

**PRINCIPAL COMPONENTS
IN THE PROBLEM OF MULTICOLLINEARITY**

109572

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of
Dokuz Eylül University
in partial Fulfillment of the Requirements for
the Degree of Master of Science in Statistics

T.C. YÜKSEKÖĞRETİM KURULU
DOKÜMANTASYON MERKEZİ

by

Neslihan ORTABAŞ

December, 2001

İZMİR

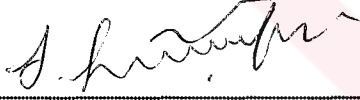
Ms.Sc. THESIS EXAMINATION RESULT FORM

We certify that we have read the thesis, entitled “**PRINCIPAL COMPONENTS IN THE PROBLEM OF MULTICOLLINEARITY**” completed by Neslihan ORTABAŞ under supervision of Prof. Dr. Serdar KURT and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Prof. Dr. Serdar KURT

Supervisor



Prof. Dr. Şevkinaz GÜMÜŞOĞLU

Committee Member



Prof. Dr. Nilgün MORALI

Committee Member

Approved by the
Graduate School of Natural and Applied Sciences



Prof. Dr. Cahit Helyacı

Director

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to my supervisor Prof. Dr. Serdar KURT for his guidance throughout course of this work.

I am also grateful to Alper VAHAPLAR and Özlem EGE for all their assistance.

Finally, I wish to express my deepest gratitude to my family and my friends for their encouragement.

Neslihan ORTABAŞ

ABSTRACT

In this study, principal components regression and ridge regression are examined among the methods used to remedy multicollinearity problem in multiple linear regression model.

One of the assumptions in multiple linear regression is that there must be no perfect linear relations among the regressors. The relationship among the regressors is called multicollinearity. In case of multicollinearity, parameter estimations by least square method have large variances and hypothesis tests result in contradictory. There are various methods for dealing with multicollinearity problem. Biased regression methods (*BRM*) are the ones that can explain the structure of multicollinearity and provide small standard errors among the methods used.

In this study two of biased regression methods; principal components regression and ridge regression are examined as theoretically and researched which methods give the best consequence by simulation.

In the application, 50 repetitions have been generated for each of the sample sizes of 40, 80 and 120. Least squares, ridge and principal components regression are used for each sample. Regression coefficients for each estimator were computed and the mean and the standard deviation of the estimates were used as statistical comparison criteria. According to comparisons among the estimators the principal components regression has been found to provide better estimates.

ÖZET

Bu çalışmada, çoklu doğrusal regresyon modelinde, çoklu doğrusal bağlantı sorununu ortadan kaldırmak için kullanılan yöntemlerden, temel bileşenler regresyon ve ridge regresyon incelenmiştir.

Çoklu doğrusal regresyon modelinin varsayımlarından biri de bağımsız değişkenler arasında tam ilişki olmamasıdır. Bağımsız değişkenler arasında önemli derecede ilişki olması, çoklu doğrusal bağlantı olarak adlandırılır. Çoklu doğrusal bağlantı olması durumunda uygulanan en küçük kareler yöntemi ile parametre tahminleri büyük standart hatalara sahip olmakta ve hipotez testleri çelişkili sonuçlar vermektedir. Bu sorunu ortadan kaldırmak için kullanılan çeşitli yöntemler vardır. Kullanılan yöntemlerden yanlı regresyon yöntemleri, hem çoklu doğrusal bağlantı yapısının açıklanabildiği hem de standart hatası daha küçük hata kareler ortalamalı tahminlerin bulunabildiği yöntemlerdir.

Çalışmada, yanlı regresyon yöntemlerinden temel bileşenler regresyon ile ridge regresyon kuramsal açıdan incelenmiş, benzetim çalışması ile hangi yöntemin daha iyi sonuç verdiği araştırılmıştır.

Benzetim çalışmasında, genişlikleri 40, 80 ve 120 olan örneklemelerin her birisi için 50 tekrar yapılmış ve bu örneklemelere en küçük kareler, ridge ve temel bileşenler regresyon uygulanarak regresyon katsayılarının tahminleri hesaplanmıştır. Tahmin ediciler arasında yapılan karşılaştırmalarda kriter olarak tahminlerin ortalaması ve tahminlerin standart hatası dikkate alınmıştır. Yapılan karşılaştırmalara göre, temel bileşenler regresyon yönteminin diğerlerinden daha iyi sonuçlar verdiği gözlemlenmiştir.

CONTENTS

	Page
Contents	vii
List of Tables	x
List of Figures	xi

Chapter One INTRODUCTION

1.1 Introduction	1
------------------------	---

Chapter Two REGRESSION ANALYSIS AND MULTICOLLINEARITY PROBLEM

2.1 Introduction	3
2.2 The Multiple Linear Regression Model	4
2.3 Sources of Multicollinearity	5
2.4 Practical Consequences of Multicollinearity	6
2.4.1 Large Variance and Covariance of OLS Estimators	7
2.4.2 Wider Confidence Intervals	9
2.4.3 “Insignificant” t Ratios	9
2.4.4 A High R^2 but Few Significant t Ratios	9
2.4.5 Sensitivity of OLS Estimators and Their Standard Errors to Small Changes in Data	9

	Page
2.5 Detection of Multicollinearity	10
2.5.1 High R^2 but Few Significant t Ratios	11
2.5.2 High Pair-wise Correlations among Predictors	11
2.5.3 Examination of Partial Correlation	11
2.5.4 Auxiliary Regressions	12
2.5.5 Eigenvalues and Condition Index	13
2.5.6 Tolerance and Variance Inflation Factor	14
2.6 Methods for Dealing with Multicollinearity	15

Chapter Three

RIDGE REGRESSION AND PRINCIPAL COMPONENTS REGRESSION

3.1 Introduction	16
3.2 Ridge Regression	17
3.2.1 Ridge Trace	19
3.2.2 Variance Inflation Factor	20
3.2.3 Hoerl, Kennard and Baldwin Method	21
3.3 Principal Components Regression	22
3.4 Comparison and Evaluation of Biased Estimators	25

Chapter Four

COMPARISON OF PRINCIPAL COMPONENTS REGRESSION WITH RIDGE REGRESSION AND LEAST SQUARES REGRESSION BY SIMULATION

4.1 Introduction	26
4.2 Generating the Population	27

	Page
4.3 An Application	28
4.4 Least Squares, Ridge and Principal Components Regression Results	34
4.4.1 Simulation Results and Interpretation for Sample Size 40	34
4.4.2 Simulation Results and Interpretation for Sample Size 80	37
4.4.3 Simulation Results and Interpretation for Sample Size 120	40

Chapter Five

CONCLUSION

5.1 Conclusions	43
References	45



LIST OF TABLES

	Page
Table 4.1 Generated Population	27
Table 4.2 Correlation Coefficients Between Variables	28
Table 4.3 Least Squares Regression Coefficients for an Application Data	30
Table 4.4 Summary Statistics of the Ridge Regression for an Application Data ..	31
Table 4.5 Summary Statistics for Principal Components Regression for an Application Data	32
Table 4.6 Comparison Least Squares, Ridge and Principal Components Regression Coefficients for an Application Data	33
Table 4.7 Least Squares Regression Coefficients for $n=40$	34
Table 4.8 Summary Statistics for the Ridge Regression when $n=40$	35
Table 4.9 Summary Statistics for Principal Components Regression when $n=40$	36
Table 4.10 Comparison Least Squares, Ridge and Principal Components Regression Coefficients when $n=40$	36
Table 4.11 Least Squares Regression Coefficients for $n=80$	37
Table 4.12 Summary Statistics for the Ridge Regression when $n=80$	38
Table 4.13 Summary Statistics for Principal Components Regression when $n=80$	39
Table 4.14 Comparison Least Squares, Ridge and Principal Components Regression Coefficients when $n=80$	39
Table 4.15 Least Squares Regression Coefficients for $n=120$	40
Table 4.16 Summary Statistics for the Ridge Regression when $n=120$	41
Table 4.17 Summary Statistics for Principal Components Regression when $n=120$	41
Table 4.18 Comparison Least Squares, Ridge and Principal Components Regression Coefficients when $n=120$	42

LIST OF FIGURES

	Page
Figure 3.1.a Sampling Distributions of Unbiased Estimators β	17
Figure 3.1.b Sampling Distributions of Biased Estimators β	17
Figure 3.2 Ridge Trace	20
Figure 4.1 Normal Probability Plot of Y Values for $n=40$	30



CHAPTER ONE

INTRODUCTION

1.1 Introduction

The term regression was introduced by Francis Galton. In a famous paper, Galton found that, although there was tendency for tall parents to have tall children and for short parents to have short children, the average height of children born of parents of a given height tended to move or “regress” toward the average height in the population as whole. Galton’s law of universal regression was confirmed by his friend Karl Pearson, who collected more than a thousand records of heights of members of family groups. He found that the average height of sons of a group of tall fathers was less than their fathers height and the average height of sons of a group of short fathers was greater than their fathers’ height, thus “regressing” tall and short sons alike toward the average height of all men. In the words of Galton, this was “regression to mediocrity.” (Gujarati,1995)

The modern mean of regression analysis is concerned with study of the one response variable, on one or more predictor variables, for the purpose of constructing models for predicting the population mean or making other inferences. In order to reach these purposes, it obtains some assumptions. One of these assumptions which often appears as a problem, is predictor variables having relationships with each other. This is called multicollinearity. Various methods are used for solving this problem. Two of these methods principal components regression and ridge regression; are discussed in this study. These methods are known as biased regression methods (BRM). They both explain the structure of multicollinearity and provide small mean square error. Principal components regression is a method uses

vertical transformation variables instead of original variables and ridge regression that is a method for minimizing the mean square error by adding a constant to the diagonal of the correlation matrix.

The purpose of this study is examining principal components regression (*PCR*) and ridge regression (*RR*) and researching which methods give the best consequence by simulation.

This thesis contains five chapters. In chapter one, a short description of the entire study is summarized. In chapter two, introduction to regression analysis and multicollinearity problem are mentioned. In chapter three, ridge regression and principal components regression are discussed. In chapter four, comparison of these methods using simulation and the solution is presented in tables. In chapter five, the conclusions are presented.

CHAPTER TWO

REGRESSION ANALYSIS AND MULTICOLLINEARITY PROBLEM

2.1 Introduction

The subject of regression analysis concerns the study of relationships among variables, for the purpose of constructing models for prediction and making other inferences. It treats two-variable (bivariate) or several variable (multivariate) data.

To obtain a useful prediction model, one should record the observations of all variables that may significantly response. These other variables may than be incorporated explicitly into the regression analysis. The name multiple regression refers to a model of relationship where the response depends on two or more predictor variables.

A response variable Y may depend on a predictor variable X but, after a straight-line fit it may turn out that the unexplained data variation is large so R^2 is small and a poor fit is indicated. At the same time, an attempt to transform one or both of the variables may fail to dramatically improve the value of R^2 . This difficulty may well be due to the fact that the response depends not just on X but on the other factors as well. When used alone, X fails to be a good predictor of Y because of the effects of those other influencing variables.

The linear multiple regression model may be written as

$$\begin{aligned}
 Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i \\
 &= \beta_0 + \sum_{j=1}^k \beta_j X_{ij} + \varepsilon_i \quad i = 1, \dots, n \quad j = 1, \dots, k
 \end{aligned} \tag{2.1}$$

where the subscript i denotes the observational unit from which the observations on Y and the k predictor variables were taken. The second subscript designates the predictor. The sample size will be denoted with n , $i=1, \dots, n$ and k will denote the number of predictors. There are $(k+1)$ parameters, when the linear model includes the intercept β_0 . For convenience, we will use $p=k+1$. The model may be more conveniently stated using matrix notation:

$$Y = X\beta + \varepsilon \tag{2.2}$$

where

Y : $n \times 1$ vector of observations on the response variable, Y_i

X : $n \times p$ matrix consisting of a column of ones, which is labeled 1 followed by the k column vectors of the observations on the predictor variables

β : $p \times 1$ vector of regression coefficients to be estimated,

ε : $n \times 1$ vector of error terms.

Note that the term “linear” refers to the fact that the model is linear in β and ε .

2.2 The Multiple Linear Regression Model

Multiple linear regression model makes several assumptions. One of these assumptions is “There is no perfect multicollinearity. That is, there are no perfect relationships among the predictors”. It states that X matrix has full columns rank equal to p , the number of columns of the X matrix are linearly independent; that is there is no exact linear relationship among the X variables. In other words there is no

multicollinearity. In scalar notation this is equivalent to saying that there exists no set of numbers $\lambda_0, \lambda_1, \lambda_2, \dots, \lambda_k$ not all zero such that

$$\lambda_0 X_{0i} + \lambda_1 X_{1i} + \dots + \lambda_k X_{ki} = 0 \quad (2.3)$$

where $X_{0i} = 1$ for all i (to allow for the column of 1's in the X matrix.) In matrix notation (2.3) can be represented as

$$\lambda'X = 0 \quad (2.4)$$

where λ' is a $1 \times p$ row vector and X is a $p \times 1$ column vector.

If an exact linear relationship such as (2.3) exists the variables are said to be collinear. If, on the other hand (2.3) holds true only if

$$\lambda_0 = \lambda_1 = \lambda_2 = \dots = 0 \quad (2.5)$$

then the X variables are said to be linearly independent.

2.3 Sources Of Multicollinearity

There are four primary sources of multicollinearity. It may be due to the following factors;

1. The data collection method employed; for example, sampling over a limited range of the values taken by the predictors in the population.
2. Constraints on the model or in the population being sampled. For instance, in the regression of electricity consumption on income (X_2) and house size (X_3) there is a physical constraint in the population in that families with higher incomes generally have larger homes than families with lower incomes.

3. Model specification; for example, adding polynomial terms to a regression model, especially when the range of the X variable is small.
4. An overdetermined model. This happens when the model has more predictors than the number of observations. This could happen in medical research where there may be a small number of patients about whom information is collected on a large number of variables.

2.4 Practical Consequences Of Multicollinearity

If multicollinearity is high (perfect), the regression coefficients of the X variables are indeterminate and their standard errors are infinite. If multicollinearity is near (less than perfect), the regression coefficients although determinate, possess large standard errors (in relation to the coefficients themselves), which means the coefficients cannot be estimated with great precision or accuracy.

In cases of near or high multicollinearity, one is likely to encounter the following consequences: (Gujarati, 1995, pp.327-332)

1. Although BLUE (Best Linear Unbiased Estimate), the OLS (Ordinary Least Squares) estimators have large variances and covariances, making precise estimation difficult.
2. Because of consequence 1, the confidence intervals tend to be much wider, leading to the acceptance of the “zero null hypothesis” (i.e. the true population coefficient is zero) more readily.
3. Also because of consequence 1, the t ratio of one or more coefficients tends to be statistically insignificant.

4. Although the t ratio of one or more coefficients is statistically insignificant, R^2 , the overall measure of goodness of fit, can be very high.
5. The OLS estimators and their standard errors can be sensitive to small changes in the data.

2.4.1 Large Variances and Covariances of OLS Estimators

Suppose that there are only two predictor variables, X_1 and X_2 . The model assuming that X_1 , X_2 and Y are scaled to unit length, is

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad (2.6)$$

and the least squares normal equations are

$$(X'X)\hat{\beta} = X'Y \quad (2.7)$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix} \quad (2.8)$$

where r_{12} is the simple correlation between X_1 and X_2 and r_{jy} is the simple correlation between X_j and Y , $i=1,2$. Now the inverse of $(X'X)$ is

$$C = (X'X)^{-1} = \begin{bmatrix} \frac{1}{(1-r_{12}^2)} & -\frac{r_{12}}{(1-r_{12}^2)} \\ -\frac{r_{12}}{(1-r_{12}^2)} & \frac{1}{(1-r_{12}^2)} \end{bmatrix} \quad (2.9)$$

and the estimates of the regression coefficients are

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{(1-r_{12}^2)} \quad \hat{\beta}_2 = \frac{r_{2y} - r_{12}r_{1y}}{(1-r_{12}^2)} \quad (2.10)$$

If there is strong multicollinearity between X_1 and X_2 , then the correlation coefficient r_{12} will be large. From equation (2.9);

$$V(\hat{\beta}_j) = C_{jj}\sigma^2 \rightarrow \infty \quad \text{and} \quad \text{cov}(\hat{\beta}_1, \hat{\beta}_2) = C_{12}\sigma^2 \rightarrow \pm\infty$$

depending on whether

$$r_{12} \rightarrow +1 \quad \text{or} \quad r_{12} \rightarrow -1.$$

Therefore strong multicollinearity between X_1 and X_2 results in large variances and covariances for the least squares estimators of the regression coefficients. This implies that different samples taken at the same X levels could lead to widely different estimates of the model parameters.

When there are more than two predictor variables multicollinearity produces similar effects. It can be shown that the diagonal elements of the $C = (X'X)^{-1}$ matrix are

$$C_{jj} = \frac{1}{1 - R_j^2} \quad j=1, 2, \dots, k \quad (2.11)$$

where R_j^2 is the coefficient of multiple determination from the regression of X_j on the remaining $(k-1)$ predictor variables. If there is strong multicollinearity between X_j and any subset of the other $(k-1)$ predictors, then the value of R_j^2 will be close to unity. Since the variance of $\hat{\beta}_j$ is $V(\hat{\beta}_j) = C_{jj}\sigma^2 = (1 - R_j^2)^{-1}\sigma^2$, strong multicollinearity implies that the variance of the least squares estimate of the regression coefficient is very large. Generally the covariance of $\hat{\beta}_i$ and $\hat{\beta}_j$ will also be large if the predictors X_i and X_j are involved in a multicollinear relationship.

2.4.2 Wider Confidence Intervals

Because of the large standard errors, the confidence intervals for the relevant population parameters tend to be larger. Therefore, in cases of high multicollinearity, the sample data may be compatible with a diverse set of hypothesis. Hence the probability of accepting a false hypothesis (i.e. type II error) increases.

2.4.3 “Insignificant” t Ratios

To test the null hypothesis that, say, $\beta_2 = 0$, we use the t ratio, that is $\hat{\beta}_2 / se(\hat{\beta}_2)$, and compare the estimated t value with the critical t value from the t table. In cases of high collinearity the estimated standard errors increase dramatically, thereby making the t values smaller. Therefore, in such cases, one will increasingly accept the null hypothesis that the relevant true population value is zero.

2.4.4 A High R^2 but Few Significant t Ratios

Consider the k -variable linear regression model From (2.1), in cases of high collinearity, it is possible to find that one or more of the partial regression coefficients are individually statistically insignificant on the basis of the t test. Yet the R^2 in such situations may be so high, say in excess of 0.9 that on the basis of the F test one can convincingly reject the hypothesis that $\beta_2 = \beta_3 = \dots = \beta_k = 0$. Indeed this is one of the signals of multicollinearity insignificant t values but a high overall R^2 (and a significant F value)

2.4.5 Sensitivity of OLS Estimators and Their Standard Errors to Small Changes in Data

As long as multicollinearity is not perfect, estimation of the regression coefficients is possible but the estimates and their standard errors become very sensitive to even the slightest change in data.

In the presence of high collinearity one cannot estimate the individual regression coefficients precisely but that linear combinations of these coefficients may be estimated more precisely.

2.5 Detection Of Multicollinerity

Having studied the nature and consequences of multicollinearity, the natural question is : “How does one know that collinearity is present in any given situation, especially in models involving more than two predictors?” It is useful to bear in mind Kmenta’s warning:

1. *Multicollinearity is a question of degree and not of kind. The meaningful distinction is not between the presence and the absence of multicollinearity, but between its various degrees.*
2. *Since multicollinearity refers to the condition of the predictors that there are assumed to be nonstochastic it is a feature of the sample and not of the population. Therefore, we do not ‘test for multicollinearity’ but can, if we wish, measure its degree in any particular sample” (Jan Kmenta, Elements of Econometrics)*

“When multicollinearity is present, we do not have one unique method of detecting it or measuring its strength. What we have are some rules of thumb, some informal and some formal, but rules of thumb all the same.” (Gujarati, 1995, pp.335-339) Some of these rules are

1. High R^2 but few significant t ratios
2. High pair-wise correlations among predictors
3. Examination of the partial correlations
4. Auxiliary regression
5. Eigenvalues and condition index
6. Tolerance and variance inflation factor

2.5.1 High R^2 but Few Significant t Ratios

This is the “classic” symptom of multicollinearity. If R^2 is high, say, in excess of 0.8, the F test in most cases will reject the hypothesis that the partial regression coefficients are simultaneously equal to zero, but the individual t tests will show that none or very few of the partial regression coefficients are statistically different from zero.

Although this diagnostic is sensible, its disadvantage is that “it is too strong in the sense that multicollinearity is considered as harmful only when all of the influences of the predictors on Y cannot be disentangled.”

2.5.2 High Pair-wise Correlations among Predictors

Another suggested rule of thumb is that if the pair-wise or zero-order correlation coefficient between two predictors is high, say, in excess of 0.8, then multicollinearity is a serious problem. The problem with this criterion is that, although high zero-order correlation may suggest collinearity, it is not necessary that they be high to have collinearity in any specific case. To put the matter somewhat technically, high zero-order correlations are a sufficient but not a necessary condition for the existence of multicollinearity because it can exist even though the zero-order or simple correlations are comparatively low (say, less than 0.50).

Therefore, in models involving more than two predictors, the simple or zero-order correlation will not provide an infallible guide to the presence of multicollinearity. Of course, if there are only two predictors, the zero-order correlation will suffice.

2.5.3 Examination of Partial Correlation

A study of the partial correlations may be useful, there is no guarantee that they will provide an infallible guide to multicollinearity, for it may happen that both R^2 and all the partial correlations are sufficiently high.

2.5.4 Auxiliary Regressions

Since multicollinearity arises because one or more of the predictors are exact or approximately linear combinations of the other predictors, one way of finding out which X variable is related to other X variables is to regress each X_i on the remaining variables and compute the corresponding R^2 , which we designate as R_i^2 , each one of these regressions is called an auxiliary regression, auxiliary to the main regression of Y on the X 's. Then following the relationship between F and R^2 established and the variable

$$F_i = \frac{R_{X_i \cdot X_2 X_3 \dots X_k}^2 / (k-1)}{(1 - R_{X_i \cdot X_2 X_3 \dots X_k}^2) / (n-k)} \quad (2.12)$$

follows the F distribution with $k-1$ and $n-k$ df. In equation (2.12) $R_{X_i \cdot X_2 X_3 \dots X_k}^2$ is the coefficient of determination in the regression of variable X_i on the remaining X variables.

If the computed F_i exceeds the critical F at the chosen level of significance, it is considered that the particular X_i is collinear with other X 's if it does not exceed the critical F . It is not collinear with other X 's. In which case we may retain that variable in the model. If F is statistically significant, we still have to decide whether the particular X_i should be dropped from the model.

Instead of formally testing all auxiliary R^2 values, one may adopt Klien's rule of thumb, which suggest that multicollinearity may be a trouble some problem only if the R^2 obtained from an auxiliary regression is greater than the overall R^2 , that is, that obtained from the regression of Y on all the predictors. Of course, like all other rules of thumb, this one should be used judiciously.

2.5.5 Eigenvalues and Condition Index

Eigenvalues and the condition index are used to diagnose multicollinearity. From these eigenvalues, however we can derive what is known as the condition number (CN) defined as

$$CN = \frac{\text{MaximumEigenvalue}}{\text{MinimumEigenvalue}} \quad (2.13)$$

and the condition index (CI) defined as

$$CI = \sqrt{\frac{\text{MaximumEigenvalue}}{\text{MinimumEigenvalue}}} = \sqrt{CN} \quad (2.14)$$

“Then we have this rule thumb. If CN is between 100 and 1000 there is moderate to strong multicollinearity and if exceeds 1000 there is severe multicollinearity. Alternatively, if the $CI(=\sqrt{CN})$ is between 10 and 30, there is moderate to strong multicollinearity and if exceeds 30 there is severe multicollinearity.” (Gujarati, 1995, p.338)

“The condition indices of the $X'X$ matrix are

$$CI = \frac{\lambda_{\max}}{\lambda_j} \quad j=1,2, \dots, k \quad (2.15)$$

clearly the largest condition index is the condition number defined in equation (2.13). The number of condition indices that are largely (say ≥ 1000) are a useful measure of the number of near linear dependencies in $X'X$.” (Montgomery & Peck, 1992, p.319)

2.5.6 Tolerance and Variance Inflation Factor

For the k -variable regression model [Y , intercept, and k predictors] as we have seen in (2.16) the variance of a partial regression coefficient can be expressed as

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum \chi_j^2} \left(\frac{1}{1-R_j^2} \right) \quad (2.16)$$

$$= \frac{\sigma^2}{\sum \chi_j^2} VIF_j \quad (2.17)$$

where β_j is the regression coefficient of the predictor X_j , R_j^2 is the R^2 in the (auxiliary) regression of X_j on the remaining k predictors and VIF_j is the variance inflation factor. As R_j^2 increases toward unity, that is as the collinearity of X_j with the other predictors increases, the VIF also increases and in the limit it can be infinite.

Another measure of tolerance to detect multicollinearity is defined as

$$TOL_j = (1 - R_j^2) = (1/VIF_j) \quad (2.18)$$

Clearly $TOL_j = 1$ if X_j is not correlated with other predictors whereas it is zero if it is perfectly related to the predictors. VIF (or tolerance) as a measure of collinearity is not free of criticism. As equation (2.17) shows, $\text{var}(\hat{\beta}_j)$ depends on three factors: σ^2 , $\sum \chi_j^2$ and VIF_j . A high VIF can be counterbalanced by a low σ^2 or a high $\sum \chi_j^2$. To put it differently a high VIF is neither necessary nor sufficient to get high variances and high standard errors. Therefore, high multicollinearity as measured by a high VIF , may not necessarily cause high standard errors. In all this discussion, the terms high and low are used in a relative sense.

2.6 Methods For Dealing With Multicollinearity

What can be done if multicollinearity is serious? As in the case of detection, there are no infallible guides because multicollinearity is essentially a sample problem. However, one can try the following rules of thumb, the success depending on the severity of the collinearity problem. Several techniques have been proposed for dealing with the problems caused by multicollinearity. (Montgomery & Peck, 1992, pp.325-358)

1. Collecting Additional Data
2. Model Respecification
3. Ridge Regression
4. Generalized Ridge Regression
5. Principal Components Regression
6. Latent Root Regression Analysis

Such methods have been suggested as a possible solution to the multicollinearity problem. In these methods Ridge Regression, Generalized Ridge Regression, Principal Components Regression, Latent Root Regression Analysis are the biased regression methods. Biased Regression Method (*BRM*) is often used as another solution for the multicollinearity problem. *BRM* is preferred to *Least Squares* because *BRM* both explains the structure of multicollinearity and provide small *MSE*.

Principal Components Regression that is a method which vertical transformation variables are used instead of original variables and Ridge Regression that is a method for minimizing the mean square error by adding a constant to the diagonal of the correlation matrix will be dealt with in chapter three.

CHAPTER THREE

RIDGE REGRESSION AND PRINCIPAL COMPONENTS REGRESSION

3.1 Introduction

In regression analysis when multicollinearity problem exists, generally it is used the way either ignoring the model or eliminating one or more variable. There are two advantages of choosing variable as follows:

- a. The regression model that has few variables, both easy to practice and economic comparing to the others.
- b. Statistically small Mean Square Error (MSE) provides perfect estimate.

The least square estimators of the regression coefficients are the best linear unbiased estimators. That is, of all possible estimators are both linear functions of the data and unbiased for the parameters begin estimated, the least squares estimators have the smallest variance. In the presence of multicollinearity, however, this minimum variance may be unacceptably large. Lightening 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 multicollinearity might be found. Biased regression refers to this class of regression methods in which unbiasedness is no longer required.

The biased regression methods prevent the multicollinearity problem by the computationally suppressing the effects of the multicollinearity. Ridge regression makes this by reducing the apparent magnitude of the correlations. Principal

components regression prevents the problem by regressing Y on the important principal components and then parceling out of the effect of the principal component variables to the original variables.

3.2 Ridge Regression

When the sample data for regression exhibit multicollinearity the least squares estimates of the β coefficients may be subject to extreme round off error as well as inflated standard errors. Since their magnitudes and signs may change considerably from sample to sample, the least squares estimates are not said to be stable. A technique developed for stabilizing the regression coefficients in the presence of multicollinearity is ridge regression.

Ridge regression is a modification of the method of the least squares to allow biased estimators of the regression coefficients. At first place, the idea of biased estimation may not seem very appealing. But consider the sampling distributions of two different estimators of a regression coefficient β , one unbiased and the other biased, showed in Figure 3.1

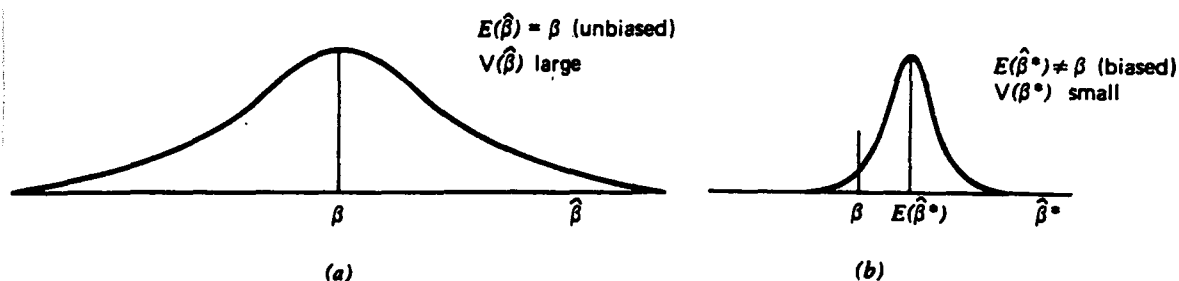


Figure 3.1 Sampling Distributions of (a) Unbiased and (b) Biased Estimators of β

(Montgomery & Peck, 1992, p.330)

Figure 3.1.a shows an unbiased estimator of β with fairly large variance. In contrast, the estimator shown in Figure 3.1.b has a slight bias but with much less variable. In this case, we would prefer the biased estimator over the unbiased estimator since it will lead to more precise estimates of the true β . One way to

measure the “goodness” of an estimator of β is to calculate the mean square error of $\hat{\beta}$, denoted by $MSE(\hat{\beta})$, where $MSE(\hat{\beta})$ defined as

$$MSE(\hat{\beta}) = (E(\hat{\beta} - \beta))^2 \quad (3.1)$$

$$= V(\hat{\beta}) + (E(\hat{\beta}) - \beta)^2 \quad (3.2)$$

the difference $E(\hat{\beta}) - \beta$ is called the bias of $\hat{\beta}$. Therefore, $MSE(\hat{\beta})$ is just the sum of the variance of $\hat{\beta}$ and the squared bias:

$$MSE(\hat{\beta}) = V(\hat{\beta}) + (\text{bias in } \hat{\beta})^2 \quad (3.3)$$

Let $\hat{\beta}_{LS}$ denote the least squares estimate of β . Then since $E(\hat{\beta}_{LS}) = \beta$, the bias is 0 and $MSE(\hat{\beta}_{LS}) = V(\hat{\beta}_{LS})$.

We have mentioned the variance of the least squares regression coefficients, and hence $MSE(\hat{\beta}_{LS})$, will be quite large in the presence of multicollinearity. The idea based on ridge regression is to introduce a small amount of bias in ridge estimator of β , denoted by $\hat{\beta}_R$, comparing its mean square error for least squares

$$MSE(\hat{\beta}_R) < MSE(\hat{\beta}_{LS}) \quad (3.4)$$

In this manner, ridge regression will lead to narrower confidence interval for the β coefficients, and hence more stable estimates.

To obtain the ridge regression coefficients, the user must be specify the value of a biasing constant c , ($c \geq 0$). In matrix notation, the ridge estimator $\hat{\beta}_R$ is calculated as follows:

$$\hat{\beta}_R = (X'X + cI)^{-1} X'Y \quad (3.5)$$

note that when $c=0$ the ridge estimator is the least squares estimator

$$\hat{\beta}_{LS} = (X'X)^{-1} X'Y \quad (3.6)$$

Ridge regression builds on the fact that a singular square matrix can be made nonsingular by adding a constant (c) to the diagonal of the matrix. That is, if $X'X$ is singular, then $(X'X + cI)$ is nonsingular, where c is some small positive constant. When the value c increases, the bias in the ridge estimates increases while the variance decreases. The idea is to choose the c so that the total mean square error for the ridge estimators is smaller than the total mean square error for the least square estimates.

Various methods for choosing the value of c have been proposed. Ridge trace, variance inflation factor and Hoerl, Kennard and Baldwin method (1975) are given next sections.

3.2.1 Ridge Trace

One commonly used graphical technique employing a ridge trace is shown in Figure 3.2. Values of the estimated ridge regression coefficients are calculated for different values of c ranging from 0 to 1 and plotted. The plots for each of the predictor variables in the model are overlaid to the ridge trace.

If multicollinearity is severe, the instability in the regression coefficients will be obvious from the ridge trace. As c is increased, some of the ridge estimates will be very dramatically change. At some value of c , the ridge estimates $\hat{\beta}_R$ will stabilize. The objective is to select a reasonably small value of c at which the ridge estimates $\hat{\beta}_R$ are stable.

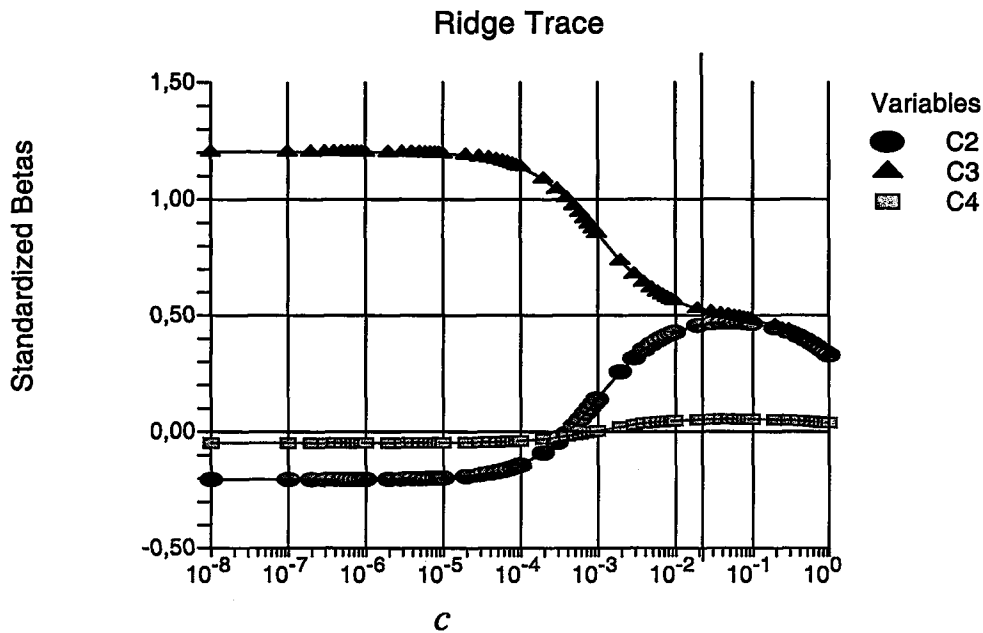


Figure 3.2 Ridge trace

In Figure 3.2 the search would be between 0.01 and 0.1. The value selected on this graph happens to be 0.066, the value obtained from the analytic search. It might be inclined to use an even smaller value of c such as 0.01. Mention that the smaller value of c , the smaller amount of bias that is included in the estimates.

3.2.2 Variance Inflation Factor

VIF_j is the method of determining biasing constant c .

$$VIF_j = \frac{1}{(1 - R_j^2)} \quad j = 1, 2, \dots, k \quad (3.7)$$

where R_j^2 is the coefficient of determination from the regression of X_j on the other independent variables. Note that VIF_j will be large when R_j^2 is large, that is, when the predictor variable X_j is strongly related to the other predictor variables.

A severe multicollinearity problem exists if the largest of the variance inflation factors for the β 's is greater than 10 or, equivalently, if the largest multiple coefficient of determination, R_j^2 , is greater than 0.90.

The ridge trace shown in Figure 3.2 is simultaneous plot of the values of the c estimated ridge standardized regression coefficients for different values of c , usually between 0 and 1. Extensive experience has indicated that the estimated regression coefficients $\hat{\beta}_R$ may fluctuate widely as c is changed slightly from 0, and some may even change signs. Gradually, however these wide fluctuation cease and the magnitudes of the regression coefficients tend to move slowly toward zero as c is increased further. At the same time, the values of $(VIF)_j$ tend to fall rapidly as c is changed from 0, and gradually the $(VIF)_j$ values also tend to change moderately as c is increased further. One therefore examines the ridge trace and the VIF values and chooses the smallest value of c where it is deemed that the regression coefficient first become stable in the ridge trace and the VIF values have become sufficiently small. The choice is thus a judgmental one.

3.2.3 Hoerl, Kennard and Baldwin Method

Hoerl, Kennard and Baldwin (1975) suggest the use of

$$c = \frac{(p-1)\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \quad (3.8)$$

where $(p-1)$ is the number of parameters excluding β_0 and $\hat{\sigma}^2$ is the mean square error estimated from the ordinary least squares regression ($c=0$). The denominator of equation (3.8) is the sum of squares of the ordinary least squares regression coefficients $\hat{\beta}$ excluding the intercept.

3.3 Principal Components Regression

Biased estimators of regression coefficients can also be obtained by using a procedure known as principal components regression. Consider the canonical form of the model,

$$Y = Z\alpha + \varepsilon \quad (3.9)$$

where

$$Z = XT \quad \alpha = T'\beta \quad T'X'XT = Z'Z = \Lambda$$

“The $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$ is a $k \times k$ diagonal matrix of the eigenvalues of $X'X$ and T is a $k \times k$ orthogonal matrix whose columns are the eigenvectors associated with $\lambda_1, \lambda_2, \dots, \lambda_k$. The columns of Z , which define a new set of orthogonal regressors, such as

$$Z = [Z_1, Z_2, \dots, Z_k]$$

are referred to as *principal components*.” (Montgomery & Peck, 1992, p.353)

The least squares estimator of α is

$$\hat{\alpha} = (Z'Z)^{-1}Z'Y = \Lambda^{-1}Z'Y \quad (3.10)$$

and the covariance matrix of $\hat{\alpha}$ is

$$V(\hat{\alpha}) = \sigma^2 (Z'Z)^{-1} = \sigma^2 \Lambda^{-1} \quad (3.11)$$

Thus a small eigenvalue of $X'X$ means that the variance of the corresponding orthogonal regression coefficient will be large.

$$Z'Z = \sum_{s=1}^k \sum_{t=1}^k Z_s Z_t' = \Lambda \quad (3.12)$$

we often refer to the eigenvalue λ_s as the variance of the s^{th} principal component. If all the λ_s are equal to unity, the original regressors are orthogonal, while if an λ_s is exactly equal to zero, this implies a perfect linear relationship between the original regressors. One or more of the λ_s near zero implies that multicollinearity is present. Note also that the covariance matrix of the standardized regression coefficients $\hat{\beta}$ is

$$V(\hat{\beta}) = V(T\hat{\alpha}) = T\Lambda^{-1}T'\sigma^2 \quad (3.13)$$

“This implies that the variance of $\hat{\beta}_s$ is $\sigma^2(\sum_{s=1}^k t_{ts}^2 / \lambda_s)$. Therefore the variance of $\hat{\beta}_s$ is a linear combination of the reciprocals of the eigenvalues. This demonstrates how one or more small eigenvalues can destroy the precision of the least squares estimate $\hat{\beta}_s$.” (Montgomery & Peck, 1992, p.354)

We have observed previously how the eigenvalues and eigenvectors of $X'X$ provide specific information on the nature of the multicollinearity. Since $Z=XT$, we have

$$Z_s = \sum_{t=1}^k t_{ts} X_t \quad (3.14)$$

where X_t is the t^{th} column of the X matrix and t_{ts} are the elements of the s^{th} column of T (the s^{th} eigenvector of $X'X$). If the variance of the s^{th} principal component (λ_s) is small, this implies that Z_s is nearly a constant, and (3.14) indicates that there is a linear combination of the original regressors that is nearly constant. This is the definition of multicollinearity. Therefore (3.14) explains why the elements of the eigenvector associated with a small eigenvalue of $X'X$ identify the regressors involved in the multicollinearity.

The principal components regression approach combats multicollinearity by using less than the full set of principal components in the model. To obtain the principal components estimator, assume that the predictors are arranged in order of decreasing eigenvalues, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k > 0$. Suppose that the last l of these eigenvalues are approximately equal to zero. In principal components regression the principal components corresponding to near-zero eigenvalues are removed from the analysis and least squares applied to the remaining components. That is,

$$\hat{\alpha}_{PC} = B\hat{\alpha} \quad (3.15)$$

where $b_1=b_2=\dots=b_{k-l}=1$ and $b_{k-l+1}=b_{k-l+2}=\dots=b_k=0$. Thus the principal components estimator is

$$\hat{\alpha}_{PC} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{k-l} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{array}{l} k-l \text{ components} \\ \\ \\ l \text{ components} \end{array} \quad (3.16)$$

or in terms of the standardized regressors (Montgomery & Peck, 1992, p.355)

$$\begin{aligned} \hat{\beta}_{PC} &= T\hat{\alpha}_{PC} \\ &= \sum_{i=1}^{k-l} \lambda_i^{-1} t_i' X' y t_i \end{aligned} \quad (3.17)$$

3.4 Comparison And Evaluation Of Biased Estimators

There is considerable evidence indicating the superiority of biased estimation to least squares if multicollinearity is present. Jeffers (1967) was the first to argue that principal components can also provide information as to which predictors should be selected. Following Jeffers' work, methods that utilize principal components to reduce the number of original variables have been suggested by several authors such as Mansfield et al. (1977).

There has also been some controversy surrounding whether the regressors and the response should be centered and scaled so that $X'X$ and $X'Y$ are in correlation form. This results in an artificial removal of the intercept from the model. Effectively the intercept in the ridge model is estimated by Hoerl, Kennard and Baldwin (1970a,b) use this approach, as Marquardt and Snee (1975) do, who note that centering tends to minimize any nonessential ill-conditioning when fitting polynomials.

In practice the procedure for selecting c with ridge trace is straightforward, easy to implement on a standard least squares computer program, and the analyst can learn to interpret the ridge trace very quickly. It is also occasionally useful to find the "optimum" value of c suggested by Hoerl, Kennard and Baldwin (1975) and compare the resulting models with the one obtained via the ridge trace.

Despite the objection noted, we believe that biased estimation methods are useful techniques that the analyst should consider when dealing with multicollinearity. Biased estimation methods certainly compare very favourably to other methods for handling multicollinearity, such as variable elimination. As Marquardt and Snee (1975) note, it is often better to use some of the information in all of the regressors, as ridge regression does, than to use all of the information in some regressors and none of the information in others, as variable elimination does. Furthermore variable elimination can be thought of as a form of biased estimation because, subset regression models often produce biased estimates of the regression coefficients. In effect, variable elimination often shrinks the vector of parameter estimates as ridge regression does.

CHAPTER FOUR

**COMPARISON OF PRINCIPAL COMPONENTS
REGRESSION WITH RIDGE REGRESSION AND
LEAST SQUARES REGRESSION BY
SIMULATION**

4.1 Introduction

In the previous chapter, the definitions of ridge regression and principal components regression methods were given. In this chapter, it will be given that how principal components regression and ridge regression methods can be used to remove multicollinearity.

In this chapter, first using a Minitab macro program a population was generated. Second, random samples were drawn from this population with sample sizes of $n=40$, 80 and 120. Least squares, ridge and principal components regression were applied to each of these samples. Regression coefficients for each of these estimators are computed and as statistical criteria to compare these estimators, the mean and the standard deviation of the estimates were used. For these purposes the following steps were followed.

1. The population is generated.
2. A sample size $n=40$ is selected from the population.
3. The least squares, ridge and principal components regression methods are applied to sampled data.
4. Returning to step 2, this process is repeated for 50 times.

5. Step 2, 3 and 4 are repeated for $n=80$ and 120.
6. The simulation values obtained for the least squares, ridge and principal components regression are compared and researched which methods give the best consequence by simulation.

4.2 Generating The Population

A Minitab macro program was written to draw a random samples from the created population. Some parameters of this population are given in the Table 4.1

Table 4.1 Generated Population

X_1	X_2	$Y_i \quad (i=1,\dots,25)$				$E(Y)$	σ_Y
10	5	46	49	...	63	58	5.0
	10	48	51	...	65	60	5.0
	15	50	53	...	67	62	5.0
	20	52	55	...	69	64	5.0
20	25	54	57	...	71	66	5.0
	30	59	59	...	75	68	5.0
	35	58	61	...	75	70	5.0
	40	60	63	...	77	72	5.0
30	45	62	65	...	79	74	5.0
	50	64	67	...	81	76	5.0
	55	66	69	...	83	78	5.0
	60	71	71	...	85	80	5.0
40	65	70	73	...	87	82	5.0
	70	72	77	...	89	84	5.0
	75	74	77	...	91	86	5.0
	80	76	79	...	93	88	5.0
50	85	78	81	...	95	90	5.0
	90	80	83	...	97	92	5.0
	95	82	85	...	99	94	5.0
	100	84	87	...	101	96	5.0
60	105	86	89	...	103	98	5.0
	110	88	91	...	105	100	5.0
	115	90	93	...	107	102	5.0
	120	92	95	...	109	104	5.0
70	125	94	97	...	111	106	5.0
	130	96	99	...	113	108	5.0
	135	98	101	...	115	110	5.0
	140	100	103	...	117	112	5.0
80	145	102	105	...	119	114	5.0
	150	104	107	...	121	116	5.0
	155	106	109	...	123	118	5.0
	160	108	111	...	125	120	5.0
90	165	110	113	...	127	122	5.0
	170	112	115	...	129	124	5.0
	175	114	117	...	131	126	5.0
	180	116	119	...	133	128	5.0
100	185	118	121	...	135	130	5.0
	190	120	123	...	137	132	5.0
	195	122	125	...	139	134	5.0
	200	124	127	...	141	136	5.0

There are two predictor variables X_1 and X_2 . X_1 has 10 different values as 10,20,30,...,100. X_2 takes 4 different values according to each X_1 value. Therefore there are 40 different groups of given X values. For each of X values there are 25 Y values and these response values are distributed normally and independently mean $E(Y)$ and constant variance $\sigma^2=25$.

The correlation matrix of Y , X_1 and X_2 is given in Table 4.2. It can be seen that there exists a high correlation between X_1 and X_2 .

Table 4.2 Correlation Coefficients Between Variables

	Y	X_1
X_1	0.97362	
X_2	0.97822	0.99530

4.3 An Application

The population was generated and a random sample of simulation 40 was selected by the following Minitab macro program 'SAM1.MTB'.

SAM1.MTB

```
RETR 'POPULATION'
SET C100
(1:25)
END
LET K100=1
SET C151
(0:975/25)K100
END
LET K1=100
EXEC 'SAM2' 40
STACK C101-C140 C150
```

```

LET C99=C150+C151
LET K1=0
LET K3=40*K100
EXEC 'SAM3' K3
ERASE C1-C9
ERASE C13-C200
NAME C1='Y' C2='X1' C3='X2'
COPY C10 C1
COPY C11 C2
COPY C12 C3
ERASE C10-C12
SAVE 'SAM01'

```

SAM2.MTB

```

LET K1=K1+1
SAMPLE K100 C100 CK1

```

SAM3.MTB

```

LET K1=K1+1
LET K2=C99(K1)
LET C10(K1)=C2(K2)
LET C11(K1)=C3(K2)
LET C12(K1)=C4(K2)

```

In the following graph an Anderson-Darling test for normality is performed and numerical results are displayed for a given sampled data. In the graph, a straight (or close to straight) line indicates normality. A lot of curvature indicates non-normal data. The null hypothesis is that the data are normal; the alternative hypothesis is that the data are not normal. A p -value greater than the cut value of our choice (0.05),

says no reject the null hypothesis. In Figure 4.1 shows that Y values have the normal distribution (p -value >0.05).

Normal Probability Plot

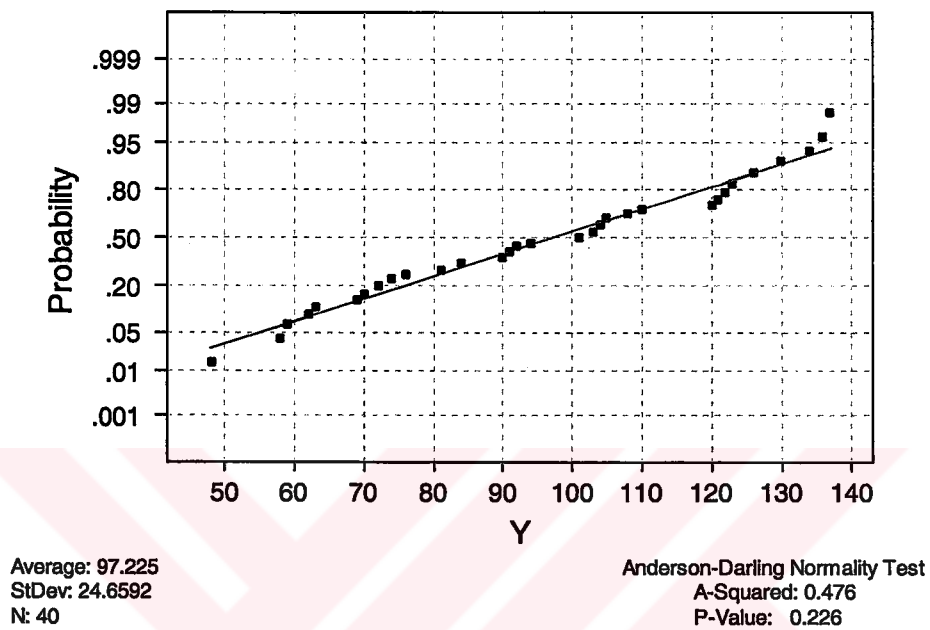


Figure 4.1 Normal Probability Plot of Y Values for $n=40$

Estimates of regression coefficients and the standard deviations of estimated regression coefficients are given in the following Table 4.3.

Table 4.3 Least Squares Regression Coefficients for an Application Data

i	$\hat{\beta}_i$	$S_{\hat{\beta}_i}$
0	53.75167	
1	0.23506	0.26627
2	0.29800	0.13251
$R^2 = 0.9658$		$S = 4.6850$

These regression coefficients are denoted by of $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$. These coefficients are calculated as $\hat{\beta}_0=53.75167$, $\hat{\beta}_1=0.23506$ and $\hat{\beta}_2=0.29800$. Standard deviations of coefficients are $S_{\hat{\beta}_1}=0.26627$ and $S_{\hat{\beta}_2}=0.13251$.

The ridge regression analysis is applied by using NCSS program. The results of ridge regression are given in Table 4.4.

Table 4.4 Summary Statistics of the Ridge Regression for an Application Data

	c	R^2	S	VIF	$c = 0.17546$	
i					$\hat{\beta}_i$	$S_{\hat{\beta}_i}$
0					56.61737	
1	0.00000	0.9658	4.6850	106.6000	0.37849	0.02494
2	0.00010	0.9657	4.6890	102.2162	0.19307	0.01241
3	0.00050	0.9655	4.7044	87.1366		
4	0.00100	0.9652	4.7231	72.5655		
5	0.00500	0.9631	4.8620	25.2257		
6	0.01000	0.9606	5.0223	11.1244		
7	0.05000	0.9418	6.1097	1.0241		
8	0.10000	0.9193	7.1933	0.4417		
9	0.17546	0.8873	8.4989	0.2841		
10	0.50000	0.7719	12.0914	0.1695		
11	1.00000	0.6430	15.1258	0.1135		

In this table for different c values R^2 , S and VIF values are given. When $c=0$ the ridge regression results are the same as least-squares results.

Since the least squares solution maximizes R -squared, the largest value of R -squared occurs when c is zero. We want to select a value of c that does not stray very much from the least squares R -squared value.

S is the square root of the mean squared error. Least squares minimizes this value, so we want to select a value of c that does not stray very much from the least squares value.

VIF is the maximum variance inflation factor. Since we are looking for that value of c which results in all VIF s being less than 10, this value is very helpful in our selection of c .

One of the methods for choosing the value of c is the Hoerl, Kennard and Baldwin method. The formula is given in equation (3.8).

Regarding all of these various statistics, from this table it is seen that the optimal c value is 0.17546 for an application data. The coefficients are $\hat{\beta}_0=56.61737$, $\hat{\beta}_1=0.37849$ and $\hat{\beta}_2=0.19307$. The standard deviations of these coefficients are $S_{\hat{\beta}_1}=0.02494$ and $S_{\hat{\beta}_2}=0.01241$.

The principal components regression analysis is applied by using NCSS program. Summary statistics for principal components regression analysis are given in the Table 4.5. Here PC 's is the number of principal components included in the regression reported in this row.

Table 4.5 Summary Statistics of the Principal Components Regression for an Application Data

PC 's	λ_i	Cum. Percent	CN	R^2	S	VIF	$\hat{\beta}_i$	$S_{\hat{\beta}_i}$
0							53.02580	
1	1.99530	99.7600	1.00000	0.96530	4.71450	0.25060	0.41694	0.01299
2	0.00470	100.0000	424.40000	0.96580	4.68500	106.60000	0.20749	0.00647

First eigenvalue is 1.99530 and the second eigenvalue is 0.00470. The first principal component accounts for 99.76% of the total variation in Y . Only first eigenvalue is greater than 1. Second eigenvalue is close to 0.

In this table *Condition Number (CN)* is largest eigenvalue divided by each corresponding eigenvalue. *CNs* greater than 1000 indicate a severe multicollinearity problem while condition numbers between 100 and 1000 indicate a mild multicollinearity problem. When all *PCs* are included the *CN* value is 424.40.

Since the least squares solution maximizes *R*-squared, the largest value of *R*-squared occurs at the bottom of the report when all *PC*'s are included.

S is the square root of mean squared error. Least squares minimizes this value, so it is wanted to select the number of *PC*'s that does not stray very much from the least squares value.

In this table *VIF* values are the maximum variance inflation factors. It is looking for the number of *PC*'s which results in all *VIF*s begin less than 10. When all *PC*'s are included the *VIF* value is 106.60.

Considering these statistics, the first principal component is chosen to use. Principal components regression with 1 component omitted results the coefficients as $\hat{\beta}_0=53.02580$, $\hat{\beta}_1=0.41694$ and $\hat{\beta}_2=0.20749$. The standard deviations of these coefficients are $S_{\hat{\beta}_1}=0.01299$ and $S_{\hat{\beta}_2}=0.00647$.

The results of least squares, ridge and principal components regression coefficients and standard deviations are summarized in Table 4.6 .

Table 4.6 Comparison Least Squares, Ridge and Principal Components Regression Coefficients for an Application Data

<i>i</i>	Least Squares		Ridge Regression		Principal Components Regression	
	$\hat{\beta}_i$	$S_{\hat{\beta}_i}$	$\hat{\beta}_i$	$S_{\hat{\beta}_i}$	$\hat{\beta}_i$	$S_{\hat{\beta}_i}$
0	53.75167		56.61737		53.02580	
1	0.23506	0.26627	0.37849	0.02494	0.41694	0.01299
2	0.29800	0.13251	0.19307	0.01241	0.20749	0.00647

As described in the literature, standard deviations of regression coefficients in the biased regression methods are smaller than the least squares regression coefficients. As it is seen from the table regression coefficients, the ridge regression and the principal components regression result similar to each other. As supposed, ridge regression method has result in smaller standard deviation values than the least squares method. Among these three methods, *PCR* gives the smallest results of standard deviation values.

4.4 Least Squares, Ridge And Principal Components Regression Results

The simulation results for $n= 40, 80$ and 120 will be given in the following sections. For every sample least squares, ridge and principal components regression will be analyzed.

4.4.1 Simulation Results and Interpretation for Sample Size 40

50 samples with $n=40$ are selected from the constructed population. The mean and the standard deviation of estimated regression coefficients are given in the following Table 4.7.

Table 4.7 Least Squares Regression Coefficients for $n=40$

i	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$
0	55.58887	1.80166
1	0.04858	0.28385
2	0.37816	0.14089
$R^2 = 0.95939$		$S = 4.93422$

Means of these coefficients are calculated as $\bar{\hat{\beta}}_0=55.58887$, $\bar{\hat{\beta}}_1=0.04858$ and $\bar{\hat{\beta}}_2=0.37816$. Standard deviations of coefficients are $S_{\hat{\beta}_0}=1.80166$, $S_{\hat{\beta}_1}=0.28385$ and $S_{\hat{\beta}_2}=0.14089$.

The results of ridge regression are given in Table 4.8. In this table for different c values R^2 , S and VIF values are given.

Table 4.8 Summary Statistics for the Ridge Regression when $n=40$

	c	R^2	S	VIF	$c = 0.22129$	
i					$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$
0					58.44952	1.72302
1	0.00000	0.95956	4.92015	104.56876	0.35609	0.01752
2	0.00010	0.95946	4.92722	99.80715	0.18514	0.01076
3	0.00050	0.95906	4.95212	85.23575		
4	0.00100	0.95855	4.98357	69.99839		
5	0.00500	0.95572	5.15456	25.43965		
6	0.01000	0.95291	5.31840	11.00959		
7	0.05000	0.93295	6.36111	1.10812		
8	0.01000	0.90964	7.38902	0.56999		
9	0.22129	0.85996	9.14367	0.44157		
10	0.50000	0.76089	12.03634	0.39422		
11	1.00000	0.63739	14.83449	0.39585		

Regarding all of these various statistics, from this table it is seen that the optimal c value is 0.22129 for $n=40$. The values of R^2 , S and VIF s are 0.85996, 9.14367 and 0.44157 when $c=0.22129$. The mean value of coefficients are $\bar{\hat{\beta}}_0=58.44952$, $\bar{\hat{\beta}}_1=0.35609$ and $\bar{\hat{\beta}}_2=0.18514$. The standard deviations of these coefficients are $S_{\hat{\beta}_0}=1.72302$, $S_{\hat{\beta}_1}=0.01752$ and $S_{\hat{\beta}_2}=0.01076$.

Summary statistics for principal components regression analysis are given in the Table 4.9.

Table 4.9 Summary Statistics for Principal Components Regression when $n=40$

PC's	λ_i	Cum. percent	CN	R^2	S	VIF	$\bar{\beta}_i$	S_{β_i}
							54.16945	1.40697
1	1.99530	99.7600	1.00000	0.95681	5.08765	0.25060	0.40423	0.01243
2	0.00470	100.0000	424.40000	0.95956	4.92015	106.60000	0.20117	0.00618

The first eigenvalue is 1.99530 and the first component accounts for 99.76% of the total variation in Y . The values of CN, R^2 , S and VIFs are 1.00000, 0.95681, 5.08765, and 0.25060 when the first component is chosen. Considering these statistics principal components regression with 1 component omitted results the mean value of coefficients as $\bar{\beta}_0=54.16945$, $\bar{\beta}_1=0.40423$ and $\bar{\beta}_2=0.20117$. The standard deviations of these coefficients are $S_{\beta_0}=1.40697$, $S_{\beta_1}=0.01243$ and $S_{\beta_2}=0.00618$.

The results of least squares, ridge and principal components regression coefficients and standard deviations are summarized in Table 4.10.

Table 4.10 Comparison Least Squares, Ridge and Principal Components Regression Coefficients when $n=40$

i	Least Squares		Ridge Regression		Principal Components Regression	
	$\bar{\beta}_i$	S_{β_i}	$\bar{\beta}_i$	S_{β_i}	$\bar{\beta}_i$	S_{β_i}
0	55.58887	1.80166	58.44952	1.72302	54.16945	1.40697
1	0.04858	0.28385	0.35609	0.01752	0.40423	0.01243
2	0.37816	0.14089	0.18514	0.01076	0.20117	0.00618

As described in the literature, standard deviation is smaller than the least squares method in both ways of biased regression. According to the regression coefficients, the ridge regression and the principal components regression result similar to each other. As supposed, ridge regression method has result in smaller standard deviation

values than the least squares method. Among these three methods, *PCR* gives the smallest results of standard deviation values.

4.4.2 Simulation Results and Interpretation for Sample Size 80

50 samples with $n=80$ are selected from the constructed population. The mean and the standard deviation of estimated regression coefficients are given in the following Table 4.11.

Table 4.11 Least Squares Regression Coefficients for $n=80$

i	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$
0	55.82650	1.27708
1	0.01337	0.16200
2	0.39580	0.08043
$R^2 = 0.96054$		$S = 4.78638$

Means of these coefficients are calculated as $\bar{\hat{\beta}}_0=55.82650$, $\bar{\hat{\beta}}_1=0.01337$ and $\bar{\hat{\beta}}_2=0.39580$. Standard deviation of coefficients are $S_{\hat{\beta}_0}=1.27708$, $S_{\hat{\beta}_1}=0.16200$ and $S_{\hat{\beta}_2}=0.08043$.

The results of ridge regression are given in Table 4.12. In this table for different c values R^2 , S and *VIF* values are given.

Tablo 4.12 Summary Statistics for the Ridge Regression when $n=80$

	c	R^2	S	VIF	$c=0.20556$	
i					$\bar{\beta}_i$	S_{β_i}
0					58.29188	1.08844
1	0.00000	0.96054	4.78638	106.60000	0.35776	0.00999
2	0.00010	0.96044	4.79244	102.21620	0.18695	0.00597
3	0.00050	0.96005	4.81540	87.13660		
4	0.00100	0.95962	4.84174	72.56550		
5	0.00500	0.95688	5.00514	25.22570		
6	0.01000	0.95408	5.16566	11.12440		
7	0.05000	0.93487	6.15940	1.02410		
8	0.10000	0.91246	7.14430	0.44170		
9	0.20556	0.86901	8.71702	0.26543		
10	0.50000	0.76611	11.68376	0.16950		
11	1.00000	0.63823	14.53194	0.11350		

Regarding all of these various statistics, from this table it is seen that the optimal c value is 0.20556 for $n=80$. The values of R^2 , S and VIF s are 0.86901, 8.71702 and 0.26543 when $c=0.20556$. The coefficients are calculated as $\bar{\beta}_0=58.29188$, $\bar{\beta}_1=0.35776$ and $\bar{\beta}_2=0.18695$. The standard deviations of coefficients are $S_{\beta_0}=1.08844$, $S_{\beta_1}=0.00999$ and $S_{\beta_2}=0.00597$.

Summary statistics for principal components regression analysis are given in the Table 4.13.

Table 4.13 Summary Statistics for Principal Components Regression when $n=80$

PC's	λ_i	Cum. percent	CN	R^2	S	VIF	$\bar{\beta}_i$	S_{β_i}
0							54.26610	1.04973
1	1.99530	99.7600	1.0000	0.95808	4.93269	0.25060	0.40436	0.00774
2	0.00470	100.0000	424.4000	0.96054	4.78638	106.60000	0.20123	0.00385

The first eigenvalue is 1.99530 and the first component accounts for 99.76% of the total variation in Y . The values of CN , R^2 , S and VIF s are 1.00000, 0.95808, 4.93269, and 0.25060 when the first component is chosen. Considering these statistics principal components regression with 1 component omitted results the mean of coefficients are $\bar{\beta}_0=54.26610$, $\bar{\beta}_1=0.40436$ and $\bar{\beta}_2=0.20123$. The standard deviations of coefficients are $S_{\beta_0}=1.04973$, $S_{\beta_1}=0.00774$ and $S_{\beta_2}=0.00385$.

The results of least squares, ridge and principal components regression coefficients and standard deviations are summarized in Table 4.14 .

Table 4.14 Comparison Least Squares, Ridge and Principal Components Regression Coefficients when $n=80$

i	Least Squares		Ridge Regression		Principal Components Regression	
	$\bar{\beta}_i$	S_{β_i}	$\bar{\beta}_i$	S_{β_i}	$\bar{\beta}_i$	S_{β_i}
0	55.82650	1.27708	58.29188	1.08844	54.26610	1.04973
1	0.01337	0.16200	0.35776	0.00999	0.40436	0.00774
2	0.39580	0.08043	0.18695	0.00597	0.20123	0.00385

As described in the literature, standard deviations of regression coefficients in the biased regression methods are smaller than the least squares regression coefficients. As it is seen from the table regression coefficients, the ridge regression and the principal components regression result similar to each other. As supposed, ridge regression method has result in smaller standard deviation values than the least

squares method. Among these three methods, *PCR* gives the smallest results of standard deviation values.

4.4.3 Simulation Results and Interpretation for Sample Size 120

50 samples with $n=120$ is selected from the constructed population. The mean and the standard deviation of estimated regression coefficients are given in Table 4.15.

Table 4.15 Least Squares Regression Coefficients for $n=120$

i	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$
0	56.31434	1.12874
1	-0.00511	0.15327
2	0.40079	0.07535
$R^2 = 0.95841$		$S = 4.84382$

Means of these coefficients are calculated as $\bar{\hat{\beta}}_0=56.31434$, $\bar{\hat{\beta}}_1= -0.00511$ and $\bar{\hat{\beta}}_2=0.40079$. Standard deviation of coefficients are $S_{\hat{\beta}_0}=1.12874$, $S_{\hat{\beta}_1}=0.15327$ and $S_{\hat{\beta}_2}=0.07535$.

The results of ridge regression are given in Table 4.16. In this table for different c values R^2 , S and VIF values are given.

Table 4.16 Summary Statistics for the Ridge Regression when $n=120$

	c	R^2	S	VIF	$c=0.21672$	
i					$\bar{\beta}_i$	S_{β_i}
0					58.87980	0.94752
1	0.00000	0.95841	4.84382	106.60000	0.35223	0.00909
2	0.00010	0.95830	4.84984	102.21620	0.18401	0.00512
3	0.00050	0.95792	4.87258	87.13660		
4	0.00100	0.95747	4.89858	72.56550		
5	0.00500	0.95466	5.05816	25.22570		
6	0.01000	0.95186	5.21315	11.12440		
7	0.05000	0.93264	6.16920	1.02410		
8	0.10000	0.91027	7.12147	0.44170		
9	0.21672	0.86237	8.80972	0.25424		
10	0.50000	0.76429	11.54607	0.16950		
11	1.00000	0.63670	14.33466	0.11350		

Regarding all of these various statistics, from this table it is seen that the optimal c value is 0.21672 for $n=120$. The values of R^2 , S and VIF s are 0.86237, 8.80972 and 0.25424 when $c=0.21672$. The mean of coefficients are calculated as $\bar{\beta}_0=58.87980$, $\bar{\beta}_1=0.35223$ and $\bar{\beta}_2=0.18401$. The standard deviation of coefficients are $S_{\beta_0}=0.94752$, $S_{\beta_1}=0.00909$ and $S_{\beta_2}=0.00512$.

Summary statistics for principal components regression analysis are given in the Table 4.17.

Table 4.17 Summary Statistics for Principal Components Regression when $n=120$

PC's	λ_i	Cum. percent	CN	R^2	S	VIF	$\bar{\beta}_i$	S_{β_i}
							54.69707	0.92746
1	1.99530	99.7600	1.0000	0.95577	4.99475	0.2506	0.40012	0.00750
2	0.00470	100.0000	424.4000	0.95841	4.84382	106.6000	0.19912	0.00373

The first eigenvalue is 1.99530 and the first component accounts for 99.76% of the total variation in Y . The values of CN , R^2 , S and VIF s are 1.00000, 0.95577, 4.99475, and 0.25060 when the first component is chosen. Considering these statistics principal components regression with 1 component omitted results the mean of coefficients are $\bar{\beta}_0=54.69707$, $\bar{\beta}_1=0.40012$ and $\bar{\beta}_2=0.19912$. The standard deviations of coefficients are $S_{\beta_0}=0.92746$, $S_{\beta_1}=0.00750$ and $S_{\beta_2}=0.00373$.

The results of least squares, ridge and principal components regression coefficients and standard deviations are summarized in Table 4.18 .

Table 4.18 Comparison Least Squares, Ridge and Principal Components

Regression Coefficients for $n=120$

i	Least Squares		Ridge Regression		Principal Components Regression	
	$\bar{\beta}_i$	S_{β_i}	$\bar{\beta}_i$	S_{β_i}	$\bar{\beta}_i$	S_{β_i}
0	56.31434	1.12874	58.87980	0.94752	54.69707	0.92746
1	-0.00511	0.15327	0.35223	0.00909	0.40012	0.00750
2	0.40079	0.07535	0.18401	0.00512	0.19912	0.00373

As described in the literature, standard deviations of regression coefficients in the biased regression methods are smaller than the least squares regression coefficients. As it is seen from the table regression coefficients, the ridge regression and the principal components regression result similar to each other. As supposed, ridge regression method has result in smaller standard deviation values than the least squares method. Among these three methods, *PCR* gives the smallest results of standard deviation values.

CHAPTER FIVE

CONCLUSION

5.1 Conclusions

In this study two of biased regression methods; principal components regression and ridge regression are examined as theoretically and investigated which methods give the best consequence by simulation.

In the application, 50 repetitions have been generated for each of the sample sizes of 40, 80 and 120. Least squares, ridge and principal components regression are used for each sample. Regression coefficients for each estimator were computed and as statistical comparison criteria, the mean and the standard deviation of the estimates were used.

As described in the literature, standard deviations of regression coefficients in the biased regression methods are found smaller than the least squares. It is seen from the comparison of least squares, ridge and principal components regression coefficients tables for each sample size.

At the same time $\bar{\hat{\beta}}_1$ and $\bar{\hat{\beta}}_2$ values of ridge regression and principal components regression are quite similar to each other. $\bar{\hat{\beta}}_0$ values of least squares regression and principal components regression are similar. In this case, it can be said that the principal components regression estimator gives closer unbiased estimates to least squares is more effective because of its small variance.

The comparison table shows that for these three sample sizes, principal components regression has given the smallest value of standard deviation of regression coefficients. The ridge regression method followed the principal components regression with small difference.

Finally, as for the regression coefficients obtained from samples of different sample sizes, principal components regression method, which gives the smallest standard deviation, can be preferred. At the same time ridge regression method, which gives the similar results to principal components regression for both regression coefficients and standard deviations of regression coefficients, can be used.



REFERENCES

- Akkaya, Ş.,& Pazarlıoğlu M.V. (1998). Ekonometri I. (4. Baskı). İzmir, Anadolu Matbaacılık.
- Aldrin, M. (1997). Length Modified Ridge Regression. *Computational Statistics & Data Analysis*, 25, 377-398.
- Boneh, S.,& Mendieta, G.R. (1994). Variable Selection in Regression Models Using Principal Components. *Commun. Statist. - Theory Meth.*, 23(1), 197-213.
- Ertek, T. (1996). Ekonometriye Giriş. (2. Baskı). İstanbul, Beta Yayınları.
- Gujarati, D.N. (1995). Basic Econometrics. (3rd ed.). New York, MCGraw-Hill, Inc.
- Hoerl, A.E., Kennard R.W.,& Baldwin K.F. (1975). Ridge Regression:Some Simulations. *Communications in Statistics*, 4(2), 105-123.
- Jackson, J.E. (1991). A User's Guide to Principal Components. Canada, John Wiley & Sons, Inc.
- Jeffers, J. N. (1967). Two Case Studies in the Application of Principal Component Analysis. *Appl. Statist.*, 22, 275-286
- Johnson, D.E. (1998) Applied Multivariate Methods for Data Analysis. California, Duxbury Press.
- Johnson R.A.,& Bhattacharyya G.K. (1992). Statistics Principles and Methods. Canada, John Wiley & Sons, Inc.

- Koçak, İ. (1998). Temel Bileşenler Analizi ve Uygulaması. T.C. İnönü Ün., Sosyal Bilimler Ens., Ekonometri Anabilim Dalı, Yüksek Lisans Tezi.
- Kotz S. & Johnson N. L. (1981). Encyclopedia of Statistical Sciences. Canada, John Wiley & Sons, Inc.
- Lawless, J.F.,& Wang, P. (1976). A Simulation Study of Ridge and Other Regression Estimators. Commun. Statis.-Theor. Meth., A5(4), 307-323
- Mansfield, E.R., Webster, J.T., Gunst, R.F. (1977). An Analytic Variable Selection Technique for Principal Component Regression. Appl. Statist., 36, 34-40
- Marquardt, D.W.,& Snee R.D. (1975). Ridge Regression in Practice. Am. Statist., 29(1), 3-20
- Mason, R.L. (1975). Regression Analysis and Problems of Multicollinearity. Communications in Statistics, 4(3), 277-292.
- Montgomery D.C.,& Peck E. (1992). Introduction to Linear Regression Analysis. Canada, John Wiley & Sons, Inc.
- Özdamar, K. (1999). Paket Programlar ile İstatistiksel Veri Analizi 2 (2.Baskı). Eskişehir, Kaan Kitabevi.
- Özen, S. (1992). Ridge Yöntemlerinde Yanlılık Parametrelerinin Seçimi. Ankara Ün., Fen Bilimleri Ens., İstatistik Anabilim Dalı, Yüksek Lisans Tezi.
- Özkan, M.M. (1989) Multicollinearity Varlığında En Uygun Regresyon Denkeleminin Belirlenmesine Yönelik Metodların Simülasyon ile Karşılaştırılması. Ankara Ün., Fen Bilimleri Ens., Zootekni Anabilim Dalı, Yüksek Lisans Tezi.
- Rawlings, J.O. (1988) Applied Regression Analysis:A Research Tool. California, Wadsworth & Brooks.