standardized independent variables, $\mathbf{Z}$, $i = 1, 2, ..., n$, $j = 1, 2, ..., p$	10
w_{ij}	element of the matrix $\mathbf{W}$, $i = 1, 2, ..., n$, $j = 1, 2, ..., p$	10

$\bar{w}_j$	mean of the independent variable w_j, $j = 1, 2, ..., p$	10
y	standardized dependent variable	10
$z_1, z_2, ..., z_p$	standardized independent variables	10
γ_j	parameter of a linear regression model with standardized variables y and $z_1, z_2, ..., z_p$, $j = 1, 2, ..., p$	10
ε	error term of linear regression model with standardized variables y and $z_1, z_2, ..., z_p$	10
$\boldsymbol{\gamma}$	$p \times 1$ vector of parameters of a linear regression model with variables $\mathbf{Y}$ and $\mathbf{Z}$	11
$\boldsymbol{\varepsilon}$	$n \times 1$ vector of errors of a linear regression model with variables $\mathbf{Y}$ and $\mathbf{Z}$, such that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$	11
σ^2	variance of error term ε	11
$\hat{\boldsymbol{\gamma}}$	the least squares estimator of the parameter $\boldsymbol{\gamma}$	11
$\hat{\gamma}_j$	the least squares estimator of the parameter γ_j, $j = 1, 2, ..., p$	12
$\hat{\phi}_j$	the least squares estimator of the parameter ϕ_j, $j = 1, 2, ..., p$	12
$\boldsymbol{\lambda}$	$p \times p$ diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{Z}'\mathbf{Z}$	12
λ_j	j-th eigenvalue of $\mathbf{Z}'\mathbf{Z}$, $j = 1, 2, ..., p$	12
λ_{max}	the largest eigenvalue	12
λ_{min}	the smallest eigenvalue	12
$\mathbf{T}$	$p \times p$ orthonormal matrix consisting of the p eigenvectors of $\mathbf{Z}'\mathbf{Z}$	12
$\mathbf{t}_j$	j-th eigenvector of $\mathbf{Z}'\mathbf{Z}$, $j = 1, 2, ..., p$	12
t_{kj}	k-th element of eigenvector $\mathbf{t}_j$, $j,k = 1, 2, ..., p$	12
$\mathbf{I}$	$p \times p$ identity matrix	12

r_{km}	correlation coefficient between independent variables w_k and w_m, $k, m = 1, 2, ..., p$	15
VIF_j	variance inflation factor for the j-th parameter, $j = 1, 2, ..., p$	16
R_j^2	multiple correlation coefficient of z_j regressed on the remaining independent variables, $j = 1, 2, ..., p$	16
CI_j	j-th condition index of the matrix $\mathbf{Z}$, $j = 1, 2, ..., p$	18
P_{kj}	k, j-th variance decomposition proportion, $j, k = 1, 2, ..., p$	20
$\mathbf{X}$	$n \times p$ matrix where $\mathbf{X} = \mathbf{ZT}$ and $\mathbf{X'X} = \boldsymbol{\lambda}$	28
$\boldsymbol{\beta}$	$p \times 1$ vector of parameters of a linear regression model with variables $\mathbf{Y}$ and $\mathbf{X}$ and $\boldsymbol{\beta} = \mathbf{T'\gamma}$	28
$\hat{\boldsymbol{\beta}}$	the least squares estimator of the parameter $\boldsymbol{\beta}$	28
$\hat{\boldsymbol{\beta}}_r$	Principal Component Regression Estimator of the parameter $\boldsymbol{\beta}$	29
$\hat{\boldsymbol{\gamma}}_r$	Principal Component Regression Estimator of the parameter $\boldsymbol{\gamma}$	29
$\mathbf{T}_r$	$p \times r$ matrix consisting of the remaining eigenvectors of $\mathbf{Z'Z}$ after having deleted $p - r$ of the columns of $\mathbf{T}$	30
$\boldsymbol{\lambda}_r$	$r \times r$ diagonal matrix whose diagonal elements are the remaining eigenvalues of $\mathbf{Z'Z}$ after having deleted $p - r$ of the eigenvalues of $\mathbf{Z'Z}$	30
$\hat{\boldsymbol{\beta}}_s$	Shrunken Estimator	30
$\hat{\boldsymbol{\beta}}_{m,\delta}$	Iteration Estimator	30
$\hat{\boldsymbol{\beta}}_{\mathbf{K}}$	Generalized Ridge Regression Estimator	31
$\mathbf{K}$	a diagonal matrix of biasing factors k_i where $\mathbf{K} = \text{diag}(k_i)$, $k_i > 0, i = 1, 2, ..., p$	31
$\hat{\boldsymbol{\beta}}_k$	Ordinary Ridge Regression Estimator	31

$\tilde{\boldsymbol{\beta}}_{\mathbf{K}}^{*}$	Almost Unbiased Generalized Ridge Regression Estimator	32
$\tilde{\boldsymbol{\beta}}_{k}^{*}$	Almost Unbiased Ridge Regression Estimator	32
$\mathbf{b}(k,\mathbf{b}^{*})$	Modified Ridge Regression Estimator	33
$\boldsymbol{\beta}^{*}(k)$	Restricted Ridge Regression Estimator	33
$\hat{\boldsymbol{\beta}}_{k,d}$	Liu-type Estimator	34
$\hat{\boldsymbol{\beta}}_{r}(k)$	$r-k$ Class Estimator of the parameter $\boldsymbol{\beta}$	35
$\hat{\boldsymbol{\gamma}}_{r}(k)$	$r-k$ Class Estimator of the parameter $\boldsymbol{\gamma}$	35
$\hat{\boldsymbol{\beta}}_{d}$	Liu Estimator	36
$\hat{\boldsymbol{\beta}}_{\mathbf{D}}$	Generalized Liu Estimator	36
$\mathbf{D}$	a diagonal matrix of biasing factors d_i where $\mathbf{D}=\text{diag}(d_i)$, $0<d_i<1$, $i=1,2,...,p$	36
$\tilde{\boldsymbol{\beta}}_{\mathbf{D}}^{*}$	Almost Unbiased Generalized Liu Estimator	37
$\tilde{\boldsymbol{\beta}}_{d}^{*}$	Almost Unbiased Liu Estimator	37
$\hat{\boldsymbol{\beta}}_{rd}$	Restricted Liu Estimator	37
$\hat{\boldsymbol{\beta}}_{r}(d)$	$r-d$ Class Estimator of the parameter $\boldsymbol{\beta}$	38
$\hat{\boldsymbol{\gamma}}_{r}(d)$	$r-d$ Class Estimator of the parameter $\boldsymbol{\gamma}$	38
β_j	j-th element of the vector of parameters $\boldsymbol{\beta}$, $j=1,2,...,p$	52
$\hat{\beta}_j$	the least squares estimator of the parameter β_j, $j=1,2,...,p$	52
$\hat{\boldsymbol{\beta}}_{c}$	the special case of the Liu-type Estimator of the parameter $\boldsymbol{\beta}$	53
$(\hat{\beta}_c)_j$	the special case of the Liu-type Estimator of the parameter β_j, $j=1,2,...,p$	55
$\tilde{\boldsymbol{\beta}}_{c}$	the first new estimator of the parameter $\boldsymbol{\beta}$	62
$(\tilde{\beta}_c)_j$	the first new estimator of the parameter β_j, $j=1,2,...,p$	63

$\hat{\boldsymbol{\gamma}}_c$	the special case of the Liu-type Estimator of the parameter $\boldsymbol{\gamma}$	68
$\hat{\boldsymbol{\gamma}}_r(c)$	*r-c* Class Estimator of the parameter $\boldsymbol{\gamma}$	69
$\mathbf{I}_r$	identity matrix of dimension $r \times r$	69
$\mathbf{T}_{p-r}$	$p \times (p-r)$ matrix consisting of the remaining eigenvectors of $\mathbf{Z}'\mathbf{Z}$ after having deleted first r columns of $\mathbf{T}$	70
$\hat{\sigma}^2$	estimated value of σ^2	114

LIST OF ABBREVIATIONS

		Page
OLSE	Ordinary Least Squares Estimator	4
PCRE	Principal Component Regression Estimator	29
ORRE	Ordinary Ridge Regression Estimator	31
GRRE	Generalized Ridge Regression Estimator	31
AUGRRE	Almost Unbiased Generalized Ridge Regression Estimator	32
AURRE	Almost Unbiased Ridge Regression Estimator	32
MRRE	Modified Ridge Regression Estimator	33
RRRE	Restricted Ridge Regression Estimator	33
GLE	Generalized Liu Estimator	35
AUGLE	Almost Unbiased Generalized Liu Estimator	36
AULE	Almost Unbiased Liu Estimator	37
RLE	Restricted Liu Estimator	37

LIST OF APPENDIX

LIST OF PUBLICATIONS & PRESENTATIONS

1. Ng, S.F., Low, H.C. and Quah, S.H. (2007). Evaluation of a new parameter estimator. *Pertanika Journal of Science and Technology*, *in press*.

2. Ng, S.F., Low, H.C. and Quah, S.H. (2007). An Overview of Biased Estimators. *Journal of Physical Science*, *in press*.

3. Ng, S.F., Low, H.C. and Quah, S.H. (2007). Mean squared error – A tool to evaluate the accuracy of parameter estimators in regression. *Journal of Quality Measurement and Analysis*, *in press*.

4. Ng, S.F., Low, H.C. and Quah, S.H. (2007). A study on the Liu-type estimator. Proceedings of the Third IMT-GT Regional Conference on Mathematics, Statistics and Applications: Strengthening Regional Cooperation through the Mathematical Sciences. Universiti Sains Malaysia, Penang, Dec 5 – 6.

5. Ng, S.F., Low, H.C. and Quah, S.H. (2006). Regression analysis using a biased estimator. Proceedings of the International Conference on Science & Technology-Applications in Industry & Education: Moving towards the Innovative Technology Era. Universiti Teknologi MARA, Penang, Dec. 8 – 9.

6. Ng, S.F. and Low, H.C. (2006). A study on how sum of squares can be used to detect multicollinearity problem. Proceedings of the Second IMT-GT Regional Conference on Mathematics, Statistics and Applications: Strengthening Regional Cooperation through the Mathematical Sciences. Universiti Sains Malaysia, Penang, June 13 – 15.

PENGANGGAR PINCANG BARU UNTUK MENANGANI MASALAH KEKOLINEARAN

ABSTRAK

Analisis regresi merupakan satu kaedah statistik yang sering digunakan dalam bidang ekonomi, teknologi, sains sosial dan kewangan. Model regresi linear menerangkan hubungan antara satu pembolehubah sambutan dengan satu atau lebih pembolehubah tak bersandar. Kekolinearan ditakrifkan sebagai kewujudan hubungan yang hampir linear antara pembolehubah-pembolehubah tak bersandar. Kewujudan kekolinearan yang serius akan mengurangkan kejituan anggaran parameter dalam model regresi linear. Penganggar Kaedah Kuasa Dua Terkecil (Ordinary Least Squares Estimator) adalah satu anggaran saksama yang digunakan untuk menganggar parameter-parameter anu dalam model regresi linear. Nilai varians bagi Penganggar Kaedah Kuasa Dua Terkecil adalah amat besar dalam kes kekolinearan. Oleh itu, penganggar pincang dicadangkan sebagai alternatif bagi Penganggar Kaedah Kuasa Dua Terkecil. Dua penganggar pincang baru serta sifat-sifat penganggar tersebut seperti pincang (bias), varians dan min ralat kuasa dua telah diterbitkan dari teori dalam kajian ini. Dua cara telah digunakan untuk mendapat penganggar-penganggar pincang baru ini. Dengan cara pertama, satu penganggar pincang baru diterbitkan dengan mengurangkan pincang Penganggar Liu-type khas (special case of the Liu-type Estimator). Dengan cara kedua, satu penganggar pincang baru diterbitkan dengan menggabungkan Penganggar Regresi Komponen Prinsipal (Principal Component Regression Estimator) dan Penganggar Liu-type khas. Penganggar pincang baru ini dinamakan '*r-c* Class Estimator'. Prestasi penganggar-penganggar baru ini dinilai dengan

membandingkan min ralat kuasa dua mereka dengan penganggar-penganggar lain. Penganggar pincang baru pertama dibandingkan dengan Penganggar Kaedah Kuasa Dua Terkecil, Penganggar Liu-type khas, Penganggar Regresi Ridge (Ordinary Ridge Regression Estimator) dan Penganggar Liu (Liu Estimator). Penganggar pincang baru kedua iaitu '*r-c* Class Estimator' dibandingkan dengan Penganggar Kaedah Kuasa Dua Terkecil, Penganggar Regresi Komponen Prinsipal dan penganggar Liu-type khas. Teknik yang berlainan telah digunakan untuk membandingkan penganggar-penganggar ini dalam situasi yang berlainan. Perbandingan ini adalah dijalankan dari segi teori. Satu set data juga digunakan untuk menjalankan perbandingan antara penganggar-penganggar tersebut. Perbandingan ini telah menunjukkan bahawa min ralat kuasa dua penganggar-penganggar pincang baru ini adalah lebih kecil daripada min ralat kuasa dua penganggar-penganggar lain di bawah situasi tertentu. Maka, kejituan anggaran parameter dapat dipertingkatkan dengan menggunakan penganggar-penganggar pincang baru ini dalam kekolinearan. Secara tidak langsung, kelemahan kekolinearan dikurangkan. Oleh itu, penganggar-penganggar pincang baru ini adalah lebih berkesan dan boleh dipertimbangkan sebagai anggaran parameter dalam model regresi linear demi menghasilkan model regresi linear yang lebih bagus.

NEW BIASED ESTIMATORS TO HANDLE THE PROBLEM OF MULTICOLLINEARITY

ABSTRACT

Regression analysis is a statistical method widely used in many fields such as economics, technology, social sciences and finance. A linear regression model is constructed to describe the relationship between the dependent variable and one or several independent variables. Multicollinearity is defined as the existence of nearly linear dependency among the independent variables. The presence of serious multicollinearity would reduce the accuracy of the parameter estimate in a linear regression model. The Ordinary Least Squares Estimator is an unbiased estimator that is used to estimate the unknown parameters in the model. The variance of the Ordinary Least Squares Estimator would be very large in the presence of multicollinearity. Therefore, biased estimators are suggested as alternatives to the Ordinary Least Squares Estimator. In this study, two new biased estimators are proposed from theory and their properties such as the bias, variance and mean squared error are derived. Two approaches are used to obtain the new biased estimators. The first approach reduces the bias of the special case of the Liu-type Estimator. Thus, a new estimator is obtained. The second approach combines the Principal Component Regression Estimator and the special case of the Liu-type Estimator. Thus, another new estimator named as the *r-c* Class Estimator, is obtained. The performance of these estimators is evaluated by comparing these estimators with some existing estimators in terms of mean squared error. The first new estimator is compared with the Ordinary Least Squares Estimator, the special case of the Liu-type Estimator, the Ordinary Ridge Regression Estimator

and the Liu Estimator. The second new estimator, that is the *r-c* Class Estimator, is compared with the least squares estimator, the Principal Component Regression Estimator and the special case of the Liu-type Estimator. Different techniques have been established to perform the comparisons depending on different situations. These comparisons between the estimators are done from a theoretical basis. In addition, numerical comparisons between these estimators are also done by using a data set. The comparisons show that these new estimators are superior to other estimators in terms of a reduction in the mean squared error when certain conditions are satisfied. Hence, the accuracy of the parameter estimate increases. Indirectly, the impact of multicollinearity is reduced. Therefore, the proposed new estimators can be considered in the linear regression model in order to obtain a better regression equation.

CHAPTER ONE

INTRODUCTION

1.1 Introduction

Multicollinearity is one of the problems we face in regression analysis. In the presence of multicollinearity, biased estimators have been suggested as alternatives to the least squares estimator to improve the accuracy of the parameter estimates in the linear regression model. In this study, new biased estimators to handle the problem of multicollinearity are proposed. The background of the study, problem identification, objectives of the study, significance of the study, scope of the study and organization of the thesis are presented in this chapter.

1.2 Background of the Study

A linear regression model is constructed to describe the relationship between the dependent variable and the related independent variables. The linear regression model with p independent variables, $w_1, w_2, ..., w_p$, and a dependent variable, y^*, is generally written as

$$y^* = \phi_0 + \phi_1 w_1 + \phi_2 w_2 + ... + \phi_p w_p + \varepsilon^*, \tag{1.1}$$

where ϕ_j, $j = 0, 1, 2, ..., p$, is a parameter and ε^* is the error term.

Suppose there are n observations in the data, the linear regression model can be written in the matrix form

$$\mathbf{Y}^* = \mathbf{W}\boldsymbol{\varphi} + \boldsymbol{\varepsilon}^*, \tag{1.2}$$

where $\mathbf{Y}^*$ is an $n\times 1$ vector of the observed random variables, $\mathbf{W}$ is an $n\times(p+1)$ matrix of the known independent variables, $\boldsymbol{\varphi}$ is a $(p+1)\times 1$ vector of parameters, $\boldsymbol{\varepsilon}^*$ is an $n\times 1$ vector of errors such that $\boldsymbol{\varepsilon}^* \sim \mathrm{N}(\mathbf{0}, \sigma_*^2\mathbf{I}_n)$ and $\mathbf{I}_n$ is an identity matrix of dimension $n\times n$. The matrix $\mathbf{W}$ is given by

$$\mathbf{W} = \begin{pmatrix} 1 & w_{11} & \cdots & w_{1p} \\ 1 & w_{21} & \cdots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & w_{n1} & \cdots & w_{np} \end{pmatrix}.$$

One of the problems in regression analysis is the linear dependencies among the independent variables. This problem is known as multicollinearity. Multicollinearity represents a near exact relationship between two or more variables. Assume that the relationship between $w_1, w_2, ..., w_p$ can be written as

$$a_1 w_1 + a_2 w_2 + ... + a_p w_p \approx k, \tag{1.3}$$

where a_j, $j = 1, 2, ..., p$, and k are constants.

The regressors $w_1, w_2, ..., w_p$ with nonzero constants are multicollinear (Ryan, 1997). Note that the term multicollinearity is used interchangeably with collinearity (Belsley, 1991).

Data with multicollinearity frequently arise and cause problems in many applications of linear regression such as in econometrics, oceanography, geophysics and other fields that rely on nonexperimental data. Multicollinearity is a natural flaw in the data set due to the uncontrollable operations of the data-generating mechanism (Belsley, 1991).

In estimating the parameters in the regression model, it is often stated that multicollinearity can cause the signs of the parameter estimator to be wrong. The presence of multicollinearity will also mislead with the significance test telling us that some important variables are not needed in the model (Belsley, 1991; Rawlings *et al*., 1998). Multicollinearity causes a reduction of statistical power in the ability of statistical tests.

Furthermore, the unique solution for the parameter estimator is very unstable. The parameter estimators would change drastically when small changes occur in the dependent or independent variables (Rawlings *et al.,* 1998). This also relates to the high variances in the parameter estimators. The variances of the parameter estimators for the independent variables involved in multicollinearity would be very large. A consequence of having large variances is that the width of the confidence intervals for the parameters will also be inflated. Therefore, the impact of multicollinearity is serious if the primary interest of a study is in estimating the parameters and identifying the important variables in the process.

1.3 Problem Identification

In order to estimate the unknown parameter, $\boldsymbol{\varphi}$, in the linear regression model (1.2), the method of least squares is used. The least squares estimation procedure uses the criterion that the solution must give the smallest possible sum of squared deviations of the observed dependent variable from the estimates of their true means provided by the solution. The least squares principle chooses an estimator that minimizes the sum of squares of the residuals, that is, $(\boldsymbol{\varepsilon}^*)'\boldsymbol{\varepsilon}^*$ (Rawlings *et al.*, 1998). Let $\hat{\boldsymbol{\varphi}}$ be the least squares estimator of the parameter $\boldsymbol{\varphi}$. The estimator, $\hat{\boldsymbol{\varphi}}$, is given by

$$\hat{\boldsymbol{\varphi}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'(\mathbf{Y}^*). \tag{1.4}$$

The least squares estimator as shown in Equation (1.4) is often called the Ordinary Least Squares Estimator (OLSE) of the parameter $\boldsymbol{\varphi}$ (Belsley, 1991). The Ordinary Least Squares Estimator, $\hat{\boldsymbol{\varphi}}$, is an unbiased estimator of $\boldsymbol{\varphi}$ because the expected value of the estimator $\hat{\boldsymbol{\varphi}}$ is equal to the parameter $\boldsymbol{\varphi}$. The Ordinary Least Squares Estimator is also the best linear unbiased estimator because it has the minimum variance of all possible estimators that are both linear functions of the data and unbiased for the parameter (Rawlings *et al.*, 1998). However, the minimum variance may be unacceptably large in the presence of multicollinearity (Rawlings *et al.*, 1998).

The mean squared error of an estimator is a measure of the goodness of the estimator. The mean squared error is equal to the variance of the estimator plus the square of its bias. The unbiased estimator, Ordinary Least Squares

Estimator, has no bias, but the high variance would cause the mean squared error of the estimator to become very large in the presence of multicollinearity. As a result, the accuracy of the parameter estimate is in question.

The unbiased estimator has no bias from the parameter to be estimated while the biased estimator has a bias from the parameter. Some biased estimators have been suggested as a means to improve the accuracy of the parameter estimate in the model where multicollinearity exists. Although the biased estimator has a certain amount of bias, it is possible for the variance of a biased estimator to be sufficiently smaller than the variance of the unbiased estimator to compensate for the bias introduced. Thus, the biased estimator has a smaller mean squared error and the accuracy of the parameter estimate is improved. The biased estimators that have been proposed are the Ordinary Ridge Regression Estimator (Hoerl and Kennard, 1970a,b), the Liu Estimator (Liu, 1993), the Principal Component Regression Estimator (Massy, 1965; Marquardt, 1970; Hawkins, 1973; Greenberg, 1975) and the Iteration Estimator (Trenkler, 1978).

Some comparisons between the biased estimators have been done. From most of the comparisons between the biased estimators (Trenkler, 1980; Nomura, 1988; Akdeniz and Kaciranlar, 1995; Sakallioglu *et al*., 2001; Akdeniz and Erol, 2003), which estimator is better depends on the unknown parameters and the variance of the error term in the linear regression model as well as the choice of the biasing factors in biased estimators. Therefore, there is still room for

improvement where new biased estimators could be developed in order to provide a better solution.

1.4 Objectives of the Study

The objectives of the study are:

(1) to develop new biased estimators to improve the accuracy of the parameter estimator in the model when multicollinearity exists.

(2) to investigate the performance of the new biased estimators by comparing the new biased estimators with existing estimators in terms of mean squared error.

1.5 Significance of the Study

Linear regression model is widely used in many applications. Multicollinearity is a problem in regression analysis. The variance of the unbiased estimator, Ordinary Least Squares Estimator, is often unacceptably large in the presence of multicollinearity. Thus, the accuracy of the parameter estimator in the model is reduced and the large variance has had a serious impact on the linear regression model. By introducing more efficient new biased estimators to handle the problem of multicollinearity, the impact of multicollinearity would be reduced and hence provide a more meaningful regression model.

1.6 Scope of the Study

In this study, the problem of multicollinearity in the linear regression model is studied. A linear regression model has a linear function of the parameters as

shown in Equation (1.1). The problem of multicollinearity is handled by using biased estimators in regression analysis.

1.7 Organization of Thesis

An overview on the multicollinearity diagnostics and the methods to handle the problem of multicollinearity is given in Chapter 2. The existing biased estimators and the unbiased estimator are also reviewed in detail in Chapter 2. In this study, new biased estimators are developed from theory. The development of these new biased estimators and their properties are presented in Chapter 3. In addition, some comparisons between these new estimators and other estimators are performed in order to investigate the performance of these new biased estimators. The comparisons between the estimators from theory are presented in Chapter 4. A numerical comparison between the estimators is also performed and it is presented in Chapter 5. Chapter 6 gives the summary and conclusions of the study.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this study, the problem of multicollinearity in the linear regression model is studied. A linear regression model that has a linear function of the parameters is defined in this chapter. An overview on the multicollinearity diagnostics is also presented. Several clues that are used to detect the presence of multicollinearity are discussed. Some issues related to multicollinearity diagnostics are also presented. Besides, an overview on the methods to handle the problem of multicollinearity is presented. In particular, some biased estimators have been suggested as a means to improve the accuracy of the parameter estimate in the model when multicollinearity exists. The rationale for using biased estimators instead of unbiased estimator for the model when multicollinearity exists is given. The details of a list of biased estimators reviewed are presented in this chapter.

2.2 Linear Regression Model

A linear regression model with p independent variables, $w_1, w_2, ..., w_p$, and a dependent variable, y^*, is generally written as

$$y^* = \phi_0 + \phi_1 w_1 + \phi_2 w_2 + ... + \phi_p w_p + \varepsilon^*, \tag{2.1}$$

where ϕ_j, $j = 0, 1, 2, ..., p$, is a parameter and ε^* is the error term.

The regression model as shown in Equation (2.1) is a linear regression model because it is a linear function of the parameters. In this study, the problem of multicollinearity in the linear regression model is studied.

Suppose there are n observations in the data, the linear regression model can be written in the matrix form (see page 2 for details)

$$\mathbf{Y}^* = \mathbf{W}\boldsymbol{\varphi} + \boldsymbol{\varepsilon}^*. \tag{2.2}$$

Let $\hat{\boldsymbol{\varphi}}$ be the least squares estimator of the parameter $\boldsymbol{\varphi}$. The estimator, $\hat{\boldsymbol{\varphi}}$, is given by (Belsley, 1991)

$$\hat{\boldsymbol{\varphi}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'(\mathbf{Y}^*). \tag{2.3}$$

The least squares estimator as shown in Equation (2.3) is often called the Ordinary Least Squares Estimator of the parameter $\boldsymbol{\varphi}$.

Suppose standardization is done on the dependent variable and independent variables so that the length of each vector is one. The standardization refers to the centering and scaling process. Let $\mathbf{Y}$ and $\mathbf{Z}$ be the vector of standardized dependent variables and the matrix of standardized independent variables, respectively. The element of the vector of standardized dependent variables, $\mathbf{Y}$, is given by (Ryan, 1997)

$$y_i = \frac{y_i^* - \bar{y}^*}{\sqrt{\sum_{i=1}^{n}(y_i^* - \bar{y}^*)^2}}, \tag{2.4}$$

where y_i, $i = 1, 2, ..., n$, is the element of the vector of standardized dependent variable, $\mathbf{Y}$,

y_i^*, $i = 1, 2, ..., n$, is the element of the vector $\mathbf{Y}^*$, and

$\bar{y}^* = \frac{\sum_{i=1}^{n} y_i^*}{n}$ is the mean of the dependent variable y^*.

The element of the matrix of standardized independent variables, $\mathbf{Z}$, is given by (Ryan, 1997)

$$z_{ij} = \frac{w_{ij} - \bar{w}_j}{\sqrt{\sum_{i=1}^{n}(w_{ij} - \bar{w}_j)^2}} \tag{2.5}$$

where z_{ij}, $i = 1, 2, ..., n$, $j = 1, 2, ..., p$, is the element of the matrix of standardized independent variables, $\mathbf{Z}$,

w_{ij}, $i = 1, 2, ..., n$, $j = 1, 2, ..., p$, is the element of the matrix $\mathbf{W}$, and

$\bar{w}_j = \frac{\sum_{i=1}^{n} w_{ij}}{n}$ is the mean of the independent variable w_j, $j = 1, 2, ..., p$.

Thus, the linear regression model with p standardized independent variables, $z_1, z_2, ..., z_p$, and a standardized dependent variable, y, is generally written as

$$y = \gamma_1 z_1 + \gamma_2 z_2 + ... + \gamma_p z_p + \varepsilon, \tag{2.6}$$

where γ_j, $j = 1, 2, ..., p$, is a parameter and ε is the error term.

When standardization is done on the dependent variable and independent variables, the linear regression model with standardized variables has no intercept because the mean of the standardized dependent variable, y, is equal to zero.

Suppose there are n observations in the data. A linear regression model with standardized variables can be written in the matrix form

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{2.7}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of standardized dependent variables, $\mathbf{Z}$ is an $n \times p$ matrix of standardized independent variables, $\boldsymbol{\gamma}$ is a $p \times 1$ vector of parameters, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors such that $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ and $\mathbf{I}_n$ is an identity matrix of dimension $n \times n$.

Let $\hat{\boldsymbol{\gamma}}$ be the least squares estimator of the parameter $\boldsymbol{\gamma}$. The estimator, $\hat{\boldsymbol{\gamma}}$, is given by (Belsley, 1991)

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}. \tag{2.8}$$

The least squares estimator as shown in Equation (2.8) is often called the Ordinary Least Squares Estimator of the parameter $\boldsymbol{\gamma}$.

Let $\hat{\gamma}_j$ and $\hat{\phi}_j$, $j = 1, 2, ..., p$, be the least squares estimators of the parameters, γ_j and ϕ_j, respectively. The relationship between the least squares estimators, $\hat{\gamma}_j$ and $\hat{\phi}_j$, is given by (Ryan, 1997)

$$\hat{\gamma}_j = \sqrt{\frac{\sum_{i=1}^{n}(w_{ij} - \bar{w}_j)^2}{\sum_{i=1}^{n}(y_i^* - \bar{y}^*)^2}}\ \hat{\phi}_j, \tag{2.9}$$

where w_{ij}, $i = 1, 2, ..., n$, $j = 1, 2, ..., p$, is an element of the matrix $\mathbf{W}$,

$\bar{w}_j = \frac{\sum_{i=1}^{n} w_{ij}}{n}$ is the mean of the independent variable w_j, $j = 1, 2, ..., p$,

y_i^*, $i = 1, 2, ..., n$, is an element of the vector $\mathbf{Y}^*$, and

$\bar{y}^* = \frac{\sum_{i=1}^{n} y_i^*}{n}$ is the mean of the dependent variable y^*.

Let the matrix $\boldsymbol{\lambda}$ be a $p \times p$ diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{Z}'\mathbf{Z}$. The matrix $\boldsymbol{\lambda}$ is given by

$$\boldsymbol{\lambda} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p),$$

where λ_j, $j = 1, 2, ..., p$, is the j-th eigenvalue of $\mathbf{Z}'\mathbf{Z}$ and $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p = \lambda_{\min} > 0$.

Let the matrix $\mathbf{T}$ be a $p \times p$ orthonormal matrix consisting of the p eigenvectors of $\mathbf{Z}'\mathbf{Z}$. The matrix $\mathbf{T}$ is given by

$$\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_p],$$

where $\mathbf{t}_j$, $j = 1, 2, ..., p$, is the *j*-th eigenvector of $\mathbf{Z'Z}$ and $\mathbf{t}_j$ is a $p \times 1$ vector where the *k*-th element of eigenvector $\mathbf{t}_j$ is denoted by t_{kj}, $j, k = 1, 2, ..., p$.

Note that the matrix $\mathbf{T}$ and the matrix $\boldsymbol{\lambda}$ satisfy $\mathbf{T'Z'ZT} = \boldsymbol{\lambda}$ and $\mathbf{T'T} = \mathbf{TT'} = \mathbf{I}$, where $\mathbf{I}$ is a $p \times p$ identity matrix.

The eigenanalysis for the purpose of multicollinearity diagnostics typically is done on $\mathbf{Z'Z}$, where $\mathbf{Z}$ is the matrix of standardized independent variables. The standardization is necessary to prevent the eigenanalysis from being dominated by one or two of the independent variables. If standardization is not performed, the independent variables in their original units of measure would contribute unequally to the eigenvalues (Rawlings *et al.*, 1998).

2.3 Methods for Detecting Multicollinearity

The problem of multicollinearity can be avoided in regression analysis by conducting a designed experiment. The independent variables in a well designed experiment should be uncorrelated. Unfortunately, time and cost constraints may prevent the researchers from collecting data in this manner. Therefore, much of the data collected are observational (Mendenhall and Sincich, 2003). Since observational data frequently consists of correlated independent variables, we should detect the presence of multicollinearity in the data. Then, corrective action should be taken if necessary in order to reduce the impact of multicollinearity on the regression analysis. Therefore, multicollinearity

diagnostics should be carried out in order to detect the presence of multicollinearity.

A review on the multicollinearity diagnostics is presented in Section 2.3.1 while some issues related to multicollinearity diagnostics are presented in Section 2.3.2.

2.3.1 A Review of Multicollinearity Diagnostics

There are several clues that could be used as a guide for multicollinearity diagnostics. The clues that are employed to detect the presence of multicollinearity are as follows:

(a) Wrong sign of the parameter estimate

In the presence of multicollinearity in the data, the sign of the parameter estimates can differ from the sign of the parameters based on the basic prior information about the true parameters (Belsley, 1991; Rawlings *et al.*, 1998).

(b) Important variables appearing as unimportant from the significance test

The variances of the parameter estimates involved in the multicollinearity become very large. Thus, some known important variables have low *t*-statistics and hence the variables appear as unimportant from the significance test. These known important variables in the model are replaced with incidental variables that are involved in the multicollinearity (Belsley, 1991; Rawlings *et al.*, 1998).

(c) High correlations between the independent variables which are shown in the correlation matrix

The correlation matrix is simply the matrix $\mathbf{Z'Z}$ (Belsley, 1991). The matrix $\mathbf{Z'Z}$ is a square symmetric matrix where each element in the matrix is the correlation coefficient between the independent variables. The correlation matrix is given by

$$\mathbf{Z'Z} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix},$$

where $r_{km} = \dfrac{\sum_{i=1}^{n}(w_{ik} - \bar{w}_k)(w_{im} - \bar{w}_m)}{\sqrt{\sum_{i=1}^{n}(w_{ik} - \bar{w}_k)^2}\sqrt{\sum_{i=1}^{n}(w_{im} - \bar{w}_m)^2}}$, $k, m = 1, 2, ..., p$, represents the correlation coefficient between independent variables w_k and w_m,

$\bar{w}_k = \dfrac{\sum_{i=1}^{n} w_{ik}}{n}$, $\bar{w}_m = \dfrac{\sum_{i=1}^{n} w_{im}}{n}$ are the means of the independent variables w_k and w_m, respectively, $k, m = 1, 2, ..., p$, and

w_{ik}, w_{im}, $i = 1, 2, ..., n$, $k, m = 1, 2, ..., p$, are the elements of the matrix $\mathbf{W}$.

High correlation coefficient values between the independent variables in the correlation matrix indicate multicollinearity in the data. However, this clue is suitable for the linear regression model which consists of only two independent variables. In the case where the linear regression model consists of more than two independent variables, looking at the value of the correlation coefficient in the correlation matrix would not be sufficient because large values of the

correlation coefficients between independent variables will only identify multicollinearity involving two variables but may miss those involving more than two variables. Thus, the correlation matrix is unable to reveal the presence or number of several coexisting collinear relations (Belsley, 1991).

(d) Large value of the variance inflation factor

The variance inflation factor for the *j*-th parameter, VIF_j, is the *j*-th diagonal element of matrix $(\mathbf{Z}'\mathbf{Z})^{-1}$, where $j = 1, 2, ..., p$ and $\mathbf{Z}$ is the matrix of standardized independent variables. The term variance inflation factor comes from the fact that the variance of the *j*-th parameter estimate is directly proportional to VIF_j (Belsley *et al.*, 1980; Ryan, 1997).

A large value of variance inflation factor is another clue for detecting the problem of multicollinearity. A value of $VIF_j > 10$ is a guideline for serious multicollinearity (Rawlings *et al.*, 1998). The presence of multicollinearity would result in having inflated variances of the parameter estimates. Thus, the width of the confidence intervals for the parameters will also be inflated, perhaps even to the point of rendering one or more intervals useless (Ryan, 1997).

A large value of VIF_j indicates there is multicollinearity involving z_j and the other independent variables. This is due to the fact that $VIF_j = \frac{1}{1 - R_j^2}$, where R_j^2 is the multiple correlation coefficient of z_j regressed on the remaining

independent variables. A high VIF_j indicates an R_j^2 near unity (Farrar and Glauber, 1967; Belsley, 1991).

The variance inflation factor has the following weaknesses: like any correlation based measure, large variance inflation factors are sufficient to collinearity but not necessary to it; variance inflation factors do not reveal the number of near dependencies that are involved in the multicollinearity (Belsley, 1991).

A procedure related to the variance inflation factor has been proposed, i.e. collinearity indices. The collinearity indices are the square roots of the variance inflation factors. The collinearity index measures the closeness of the regression matrix to one, that is, exactly collinear (Steward, 1987; Belsley, 1991).

(e) High condition index

Let the matrix $\boldsymbol{\lambda} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ be a $p \times p$ diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{Z}'\mathbf{Z}$. Let the matrix $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_p]$ be a $p \times p$ orthonormal matrix consisting of the p eigenvectors of $\mathbf{Z}'\mathbf{Z}$. Note that the matrix $\mathbf{T}$ and the matrix $\boldsymbol{\lambda}$ satisfy $\mathbf{T}'\mathbf{Z}'\mathbf{Z}\mathbf{T} = \boldsymbol{\lambda}$ and $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}$, where $\mathbf{I}$ is a $p \times p$ identity matrix.

The condition number of the matrix $\mathbf{Z}$ is defined as

$$\text{condition number} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}. \qquad (2.10)$$

Note that the condition number ≥ 1, and λ_{max} and λ_{min} refer to the largest eigenvalue and smallest eigenvalue, respectively (Rawlings *et al.*, 1998).

A large condition number indicates the presence of multicollinearity. When multicollinearity does not exist in the data, the condition number is 1. That is, the condition number of a matrix is unity when all the columns are pairwise orthogonal and scaled to have unit length. Furthermore, all eigenvalues are also equal to 1 when multicollinearity does not exist in the data.

Extending the concept of condition number, the *j*-th condition index of the matrix $\mathbf{Z}$, CI_j, is defined as

$$CI_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}, \qquad (2.11)$$

where $CI_j \geq 1$, λ_{max} and λ_j refer to the largest eigenvalue and the *j*-th eigenvalue, respectively, $j = 1, 2, ..., p$ (Rawlings *et al.*, 1998).

The largest condition index is also the condition number. Condition indices around 10 indicate weak dependencies among the independent variables. Condition indices between 30 and 100 indicate moderate to strong dependencies and condition indices larger than 100 indicate serious multicollinearity. The number of condition indices greater than 30 represents the number of near-dependencies contributing to the problem of multicollinearity (Belsley *et al.*, 1980; Rawlings *et al.*, 1998).

(f) High variance-decomposition proportions

The presence of multicollinearity would result in large variance inflation factors and inflated variances of the parameter estimates. A more useful value to identify the proportion of variance of the parameter estimator that results from multicollinearity is the variance decomposition proportion.

Let $\hat{\boldsymbol{\gamma}}$ be the least squares estimator of the parameter $\boldsymbol{\gamma}$. The estimator, $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$, is given by Equation (2.8).

The variance-covariance matrix of $\hat{\boldsymbol{\gamma}}$ is given by

$$\text{Var}(\hat{\boldsymbol{\gamma}}) = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}. \tag{2.12}$$

Let $\hat{\gamma}_k$, $k = 1, 2, ..., p$, be the least squares estimator of the parameter γ_k. Thus, the variance of $\hat{\gamma}_k$, $k = 1, 2, ..., p$, is given by

$$\text{var}(\hat{\gamma}_k) = \sigma^2 \sum_{j=1}^{p} \frac{t_{kj}^2}{\lambda_j}, \tag{2.13}$$

where λ_j, $j = 1, 2, ..., p$, is the j-th eigenvalue of $\mathbf{Z}'\mathbf{Z}$, and t_{kj}, $j, k = 1, 2, ..., p$, is the k-th element of eigenvector $\mathbf{t}_j$.

Note that Equation (2.13) decomposes $\text{var}(\hat{\gamma}_k)$ into a sum of components, each associated with one of the eigenvalues λ_j. Thus, the k, j-th variance decomposition proportion is defined as the proportion of the variance of the k-th

parameter estimate associated with the λ_j. The k, j-th variance decomposition proportion, P_{kj}, $j,k = 1, 2, ..., p$, is given by (Belsley *et al.*, 1980)

$$P_{kj} = \frac{t_{kj}^2 / \lambda_j}{\sum_{j=1}^{p} t_{kj}^2 / \lambda_j}, \tag{2.14}$$

where λ_j, $j = 1, 2, ..., p$, is the j-th eigenvalue of $\mathbf{Z}'\mathbf{Z}$, and t_{kj}, $j,k = 1, 2, ..., p$, is the k-th element of eigenvector $\mathbf{t}_j$.

An interpretation of the variance decomposition proportions requires the following two conditions for the result to be an indication of serious multicollinearity (Belsley *et al.*, 1980):

1. The condition index CI_j is greater than 30.
2. High variance decomposition proportions for two or more variances of the parameter estimator, that is, $P_{kj} > 0.5$ for two or more $\text{var}(\hat{\gamma}_k)$, where $k = 1, 2, ..., p$.

2.3.2 Issues Related to Multicollinearity Diagnostics

Several procedures for multicollinearity diagnostics that have been proposed in the literature focus on the linear regression model (Lee and Weissfeld, 1996). A linear regression model has a linear function of the parameters while a nonlinear regression model has a nonlinear function of the parameters. However, many models developed from principles of behaviour of the system are nonlinear in the parameters. There are some recent researches focusing on

multicollinearity diagnostics for nonlinear regression (Weissfeld, 1989; Lee and Weissfeld, 1996; Weissfeld and Sereika, 1991). In addition, Belsley (1991) noted that not all collinearity is harmful. A procedure for determining a harmful collinearity is proposed by Belsley (1991). The details of some issues related to multicollinearity diagnostics are as follows:

(i) Multicollinearity Diagnostics for Nonlinear Regression

Belsley *et al.* (1980) developed a method for detecting multicollinearity in a linear regression model based on a set of condition indices and variance decomposition proportions. Based on Belsley *et al.* (1980), a number of multicollinearity diagnostics were proposed for nonlinear regression models such as the parametric censored data models (Weissfeld, 1989), the Cox model with time dependent covariates (Lee and Weissfeld, 1996) and the generalized linear models, particularly, the binary logistic and proportional odds regression models (Weissfeld and Sereika, 1991).

(ii) Procedure for Determining Harmful Collinearity

Belsley (1991) pointed out that diagnosing the presence of collinear relations is one thing; determining whether those collinear relations are causing statistical harm to a regression analysis is another. Thus, a procedure for determining harmful collinearity is proposed by applying the multicollinearity diagnostics followed by a test for adequate signal-to-noise. To determine whether any particular variance of the parameter estimate is large or small is necessarily relative. The signal-to-noise gives the ratio of the parameter estimate to the variance of the parameter estimate. Hence, if the variance of the parameter

estimate were large relative to the parameter estimate, the situation would be harmful. Harmful collinearity is defined for the joint occurrence of multicollinearity and adequate signal-to-noise.

2.4 Methods for Combating Multicollinearity

When serious multicollinearity is detected in the data, some corrective actions should be taken in order to reduce its impact. The remedies for the problem of multicollinearity depend on the objective of the regression analysis. Multicollinearity causes no serious problem if the objective is prediction. However, multicollinearity is a problem when our primary interest is in the estimation of parameters (Rawlings *et al.*, 1998). The variances of parameter estimates, when multicollinearity exists, can become very large. Hence, the accuracy of the parameter estimates is reduced.

One suggestion that has been frequently made in trying to overcome the problem of multicollinearity is to collect new data (Ryan, 1997). Sometimes, the problem of multicollinearity occurs due to inadequate or erroneous data. Unfortunately, this is not always possible since some analysis must be based on the available data. Furthermore, this solution is not possible when the presence of multicollinearity is the result of internal constraints of the system being studied (Rawlings *et al.*, 1998).

Another obvious solution is to eliminate the regressors that are causing the multicollinearity. However, selecting regressors to delete for the purpose of removing or reducing multicollinearity is not a safe strategy. Even with extensive

examination of different subsets of the available regressors, one might still select a subset of regressors that is far from optimal. This is because a small amount of sampling variability in the regressors or the dependent variable in a multicollinear data can result in a different subset being selected (Ryan, 1997).

An alternative to regressor deletion is to retain all of the regressors, but to use a biased estimator instead of a least squares estimator in the regression analysis. The least squares estimator is an unbiased estimator that is frequently used in regression analysis. When the primary interest of the regression analysis is in parameter estimation, some biased estimators have been suggested as a means for improving the accuracy of the parameter estimate in the model when multicollinearity exists.

An introduction of the unbiased estimator is presented in Section 2.4.1 while an overview of biased estimators is presented in Section 2.4.2. Some hybrids of the biased estimators are presented in Section 2.4.3. A review on the comparisons between the biased estimators is presented in Section 2.4.4.

2.4.1 Unbiased Estimator

The Ordinary Least Squares Estimator, $\hat{\boldsymbol{\gamma}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$ (Equation (2.8)), is an unbiased estimator of $\boldsymbol{\gamma}$ because the expected value of $\hat{\boldsymbol{\gamma}}$ is equal to $\boldsymbol{\gamma}$, that is,

$$E(\hat{\boldsymbol{\gamma}}) = \boldsymbol{\gamma}. \qquad (2.15)$$

The least squares estimator $\hat{\boldsymbol{\gamma}}$ is the best linear unbiased estimator of the parameter, $\boldsymbol{\gamma}$. The least squares estimator has the smallest variance of all possible estimators that are both linear functions of the data and unbiased for the parameter. The minimum variance of the least squares estimator may be very large in the presence of multicollinearity. A number of biased estimators have been recommended to estimate the parameter when multicollinearity is detected in the data. Although the biased estimators have a certain amount of bias, they may be preferable to the least squares estimator in terms of a reduction in variance. The rationale for using a biased estimator instead of the least squares estimator is further explained by Rawlings *et al.* (1998):

> "Relaxing the least squares condition that estimators be unbiased opens for consideration a much larger set of possible estimators from which one with better properties in the presence of collinearity might be found. Biased regression refers to this class of regression methods in which unbiasedness is no longer required. Such methods have been suggested as a possible solution to the collinearity problem. The motivation for biased regression methods rests in the potential for obtaining estimators that are closer, on average, to the parameter being estimated than are the least squares estimators."

2.4.2 Biased Estimators

Instead of using the least squares estimator, biased estimators are considered in the regression analysis in the presence of multicollinearity. When the expected value of the estimator is equal to the parameter which is supposed to