https://doi.org/10.26219/heal.aueb.6167

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

RIDGE REGRESSION ANALYSIS OF COLLINEAR DATA

By

Evangelia D. Mitsaki

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
2006

https://doi.org/10.26219/heal.aueb.6167

https://doi.org/10.26219/heal.aueb.6167

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

RIDGE REGRESSION ANALYSIS OF COLLINEAR DATA

By

Evangelia D. Mitsaki

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
May 2006

https://doi.org/10.26219/heal.aueb.6167

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΑΝΑΛΥΣΗ ΣΥΣΧΕΤΙΣΜΕΝΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ RIDGE ΠΑΛΙΝΔΡΟΜΗΣΗ

Ευαγγελία Δ. Μητσάκη

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Μάιος 2006

https://doi.org/10.26219/heal.aueb.6167

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

A Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science

RIDGE REGRESSION ANALYSIS OF COLLINEAR DATA

Evangelia D. Mitsaki

Thesis Supervisor:
John Panaretos
Professor

Athens, May 2006

Michael Zazanis, Professor
Director of the Graduate Program

https://doi.org/10.26219/heal.aueb.6167

DEDICATION

To my daughter

https://doi.org/10.26219/heal.aueb.6167

ACKNOWLEDGEMENTS

I would like to express my appreciation to several people for their support during my thesis work. More specifically, I would like to thank my supervisor, professor J.Panaretos for his guidance and patience. I would also like to thank Dimitris Karlis for his help and valuable advice. Finally, I owe my friend Vassiliki for providing me with most of the articles for my thesis.

https://doi.org/10.26219/heal.aueb.6167

https://doi.org/10.26219/heal.aueb.6167

VITA

I was born in Athens on March 4, 1975. In 1997 I graduated from University of Pireaus with a bachelor of Statistics and Actuarial Science. I started my Master of Science degree in 1998 at the Athens University of Economics and Business and attended the final semester at the Katholieke Universiteit of Leuven in Belgium.

https://doi.org/10.26219/heal.aueb.6167

https://doi.org/10.26219/heal.aueb.6167

ABSTRACT

Evangelia Mitsaki

Ridge regression analysis of collinear data

05/2006

Multicollinearity is a commonly occurring problem in regression analysis. It is the situation in which two or more explanatory variables are strongly (but not perfectly) correlated to one other, making it difficult to interpret the strength of the effect of each variable. This thesis deals with the theory of multicollinearity as well as with ways that have been proposed to detect and correct it. In order to cope with collinear data we present several remedial measures such as principal components, variable selection and biased estimation.

The focus of the thesis is on ridge regression. Since the seminal work of Hoerl and Kennard ridge regression has proven to be a useful technique to tackle the multicollinearity problem in the linear regression model. The thesis presents the ridge estimator (ordinary and generalized) and its properties and also ways for selecting the ridge constant. Different interpretations of ridge regression are also discussed as well as applications of ridge regression to cases other than multiple linear regression. A recent advance concerning influence analysis is also presented. Illustrative examples are given where necessary.

In order to demonstrate the application of the ridge regression model to data, Monte Carlo simulations will be primarily used. The simulations are intended to give some insight to the behaviour of the ridge estimators, i.e. to compare their characteristics and performance.

https://doi.org/10.26219/heal.aueb.6167

https://doi.org/10.26219/heal.aueb.6167

ΠΕΡΙΛΗΨΗ

Ευαγγελία Μητσάκη

Ανάλυση συσχετισμένων δεδομένων με ridge παλινδρόμηση

05/2006

Η πολυσυγραμμικότητα είναι σύνηθες φαινόμενο στην ανάλυση παλινδρόμησης κατά το οποίο δύο ή περισσότερες ερμηνευτικές μεταβλητές παρουσιάζουν υψηλή συσχέτιση (αλλά όχι πλήρη) μεταξύ τους, κάνοντας έτσι δύσκολη την ερμηνεία της κάθε μεταβλητής. Σε αυτή την εργασία ασχολούμαστε με τη θεωρία της πολυσυγγραμικότητας καθώς και τους τρόπους που έχουν προταθεί για να αναγνωρίζουμε την παρουσία της. Προκειμένου να χειριστούμε τα πολυσυγγραμικά δεδομένα παρουσιάζουμε διάφορες μεθόδους όπως τη μέθοδο των κυρίων συνιστωσών (principal components), μεθόδους επιλογής μεταβλητών (variable selection) και τη μέθοδο της μεροληπτικής εκτίμησης (biased estimation).

Ωστόσο αυτή η εργασία επικεντρώνεται στη ridge παλινδρόμηση. Προτεινόμενη αρχικά από τους Hoerl and Kennard, η ridge παλινδρόμηση έχει αποδειχθεί χρήσιμη μέθοδος αντιμετώπισης του προβλήματος της πολυσυγγραμικότητας στα πλαίσια του γράμμικου μοντέλου παλινδρόμησης. Παρουσιάζουμε τους ridge εκτιμητές και τις ιδιότητές τους καθώς και τρόπους επιλογής της ridge σταθεράς. Θα αναφερθούμε επίσης σε διαφορετικές προσεγγίσεις της ridge παλινδρόμησης καθώς και εφαρμογές της σε περιπτώσεις άλλες εκτός της γραμμικής παλινδρόμησης. Ακόμα, παρουσιάζουμε τη χρήση της influence analysis σε συσχετισμένες μεταβλητές. Δίνουμε παραδείγματα με πραγματικά δεδομένα όπου κρίνεται αναγκαίο. Τέλος, θέλοντας να δείξουμε την εφαρμογή της ridge παλινδρόμησης σε δεδομένα και να συγκρίνουμε τα χαρακτηριστικά των ridge εκτιμητών χρησιμοποιούμε προσομοιώσεις (Monte Carlo simulations).

https://doi.org/10.26219/heal.aueb.6167

https://doi.org/10.26219/heal.aueb.6167

TABLE OF CONTENTS

LIST OF TABLES

https://doi.org/10.26219/heal.aueb.6167

https://doi.org/10.26219/heal.aueb.6167

LIST OF FIGURES

https://doi.org/10.26219/heal.aueb.6167

https://doi.org/10.26219/heal.aueb.6167

CHAPTER 1

INTRODUCTION

Multiple linear regression is a widely used statistical technique that allows us to estimate models that describe the distribution of a response variable with the help of a number of other variables (explanatory). The use of multiple regression mainly regards the interpretation of the regression coefficients. In case of independent coefficients the least-squares solution gives stable estimates and useful results.

However, data are not always "well behaved". We often come across cases where the regressors (variables) are nearly collinear. This condition is called multicollinearity and is one of the most oftenly encountered in econometrics. The major problem with multicollinearity is that it leads to estimates with inflated variances in the estimation of regression coefficients and thus unacceptably large prediction intervals. High estimated variances (and therefore high estimated standard errors) also mean small observed test statistics. That is the analyst will accept too many null hypotheses. Estimates of standard errors and parameters tend to be sensitive to changes in the data and the specification of the model. In addition, the least-squares estimates are usually inflated with wrong signs-though they remain the best linear unbiased estimates (BLUE).

Note that, if the aim of the analyst is to generate forecasts, and if it is assumed that the multicollinearity problem will not be different for the forecast period, then multicollinearity may be considered not to be a problem at all. This is because multicollinearity will not affect the forecasts of a model but only the estimation of the coefficients (Koutsoyiannis, 1977).

In order to detect the presence of collinear variables many diagnostics have been proposed in the literature as the condition number, variance inflation factors, variance decomposition proportions etc. Approaches to remedy the problem of multicollinearity

have also been proposed. Model respecification, variable selection, and biased estimation are some of them.

In 1970 Arthur Hoerl and Robert Kennard published a paper on ridge regression, also known as the biased estimation method, that became the most commocly used method for the remedy of multicollinearity. Hoerl and Kennard's method was in fact a crude form of regularization, a technique developed by Andre Tikhonov (Tikhonov and Arsenin, 1977). Ridge regression involves the introduction of some bias into the regression equation in order to reduce the variance of the estimators. The bias is introduced by the use of a ridge constant which controls the extent to which ridge estimates differ from the least squares estimates. Depending on the method used to calculate the ridge constant different ridge estimators are defined. Ridge estimators have been proposed by many authors. McDonald and Galarneau (1975) proposed an estimator whose squared length equals an estimated squared length of β. Based on the mean square error property of the ridge estimator Hoerl, Kennard and Baldwin (1975), Guilkey and Murphy (1975), Goldstein and Smith (1974) and others have also proposed ridge estimators. Furthermore, considering the Bayesian approach, Lindley and Smith (1972), Lawless and Wang (1976) and others have also introduced ridge estimators.

The purpose of this thesis is to present the properties of ridge regression as a way to tackle the multicollinearity problem. More specifically, chapter 2 is devoted to the description of multicollinearity. First, we recall the fundamentals of linear regression and provide a description of multicollinearity and its effects. Furthermore, this chapter deals with the diagnostics proposed to detect multicollinearity as well as the remedial measures available, while for better illustration an example is provided. In chapter 3 we concentrate on ridge regression. Beginning with the reasoning given by Hoerl and Kennard, we proceed to the presentation of some properties of the ridge estimator as well as existence theorems that ensure that the ridge constant always exists. In addition, part of this chapter presents the available methods for calculating the ridge constant as well as the results of ridge regression applied to real data. Chapter 4 deals with different interpretations of ridge regression as well as its use in cases different from the multiple linear regression model. For example, the use of ridge regression in the logistic model or in simple linear regression with a small number of observations or in the context of generalized linear

https://doi.org/10.26219/heal.aueb.6167

models. Moreover, we discuss the effects of collinearity when, in addition, there are influential cases in the data set, a rather usual case. Finally, chapter 5 provides a simulation experiment for the comparison of certain ridge estimators.

https://doi.org/10.26219/heal.aueb.6167

https://doi.org/10.26219/heal.aueb.6167

CHAPTER 2

THE PROBLEM OF MULTICOLLINEARITY

2.1 *Introduction*

Regression analysis examines the relationship between a dependent variable y and one or more independent variables $X_1, X_2, ..., X_p$. Such an analysis assumes the use of a model with a specified set of independent variables. But in many cases we do not know exactly what variables should be included in a model. Hence, one may propose an initial model, often containing a large number of independent variables, and proceed with a statistical analysis aiming at revealing the correct model.

The inclusion of a large number of variables in a regression model often results in multicollinearity. The term multicollinearity refers to high correlation among the independent variables. This occurs when too many variables have been put into the model and a number of different variables measure similar phenomena. The existence of multicollinearity affects the estimation of the model as well as the interpretation of the results.

In this chapter we will give some preliminary material on:

1. The general regression situation.
2. Multicollinearity and how to detect it.
3. Strategies for coping with collinear data.

2.2 *The General Regression Situation*

The following definitions and proofs concerning multiple regression are based on Draper and Smith (1981) as well as on Rao and Toutenburg (1999).

Suppose we have a model under consideration, which can be written in the form

https://doi.org/10.26219/heal.aueb.6167

$$y = X\beta + u \tag{2.2.1}$$

where y is an $(T \times 1)$ vector of observations on a random variable, X is an $(T \times p)$ matrix of observations of the p independent variables, β is a $(p \times 1)$ vector of unobserved parameters, u is an $(T \times 1)$ vector of errors. We often use the following assumptions:

a) X is a fixed matrix of regressors (nonrandom),

b) The rank of X is p

c) Normality of the errors, i.e. the errors follow a normal distribution with zero mean, $u \sim \mathrm{N}(0, \sigma^2 \mathrm{I})$. This assumption is required for tests of significance and also for confidence and prediction intervals. It implies that the errors are homoscedastic, i.e. $V(u_t) = \sigma^2$ for all $\mathrm{t} = 1,\ldots,\mathrm{T}$, and that they are independent, i.e. $\mathrm{cov}(u_t, u_{t'}) = 0$ for all $\mathrm{t} \neq \mathrm{t}' = 1,\ldots,\mathrm{T}$.

Let us now consider the least squares method which is the most common method of estimating the parameters of the model. Since the error u is equal to $y - X\beta$ we shall estimate it with the residual which is defined as $\hat{u} = y - X\beta_0$, where β_0 is an arbitrary choice for β. The least squares coefficient vector minimizes the sum of squared residuals:

$$\begin{aligned} \hat{u}'\hat{u} &= (y - X\beta_0)'(y - X\beta_0) \\ &= y'y - \beta_0' X'y - y'X\beta_0 + \beta_0' X'X\beta_0 \\ &= y'y - 2\beta_0' X'y + \beta_0' X'X\beta_0 \end{aligned} \tag{2.2.2}$$

It can be determined by differentiating (2.2.2), with respect to β_0, and setting the resulting matrix equation equal to zero. Let $\hat{\beta}$ be the solution, then $\hat{\beta}$ satisfies the least squares normat equations

$$(X'X)\hat{\beta} = X'y. \tag{2.2.3}$$

If X is not of full rank, $X'X$ is singular, (2.2.3) has a set of solutions

https://doi.org/10.26219/heal.aueb.6167

$$\hat{\beta} = (X'X)^{-} X'y + \left(I - (X'X)^{-} X'X\right)\omega,$$

where $(X'X)^{-}$ is a generalized inverse (see Appendix A) of $X'X$ and ω is an arbitrary vector. Then either the model should be expressed in terms of fewer parameters or additional restrictions on the parameters must be given or assumed.

If the normal equations are independent, $X'X$ is nonsingular, and its inverse exists. In this case the solution of the normal equations is unique

$$\hat{\beta} = (X'X)^{-1} X'y. \tag{2.2.4}$$

Once β has been estimated by $\hat{\beta}$, we can write the residual as

$$\hat{u} = y - X\hat{\beta} = y - X(X'X)^{-1} X'y = (I - H)y,$$

where $H = X(X'X)^{-1} X'$ and I is the identity matrix. Further, the sum of squares of residuals divided by *T-p*,

$$s^2 = \frac{\hat{u}'\hat{u}}{T-p}, \tag{2.2.5}$$

can be shown to be a consistent and unbiased estimator of σ^2. The estimated regression is $y = X\hat{\beta} + \hat{u}$ and since $X'\hat{u} = 0$ the total sum of squares is

$$y'y = \hat{\beta}'X'X\hat{\beta} + \hat{u}'\hat{u} \tag{2.2.6}$$

where $\hat{\beta}'X'X\hat{\beta}$ is the sum of squares due to regression, and $\hat{u}'\hat{u}$ is the sum of squares due to errors. The multiple correlation coefficient, which measures the goodness of fit, is then defined as

$$R^2 = \frac{\hat{\beta}'X'X\hat{\beta}}{y'y} = 1 - \frac{\hat{u}'\hat{u}}{y'y} \tag{2.2.7}$$

R^2 tends to overestimate the true value of the coefficient. The following formula, which gives the multiple correlation coefficient adjusted by the degrees of freedom and is therefore unbiased, can be used instead:

https://doi.org/10.26219/heal.aueb.6167

$$R^2{}_{\text{adj}} = 1-(1-R^2)\frac{T-1}{T-p-1}.$$

The least squares estimate has some well-known properties (see e.g. in Seber, 1977):

1. It is an estimate of β, which minimizes the residual sum of squares, irrespective of any distribution properties of the errors.
2. Under the assumption of normality of the errors, $\hat{\beta}$ is the maximum likelihood estimate of β.
3. The elements of $\hat{\beta}$ are linear functions of the observations $y_1, y_2, ..., y_n$, and provide unbiased estimates of the elements of β which have the minimum variances, irrespective of any distributional properties of the errors (BLUE).

It can be deduced that since $E(u)=0$ then

$$\begin{aligned} E(\hat{\beta}) &= (X'X)^{-1}X'E(y) \\ &= (X'X)^{-1}X'X\beta \\ &= \beta \end{aligned} \tag{2.2.8}$$

and $\hat{\beta}$ is an unbiased estimate of β. If we further assume that the u_t are uncorrelated and have the same variance then $V(u)=\sigma^2 I_n$ and $V(y)=V(u)$. Hence the variance covariance matrix of $\hat{\beta}$ is given by

$$\begin{aligned} V(\hat{\beta}) &= V((X'X)^{-1}X'y) \\ &= (X'X)^{-1}X'V(y)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned} \tag{2.2.9}$$

https://doi.org/10.26219/heal.aueb.6167

2.3 *Multicollinearity*

In order to study the relationships among variables we collect data either from observational studies or designed experiments. It is not always possible, however, to carefully design controlled experiments in order to ensure that sufficient sample information is available. Observational studies are used instead and as the name implies, observe the variables and simply record them. Therefore some or most of the explanatory variables will be random hence the existence of general interrelationships among them is possible. Once detected, the best and obvious solution to the problem is to obtain and incorporate more information. Unfortunately, the researcher is usually not able to do so. Other procedures have been developed instead, for instance, biased estimation, and various variable selection procedures.

Recall that one of the assumptions for the model (2.2.1) was that X is of full rank, i.e. $|X'X| \neq 0$. This requirement says that no column of X can be written as exact linear combination of the other columns. If X is not of full rank, then $|X'X| = 0$, so that a)the ordinary least squares (OLS) estimate $\hat{\beta} = (X'X)^{-1}X'y$ is not uniquely defined and b)the sampling variance of the estimate is infinite. However, if the columns of X are nearly collinear although not exactly then $|X'X|$ is close to zero and the least squares coefficients become unstable since $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ can be too large. Multicollinearity among the columns can exist in varying degrees. One extreme situation is where the columns of X are pairwise orthogonal (that is, $x_i'x_j = 0$ for all i and j, $i \neq j$), so that there is a complete lack of multicollinearity; at the other extreme is the case of perfect linear relationship among the X's, that is, there exist nonzero constants c_i $(i = 1,\ldots, p)$, such that $c_1X_{j1} + c_2X_{j2} + \ldots + c_kX_{jk} = 0$ (see e.g. Huang, 1970).

In practice neither of the above extreme cases is often met. In most cases there is some degree of intercorrelation among the explanatory variables. It should be noted that multicollinearity is connected with time series as well as regression analysis. It is also quite frequent in cross-section data (Koutsoyiannis, 1977).

https://doi.org/10.26219/heal.aueb.6167

We now turn to an example to illustrate our discussion of multicollinearity.

Example: The following data set gives the merchandise imports of goods, Gross National Product (GNP) and the consumer price index (CPI) for the U.S. over the period 1970-1983.

Import	Year (X_1)	GNP (X_2)	CPI (X_3)
39,866	1970	992.7	116.3
45,579	1971	1077.6	121.3
55,797	1972	1185.9	125.3
70,499	1973	1326.4	133.1
103,811	1974	1434.2	147.7
98,185	1975	1549.2	161.2
124,228	1976	1718.0	170.5
151,907	1977	1918.3	181.5
176,020	1978	2163.9	195.4
212,028	1979	2417.8	217.4
249,781	1980	2631.7	246.8
265,086	1981	2957.8	272.4
247,667	1982	3069.3	289.1
261,312	1983	3304.8	298.4

Table 2.1: Example Data

Consider the following regression: $\text{Import}_t = \beta_0 + \beta_1 \text{Year}_t + \beta_2 \text{GNP}_t + \beta_3 \text{CPI}_t + e_t$

The correlation matrix for the predictors is given by

	X_1	X_2	X_3
X_1	1.000	**0.987**	**0.978**
X_2	0.987	1.000	**0.996**
X_3	0.978	0.996	1.000

There is collinearity among the regressors and the determinant for this table is $|X'X| = 0.000164$, which is very close to zero.

Then calculate $(X'X)^{-1}$,

47.929	-79.686	32.476
-79.686	259.848	-180.854
32.476	-180.854	149.366

These large numbers will give large coefficients and large estimated values for the variance of these coefficients.

https://doi.org/10.26219/heal.aueb.6167

2.3.1 Effects of Collinearity

The principles of least squares are not invalidated by the existence of multicollinearity since we still obtain the best linear unbiased estimates. The fact is that the data will simply not allow any method to distinguish between the effects of collinear variables on the dependent variable.

The consequences of collinearity in the case of several variables are:

- High estimated variance of $\hat{\beta}$

The existence of multicollinearity tends to inflate the estimated variances of the parameter estimates, which means that the confidence intervals for the parameters will be wide, and thus increasing the likelihood of not rejecting a false hypothesis. Since the regression coefficient measures the effect of the corresponding independent variable, holding constant all other variables, the existence of high correlation with other independent variables makes the estimation of such a coefficient difficult. Inflated variances are quite harmful to the use of regression analysis for estimation and hypothesis testing.

- High estimated variance of $\hat{y}$

The existence of multicollinearity tends to inflate the estimated variances of predicted values, that is, predictions of the response variable for sets of x values, especially when these values are not in the sample. The estimated variance of the predicted values is given by: $V(\hat{y}) = V(X\hat{\beta}) = XV(\hat{\beta})X' = \sigma^2 X(X'X)^{-1}X'$. Therefore, correlated X's correspond to large values of $(X'X)^{-1}$ and inflated estimated variances for $\hat{y}$.

- Unstable regression coefficients

The parameter estimates and their standard errors become extremely sensitive to slight changes in the data points.

- Wrong signs for regression coefficients

Coefficients will have wrong signs or an implausible magnitude (e.g. in econometric models there are coefficients that must have positive sign. Multicollinearity may lead to a coefficient with negative sign).

https://doi.org/10.26219/heal.aueb.6167

- Effect on specification

Given the above, variables may be dropped from the analysis, not because they have no effect but simply because the sample is inadequate to isolate the effect precisely.

2.4 Detecting Collinearity

Many diagnostics have been proposed in order to determine whether there is multicollinearity among the columns of *X*. Some of them will be discussed and better illustrated through an example.

2.4.1 Correlation Coefficients

A simple method for detecting multicollinearity is to calculate the correlation coefficients between any two of the explanatory variables. If these coefficients are greater than 0.80. or 0.90 then this is an indication of multicollinearity. A more elaborate rule is the following: if r_{ij} is the correlation coefficient between X_i and X_j,

$$r_{ij} = \frac{\sum_{k=1}^{T}\left(X_{ki} - \bar{X}_i\right)\left(X_{kj} - \bar{X}_j\right)}{\sqrt{\sum_{k=1}^{T}\left(X_{ki} - \bar{X}_i\right)^2}\sqrt{\sum_{k=1}^{T}\left(X_{kj} - \bar{X}_j\right)^2}}$$

and R^2 is the multiple correlation as defined in (2.2.7) between dependent and independent variables, multicollinearity is said to be "harmful" if $r_{ij} \geq R^2$ (Huang, 1970). Such simple correlation coefficients are sufficient but not necessary condition for multicollinearity. In many cases there are linear dependencies, which involve more than two explanatory variables, that this method cannot detect (Judge et al., 1985).

We can extend the concept of simple correlation between independent variables to multiple correlation within an independent variables set. A variable X_i then, would be harmfully multicollinear only if its multiple correlation with other members of the

https://doi.org/10.26219/heal.aueb.6167

independent variable set, R_i^2, were greater than the dependent variable's multiple correlation with the entire set, R^2 (Greene, 1993).

2.4.2 Calculation of $|X'X|$

A test which is most commonly used relies on the property that the determinant of a singular matrix is zero. Defining a small, positive test value, $\varepsilon > 0$, a solution is attempted only if the determinant based on a normalized correlation matrix is larger than this value, i.e. $|X'X| > \varepsilon$; Recall that the position of such a determinant on the scale is $0 \le |X'X| \le 1$. The closer $|X'X|$ is to 0, the greater the severity of multicollinearity and the closer $|X'X|$ is to 1, the less the degree of multicollinearity. Note that, in practice $|X'X|$ is rarely greater than 0.1.

Near singularity may result from strong, sample pairwise correlation between independent variables, or from a more complex relationship between several members of a set. The determinant gives no information about this interaction.

2.4.3 Leamer's Method

Leamer (in Greene, 1993) have suggested the following measure of the effect of multicollinearity for the jth variable:

$$c_j = \left\{ \frac{\left(\sum_i (X_{ij} - \bar{X}_j)^2\right)^{-1}}{(X'X)_{jj}^{-1}} \right\}^{1/2},$$

where $(X'X)_{jj}^{-1}$ is the *jj*-th element of the matrix $(X'X)^{-1}$. This measure is the square root of the ratio of the variances of $\hat{\beta}_j$ when estimated without and with the other variables. If X_j was uncorrelated with the other variables, c_j would be 1. Otherwise, c_j is equivalent to $(1 - R_j^2)^{1/2}$, where R_j^2 is *the multiple correlation of the variable X_j as dependent with the other members of the independent variable set as predictors.*

https://doi.org/10.26219/heal.aueb.6167

2.4.4 The Condition Number

Another way to test the degree of multicollinearity is the magnitude of the eigenvalues of the correlation matrix of the regressors. Large variability among the eigenvalues indicates a greater degree of multicollinearity. Two features of these eigenvalues are of interest:

- Eigenvalues of zero indicate exact collinearities. Therefore, very small eigenvalues indicate near linear dependencies or high degrees of multicollinearity.
- The square root of the ratio of the largest to the smallest eigenvalue $K = \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{1/2}$, called the condition number, is a commonly employed index of the "instability" of the least-squares regression coefficients. A large condition number (say, 10 or more) indicates that relatively small changes in the data tend to produce large changes in the least-squares estimate. In this event, the correlation matrix of the regressors is said to be ill conditioned (Greene, 1993). Observe the following simple situation where we have a two regressors model: the condition number is

$$K = \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{1/2} = \left(\frac{1+\sqrt{r_{12}^2}}{1-\sqrt{r_{12}^2}}\right)^{1/2}.$$

Setting K equal to 10 corresponds to $r_{12}^2 = 0.9608$ (Fox, 1997).

2.4.5 Variance Inflation Factors

A consequence of multicollinearity is the inflation of variation. For the jth independent variable, ***the variance inflation factor is defined as*** $\frac{1}{\left(1 - R_j^2\right)}$, see section 2.4.3. These factors are useful in determining which variables may be involved in the multicollinearities.

The sampling variance of the jth coefficient $\hat{\beta}_j$ is

https://doi.org/10.26219/heal.aueb.6167

$$V(\hat{\beta}_j) = \frac{1}{(1-R_j^2)} \frac{\sigma^2}{(T-1)S_j^2}$$

where $S_j^2 = \frac{\sum_{i=1}^{T}(X_{ij} - \bar{X}_j)^2}{T-1}$ is the variance of X_j and σ^2 the error variance (Fox, 1997).

The term $\frac{1}{1-R_j^2}$, indicates the impact of collinearity on the precision of the estimate $\hat{\beta}_j$.

It can be interpreted as the ratio of the variance of $\hat{\beta}_j$ to what that variance would be if X_j were uncorrelated with the remaining X_i. The inverse of VIF (i.e 1-R_j^2) is called tolerance.

It is better to examine the square root of the VIF than the VIF itself because the precision of estimation of β_j is proportional to the standard error of $\hat{\beta}_j$ (not its variance). Because of its simplicity and direct interpretation, the VIF (or its square root) is the principal diagnostic for desribing the sources of imprecision. There are no formal criteria for determining the magnitude of variance inflation factors that cause poorly estimated coefficients. Some authors state that values exceeding 10 may be cause for concern, but this value is arbitrary (Fox, 1997). A VIF equal to 10 implies that the R_j^2 is 0.9.

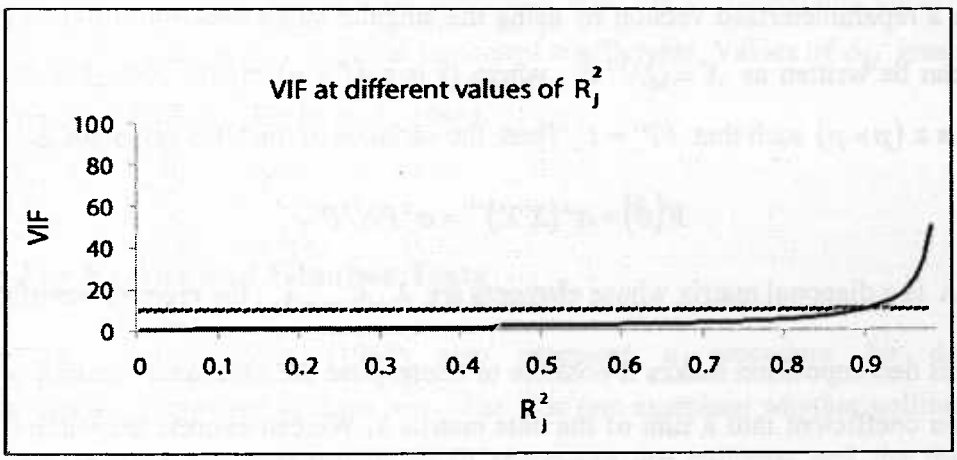


Figure 2.1 Variance inflation factor

https://doi.org/10.26219/heal.aueb.6167

Graphically speaking in a Venn diagram (after John Venn, an English mathematician), *VIF* is shown by many overlapping circles. In the following figure, the circle at the center represents the explanatory variable and all surrounding ones represent the independent variables. The area covered by the surrounding circles denotes the variance explained. In this case where too many variables are included in the model the explanatory variable is almost entirely covered by many inter-related *X*'s. While the variance explained is high the model is over-specified and most likely useless.

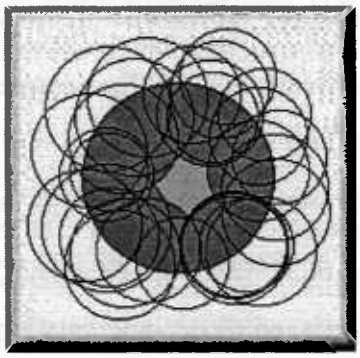


Figure 2.2 Venn diagram

2.4.6 Variance Decomposition Proportions

Consider again the linear model $y = X\beta + u$, X is a $(T \times p)$ design matrix. Now consider a reparameterized version by using the singular value decomposition of *X*. The matrix can be written as $X = Q\Lambda^{1/2}P'$, where Q is a $(T \times p)$ matrix such that $Q'Q = I$ and P' is a $(p \times p)$ such that $PP' = I$. Thus, the variance of the OLS estimator is

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} = \sigma^2 P\Lambda^{-1}P',$$

where Λ is a diagonal matrix whose elements are $\lambda_1, \lambda_2, ..., \lambda_p$, the eigenvalues of $X'X$.

Using this decomposition makes it possible to decompose the estimated variance of each regression coefficient into a sum of the data matrix *X*. We can express the variance of a single coefficient as

$$\mathrm{V}(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^{p} \frac{p_{kj}^2}{\lambda_j},$$

https://doi.org/10.26219/heal.aueb.6167

where p_{kj} denotes the $(k, j)^{th}$ element of the matrix P. Consequently, the proportion of $V(\hat{\beta}_k)$ associated with any single eigenvalue is

$$\phi_{kj} = \frac{p_{kj}^2/\lambda_j}{\sum_{j=1}^{p} p_{kj}^2/\lambda_j}.$$

It is useful to view these values as in table 2.2:

Eigenvalue	$V(\hat{\beta}_1)$	$V(\hat{\beta}_2)$	.....	$V(\hat{\beta}_k)$	.....	$V(\hat{\beta}_p)$
λ_1	ϕ_{11}	ϕ_{21}	.	ϕ_{k1}	.	ϕ_{p1}
λ_2	ϕ_{12}	ϕ_{22}	.	ϕ_{k2}	.	ϕ_{p2}
.	.	.	.		.	
λ_p	ϕ_{1p}	ϕ_{2p}	.....	ϕ_{kp}	.....	ϕ_{pp}

Table 2.2 Variance-Decomposition Proportions

The columns in the table sum to one. The presence of two or more large values of ϕ_{kj} in a row indicates that linear dependence associated with the corresponding eigenvalue is adversely affecting the precision of the associated coefficients. Values of ϕ_{kj} greater than 0.50 are considered large (Judge et al., 1985).

2.4.7 The Farrar and Glauber Tests

Farrar and Glauber (1967) also proposed a procedure for detecting multicollinearity comprised of three tests. The first one examines whether collinearity is present, the second one determines which regressors are collinear and the third one determines the form of multicollinearity. Based on the assumption that X is multivariate normal the authors propose the following:

https://doi.org/10.26219/heal.aueb.6167

- The chi-square test for the presence of multicollinearity

The null hypothesis is that the X's are orthogonal. A statistic based on the determinant $|X'X|$ could provide a useful first measure of the presence of multicollinearity within the independent variables. Bartlett obtained a transformation of $|X'X|$,

$$\chi_*^2 = -\left[T - 1 - \frac{1}{6}(2p+5)\right]\ln|X'X|,$$

that is distributed approximately as chi square with $v = \frac{1}{2}p(p-1)$ degrees of freedom; p is the number of independent variables. This is the well known Bartlett's sphericity test. From the sample data we obtain the empirical value χ_*^2. If this value is greater than the tabulated value of χ_v^2, we reject the assumption of orthogonality.

- The F-test for the determination of collinear regressors

The null hypothesis that R_i^2 is equal to zero. Consider the variable Z_i, which is equal to $1 - R_i^2$ and the new variate,

$$\omega_i = \left(\frac{1}{Z_i} - 1\right)\left(\frac{T-p}{p-1}\right) = \frac{R_i^2}{1-R_i^2}\left(\frac{T-p}{p-1}\right).$$

The distribution of ω_i is the F-distribution with T-p and p-1 degrees of freedom since $\frac{R_i^2}{p-1}$ (and $\frac{1-R_i^2}{T-p}$) is distributed as a chi-square with p-1 (and T-p respectively) degrees of freedom under the null hypothesis. Since R_i^2 is the multiple correlation coefficient between X_i and the other members of X, ω_i is the ratio of explained to unexplained variance. If the observed value $\omega_i > F$, we accept that the variable X_i is multicollinear.

- The t-test for the pattern

To understand the form of collinearity in X, the authors use the partial correlation coefficients between X_i and X_j, which describe the relationship of X_i and X_j when all other members of X are held fixed, namely $r_{ij.12..p}$. The basic hypothesis here is that $r_{ij.12..p} = 0$. To test this hypothesis we are based in the following statistic

https://doi.org/10.26219/heal.aueb.6167

$$t^*_v = \frac{r_{ij.12..p}\sqrt{T-p}}{\sqrt{1-r^2_{ij.12..p}}}$$

which is distributed as Student's with $v = T - p$ degrees of freedom. If $t_v^* > t$, where t is the theoretical value of the Student's distribution with v degrees of freedom, then we accept that the variables X_i and X_j are responsible for the multicollinearity. Therefore if the i_{th} variable is detected collinear by the F-test presented above and the null hypothesis based on the partial correlation coefficient between X_i and X_j is rejected then we can conclude that the j_{th} variable is responsible for the multicollinearity of the i_{th} variable.

These tests have been greatly criticized. Robert Wichers (1975) claims that the third test, where the authors use the partial-correlation coefficients $r_{ij.12..p}$, is ineffective while O'Hagan and McCabe (1974) quote, "Farrar and Glauber have made a fundamental mistake in interpreting their diagnostics."

2.4.8 The Sum of λ_i^{-1}

One easy way of assessing the degree of multicollinearity is to investigate the eigenvalues $\lambda_1 > \lambda_2 > ... > \lambda_p$ and eigenvectors of the matrix $X'X$. In an orthogonal system

$$\sum_{i=1}^{p} \lambda_i^* = \sum_{i=1}^{p} \lambda_i^{*-1} = p,$$

where λ_i^* correspond to the p eigenvalues of the correlation matrix $R^* = I$. Thus for a sample-based correlation matrix R with eigenvalues λ_i, $i = 1,2,\ldots,p$, we can compare

$$p \quad \text{vs} \quad \sum_{i=1}^{p} \lambda_i^{-1}.$$

Large values of $\sum_{i=1}^{p} \lambda_i^{-1}$ would indicate severe collinearity (Dillon and Goldstein, 1984).

https://doi.org/10.26219/heal.aueb.6167

2.5 *Example*

We will now provide an example to illustrate the use of the above defined diagnostics. The data were presented in Longley (1967) and have been used by many authors to present multicollinearity related topics. The data (Appendix, part 2) concern as a dependent variable the number of people employed (in thousands), yearly from 1947 to 1962. There are six explanatory variables namely:

1) Gross National Product (GNP) implicit price deflator (1954=100),
2) GNP (in millions of dollars),
3) Unemployed (in thousands),
4) Armed forces (in thousands),
5) No institutionalized polulation 14 years of age and over (in thousands), and
6) Year.

This regression is known to be highly collinear.

Parameter estimates					
	Parameter Estimate	Std. Error	Standardized Estimate	t value	*p*-value
Intercept	-3,482,259	890420	0	-3.91	0.0036
GNP deflator	15.06187	84.91493	0.04628	0.18	0.8631
GNP	-0.03582	0.03349	-1.01375	-1.07	0.3127
Unemployed	-2.02023	0.48840	-0.53754	-4.14	0.0025
Armed forces	-1.03323	0.21427	-0.20474	-4.82	0.0009
Population	-0.05110	0.22607	-0.10122	-0.23	0.8262
Year	1829.15146	455.4785	2.47966	4.02	0.0030

Multiple R-squared: 0.9955

Table 2.3: The values of the regression coefficients and the *p*-values

https://doi.org/10.26219/heal.aueb.6167

We note that some predictors (e.g. population) have large p-values though we would expect them to be significant. If we check the correlation matrix below we will find several large pairwise correlations.

	GNP deflator	GNP	Unemployed	Armed forces	Population	Year
GNP deflator	1.000					
Sig. (2-tailed)	.					
GNP	***0.992***	1.000				
Sig. (2-tailed)	.000	.				
Unemployed	0.621	0.604	1.000			
Sig. (2-tailed)	.010	.013	.			
Armed forces	0.465	0.446	-0.177	1.000		
Sig. (2-tailed)	.070	.083	.511	.		
Population	***0.979***	***0.991***	0.687	0.364	1.000	
Sig. (2-tailed)	.000	.000	.003		.	
Year	***0.991***	***0.995***	0.668	0.417	***0.994***	1.000
Sig. (2-tailed)	.000	.000	.005	.108	.000	.

Table 2.4: The correlation matrix of the predictors

Following, let us calculate some of the multicollinearity diagnostics presented in section 2.4. Specifically, the variance inflation factors, the coefficients of determination R_i^2, and Leamer's measure.

The predictors	VIF	R_i^2	Leamer's c_i
GNP deflator	135.532	0.993	0.086
GNP	1788.513	0.999	0.024
Unemployed	33.619	0.970	0.173
Armed forces	3.589	0.721	0.528
Population	399.151	0.997	0.050
Year	758.981	0.998	0.036

Table 2.5: The multicollinearity diagnostics

https://doi.org/10.26219/heal.aueb.6167

The variance inflation factors are large, namely 399.151 for "population", 758.981 for "year" and up to 1788.513 for the "GNP" regressor. Considering that the VIF for the orthogonal predictors is 1 we see that there is considerable variance inflation. Consider next R_i^2, the multiple correlation of the variable X_i as dependent with the other members of the independent variable set as predictors. These values vary from 0.721 to 0.999 suggesting that GNP for instance is well explained by the remaining independent variables. Next we present the eigenvalues and the variance decomposition proportions

		Variance Proportions						
Dimension	Eigenvalue	(Constant)	GNP deflator	GNP	Unemployed	Armed Forces	Population	Year
1	6.861	.00	.00	.00	.00	.00	.00	.00
2	.008	.00	.00	.00	.01	.09	.00	.00
3	.046	.00	.00	.00	.00	.06	.00	.00
4	.000	.00	.00	.00	.06	.43	.00	.00
5	.002	.00	.46	.02	.01	.12	.01	.00
6	.000	.00	.50	.33	.23	.00	.83	.00
7	.000	1.00	.04	.65	.69	.30	.16	1.00

Table 2.6: Eigenvalues and variance proportions

The eigenvalues vary from 6.8614 to 0.00000000366, while the condition number - $K = \left(\frac{6.6814}{0.000376}\right)^{1/2}$ = 43,275- is quite large. Two more diagnostics are the determinant of the correlation matrix and the sum of λ_i^{-1}. The values are $|X'X| = 0.157 \times 10^{-7}$, which is very close to zero and $\sum \lambda_i^{-1}$ =3119.4, which is very large.

We can also calculate the tests proposed by Farrar and Glauber. The chi-square statistic that measures the presence and severity of multicollinearity is $\chi_*^2 = 218.56$. This value is greater than the tabulated value of $\chi_{15}^2 = 25$ so we reject the assumption of orthogonality. Continuing with Farrar and Glauber's tests we find: all ω_i are greater than $F_{10,5} = 4.74$ and most of t^* are greater than $t_{10} = 2.24$ that is, there is multicollinearity.

https://doi.org/10.26219/heal.aueb.6167

	ω_i test for localization	t^* statistic for the pattern				
		GNP deflator	GNP	Unemployed	Armed forces	Population
GNP deflator	269.065					
GNP	3575.027	*24.228*				
Unemployed	65.238	*2.503*	*2.398*			
Armed forces	5.178	1.660	1.578	-0.570		
Population	796.302	*15.248*	*23.531*	*2.986*	1.237	
Year	1515.961	*23.610*	*32.409*	*2.841*	1.452	*28.624*

Table 2.7: Farrar and Glauber's diagnostics for localization and pattern

2.6 Remedial Measures

2.6.1 Model Respecification

One approach to the problem of multicollinearity is to respecify the model. Perhaps it may be useful to implement multivariate techniques to study the structure of multicollinearity and consequently to provide a better understanding of the regression relationships. One such multivariate method is principal components, developed in the early part of the 20th century.

Principal component analysis is a multivariate technique that attempts to describe interrelationships among a set of variables. Starting with a set of observed values on a set of p variables, the method uses linear transformations to create a new set of variables, called the principal components, which have the following properties:

- *The principal component variables, or simply the components, are jointly* ***uncorrelated.***
- *The first principal component has the largest variance of any linear function of the original variables. The second component has the second largest variance, and so on.*

https://doi.org/10.26219/heal.aueb.6167

We shall describe the method briefly following Jackson (1991):

The principal components of the p standardized regressors are a new set of p variables derived from X by a linear transformation: $W = XA$, where A is the $(p \times p)$ transformation matrix. The transformation A is selected so that the columns of W are orthogonal-that is, the principal components are uncorrelated. In addition, A is constructed so that the first component accounts for maximum variance in the X's; the second for maximum variance under the constraint that it is orthogonal to the first; and so on. The principal components therefore partition the variance of the X's.

The transformation matrix A contains (by columns) normalized eigenvectors of the correlation matrix of the regressors $X'X = R_X$. The columns of A are ordered by their corresponding eigenvalues: the first column corresponds to the eigenvector of the largest eigenvalue, and the last column to the smallest. The eigenvalue λ_j associated with the *jth* component represents the variance attributable to that component. If there are perfect collinearities in X, then some eigenvalues of R_X will be 0, and there will be fewer than p principal commponents, the number of components corresponding to rank(X)=rank(R_X).

As we showed earlier the VIF_j is equal to $1/(1-R_j^2)$. It can also be shown that the VIF_j is the ***diagonal entry of*** R_X^{-1} and since

$$R_X^{-1} = A\Lambda^{-1}A',$$

where $\Lambda = diag(\lambda_1,...,\lambda_p)$ is the matrix of eigenvalues of R_X, the VIF_j can be expressed as function of the eigenvalue of R_X and the principal components;

$$VIF_j = \sum_{i=1}^{p} \frac{A_{ji}^2}{\lambda_i}. \qquad (2.6.1)$$

Thus, it is only the small eigenvalues that contribute to large sampling variance, but only for those regressors that have large coefficients associated with the corresponding "short" principal components.

https://doi.org/10.26219/heal.aueb.6167

Principal components analysis is considered a remedy for multicollinearity since we can calculate one or several principal components on the set of collinear variables and use the components in the regression instead of the original variables. A possible problem, though, is a possible lack of interpretability of the components.

2.6.2 Variable Selection

When a number of variables in a regression analysis do not appear to contribute significantly to the predictive power of the model, or when the regressors are highly correlated, it is natural to try to find a suitable subset of important or useful variables. An optimum subset model is one that, for a given number of variables, produces the minimum error sum of squares, or, equivalently, the maximum R^2. The only way to ensure finding optimum subsets is to examine all possible subsets. Fortunately, high-speed computing capabilities make such a procedure feasible for models with a moderate number of variables.

When the examination of the R-square values does not reveal any obvious choices for selecting the most useful model, we can use instead a number of other statistics. Among these, the most frequently used is the C_p statistic, proposed by Mallows (see Fox, 1997). This statistic is a measure of total squared error for a subset model containing p independent variables. The total squared error is a measure of the error variance plus the bias introduced by not including important variables in a model. It may, therefore, indicate whether variable selection is deleting too many variables. The C_p statistic is computed as follows:

$$C_p = \frac{SSE(p)}{MSE} - (T - 2p) + 1$$

where MSE is the error mean square for the full model, $SSE(p)$ is the error sum of squares for the subset model containing p independent variables (not including the intercept), and T is the total sample size. It is recommended that C_p be plotted against

https://doi.org/10.26219/heal.aueb.6167

p, and further select that subset size where the minimum C_p first approaches $(p+1)$, starting from the full model.

A number of other statistics are available to assist in choosing subset models. Some are relatively obvious, such as the residual mean square or standard deviations, while others are related to R^2, with some providing adjustments for degrees of freedom. Subset techniques have the advantage of revealing alternative, nearly equivalent models, and thus avoid the misleading appearance of producing a uniquely "correct" result.

Popular alternatives to the guaranteed optimum subset selection are the stepwise procedures that add or delete variables one at a time until, by some criterion based on R^2, a reasonable stopping point is reached. These selection methods do not guarantee finding optimum subsets, but they work quite well in many cases and are especially useful for models with many variables. Such selection methods are:

Backward selection; Starts with a full regression equation that includes all the independent variables. The R^2 induced from deleting each independent variable, or the partial F test value for each independent variable treated as though it were the last variable to enter the regression equation, is calculated. The lowest partial F test value (which corresponds to the variable that contributes least to the fit of the model) is compared with a predetermined critical tabulated F-value. If this partial F value is smaller than the tabulated F-value we delete it and examine the regression with the remaining independent variables. The procedure stops when all coefficients remaining in the model are statistically significant. Note, that the decision rule is irreversible; once a variable has been deleted, it is deleted permanently

Forward selection: The process begins with the inclusion of the variable with the largest correlation with the dependent variable. Next, variables are entered according to their squared partial correlation controlling for those variables already in the model. The process continues until no variable considered for addition to the model provides a reduction in sum of squares considered statistically significant at the predetermined level. An important feature of this method is that once a variable has been selected, it stays in the model

https://doi.org/10.26219/heal.aueb.6167

Stepwise selection: It begins like forward selection but differs in that the decision to include a predictor is not irreversible.
For more information see Dillon and Goldstein (1984). A technical objection to stepwise methods is that they can fail to return the optimal subset of regressors of a given size.

In applying variable selection, it is essential to keep in mind the following: Variable selection is a good strategy when the variables are orthogonal or nearly so. On the contrary when the variables are highly correlated or include curvilinear effects of other variables this is not a promising method. In these cases biased estimation has proven to be a good solution as it is better to use a part of all the variables than all of some variables and none of the remaining ones.

2.6.3 Biased Estimation

Another approach to collinear data is biased estimation. Least-squares estimators provide unbiased estimates of parameters. The essential idea here is to trade a small amount of bias in the coefficient estimates for a substantial reduction in coefficient sampling variance. The precision of a biased estimate, called the mean squared error, is the square of the bias plus the variance. The hoped-for result is a smaller mean-squared error of estimation of the β's than is provided by the least-squares estimates.

Ridge regression

The most common biased estimation method is **ridge regression**. Hoerl and Kennard (1970) proved that it is always possible to choose a positive value of a constant, namely the ridge constant, so that the mean-squared error of the ridge estimator is less than the mean-squared error of the least-squares estimator. Their equation of the ridge estimate is

$$\hat{\beta}(k) = (X'X + kI)^{-1} X'y, \qquad (2.6.2)$$

where $k \geq 0$ is the nonstochastic quantity, ($\hat{\beta}(0) = \beta$ is the ordinary LS estimator) and I is the identity matrix.

https://doi.org/10.26219/heal.aueb.6167

Ridge regression involves the arbitrary selection of a "ridge constant" which controls the extent to which ridge estimates differ from the least-squares estimates: the larger the ridge constant, the greater the bias and the smaller the variance of the ridge estimator. The vital issue therefore is to find a value of k for which the trade-off of bias against variance is favorable. Unfortunately, to pick the optimal ridge constant generally requires knowledge about the unknown β's that we are trying to estimate (Fox, 1997).

A number of methods have been proposed for selecting the constant k. One very popular method is to compute ridge regression estimates for a set of values of k starting with $k = 0$ (the unbiased estimate) and to plot these coefficients against k (Ridge Trace). As the value of k increases from zero, the coefficients involved in multicollinearities change rapidly. However, as k increases further, these coefficients change more slowly. The selection of k is done by examining such a plot and choosing that value of k where the coefficients settle down. We will present analytically ridge regression in the next chapter.

Shrinkage estimators

Shrinkage estimators are of the form

$$\hat{\beta}_{SH} = s\hat{\beta} \tag{2.6.3}$$

where $0 \le s \le 1$ is a deterministically or stochastically chosen constant. The only known shrinkage estimator (SH) with a stochastically chosen value of c possessing any optimal properties is the estimator due to James and Stein (see Gunst and Mason, 1977). Provided $p \ge 3$ and $X'X = I$, the SH estimator is given by (2.6.3) with

$$s = \max\left\{0, \left(1 - \frac{c\hat{u}'\hat{u}}{\hat{\beta}'\hat{\beta}}\right)\right\} \tag{2.6.4}$$

where $0 < c < 2(p-2)/(v+2)$, $\hat{u}'\hat{u}$ is the residual sum of squares using $\hat{\beta}$ to predict the response and v is the number of degrees of freedom on which $\hat{u}'\hat{u}$ is based. The estimator $\hat{\beta}_{SH}$ with s given by (2.6.4) has smaller MSE than LS. Moreover, the $MSE(\hat{\beta}_{SH})$ is minimized for s given by (2.6.4) if $c = (p-2)/(v+2)$.

https://doi.org/10.26219/heal.aueb.6167

The drawbacks of SH are the requirements that $p \geq 3$ and $X'X = I$; and as Gunst and Mason comment this eliminates most of the cases met in practice.

Generalized inverse estimators (Marquardt, 1970)

Since the matrix $X'X$ is singular, an option is to invert it by means of a generalized inverse. Let the diagonalized matrix be denoted *D*, with ordered diagonal elements $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ and the eigenvector matrix that transforms $X'X$ into *D* be denoted *S*.

Thus,

$$S'X'XS = D$$

where $S'S = I$. Then

$$(X'X)^{-1} = SD^{-1}S'.$$

Suppose $X'X$ is of rank *r*, so that the last $(p-r)$ ordered elements of *D* are zero (or nearly so; if $X'X$ is only "nearly singular"). Partition *S* as follows:

$$S = (S_r : S_{p-r})$$

where S_r is $(p \times r)$; S_{p-r} is $(p \times (p-r))$ and then partition *D* as

$$D = \begin{bmatrix} D_r & \vdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \vdots & D_{p-r} \end{bmatrix}$$

where D_r is $(r \times r)$; D_{p-r} is $((p-r) \times (p-r))$.

Now, by assumption, D_{p-r} is zero, so that $D_{p-r}^{-1} = 0$ by definition. Thus, the inverse becomes

$$(X'X)_r^+ = S_r D_r^{-1} S_r'. \tag{2.6.5}$$

A class of generalized inverse regression estimators is defined by

$$\hat{\beta}^+ = (X'X)_r^+ X'Y. \tag{2.6.6}$$

In general, there is an "optimum" value for *r* for any problem, but it is desirable to examine the generalized inverse solution for a range of admissible values for *r* (see Rao, C.R. and Toutenburg, H. 1999)

https://doi.org/10.26219/heal.aueb.6167

2.6.4 Prior Information about the Regression Coefficients

Another approach to estimation with collinear data is to introduce additional prior information that reduces the ambiguity produced by collinearity. In a Bayesian framework the incorporation of prior information is achieved as usual by the use of a prior density function upon the parameter vector β. For the Bayesian, a singular or near-singular $X'X$ matrix causes no special problems. The difficulty that Bayesians have when the data are collinear is that the posterior distribution becomes very sensitive to changes in the prior distribution (Judge et al., 1985).

2.6.5 Partial Least Squares

Partial Least Squares (PLS) is a method for constructing predictive models when the variables are too many and highly collinear (Tobias, 1999). Like principal component analysis, the basic idea of PLS is to extract several latent factors and responses from a large number of observed variables. More specifically, the aim is to predict the response by a model that is based on linear transformations of the explanatory variables. The regression models are of the following type

$$\hat{Y} = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p, \tag{2.6.7}$$

where the Z_i are linear combinations of the explanatory variables $X_1, X_2, \ldots, X_k$ such that the sample correlation coefficient for any pair Z_i, Z_j $i \neq j$ is 0. The simple consequence of this feature is that parameters β_k in equation (2.6.7) may be estimated by simple univariate regressions of Y against Z_k (Rao and Toutenburg, 1999). It is important to note that in PLS the emphasis is on prediction rather than explaining the underlying relationships between the variables. Note also that unlike an ordinary least squares regression, PLS can accept multiple dependent variables.

https://doi.org/10.26219/heal.aueb.6167

2.7 *Multicollinearity with Stochastic Regressors*

Consider the linear regression model with stochastic regressors

$$y = X\beta + u,$$

where y is a $(T \times 1)$ vector of observations, X is now a $(T \times p)$ **stochastic matrix**, β is a $(p \times 1)$ vector of unknown parameters, and u is a $(T \times 1)$ vector of errors that is distributed independently of X so that $E(u|X) = 0$ and $E(uu'|X) = \sigma^2 I$.

First, we could analyze the sample design as if X were nonstochastic with all results conditional on the values of the sample actually drawn. Multicollinearity can then be properly analyzed as a feature of the sample, not the population. This is usually the approach followed.

On the other hand, if we are willing to assume that the X_i are normally and independently distributed, the tests of Farrar and Glauber are available and confidence statements can be made. Wichers (see Farrar and Glauber, 1976) proposes a modification of the Farrar and Glauber tests designed to identify the nature of the linear dependencies. Alternatively, we may test hypotheses about the characteristic roots, which are now themselves stochastic. Note that we are not testing for the singularity or nonsingularity of X, for if exact linear constraints were obeyed in the population, the sample would obey those constraints with probability one and $X'X$ would be singular. Thus, we are testing only whether or not there is little independent variation within a set of explanatory variables.

Given the assumption of the stochastic regressor model, the search for improved estimators becomes difficult. Although little has been done in this area some sampling experiments indicate that the Stein-like estimators may do well when the covariance matrix is estimated rather than known. Consider the situation where y and X are jointly normal. Under this model and if the loss function is the mean square error of prediction, an estimator was found that dominates the usual maximum likelihood estimator (Judge et al., 1985).

https://doi.org/10.26219/heal.aueb.6167

2.8 *Multicollinearity and Prediction*

In general, regression models are used for the related purposes of description and estimation, i.e. the description of the relationship between y and X and the accurate estimation of the value of individual coefficients, or the purpose of prediction, i.e. the prediction of the value of the dependent variable in a future period.

Multicollinearity is a problem if we are using regression for description or estimation. When multicollinearity is present one cannot examine the individual effects of each explanatory variable. If the purpose is the estimation of individual coefficients, either the inclusion or the exclusion of intercorrelated variables will not help, because the estimates in both cases will most probably be imprecise. In this case the only real improvement in the estimate is to use additional information, for example extraneous estimates, larger samples, and so on.

If the purpose of the estimation is to predict the values of the dependent variable, then we may include the intercorrelated variables and ignore the problems of multicollinearity, provided that we are certain that the same pattern of intercorrelation of the explanatory variables will continue in the period of prediction (Koutsoyiannis, 1977). This is because multicollinearity will not affect the forecasts of a model but only the weights of the explanatory variables in the model.

https://doi.org/10.26219/heal.aueb.6167

CHAPTER 3

RIDGE REGRESSION

3.1 Introduction

The ridge regression procedure is based on the matrix $(X'X + kI)$, I denoting the identity matrix and k being a positive scalar parameter. It is a procedure that can be used in "ill-condition" situations where correlations between the various predictors in the model cause the $X'X$ matrix to be close to singular. In particular, we can obtain a point estimate with a smaller mean square error.

Hoerl and Kennard (1970) suggested that in order to control inflation and general instability associated with the least squares estimates, one can use

$$\hat{\beta}(k) = (X'X + kI)^{-1} X'y; \quad k \geq 0 \tag{3.1.1}$$

Note that the LS estimator is a member of this family with $k = 0$.

The ridge estimator, though biased, has lower mean square error than the BLUE (best linear unbiased estimator). Unfortunately, this mean-squared error is a function of the unknown parameters that we are trying to estimate. Let us denote the mean square error (MSE) of a biased estimator $\hat{\beta}^*$ of β as:

$$MSE(\hat{\beta}^*) = E(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta) \tag{3.1.2}$$

Since the squared Euclidean distance between $\hat{\beta}^*$ and β is

$$L^2 = (\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta), \tag{3.1.3}$$

https://doi.org/10.26219/heal.aueb.6167

the $MSE(\hat{\beta}^*)$ can be interpreted as the mean squared Euclidean distance between the vectors $\hat{\beta}^*$ and β (Koutsoyiannis, 1977). Thus, an estimator with low MSE will be close to the true parameter.

One property of the least squares estimator $\hat{\beta}$ that is frequently noted in the ridge regression literature is (Judge et al., 1985)

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 tr(X'X)^{-1} > \beta'\beta + \frac{\sigma^2}{\lambda_k}, \tag{3.1.4}$$

where λ_k is the minimum eigenvalue of $X'X$. Thus, if the data are collinear, and λ_k is small, this implies that the expected squared length of the least squares coefficient vector is greater than the squared length of the true coefficient vector. In addition, the smaller the λ_k, the greater the difference.

3.2 *The Reparameterized model*

Let us begin with the linear regression model as given in (2.2.1). We assume that the data are in standardized form and compute the correlation matrix, and the correlation coefficients between the dependent variable and the predictors, i.e. we compute $X'y$. A parameterization that is popular in ridge regression is the one that is based on the singular value decomposition of X. The matrix X can be written as

$$X = Q\Lambda^{1/2}P', \tag{3.2.1}$$

where Q is a $(T \times p)$ matrix of the coordinates of the observations along the principal axes of X standardized in the sense that $Q'Q = I$. The matrix Λ is a diagonal matrix of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, that is,

https://doi.org/10.26219/heal.aueb.6167

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ 0 & 0 & 0 & \dots & \lambda_p \end{bmatrix}.$$

The P matrix is the $(p \times p)$ matrix of eigenvectors satisfying $X'X = P\Lambda P'$, and $P'P = I$. Then the regression model can be written as follows:

$$y = X\beta + u = XP'P\beta + u = X^{*}\alpha + u, \tag{3.2.2}$$

which defines a parameter vector $\alpha = P'\beta$, and $X^{*} = XP$. The OLS estimate of α is denoted by $\hat{\alpha}$ and is

$$\hat{\alpha} = \left(X^{*'}X^{*}\right)^{-1} X^{*'}y = (P'X'XP)^{-1}P'X'y = (P'X'XP)^{-1}P'X'X\hat{\beta}$$

$$= (P'X'XP)^{-1}P'X'XPP'\hat{\beta} = (P'X'XP)^{-1}(P'X'XP)P'\hat{\beta}$$

$$= P'\hat{\beta}. \tag{3.2.3}$$

As we showed in chapter 1 the variance of the OLS estimator is

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} = \sigma^2 P\Lambda^{-1}P',$$

while

$$V(\hat{a}) = P'V(\hat{\beta})P = \sigma^2 P'P\Lambda^{-1}P'P$$

$$= \sigma^2\Lambda^{-1}. \tag{3.2.4}$$

The elements $\hat{\alpha}_i$ are called "uncorrelated components" because $V(\hat{a}) = \sigma^2\Lambda^{-1}$ is diagonal. Since the ridge estimator of α is

$$\hat{\alpha}(k) = (\Lambda + kI)^{-1}P'X'y, \tag{3.2.5}$$

we can easily obtain the relationship

$$\hat{\beta}(k) = (X'X + kI)^{-1}X'y = (P\Lambda P' + kI)^{-1}X'y = (P\Lambda P' + kPP')^{-1}X'y$$

https://doi.org/10.26219/heal.aueb.6167

$$= P(\Lambda + kI)^{-1} P'X'y = P\hat{\alpha}(k). \tag{3.2.6}$$

We can also find from the above the relationship between the ridge and the ordinary estimate

$$\hat{\beta}(k) = P(\Lambda + kI)^{-1} P'X'y = P(\Lambda + kI)^{-1} P'X'X\hat{\beta}$$

$$= P(\Lambda + kI)^{-1} \Lambda P'\hat{\beta} = P\Delta P'\hat{\beta}, \tag{3.2.7}$$

where $\Delta = diag(\delta_i); \delta_i = \lambda_i(\lambda_i + k)^{-1}$, i = 1,..., p is a diagonal matrix of "shrinkage factors".

We must warn the user of ridge regression that the direct ridge estimators based on the model before standardization do not coincide with their unstandardized counterparts based on model (2.2.1) (Vinod, 1978).

3.3 *Hoerl and Kennard's Reasoning*

If B is an estimate of the vector β, the residual sums of squares is given by

$$\phi = (y - XB)'(y - XB)$$

$$= \left(y - X\hat{\beta} + X(\hat{\beta} - B)\right)'\left(y - X\hat{\beta} + X(\hat{\beta} - B)\right)$$

$$= \left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right) + \left(\hat{\beta} - B\right)' X'X\left(\hat{\beta} - B\right)$$

$$= \phi_{\min} + \phi(B), \tag{3.3.1}$$

since $2\left(y - X\hat{\beta}\right)' X\left(\hat{\beta} - B\right) = 2y'\left(I - X(X'X)^{-1}X'\right)X\left(\hat{\beta} - B\right) = 0$; $\phi_{\min}$ is the residual sums of squares of the OLS.

Let $\phi_0 > 0$ be a fixed value for the error sum of squares. Then there is a set of values of B_0 that will satisfy the relationship $\phi = \phi_{\min} + \phi_0$. In this set we look for the estimate that has the minimum length. This can be stated as minimize $B'B$

https://doi.org/10.26219/heal.aueb.6167

$$\text{subject to } \left(B-\hat{\beta}\right)' X'X\left(B-\hat{\beta}\right)=\phi_0 . \tag{3.3.2}$$

As a Lagrangian problem this is

$$\text{minimize } F = B'B + (1/k)\left[\left(B-\hat{\beta}\right)' X'X\left(B-\hat{\beta}\right)-\phi_0\right]$$

where $(1/k)$ is the multiplier. Then

$$\frac{\partial F}{\partial B} = 2B + (1/k)\left[2(X'X)B - 2(X'X)\hat{\beta}\right] = 0 .$$

Solving for B we obtain $B = \hat{\beta}(k) = (X'X + kI)^{-1} X'y$; k is determined so that (3.3.2) is fulfilled. From (3.3.1) and the relationship $\hat{\beta}(k) = (X'X + kI)^{-1} X'X\hat{\beta}$ the residual sum of squares $\hat{\beta}(k)$ is equal to $\phi(k) = \left(y - X\hat{\beta}(k)\right)'\left(y - X\hat{\beta}(k)\right) =$

$$\left(y - X\hat{\beta} + \left(X\left(\hat{\beta}\right) - \hat{\beta}(k)\right)\right)'\left(y - X\hat{\beta} + \left(X\left(\hat{\beta}\right) - \hat{\beta}(k)\right)\right) =$$

$$\phi_{\min} + \left(\hat{\beta} - \hat{\beta}(k)\right)' X'X\left(\hat{\beta} - \hat{\beta}(k)\right),$$

which after simple calculations becomes equal to $\phi_{\min} + k^2 \hat{\beta}(k)' (X'X)^{-1} \hat{\beta}(k)$ (Hoerl and Kennard, 1970).

3.4 *Properties of the Ridge Estimator*

As shown above, Hoerl and Kennard's definition of the ridge estimate is

$$\hat{\beta}(k) = (X'X + kI)^{-1} X'y ,$$

with $k \geq 0$ being the ridge parameter. Using the abbreviation $G_k = (X'X + kI)^{-1}$ and $Z_k = G_k X'X$, we can write the ridge estimate as

$$\hat{\beta}(k) = G_k X'y = G_k X'X\hat{\beta} = Z_k \hat{\beta} . \tag{3.4.1}$$

https://doi.org/10.26219/heal.aueb.6167

Following, we present some properties of the ridge estimator.

A) Let $\xi_i(G_k)$ and $\xi_i(Z_k)$ be the eigenvalues of G_k and Z_k, respectively. Then

$$\xi_i(G_k) = 1/(\lambda_i + k) \tag{3.4.2}$$

$$\xi_i(Z_k) = \lambda_i/(\lambda_i + k). \tag{3.4.3}$$

B) The ratio of the largest characteristic root of the design matrix $(X'X + kI)$ to the smallest root is $(\lambda_1 + k)/(\lambda_p + k)$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ are the ordered roots of $X'X$, and is a decreasing function of k.

C) $\hat{\beta}(k)$ for $k \neq 0$ is shorter than $\hat{\beta}$, i.e.

$$\left(\hat{\beta}(k)\right)'\left(\hat{\beta}(k)\right) < \hat{\beta}'\hat{\beta}. \tag{3.4.4}$$

Recall (3.4.1) and since Z_k is symmetric positive definite the following holds (Hoerl and Kennard, 1970):

$$\left(\hat{\beta}(k)\right)'\left(\hat{\beta}(k)\right) \leq \xi^2_{\max}(Z_k)\hat{\beta}'\hat{\beta}.$$

Since $\xi_{\max}(Z_k) = \lambda_1/(\lambda_1 + k)$ then (3.4.4) is verified (Hoerl and Kennard, 1970).

For $\hat{\beta}(k)$ the residual sum of squares can be written as

$$\phi(k) = \left(y - X\hat{\beta}(k)\right)'\left(y - X\hat{\beta}(k)\right) = \left(y' - \left(\hat{\beta}(k)\right)' X'\right)\left(y - X\hat{\beta}(k)\right) =$$

$$y'y - \left(\hat{\beta}(k)\right)' X'y - \left(y'X - \left(\hat{\beta}(k)\right)' X'X\right)\hat{\beta}(k).$$

From the definition of the ridge estimator we can replace the quantity $y'X$ above with $\left(\hat{\beta}(k)\right)'(X'X + kI)$ so the residual sum of squares becomes

$$\phi(k) = y'y - \left(\hat{\beta}(k)\right)' X'y - k\left(\hat{\beta}(k)\right)'\left(\hat{\beta}(k)\right).$$

This way the residual sum of squares can be described as the total sum of squares minus the "regression" sum of squares for $\hat{\beta}(k)$ with a modification analogous to the squared length of $\hat{\beta}(k)$.

https://doi.org/10.26219/heal.aueb.6167

Mean, bias and variance

The *mean* of the ridge estimator is given by

$$E\left(\hat{\beta}(k)\right) = G_k X'E(y)$$

$$= G_k X'X\beta = Z_k \beta . \qquad (3.4.5)$$

Note that when $k = 0$ then $Z_k = I$ and hence $E\left(\hat{\beta}(k)\right) = \beta$, but when $k \neq 0$, $\hat{\beta}(k)$ provides a biased estimate of β.

The *bias* of the estimator $\hat{\beta}(k)$ is given by

We know that the bias of an estimator b^* is defined as

$$Bias\left(b^*\right) = E\left(b^*\right) - \beta .$$

So the bias of $\hat{\beta}(k)$ is

$$Bias\left(\hat{\beta}(k)\right) = E\left(\hat{\beta}(k)\right) - \beta = Z_k \beta - \beta$$

$$= \left[(X'X + kI)^{-1} X'X - I\right]\beta$$

$$= (X'X + kI)^{-1}\left[X'X - (X'X + kI)\right]\beta$$

$$= -kG_k \beta \qquad (3.4.6)$$

or alternatively from the relationship between the ridge and the ordinary estimate (3.2.7)

$$Bias\left(\hat{\beta}(k)\right) = E\left(\hat{\beta}(k)\right) - \beta$$

$$= P(\Lambda + kI)^{-1}\Lambda P'\beta - \beta$$

$$= \left[P(\Lambda + kI)^{-1}\Lambda P' - I\right]\beta$$

$$= \left[P(\Lambda + kI)^{-1}\Lambda P' - PP'\right]\beta$$

$$= P\left[(\Lambda + kI)^{-1}\Lambda - I\right]P'\beta$$

$$= P\left[(\Lambda + kI)^{-1}\Lambda - (\Lambda + kI)^{-1}(\Lambda + kI)\right]P'\beta$$

$$= P(\Lambda + kI)^{-1}(\Lambda - \Lambda - kI)P'\beta$$

$$= -kP(\Lambda + kI)^{-1}P'\beta \qquad (3.4.7)$$

https://doi.org/10.26219/heal.aueb.6167

Now it useful to give the squared bias (in its matrix version)

$$Bias\left(\hat{\beta}(k)\right)Bias\left(\hat{\beta}(k)\right)' = k^2 P(\Lambda + kI)^{-1} P'\beta\beta' P(\Lambda + kI)^{-1} P' \tag{3.4.8}$$

The *variance-covariance matrix* for the ridge regression estimators is

$$\operatorname{cov}\left(\hat{\beta}(k)\right) = \operatorname{cov}\left(Z_k\hat{\beta}\right) = Z_k \operatorname{cov}\left(\hat{\beta}\right) Z_k'$$

$$= \sigma^2 Z_k (X'X)^{-1} Z_k' = \sigma^2 Z_k G_k . \tag{3.4.9}$$

Alternatively, we can write (3.4.9) using the matrices P and Λ. Since

$$\hat{\beta}(k) = P(\Lambda + kI)^{-1}\Lambda P'\hat{\beta}, \text{ then}$$

$$\operatorname{cov}\left(\hat{\beta}(k)\right) = P(\Lambda + kI)^{-1}\Lambda P'\sigma^2 (X'X)^{-1} P\Lambda(\Lambda + kI)^{-1} P'$$

$$= \sigma^2 P(\Lambda + kI)^{-1}\Lambda(\Lambda + kI)^{-1} P' \tag{3.4.10}$$

$$= \sigma^2 P \begin{bmatrix} \frac{\lambda_1}{(\lambda_1 + k)^2} & 0 & \cdots & 0 \\ 0 & \frac{\lambda_2}{(\lambda_2 + k)^2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{\lambda_p}{(\lambda_p + k)^2} \end{bmatrix} P' .$$

3.5 *Mean Squared Error Properties*

We denoted in (3.1.2) the MSE of an estimator as the mean Euclidean distance between the estimator and the true value. MSE is also defined as the trace of the mean dispersion error matrix (Rao and Toutenburg 1999). The mean dispersion error matrix is

$$M\left(\hat{\beta}, \beta\right) = E\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'$$

$$= E\left(\hat{\beta} - E\left(\hat{\beta}\right) + E\left(\hat{\beta}\right) - \beta\right)\left(\hat{\beta} - E\left(\hat{\beta}\right) + E\left(\hat{\beta}\right) - \beta\right)'$$

$$= V\left(\hat{\beta}\right) + Bias\left(\hat{\beta}\right)Bias\left(\hat{\beta}\right)' \tag{3.5.1}$$

https://doi.org/10.26219/heal.aueb.6167

Therefore, $MSE(\hat{\beta}) = tr\{M(\hat{\beta},\beta)\} = tr[V(\hat{\beta})] + [Bias(\hat{\beta})]'[Bias(\hat{\beta})]$. (3.5.2)

For instance, recalling (3.1.2) the MSE of the OLS estimator is:

$$\text{MSE}= E(L^2) = tr(V(\hat{\beta})) = \sigma^2 tr(X'X)^{-1}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{3.5.3}$$

where λ_i is the ith eigenvalue of $X'X$. In the case of the ridge estimator we have from (3.4.8) and (3.4.10)

$$MSE(\hat{\beta}(k)) = tr\left\{V(\hat{\beta}(k)) + Bias(\hat{\beta}(k))Bias(\hat{\beta}(k))'\right\}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\beta_i^2}{(\lambda_i + k)^2}$$

or

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \beta'(X'X + kI)^{-2}\beta$$

$$= \gamma_1(k) + \gamma_2(k). \tag{3.5.4}$$

Hoerl and Kennard (1970) proved that $\gamma_1(k)$ is a monotonic decreasing function of k, while $\gamma_2(k)$ is monotonic increasing. In addition, $\gamma_2(k)$ can be considered the square of a bias introduced when $\hat{\beta}(k)$ is used instead of $\hat{\beta}$ while $\gamma_1(k)$ can be shown to be the sum of the variances of the parameter estimates. The sum of the variances of all $\hat{\beta}_i(k)$'s is the sum of the diagonal elements of (3.4.10). Note that since $X'X = P\Lambda P'$ then $\gamma_2(k)$ can be written as

$$\gamma_2(k) = k^2 \sum_{1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \tag{3.5.5}$$

where $\alpha = P'\beta$.

https://doi.org/10.26219/heal.aueb.6167

3.6 Existence Theorems

The main justification for ridge regression by Hoerl and Kennard is their theorem that there always exists a $k > 0$ such that $E[L^2(k)] < E[L^2(0)] = \sigma^2 \sum_1^p (1/\lambda_i)$, where $L^2(k)$ is the Euclidean distance between the ridge estimator and β while $L^2(0)$ is the Euclidean distance between the OLS and β. To see this from (3.5.3) (3.5.4) and (3.5.5)

$$\frac{dE[L^2(k)]}{dk} = \frac{d\gamma_1(k)}{dk} + \frac{d\gamma_2(k)}{dk}$$

$$= -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^3} + 2k \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3} \qquad (3.6.1)$$

As mentioned in the previous paragraph $\gamma_1(k)$ and $\gamma_2(k)$ are monotonically decreasing and increasing and thus their first derivatives are always non-positive and non-negative, respectively. So the result can be proved if we can show that there always exists a $k > 0$ such that $\frac{dE[L^2(k)]}{dk} < 0$. And this holds when

$$k < \sigma^2 / \alpha_{\max}^2 \qquad (3.6.2)$$

where $\alpha_{\max}^2$ is the squared value of the larger α_i. In most applications, interesting values of k usually lie in the range (0, 1). For standardized variables, this is always the case.

The difficulty in the above result is that k depends on σ^2 and β, neither of which is known. Thus although k exists, we do not know whether or not we have attained a value for k which provides a lower MSE than that of LS in a specific practical problem (Draper and Smith, 1981).

In Hoerl and Kennard's existence theorem the mean square error of $\hat{\beta}(k)$ has been compared with $\sigma^2 tr(X'X)^{-1} = \sigma^2 \sum_{i=1}^{p} (1/\lambda_i)$. Banerjee and Carr (1971) suggested comparing it with

https://doi.org/10.26219/heal.aueb.6167

$$\sigma^2 tr(X'X + kI)^{-1} = \sigma^2 \sum_{1}^{p} 1/(\lambda_i + k), \tag{3.6.3}$$

and not to the larger quantity $\sigma^2 \sum_{i=1}^{p} (1/\lambda_i)$. In order to explain their suggestion, Banerjee and Carr (1971) introduced (see appendix B) the augmented model:

$$\begin{bmatrix} y_X \\ \cdots \\ y_A \end{bmatrix} = \begin{bmatrix} X \\ \cdots \\ k^{1/2} I_p \end{bmatrix} \begin{bmatrix} \beta \end{bmatrix} + u, \tag{3.6.4}$$

where y_x is the original y, y_A is a $(p \times 1)$ observation vector corresponding to the augmented part, I_p is a $(p \times p)$ identity matrix, and u is $(n+p) \times 1$ error vector. In addition, we have $E(y_X) = X\beta$ and $E(y_A) = \sqrt{k}\beta$. The least squares estimate of β in the augmented model is

$$\begin{aligned} \hat{\beta}_A &= (X'X + kI)^{-1}(X'y + \sqrt{k} y_A) \\ &= \hat{\beta}(k) + \sqrt{k}(X'X + kI)^{-1} y_A. \end{aligned}$$

For the augmented model, the authors have proved a corresponding "existence theorem". There always exists a $k > 0$ such that $E[L^2(k)] < \sigma^2 \sum_{i=1}^{p} 1/(\lambda_i + k)$. For the proof see details in Banerjee and Carr (1971). It is interesting to note that the same condition for k was obtained in the augmented model, namely $k < \sigma^2/\alpha_{\max}^2$, where $\alpha_{\max}^2$ is the largest component of α.

Coniffe and Stone (1973) comment that only if the appropriate value of k is assumed known is the proof of Hoerl and Kennard's existence theorem valid. What is important is whether the estimator with *estimated k* has better mean square error properties than least squares estimators. They also note that mean square error is not the only criterion that determines the quality of a particular estimator. Other criteria, such as that of having a tractable distribution, are also important.

https://doi.org/10.26219/heal.aueb.6167

3.7 Generalized Ridge Estimator

In Vinod and Ullah (1981) one can find the definition of a generalized ridge estimator (GRE) of α (as given in (3.2.3)). It is obtained by augmenting the i_{th} diagonal element of Λ by a positive constant k_i, and using the singular value decomposition of X

$$\alpha_K = (\Lambda + K)^{-1} \Lambda^{1/2} Q'y, \tag{3.7.1}$$

where $K = diag(k_i)$ is a diagonal matrix. The GRE of β in (2.2.1) can be written as

$$\begin{aligned} b_K &= P\alpha_K = P(\Lambda + K)^{-1} \Lambda^{1/2} Q'y \\ &= P(P'X'XP + K)^{-1} \Lambda^{1/2} Q'y \\ &= P(P'X'XP + P'PKP'P)^{-1} \Lambda^{1/2} Q'y \\ &= (X'X + PKP')^{-1} X'y. \end{aligned}$$

Alternatively it can be written as

$$b_K = (X'X + PKP')^{-1} X'y = P\Delta P'\hat{\beta}, \tag{3.7.2}$$

where $\Delta = diag(\delta_i)$, the diagonal matrix of $\delta_i = \lambda_i(\lambda_i + k_i)^{-1}$.

Guilkey and Murphy (1975) considered a modification of the GRE which they called "Direct Ridge Estimator" (DRE). They suggest that only the diagonal elements of Λ corresponding to relatively small eigenvalues (λ_i is defined as small if $\lambda_i < 10^{-c}\lambda_{\max}$ where $\lambda_{\max}$ is the largest eigenvalue of $X'X$ and c arbitrary constant) of $X'X$ should be augmented by a k_i value. This DRE will result in an estimate of β, that is less biased than b_K, and in cases with severe multicollinearity DRE will have a smaller MSE than the GRE.

3.8 Ridge Trace

Hoerl and Kennard (1970) claimed that a method to select the "right" value of k is the ridge trace. The ridge trace is a two-dimensional plot of $\hat{\beta}_i(k)$ against k, where $\hat{\beta}_i(k)$ is the ridge estimate of β_i obtained using the fixed value k; it usually includes a plot of

https://doi.org/10.26219/heal.aueb.6167

$RSS(\hat{\beta}_k)$ against k. As k is increased, the estimates become smaller in absolute value, tending to zero as k tends to infinity. Hoerl and Kennard propose to choose the value where the "system" stabilizes.

Below we present the ridge trace for the Longley data (Appendix, Part 2); the lines present the ridge coefficients for values of $k = 0$ to $k = 0.1$

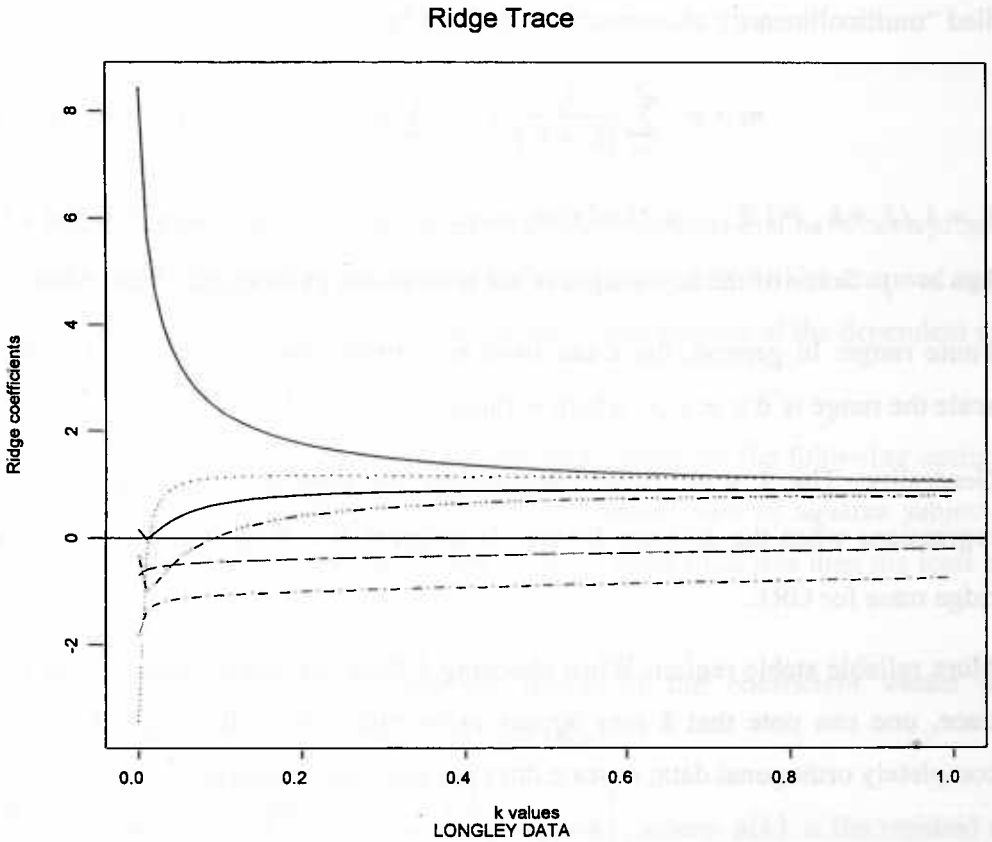


Figure 3.1 Ridge Trace

Hoerl and Kennard claimed that the *ridge trace* is a diagnostic tool that can help the analyst to estimate the value of k. However, since this procedure is based on the user's personal judgment it may be considered unreliable. Judge et. al. (1985) seem to doubt ridge trace as this "visual inspection" will lead to estimates of unknown properties. In addition, the ridge trace leads to a k which is a random variable and therefore the bias

https://doi.org/10.26219/heal.aueb.6167

introduced complicates the confidence intervals. They accept however, that one can learn from the data using the ridge trace.

3.8.1 An alternative Scaling for the Ridge Trace

Vinod (1976) has choosen another scaling on the horizontal axis for the ridge trace called "multicollinearity allowance", *m*, defined by

$$m = p - \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k_i)} = p - \sum_{i=1}^{p} \delta_i \qquad (3.8.1)$$

where $\delta_i = \lambda_i / \lambda_i + k$, i=1,2,..., p. Note that, when $k = k_1 = ... = k_p = 0$, $m = 0$ and when $k = \infty$ then $m = p$. Some of the advantages of the *m* scale are (Vinod and Ullah, 1981):

- Finite range: In general, the *k* can have an infinite range $0 \le k \le \infty$. For the *m* scale the range is $0 \le m \le p$, which is finite.
- Generality: The *k* scale ridge trace cannot be plotted for generalized ridge regressions when the k_i's are distinct. In contrast, it is simple to plot an *m*-scale ridge trace for GRE.
- More reliable stable region: When choosing *k* from the stable region of the ridge trace, one can note that *k* may appear to be more stable for larger *k* even for completely orthogonal data; *m* scale does not have this property.

3.8.2 Quantification of the concept of a Stable region

As discussed above the ridge trace may appear to be more stable for larger *k* even for completely orthogonal data; this is not the case for the *m* scale which will not give greater stability at larger *m*. It is this property of the *m* scale that suggested a numerical measure called Index of Stability of Relative Magnitudes (ISRM), defined for $m < p$

https://doi.org/10.26219/heal.aueb.6167

$$ISRM = \sum_i \left[\left(p\delta_i^2 / \bar{S}\lambda_i\right) - 1\right]^2, \tag{3.8.2}$$

where $\bar{S} = \frac{dm}{dk} = \sum \frac{\lambda_i}{(\lambda_i + k)^2}$. For completely orthogonal systems ISRM is equal to zero. It is possible to compute ISRM for each m ($< p$) and choose m where *ISRM* is the smallest. Vinod notes an important advantage of *ISRM*, that is not stochastic. The $\hat{\beta}_i(k)$ plotted in a ridge trace are stochastic and therefore k is a random variable.

3.9 Selecting k

In this section our aim is to bring together the methods that have been proposed in the literature and employed in practice for the selection of k. It will be assumed again that $X'X$ is in correlation form and $X'y$ is the vector of correlations of the dependent variable with each explanatory variable.

First we present two methods that are partly based on the following optimization problem: Ridge estimators should minimize the residual sum of squares subject to the constraint that the length of the coefficient vector is something less than the least squares length.

1) Hoerl (1962) proposed reducing the length of the coefficient vector without increasing the residual sum of squares. We take

$\frac{d^2\left(\phi(k)^{1/2}\right)}{dC^2} = \phi(k)^{-1/2}\left\{\frac{-(kC)^2}{\phi(k)} + \frac{C^2}{\left[\hat{\beta}'(k)G_k\hat{\beta}(k)\right]} - k\right\}$, where $\phi(k)$ is the residual sum of squares and $C^2 = \left(\hat{\beta}(k)\right)'\hat{\beta}(k)$ is the squared length of the vector. We choose the value k that yields the maximum value for the above derivative (Gibbons, 1981).

2) McDonald and Galarneau (1975). The choice of k is made in such a way that the squared length of the corresponding ridge estimator equals an estimated squared length of β.

$$Q = \hat{\beta}'\hat{\beta} - s^2 \sum_{j=1}^{p} \lambda_j^{-1} \tag{3.9.1}$$

where $s^2 = \dfrac{(y - X\hat{\beta})'(y - X\hat{\beta})}{(T - p - 1)}$. Choose k such that $(\hat{\beta}(k))'\hat{\beta}(k) = Q$, if $Q > 0$; choose $k = 0$ otherwise.

The next six estimators are based on the MSE property of ridge estimators:

3) Consider the general ridge estimator (Hoerl, Kennard and Baldwin, 1975) as given in (3.7.2), i.e.

$$b_K = (X'X + PKP')^{-1} X'y,$$

where $K = diag(k_1, k_2, ... k_p)$. The MSE function is minimized at $k_i = \sigma^2 / \alpha_i^2$ where $\alpha = P'\beta$. This optimal choice for k_i was also presented by Hoerl and Kennard (1970) and Goldstein and Smith (1974).

Hoerl et al. (1975) propose the use of the harmonic mean of these k_i to obtain a single value, namely k_h is given by

$$k_h = p\sigma^2 / \beta'\beta. \tag{3.9.2}$$

And using the estimates of σ^2 and β for the calculation of (3.9.2) we obtain

$$k_{HKB} = ps^2 / \hat{\beta}'\hat{\beta}. \tag{3.9.3}$$

4) Hoerl, Kennard, Baldwin, Thisted rule (see Lin and Kmenta, 1982)

$$k_{HKBM} = (p - 2)s^2 / \hat{\beta}'\hat{\beta}. \tag{3.9.4}$$

This estimator was suggested because the Hoerl, Kennard and Baldwin (HKB) estimator seems to overshrink towards zero.

5) Dwividi and Srivastava (1978) select k in a way similar to the one of Hoerl, Kennard and Baldwin by using

https://doi.org/10.26219/heal.aueb.6167

$$k = \frac{s^2}{\hat{\beta}'\hat{\beta}} \tag{3.9.5}$$

6) The optimal value of k_i for which the MSE of the almost unbiased generalized ridge regression estimator (AUGRR) proposed by Singh, Chaubey and Dwivedi (1986) is minimum is

$$k_i^* = \frac{\left\{\sigma^2 + \left(\sigma^4 + \sigma^2 \lambda_i \alpha_i^2\right)^{1/2}\right\}}{\alpha_i^2}$$

$$= \frac{\sigma^2}{\alpha_i^2}\left[1 + \left\{1 + \lambda_i\left(\frac{\alpha_i^2}{\sigma^2}\right)\right\}^{1/2}\right]. \tag{3.9.6}$$

In the case of the almost unbiased ordinary ridge regression estimator (AUORR) estimator where $k = k_1 = k_2 = \ldots = k_p$, we can obtain k by considering the harmonic mean of k_i^* in (3.9.6). It is given by

$$k^h = p\sigma^2 \Big/ \sum_{i=1}^{p}\left(\alpha_i^2 \Big/ \left\{1 + \left(1 + \lambda_i\left(\alpha_i^2/\sigma^2\right)\right)^{1/2}\right\}\right). \tag{3.9.7}$$

Since (3.9.7) depends on the unknown α and σ^2, we replace them by their OLS estimates. Therefore the parameter in (3.9.7) becomes

$$k_{HMO} = p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p}\left(\hat{\alpha}_i^2 \Big/ \left\{1 + \left(1 + \lambda_i\left(\hat{\alpha}_i^2/\hat{\sigma}^2\right)\right)^{1/2}\right\}\right), \tag{3.9.8}$$

where $\hat{\sigma}^2 = \dfrac{\left(Y - X^*\hat{\alpha}\right)'\left(Y - X^*\hat{\alpha}\right)}{(T-p)}$ and $\hat{\alpha}$ as given in (3.2.3).

7) If we assume that all k_i are equal to k, then the MSE function is minimized when $\sum \lambda_i\left(k\alpha_i^2 - \sigma^2\right)/(\lambda_i + k)^3 = 0$ (Dempster, Schatzoff, and Wermuth, 1977). The algorithm evaluates

$$\left|\sum \lambda_i\left(k\hat{\alpha}_i^2(k) - s^2\right)/(\lambda_i + k)^3\right|,$$

for values of k and selects that value of k that gives the observed minimum.

https://doi.org/10.26219/heal.aueb.6167

8) Consider the MSE of the ridge estimator as function of k, λ, α and σ, i.e.

$$MSE\left(\hat{\beta}(k)\right) = \gamma(k,\lambda,\alpha,\sigma) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2}$$

To get the best k Nordberg (1982) suggest to use "the empirical MSE-function" $\gamma(k,\lambda,\hat{\alpha},\hat{\sigma})$ where $\hat{\alpha}$ and $\hat{\sigma}$ are "good" estimators of α and σ. The procedure is the following:

(i) Choose a preliminary $k = k_0 \geq 0$.

(ii) Set $\hat{\alpha} = P'\hat{\beta}(k_0)$.

(iii) Set $\hat{\sigma}^2 = \frac{1}{T-p}\left|Y - X\hat{\beta}(0)\right|^2$,

(iv) Compute the k-value which minimizes the function $\gamma = (k,\lambda,\hat{\alpha},\hat{\sigma})$.

Denote by $K^*(k_0)$ the k-value obtained by the above procedure with k_0 as a "start value". By setting $k_0 = 0$ and by iterating the above procedure by $k_{j+1} = K^*(k_j)$, $j = 0,1,2,\ldots$ until it "stabilizes", i.e. until $k_{j+1} \approx k_j$ we can obtain good k-values.

9) As already shown the GRE is given by $\alpha_K = (\Lambda + K)^{-1}\Lambda^{1/2}Q'y$. Minimizing the MSE of the GRE term-by-term i.e., minimizing the diagonal elements of the mean squared error matrix

$$\sigma^2 \sum_{i=1}^{p} \lambda_i/(\lambda_i + k_i)^2 + k_i^2 \sum_{1}^{p} \alpha_i^2/(\lambda_i + k_i)^2 \qquad (3.9.9)$$

with respect to k_i yields the optimum value

$$k_{i(opt)} = \frac{\sigma^2}{\alpha_i^2} \quad (i = 1,2,\ldots,p). \qquad (3.9.10)$$

Hoerl and Kennard suggested to start with $\frac{S^2}{\hat{\beta}_i^2} = \hat{k}_i$ where $\hat{\beta}_i$ is the *ith* element of the least squares estimator and S^2 is an unbiased estimator of σ^2.

https://doi.org/10.26219/heal.aueb.6167

Replacing k_i in K by $\hat{k}_i$ to form $\hat{K}$ and substituting it in place of K in (3.9.9) leads to an adaptive estimator of α (Dwivedi et al., 1980): $\alpha_{ad} = \left(\Lambda + \hat{K}\right)^{-1} X^{*'} y$.

10) Obenchain (see Gibbons, 1981) considers a family of two-parameter estimators $\beta^*(k,q) = \left[(X'X)^{-q+1} + kI\right]^{-1} (X'X)^{-q} X'y$. For $q = 0$ we obtain the ridge estimator. In order to obtain the minimum mean squared error we choose q so as to maximize

$$C(q) = \frac{\sum_{i=1}^{p} |r_i| \lambda_i^{1-q/2}}{\left[\sum_{i=1}^{p} r_i^2 \sum_{i=1}^{p} \lambda_i^{(1-q)}\right]^{1/2}},$$

where $r = \Lambda^{-1/2} P'X'y$. The parameter k is then chosen so as to minimize $\tilde{L} = n \ln(2\pi e \tilde{\sigma}) + \xi'\xi - (r'\xi)/\tilde{\sigma}$, where $\xi_i = sign(r_i)[\delta_i/(1-\delta_i)]^{1/2}$, $\delta_i = \lambda / \left(\lambda_i + k\lambda_i^q\right)$ and $\tilde{\sigma} = 2\left\{ \left[(r'\xi)^2 + 4n\right]^{1/2} + (r'\xi)\right\}^{-1}$. Goldstein and Smith (1974) have considered an equivalent two-parameter estimator where the parameter $m=1-q$ is an integer.

Next we consider Bayesian approaches to the selection of k.

11) Lindley and Smith (1972) showed that if $Y \sim N\left(X\beta, \sigma^2 I\right)$ and the prior for β is $\beta \sim N\left(0, \sigma_\beta{}^2 I\right)$, then $\hat{\beta}(k)$ is the Bayes estimator where $k = \frac{\sigma^2}{\sigma_\beta^2}$. Since σ^2, the residual regression variance, and $\sigma^2{}_\beta$, the variance of the regression coefficients are usually both unknown we should estimate them and calculate k as follows:

$$k_{LS} = \frac{s^2}{s_\beta^2}. \tag{3.9.11}$$

12) Lawless and Wang (1976) also adopt a Bayesian approach and estimate the variance ratio by

$$k_{LW} = \frac{ps^2}{\sum \lambda_i \hat{\alpha}_i^2} \tag{3.9.12}$$

13) Dempster, Schatzoff, and Wermuth (1977) in a large simulation study suggested an estimator RIDGM, which is motivated by the Bayesian interpretation and is similar to the McDonald-Galarneau estimator. Given $\alpha \sim N(0, \sigma^2 I)$ it follows that

$$\sum_{i=1}^{p} \hat{\alpha}_i^2 / \sigma^2 [(1/k) + (1/\lambda_i)] \sim \chi_p^2 \tag{3.9.13}$$

where $k = \sigma^2/\omega^2$. Replacing σ^2 by s^2 and using the fact that $E(\chi_p^2) = p$, i.e.

$$\sum_{i=1}^{p} \hat{\alpha}_i^2 / \hat{\sigma}^2 [(1/k) + (1/\lambda_i)] = p . \tag{3.9.14}$$

The authors propose to use that value of k that satisfies (3.9.14).

Table 3.1 summarizes the different criteria.

https://doi.org/10.26219/heal.aueb.6167

aa	Criterion	Function to minimize-maximize	Reference
1		Choose k that yields the observed maximum of $\frac{d^2\left(\phi(k)^{1/2}\right)}{dC^2} = \phi(k)^{-1/2}\left\{\frac{-(kC)^2}{\phi(k)} + \frac{C^2}{\left[\hat{\beta}'(k)G_k\hat{\beta}(k)\right]} - k\right\}$	Hoerl, 1962 (in Gibbons, 1981)
2		Choose k such that $\left(\hat{\beta}(k)\right)'\hat{\beta}(k) = Q = \hat{\beta}'\hat{\beta} - s^2\sum_{j=1}^{p}\lambda_j^{-1}$, if $Q > 0$; choose $k = 0$ otherwise	McDonald and Galarneau (1975)
3	$k_{HKB} = ps^2/\hat{\beta}'\hat{\beta}$		Hoerl, Kennard and Baldwin (1975)
4	$k_{HKBM} = (p-2)s^2/\hat{\beta}'\hat{\beta}$		Hoerl, Kennard, Baldwin, Thisted (in Lin and Kmenta, 1982)
5	$k_{DS} = \frac{s^2}{\hat{\beta}'\hat{\beta}}$		Dwividi and Srivastava (1978)
6	$k_{HMO} = p\hat{\sigma}^2 \Big/ \sum_{i=1}^{p}\left(\hat{\alpha}_i^2 / \left\{1 + \left(1 + \lambda_i\left(\hat{\alpha}_i^2/\hat{\sigma}^2\right)\right)^{1/2}\right\}\right)$		Singh, Chaubey and Dwivedi (1986)

https://doi.org/10.26219/heal.aueb.6167

(continued from previous page)

7		Choose k such that $\left\lvert \sum \lambda_i \left(k\hat{\alpha}_i^2(k) - s^2\right) / (\lambda_i + k)^3 \right\rvert$ is minimum	Dempster, Schatzoff, and Wermuth (1977)
8		Choose k by minimising $MSE\left(\hat{\beta}(k)\right) = \gamma(k, \lambda, \alpha, \sigma)$ $= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2}$ using an algorithm which gives values to k then to a and σ.	Nordberg (1982)
9	For the general ridge regression estimator $\frac{S^2}{\hat{\beta}_i^2} = \hat{k}_i$, where S^2 is an unbiased estimator of σ^2		Hoerl and Kennard (in Dwivedi et al., 1980)
10		For the two-parameter estimator $\beta^*(k,q) = \left[(X'X)^{-q+1} + kI\right]^{-1} (X'X)^{-q} X'y$, choose q so as to maximize $C(q) = \left[\sum \lvert r_i \rvert \lambda_i^{((1-q)/2)}\right] \Big/ \left[\left(\sum r_i^2\right)\left(\sum \lambda_i^{(1-q)}\right)\right]^{1/2}$, choose k so as to minimize $\tilde{L} = n \ln(2\pi e \tilde{\sigma}) + \xi'\xi - (r'\xi)/\tilde{\sigma}$	Obenchain (in Gibbons, 1981)

https://doi.org/10.26219/heal.aueb.6167

(continued from previous page)

11	$k = s^2 / s_\beta^2$		Lindley and Smith (1972)
12	$k = ps^2 / \sum \lambda_i \hat{\alpha}_i^2$		Lawless and Wang (1976)
13		k is obtained by solving $\sum_{i=1}^{p} \hat{\alpha}_i^2 / \hat{\sigma}^2 [(1/k) + (1/\lambda_i)] = p$	Dempster, Schatzoff, and Wermuth (1977)

Table 3.1: Selection Criteria

https://doi.org/10.26219/heal.aueb.6167

3.10 Illustration to Real Data

In order to illustrate the use of ridge regression we applied the method to a real data set.

3.10.1 Bodyfat data

The data are the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men. The dependent variable (y) is the PERCENT BODY FAT (from Siri's equation). The data were found in StatLib (Dataset Archive) and were submitted by Dr. A. Garth Fisher. More details about the data can be found in http://lib.stat. cmu.edu /datasets/ bodyfat.

The independent variables (matrix X) are

- AGE(years)
- WEIGHT(lbs)
- HEIGHT(inches)
- NECK CIRCUMFERENCE(cm)
- CHEST CIRCUMFERENCE(cm)
- THIGH CIRCUMFERENCE(cm)
- FOREARM CIRCUMFERENCE(cm)

Note: This data set included another 7 explanatory variables, which were left out for reasons of convenience and efficient data presentation.

Accurate measurement of body fat is inconvenient or costly so it is desirable to have easy methods of estimating body fat that are not inconvenient or costly. Eventually, we wish to produce a regression equation which will predict percentage body fat in terms of anatomical measurements.

3.10.2 Data analysis

The regression model is: $y = X\beta + u$. We wish to examine the inclusion of correlated variables to our model in order to illustrate collinearity diagnostics and the ridge regression solution. As one can see from the next figure (contains all the scatterplots of one variable against another) the explanatory variables are higly correlated.

https://doi.org/10.26219/heal.aueb.6167

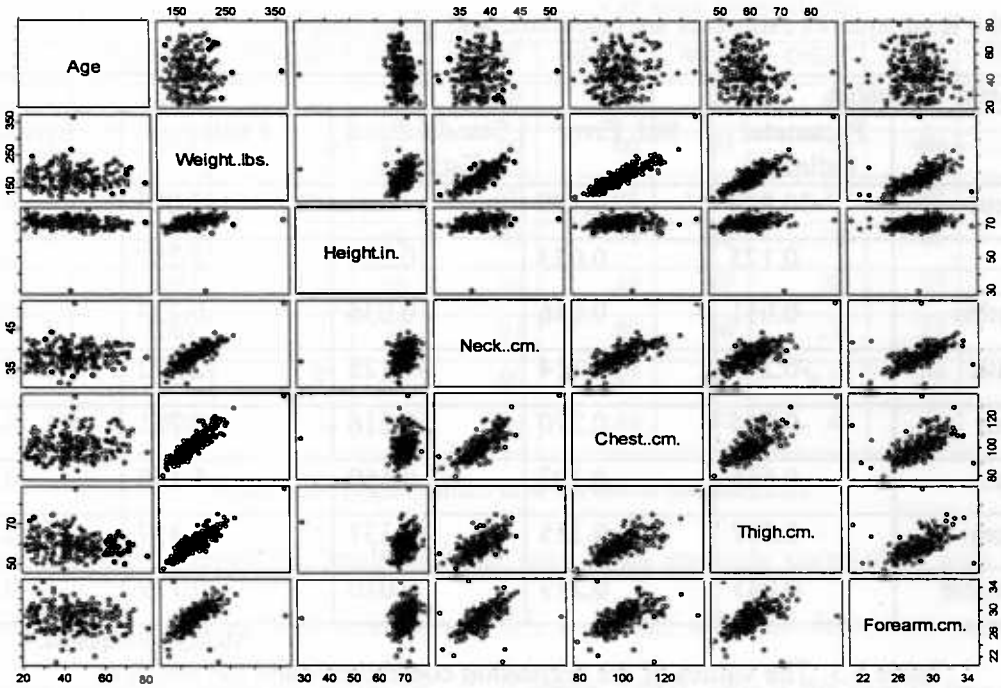


Figure 3.2: Correlation matrix

Checking the correlation coefficients as well we find several large pairwise correlations. For example, the correlation between chest circumference and weight is 0.894 which is rather large. In addition, we check whether $r_{ij} \geq R^2$ where $R^2 = 0.5894$. This holds in 7 cases (with bold, italics in Table 3.3) so we can say that multicollinearity is present.

	Age	Weight lbs	Height in	Neck cm	Chestcm	Thigh cm	Forearm cm
Age Sig. (2-tailed)	1.000 .						
Weightlbs Sig. (2-tailed)	-0.013 .840	1.000 .					
Heightin Sig. (2-tailed)	-0.172 .006	0.308 .000	1.000 .				
Neckcm Sig. (2-tailed)	0.114 .072	***0.831*** .000	0.254 .000	1.000 .			
Chestcm Sig. (2-tailed)	0.176 .005	***0.894*** .000	0.135 .032	***0.785*** .000	1.000 .		
Thighcm Sig. (2-tailed)	-0.200 .001	***0.869*** .000	0.148 .018	***0.696*** .000	***0.730*** .000	1.000 .	
Forearmcm Sig. (2-tailed)	-0.085 .178	0.630 .000	0.229 .000	***0.624*** .000	0.580 .000	0.567 .000	1.000 .

Table 3.2: Correlation coefficients of the predictors

https://doi.org/10.26219/heal.aueb.6167

Then the least squares estimates are calculated and given below:

Parameter estimates					
	Parameter Estimate	Std. Error	Standardized Estimate	t value	p-value
Intercept	-30.808	14.769		-2.086	0.038
Age	0.175	0.033	0.264	5.287	0.000
Weightlbs	0.011	0.046	0.036	0.223	0.824
Heightin	-0.293	0.114	-0.128	-2.572	0.011
Neckcm	-0.744	0.270	-0.216	-2.751	0.006
Chestcm	0.555	0.107	0.559	5.168	0.000
Thighcm	0.537	0.155	0.337	3.457	0.001
Forearmcm	0.041	0.229	0.010	0.179	0.858

Multiple R-squared: 0.5894

Table 3.3: The values of the regression coefficients and the *p*-values

Only two predictors (weightlbs and forearmcm) have large *p*-values i.e. they are not significant.

To decide for multicollinearity we calculate some diagnostics (see next table):

The predictors	VIF	R_i^2	Leamer's c_i
Age	1.482	0.325	0.822
Weightlbs	15.292	0.935	0.256
Heightin	1.474	0.322	0.824
Neckcm	3.668	0.727	0.522
Chestcm	6.960	0.856	0.379
Thighcm	5.644	0.823	0.421
Forearmcm	1.817	0.445	0.742

Table 3.4: The multicollinearity diagnostics

https://doi.org/10.26219/heal.aueb.6167

		Variance Proportions							
Dimension	Eigenvalue	(Constant)	AGE	WEIGHT	HEIGHT	NECK	CHEST	THIGH	FOREARM
1	7.906	.00	.00	.00	.00	.00	.00	.00	.00
2	.070	.00	.61	.00	.00	.00	.00	.00	.00
3	.017	.01	.04	.06	.02	.00	.00	.00	.00
4	.003	.00	.00	.04	.31	.00	.00	.01	.48
5	.002	.03	.00	.04	.09	.00	.01	.27	.37
6	.001	.01	.34	.01	.06	.06	.38	.27	.10
7	.001	.01	.01	.00	.03	.87	.18	.00	.05
8	.000	.95	.01	.84	.49	.07	.43	.44	.00

Table 3.5: Eigenvalues and variance proportions

A variable X_i is harmfully multicollinear only if its multiple correlation with other members of the independent variable set, R_i^2, is greater than the dependent variable's multiple correlation with the entire set, R^2 (Greene, 1993). This is the case in four cases as we can see from Table 3.4. We can reach the same conclusion using Leamer's diagnostic which is small for the same cases. The *VIF* for weightlbs is 15.292 which is quite large.

Calculating the condition number we find $K = \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{1/2} = \left(\frac{7.90631}{0.0003}\right)^{1/2} = 26{,}385.36$ which is rather large, while small eigenvalues (i.e. 0.002) indicate near linear dependencies. Another diagnostic is the determinant of the correlation matrix $|X'X|$ = 0.0038. The closer $|X'X|$ is to 0, the greater the severity of multicollinearity. Finally the sum of λ_i^{-1} = 6,273.9487. Recall that in an orthogonal system it would be 7.

Since our data suffer from multicollinearity we will try to implement ridge regression. To this aim we calculate *k* by four methods.

a) Hoerl and Kennard: $k_{HK} = s^2/\max(\hat{\alpha}_i^2) = 0.008$, where s^2 is the estimate of the variance and $\hat{\alpha}$ the least square estimate (see (2.2.4)).

b) Hoerl Kennard and Baldwin: $k_{HKB} = \frac{ps^2}{\hat{\alpha}'\hat{\alpha}} = 0.021$

https://doi.org/10.26219/heal.aueb.6167

c) Lawless and Wang: $k_{LW} = \dfrac{ps^2}{\sum \lambda_i \hat{\alpha}_i^2} = 0.020$

d) Vinod's *ISRM*: $k_{ISRM} = 0.44$

Ridge Regression Results			
	Ridge Estimate	Std. Error	Standardized Ridge Estimate
Intercept	-22.978	6.321	
Age	0.120	0.019	0.181
Weightlbs	0.046	0.006	0.160
Heightin	-0.285	0.068	-0.125
Neckcm	0.037	0.099	0.011
Chestcm	0.270	0.026	0.272
Thighcm	0.281	0.043	0.176
Forearmcm	0.116	0.128	0.028

Table 3.6: Ridge estimates

Observing the Ridge Trace we note that when $k_{ISRM} = 0.44$ the coefficients stabilize. So in Table 3.6 we give the ridge estimates using the k obtained by minimizing the Index of Stability of Relative Magnitudes (*ISRM*).

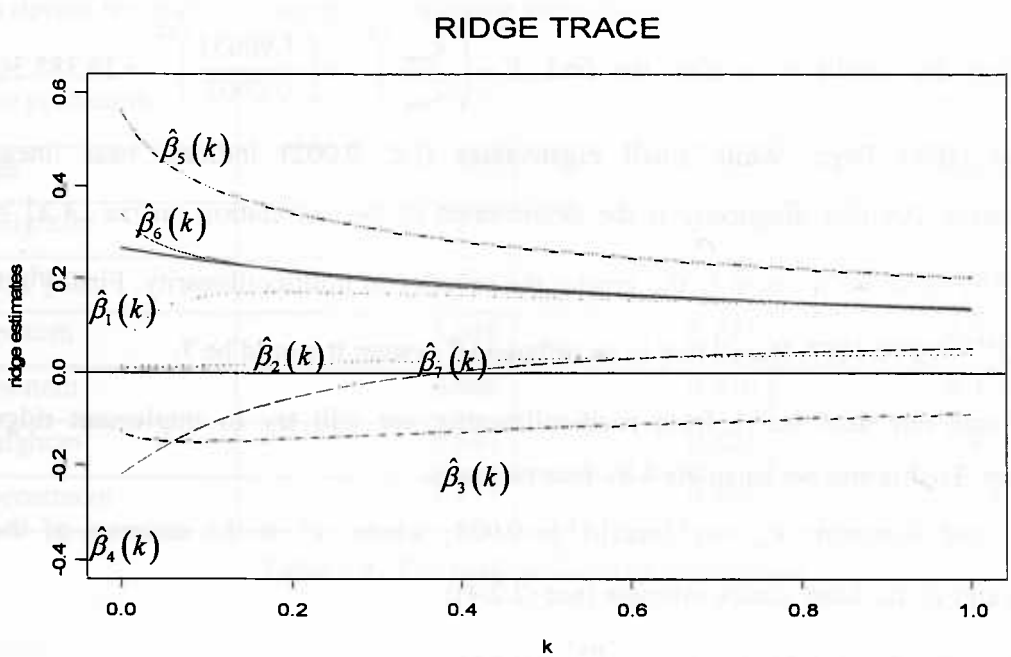


Figure 3.3: The Ridge Trace

https://doi.org/10.26219/heal.aueb.6167

3.11 A Critical View of Ridge Regression

There are a lot of controversies in the literature about the success of ridge regression. Some authors are in favour of using ridge regression while others greatly criticize it stating that it is not "always" better than least squares.

According to Draper and Van Nostrand (1979) ridge regression is a technique that enables one to assume prior information of a specific kind. So ridge regression is an appropriate multicollinearity remedy in case we consider a Bayesian formulation or in case of a restricted least squares formulation. In any other circumstances, they claim that ridge regression should not be used. They doubt the value of ridge regression, since they find that it is favored over least squares only when the ridge estimators are close to the least squares values. Marquardt and Snee (1975) express the same opinion with Draper and Van Nostrand stating that ridge regression is reasonable under a Bayesian interpretation. They also comment that since in correlation form regression coefficients rarely exceed three, one can consider bounded priors.

Recalling the fact that the ridge estimator is just the OLS estimator biased by $(\lambda_i/(\lambda_i + k))$, Pagel (1981) points out that since this fraction declines with λ_i, ridge applies the greatest shrinkage, and thus reduces the variability most, for the coefficients associated with small eigenvalues. However, it is not always right to treat the coefficients of small eigenvalues as less "important" than those of large eigenvalues. Small eigenvalues may derive from the fact that the data are inadequate for the estimation of the model parameters; or from a misspecification of the model. Ridge regression ignores such problems and tries to obtain the regression estimates.

Gunst and Mason (1977), in their evaluation on five estimators of the regression coefficients (least squares (LS), principal components (PC), ridge regression (RR), latent root (LR) and a shrunken estimator (SH)), concluded that the PC and LR estimators appeared to offer the best opportunity for large decrease in MSE over LS for the multicollinear data, while ridge regression and SH performed well for the near-orthogonal data.

Many simulations have been performed to compare ridge regression estimates to least squares estimates, in a mean square error sense. Pagel (1981) notes that based on

Monte Carlo studies, ridge regression reliably reduces the mean squared error of the estimated coefficients under conditions of multicollinearity and low signal to noise ratios. However, these simulations must be viewed with caution. Draper and Van Nostrand (1979) claim that these simulations involve restrictions on the parameter values (the situations where ridge regression is the appropriate technique theoretically). Opponents of ridge regression also cite *inconsistent findings* among studies and criticize the modeling of fixed length of the regression coefficient vectors in many simulations.

https://doi.org/10.26219/heal.aueb.6167

CHAPTER 4

FURTHER RIDGE THEORY

4.1 Other Interpretations of Ridge Regression

In this section we will present three interpretations for the use of ridge regression. The first one is analogous to Hoerl and Kennard reasoning while the second one is based on a Bayesian approach. In addition, in recent literature one new characterization for ridge regression is presented based on an optimization problem.

4.1.1 Restricted Least Squares Interpretation

Ridge regression may be viewed as least squares subject to a spherical restriction on the parameters. Suppose that the regression problem under study is in correlation form and that we perform least squares subject to the spherical restriction

$$\beta'\beta \le c^2, \tag{4.1.1}$$

where c^2 is a specified value. A restricted least squares estimator can be estimated by minimizing $(y - X\beta)'(y - X\beta)$ subject to the constraint (4.1.1). Using the method of Lagrange multipliers, we can form

$$F = (y - X\beta)'(y - X\beta) + k(\beta'\beta - c^2), \tag{4.1.2}$$

Setting $\partial F / \partial \beta = 0$ gives the equations

$$(X'X + kI)\beta = X'y, \tag{4.1.3}$$

which is the ridge solution. (Vinod and Ullah, 1981).

https://doi.org/10.26219/heal.aueb.6167

4.1.2 Bayesian Interpretation

The Bayesian approach to ridge regression is based on the assumption that we have a regression situation where

$$y \sim N(X\beta, I\sigma^2). \tag{4.1.4}$$

Consider the case where the individual regression coefficients in $\beta' = (\beta_1, ..., \beta_p)$ are exchangeable ("an assumption that may not be appropriate" as emphasized by Lindley and Smith, 1972) i.e. they are unaltered by a permutation of the suffixes (i = 1,2,…, p). Suppose further that

$$\beta_j \sim N(\xi, \sigma_\beta^2). \tag{4.1.5}$$

If we suppose vague prior knowledge for ξ, then the Bayes estimate is

$$\beta^* = \{I_p + k(X'X)^{-1}(I_p - p^{-1}J_p)\}^{-1}\hat{\beta}, \tag{4.1.6}$$

where $k = \sigma^2/\sigma^2{}_\beta$ and J is a matrix of ones. If we assume $\xi = 0$, and thus imply that β_i's are small then the Bayes estimate is given by

$$\beta^* = \{I_p + k(X'X)^{-1}\}^{-1}\hat{\beta}. \tag{4.1.7}$$

When σ^2, the residual regression variance, and $\sigma^2{}_\beta$, the variance of the regression coefficients are both unknown we can estimate them and calculate k^* as follows: $k^* = s^2/s_\beta^2$.

In the estimates above k is a variance ratio and is estimated from the data while in Hoerl and Kennard's argument k is the constant where the regression estimates stabilize. Like ridge method the Bayesian method attempts to avoid some of the problems caused by non-orthogonality in the data but in addition it has the advantage "of dispensing with the rather arbitrary choice of k and allows data to estimate it" (Lindley and Smith, 1972).

https://doi.org/10.26219/heal.aueb.6167

4.1.3 An Optimization Problem

Consider linear estimators that can be written as

$$B = JR_X B_0,$$

where J is a $p \times p$ matrix, B_0 is the ordinary LS estimator and $R_X = X'X$ (the correlation matrix). Since B is a linear transform of B_0, it is a biased estimator unless $J = R_X^{-1}$. We have $E(B) = JR_X\beta$. From (3.5.2) it can be shown that

$$MSE(B) = D(B) + \sigma^2 tr(JR_X J'),$$

where $D(B)$ is the squared bias term of B and is equal to $\|(JR_X - I)\beta\|^2$.

Ridge regression is a biased estimation method based on linear estimators. Qannari et al. (1997) present an optimization problem, which leads to the ridge estimator but from another viewpoint. They suggest keeping the total variance of the parameter estimates at an "acceptable level", whiling allowing the smallest possible bias.

Consider the inequality that holds for the Euclidean norm of a matrix $0 \le D(B) \le \|JR_X - I\|^2 \|\beta\|^2$, it seems that

(i) $D(B)$ is zero when $J = R_X^{-1}$,

(ii) or approaches zero when $\|JR_X - I\|^2$ approaches zero.

Therefore the authors, as explained earlier, suggest minimizing the bias, i.e. $\min_J \|JR_X - I\|^2$, under the constraint that the total variance is fixed, i.e. $tr(JR_X J') = c$, where c is a fixed positive scalar. Solving the Lagrangian problem we obtain

$$J = (R_X + kI)^{-1},$$

which is the ridge estimator (Qannari et al., 1997).

4.2 *Application of Ridge Regression in Special Cases*

In chapter 3 we only consider the use of ridge regression in the multivariate linear regression model. However, many authors have used ridge regression in different cases,

https://doi.org/10.26219/heal.aueb.6167

for example, in logistic regression. We will discuss some cases which we consider rather useful.

4.2.1 Rank deficient model

Let us consider the case where our model is *rank deficient*. Brown (1978) examines the ridge estimator in the context of a linear model, which may be rank deficient (X is an $(T \times p)$ given matrix of rank m $(\leq p)$). In such a case the ridge estimator, $(X'X + kI)^{-1} X'y$ is not defined at $k = 0$, so Brown (1978) suggests the following definition. Let

$$\hat{\beta}(k) = \begin{cases} (X'X + kI)^{-1} X'y & for \quad k > 0 \\ X^{+} y & for \quad k = 0 \end{cases}$$

where X^{+} denotes the Moore-Penrose pseudoinverse (Appendix A).

4.2.2 Straight line regression with a small number of observations

Carmer and Hsieh (1978) try to apply biased techniques to *straight line regression* with a small number of observations.

Having y and X in standardized form leads to a LS estimate equal to the simple correlation $\hat{r}$, between X and y; the regression sum of squares is equal to $\hat{r}^2$ and the residual mean square is $\hat{\sigma}^2 = (1 - \hat{r}^2)/(T - 2)$, where T is the number of observations.

The biased estimate of the standardized regression coefficient is $\tilde{\beta} = \tilde{r} = \hat{r}/(1 + k)$.

Farebrother (in Carmer and Hsieh, 1978) proposed for an estimate of k the following:

$$k_1 = \frac{\hat{\sigma}^2}{\hat{r}^2} = \frac{(1 - \hat{r}^2)}{(T - 2)\ \hat{r}^2}$$

The results of the simulation study of the authors showed that none of the biased procedures are recommended for use in straight line regression problems with a small number of observations. According to the authors "all the procedures rather severely

reduced the estimate of the slope, relative to least squares, and none of the procedures produced dramatic improvements in the mean square error".

4.2.3 Models with lagged effects

In *models with lagged effects* we have

$$Y_t = a + \beta_0 X_t + \beta_1 X_{t-1} + \ldots + \beta_\ell X_{t-\ell} + \varepsilon_t; \quad t = 1,2,\ldots,n \tag{4.2.1}$$

where Y_t is a dependent variable, X_t represent the matrix of regressors and ε_t the random error. As we can notice from (4.2.1) the regressors involve time series which are often autocorrelated. So using ridge regression particularly for large values of ℓ is a way to tackle this problem.

However, a problem of lagged effects model is to select an appropriate number of lagged terms, i.e. the right ℓ. Erickson (1981) deals with the topic of variable selection with ridge regression. In order to select variables he minimizes a prediction error, or at least an estimate of the prediction error based on ridge regression- Ridge Regression Prediction criterion (*RP*). *RP* depends on which observations and regressors are used and on the value of *k*- the ridge constant. Using ridge regression on some data the author shows that in order to find the "right" estimates for the number of lagged terms one should first calculate a *k* that minimizes the *RP* criterion for each value of ℓ and then find the overall minimum of ℓ's.

4.2.4 Subset selection

The ridge regression has also been used as a *subset selection technique* by Hoerl et al. (1986). They propose a ridge selection method that examines a full ridge solution and then deletes terms that are not significant. The deletion of the terms is based on a modified *t-test*, $t = \hat{\beta}(k)/S_{Ri}$, where S_{Ri}^2 is the i_{th} diagonal element of $\sigma^2 (X'X + kI)^{-1} X'X (X'X + kI)^{-1}$. This means that we are actually testing the hypothesis $H_0 : E(\hat{\beta}(k)) = 0$.

https://doi.org/10.26219/heal.aueb.6167

4.2.5 Logistic regression

Consider the *logistic regression model*:

$$\pi = \frac{1}{\left(1+e^{-X\beta}\right)} \tag{4.2.2}$$

where $\beta = \left(\beta_0, \beta_1, ..., \beta_p\right)'$ and π is is the probability that the event Y occurs, $\pi = P(Y=1)$. The unknown parameter β can be estimated by $\hat{\beta}$, the maximum likelihood estimator (MLE) of β. Schaefer et al (in Lee and Silvapulle, 1988) have derived the ridge estimator for the logistic regression model as

$$\hat{\beta}(k) = \left(X'\hat{V}X + kI\right)^{-1}\left(X'\hat{V}X\right)\hat{\beta}, \text{ where } \hat{V} = V\left(\hat{\beta}\right).$$

They have also shown that if the degree of multicollinearity is high then $MSE\left\{\hat{\beta}(k)\right\} < MSE\left\{\hat{\beta}\right\}$ for many observations and small value of *k*.

Lee and Silvapulle (1988) propose a method for the determination of *k* using Bayesian methods. They obtained the following two choices of *k*:

$$\hat{k}_a = (\pi+1)\left(\hat{\beta}'\hat{\beta}\right)^{-1}, \tag{4.2.3}$$

$$k_b = \left[tr\left(\operatorname{cov}\left(\hat{\beta}\right)\right)^{-1}\right]\left[\hat{\beta}'\left(\operatorname{cov}\left(\hat{\beta}\right)\right)^{-1}\hat{\beta}\right]^{-1}. \tag{4.2.4}$$

After a Monte Carlo study for the examination of the performance of the above estimators the authors concluded that $\hat{k}_a$ is considered the "best" choice for *k*.

4.2.6 Autocorrelated disturbances

Firinguetti (1989) studies the effect of collinearity and autocorrelated disturbances in the performance of several ridge regression estimators. The use of ridge regression in generalized linear models has been considered by other authors too. Yet it had only been discussed in cases where the error variance-covariance matrix ($\sigma^2\Omega$) were known. Firinguetti suggests that even when one has to estimate k and Ω there can be found conditions where the ordinary ridge regression estimator dominates the generalized least squares (GLS) estimator.

Consider the model $y = X\beta + u$ as described in (2.2.2), where u is a vector of T disturbances such that

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad |\rho| < 1, \quad t = 1,2,\ldots,T \tag{4.2.5}$$

and

$$\varepsilon_t \sim N(0,\sigma^2), \quad E(\varepsilon_t \varepsilon_{t'}) = 0 \quad \text{for each } t, t' \neq t. \tag{4.2.6}$$

The GLS estimator

$$b = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1} y, \tag{4.2.7}$$

where

$$\Omega = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & \rho & \ldots & \rho^{T-1} \\ \rho & 1 & \ldots & \rho^{T-2} \\ \cdot & \cdot & \cdot & \cdot \\ \rho^{T-1} & \rho^{T-2} & \ldots & 1 \end{bmatrix} \tag{4.2.8}$$

is the minimum variance unbiased estimator. Since in practice ρ is usually unknown it is estimated by $\hat{\rho} = \dfrac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=1}^{T} e_t^2}$ where $e = Y - X\hat{\beta}$, the OLS residuals. Then the GLS estimator of β becomes

$$\hat{b} = \left(X'\hat{\Omega}^{-1}X\right)^{-1} X'\hat{\Omega}^{-1} y \tag{4.2.9}$$

In case when collinearity is present in a GLRM, the author suggested considering a generalized version of some well-known ridge estimators. For example, the generalized Hoerl, Kennard and Baldwin RR (GHKB) estimator is:

$$\begin{aligned} \hat{b}(k_1) &= \left(X'\hat{\Omega}^{-1}X + k_1 I\right)^{-1} X'\hat{\Omega}^{-1} y \\ &= \left(X'\hat{\Omega}^{-1}X + k_1 I\right)^{-1} X'\hat{\Omega}^{-1} X\hat{b} \end{aligned}$$

with $k_1 = \dfrac{ps^2}{\hat{b}'\hat{b}}$ and $s^2 = \dfrac{\left(y - X\hat{b}\right)'\hat{\Omega}^{-1}\left(y - X\hat{b}\right)}{(n-p)}$.

One can also define the generalized Lawless and Wang RR (GLWR) estimators as

$$\hat{b}(k_2) = \left(X'\hat{\Omega}^{-1}X + k_2 I\right)^{-1} X'\hat{\Omega}^{-1} y$$

https://doi.org/10.26219/heal.aueb.6167

$$= \left(X'\hat{\Omega}^{-1}X + k_2 I\right)^{-1} X'\hat{\Omega}^{-1}X\hat{b}$$

with $k_2 = \dfrac{ps^2}{\hat{b}'X'\hat{\Omega}^{-1}X\hat{b}}$.

Comparing the different estimators using MSE and absolute bias the author suggests that in the presence of multicollinearity and autocorrelation the generalized ridge regression estimators can perform better than the other methods.

4.3 A Recent Advance in Ridge Regression

It is not unusual to have collinearity and influential cases simultaneously in a data set. Walker and Birch (1988) discuss about the effect that collinearity can have on the influence of any given case and propose some influence measures in case we use ridge regression. Part C in the Appendix provides a brief overview of influence analysis.

4.3.1 Influence in Ridge Regression

Using a different notation for convenience the ridge estimator of β is now denoted as

$$b^* = (X'X + kI)^{-1} X'y, \tag{4.3.1}$$

The i_{th} ridge residual is defined as $e_i^* = y_i - x_i b^*$. In order to measure the influence of a single case a version of *DFFITS* (difference in fit standardized) for RR can be used, namely

$$DFFITS^*(i) = \frac{x_i\left(b^* - b^*(i)\right)}{SE\left(x_i b^*\right)},$$

where $b^*(i)$ is the ridge estimator of β without the i_{th} case and $SE\left(x_i b^*\right)$ is an estimator of the standard error (SE) of the fitted value.

The authors also define two versions of Cook's distance D_i

https://doi.org/10.26219/heal.aueb.6167

$$D_i^* = \frac{\left(b^* - b^*(i)\right)' X'X\left(b^* - b^*(i)\right)}{ps^2} \text{ or } D_i^* = \frac{\left(\hat{y}^* - \hat{y}^*(i)\right)'\left(\hat{y}^* - \hat{y}^*(i)\right)}{ps^2}$$

and

$$D_i^{**} = \frac{\left(b^* - b^*(i)\right)'(X'X + kI)^{-1}(X'X)^{-1}(X'X + kI)^{-1}\left(b^* - b^*(i)\right)}{ps^2}.$$

For choosing the value of k the authors suggest the value of k that minimizes the following quantity

$$C_k = \left(SSR_k / s^2\right) - T + 2tr\left(H^*\right), \tag{4.3.2}$$

where SSR_k is the sum of squares of residuals from RR and $H^* = X\left(X'X + kI\right)^{-1} X'$.

As one can conclude from the definitions of *DFFITS* and Cook's distance, the influence of each case is a function of the ridge parameter k. It is interesting to note that while the influence of some cases decreases the influence of some others increases. Thus, the authors advise us to determine the value of k and then compute the influence measures for that k. If it is necessary to delete certain cases, the process described should be repeated.

4.3.2 Local change of small perturbations

Shi and Wang (1999) presented another approach in order to measure the influence of observations on the ridge estimator. Instead of examining the influence of case deletion they perform local influence analysis. In local influence analysis we try to estimate the local change of small perturbations on the variance or on the explanatory variables.

The functions used to estimate these changes are the generalized influence function (GIF) and the generalized Cook statistic (GC)

- Perturbing the variance

The variance of *the errors* becomes $\sigma^2 W^{-1}$ where $W = diag(\omega)$ with diagonal elements of $\omega = \left(\omega_1, \ldots, \omega_n\right)'$. The perturbed version of the ridge estimator is

$$b^*(\omega) = \left(X'WX + kI\right)^{-1} X'Wy. \tag{4.3.3}$$

The generalized influence function of b^* under the perturbation is given by

https://doi.org/10.26219/heal.aueb.6167

$GIF(b^*, l) = (X'X + kI)^{-1} X'D(e^*)l$, where $D(e^*) = diag(e^*)$ and l is a unit-length vector.

Again two versions of the generalized Cook statistic of b^* can be defined

$$GC_1(b^*, l) = l'D(e^*)HD(e^*)l/ps^2, \tag{4.3.4}$$

and

$$GC_2(b^*, l) = l'D(e^*)H^{*2}D(e^*)l/ps^2, \tag{4.3.5}$$

where $H^* = X(X'X + kI)^{-1}X'$ and H is the hat matrix of LS regression.

- Perturbing the explanatory variables

Similar influence measures can be defined when we have perturbation of the explanatory variables.

Finally, recall the quantity (4.3.2) and consider the perturbation of the variance. Let $C_k(\omega)$, $SSR_k(\omega)$ and $H^*(\omega)$ denote the perturbed versions of C_k, SSR_k and H^*, respectively. Then

$$C_k(\omega) = SSR_k(\omega)/s^2 - T + 2tr(H^*(\omega)). \tag{4.3.6}$$

https://doi.org/10.26219/heal.aueb.6167

CHAPTER 5

SIMULATION - APPLICATION

Several authors have conducted simulations in order to verify that ridge estimators are better than least squares in certain cases. The interested reader is referred to Gibbons (1981), Gunst and Mason (1977), Hoerl, Kennard, and Baldwin (1975), McDonald and Galarneau (1975), Wichern and Churchill (1978), and Dempster et al. (1977). In this section we will present a simulation based on the pattern of Wichern and Churchill. Our aim is to compare estimators (ridge and the LS estimator) with respect to their MSE so as to identify cases where ridge estimators provide a good alternative to the LS estimator.

5.1 Description of the Simulation

We use the five parameter model $y = X\beta + u$,

where X is a 30×5 matrix of explanatory variables, y is a 30×1 response vector, β is a $(5+1)\times 1$ vector of parameters and u is a 30×1 vector of errors.

STEP 1: Thirty observations are generated for each explanatory variable. The explanatory variables are generated by:

$$X_{ij} = \left(1-\alpha^2\right)^{1/2} Z_{ij} + \alpha Z_{i6} \qquad i=1,2,\ldots,30 \qquad j=1,2,3$$

$$X_{ij} = \left(1-\alpha_*^2\right)^{1/2} Z_{ij} + \alpha_* Z_{i6} \qquad i=1,2,\ldots,30 \qquad j=4,5$$

where $Z_{i1}, Z_{i2}, \ldots, Z_{i6}$ are independent standard normal numbers and α^2, α_*^2 are coefficients leading to the following correlation matrix:

https://doi.org/10.26219/heal.aueb.6167

Correlation matrix					
	X_1	X_2	X_3	X_4	X_5
X_1	1	α^2	α^2	$\alpha\alpha_*$	$\alpha\alpha_*$
X_2	α^2	1	α^2	$\alpha\alpha_*$	$\alpha\alpha_*$
X_3	α^2	α^2	1	$\alpha\alpha_*$	$\alpha\alpha_*$
X_4	$\alpha\alpha_*$	$\alpha\alpha_*$	$\alpha\alpha_*$	1	α_*^2
X_5	$\alpha\alpha_*$	$\alpha\alpha_*$	$\alpha\alpha_*$	α_*^2	1

Table 5.1: The correlation matrix

The explanatory variables are then standardized so that $X'X$ is in correlation form. Three different combinations of α and α_* are investigated:

CASE 1: Both α and α_* are equal to 0.99, the condition number $\lambda_1/\lambda_5 = 581$.

CASE 2: α is equal to 0.99, α_* is equal to 0.10 and $\lambda_1/\lambda_5 = 165$.

CASE 3: α is equal to 0.70, α_* is equal to 0.30 and $\lambda_1/\lambda_5 = 8$.

Case 1 is a case of extreme multicollinearity while 2 represents a mixed case. Case 3 represents a moderate situation.

STEP 2: For each design matrix we use two coefficient vectors: β_L, the normalized eigenvector corresponding to the largest eigenvalue of $X'X$ and β_S, the normalized eigenvector corresponding to the smallest eigenvalue of $X'X$. This choice seems appropriate since these eigenvectors give the maximum (for β_L) and minimum (for β_S) MSE considering however certain constraints (McDonald and Galarneau, 1975).

STEP 3: Observations on the dependent variable are determined by:

$$y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_5 X_{i5} + u_i \qquad i = 1, 2, \ldots, 30$$

where $X_{i1}, X_{i2}, \ldots, X_{i5}$ are the original unstandardized variables, $u_i \sim N(0, \sigma^2)$, and β_0 is zero. $\beta_1, \ldots, \beta_5$ are the appropriate eigenvector values. Four values of σ are investigated:

https://doi.org/10.26219/heal.aueb.6167

$\sigma = 0.1, 0.5, 1.0, 5.0$ or equivalently four signal-to-noise ratios $\rho = \beta'\beta/\sigma^2$ =100, 4, 1, 0.04. The dependent variable is standardized so that $X'y$ is a vector of correlations.

STEP 4: Additional samples of size 100 are generated; $X'X$, β remain fixed while e_i and hence the dependent variable change.

The least squares and ridge estimates are determined using the standardized variables and then the estimated coefficients are transformed back to the original model. The k values and the standard deviations are computed for the following rules:

	Rules (m)
1.Hoerl Kennard (HK)	1. $k_{HK} = s^2/\max(\hat{a}_i^2)$
2.Hoerl Kennard and Baldwin (HKB)	2. $k_{HKB} = ps^2/\hat{a}'\hat{a}$
3.Lawless Wang	3. $k_{LW} = ps^2 \Big/ \sum_{i=1}^{5} \lambda_i \hat{a}_i^2$

Table 5.2: Investigated rules

The performance of the estimators is evaluated in terms of the averaged total squared errors

$$MSE\left(\hat{\beta}_i^m(k)\right) = \frac{1}{100}\sum_{j=1}^{100}\left(\hat{\beta}_{ij}^m(k) - \beta_i\right)^2 \qquad i = 0,1,..,5\,, \quad m = 1,2,3 \text{ and}$$

$$TMSE(m) = \sum_{i=0}^{5} MSE\left(\hat{\beta}_i^m(k)\right) \qquad m = 1,2,3$$

where β_0 zero, $\beta_1,..., \beta_5$ the appropriate eigenvectors, and $\hat{\beta}_{ij}^m(k)$ the estimates in terms of the original model. In order to compare the ridge estimators with the least squares we also compute the ratio

$$R_m = \frac{TMSE(m)}{TMSE(LS)} \qquad m = 1,2,3.$$

https://doi.org/10.26219/heal.aueb.6167

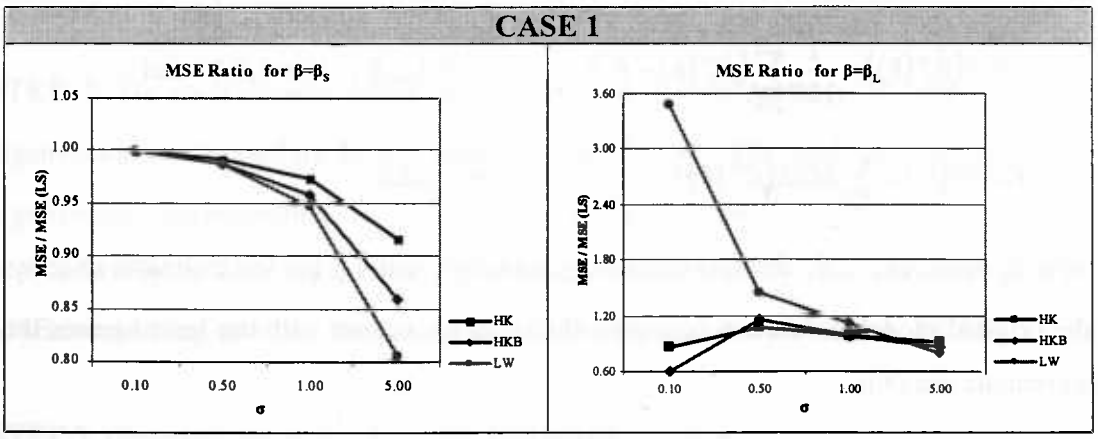
5.2 *The Simulation Results*

The results of the simulation are presented in the Appendix (Part 2, Table 2). The comments made below are based on these results.

5.2.1 Mean Squared Error (MSE)

The main theoretical justification for the construction of ridge estimators is that they have smaller MSE than the least squares estimator. Therefore in order to measure the improvement one can check the ratio of the estimated MSE for a particular ridge estimator to the estimated MSE for LS. The ratio for LS is obviously 1.

These ratios are plotted in the next figure. Each plot presents the ratio for each of the three estimators (HK, HKB, and LW) as a function of σ. The left (right) graph presents the results for $\beta = \beta_S$, ($\beta = \beta_L$). Each point plotted represents the average of 100 samples.



https://doi.org/10.26219/heal.aueb.6167

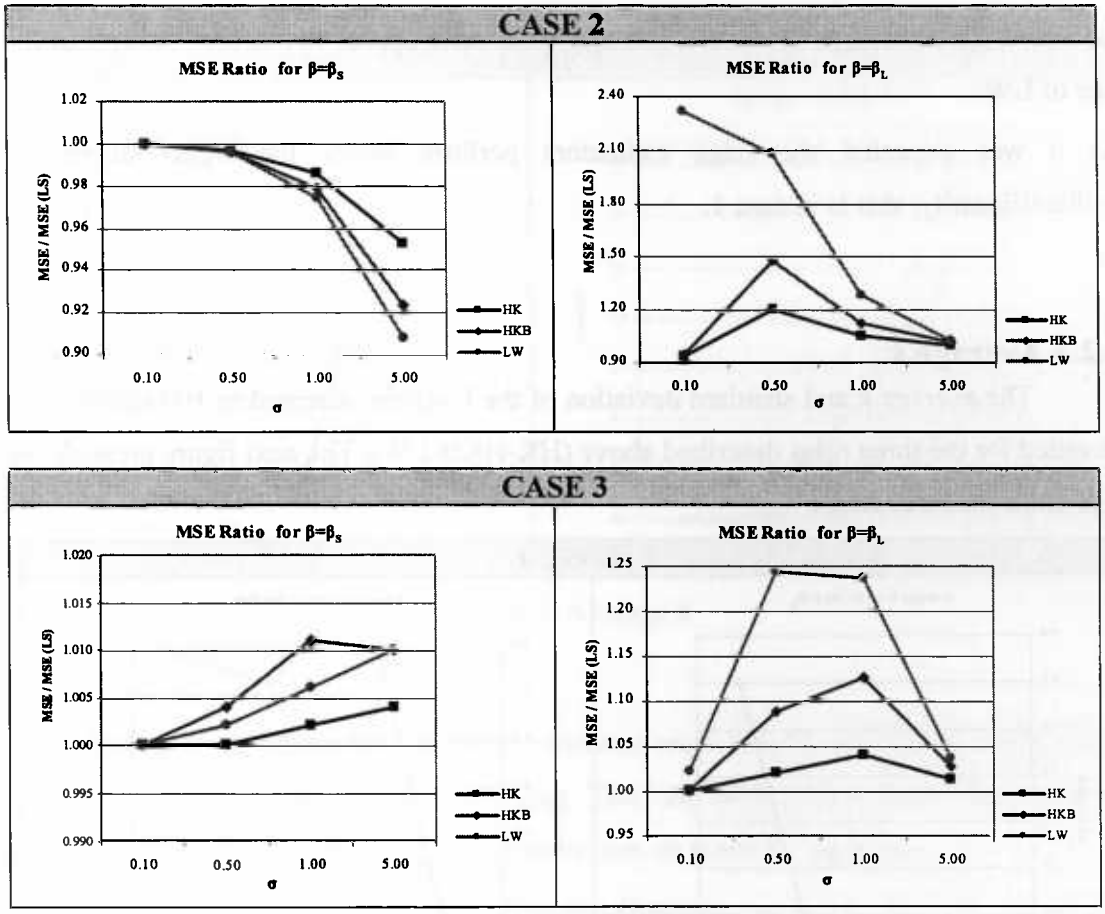


Figure 5.1: MSE Ratio

The graphs above provide the basis for these observations:

- $\beta = \beta_S$

All estimators are at least as good as the LS estimator for the first two cases, that is the MSE for those estimators is smaller than 1. For case 3 (where the correlation is smaller) the estimators are slightly worse than LS- the ratio ranges between 1 and 1.011.

The ratio decreases for larger values of σ again for the first two cases. In addition, in case 3 the estimators have almost the same MSE.

- $\beta = \beta_L$

None of the estimators is constantly better than the LS estimator. In all 3 cases the ratio increases and then decreases for larger values of σ, namely it does not appear to be a

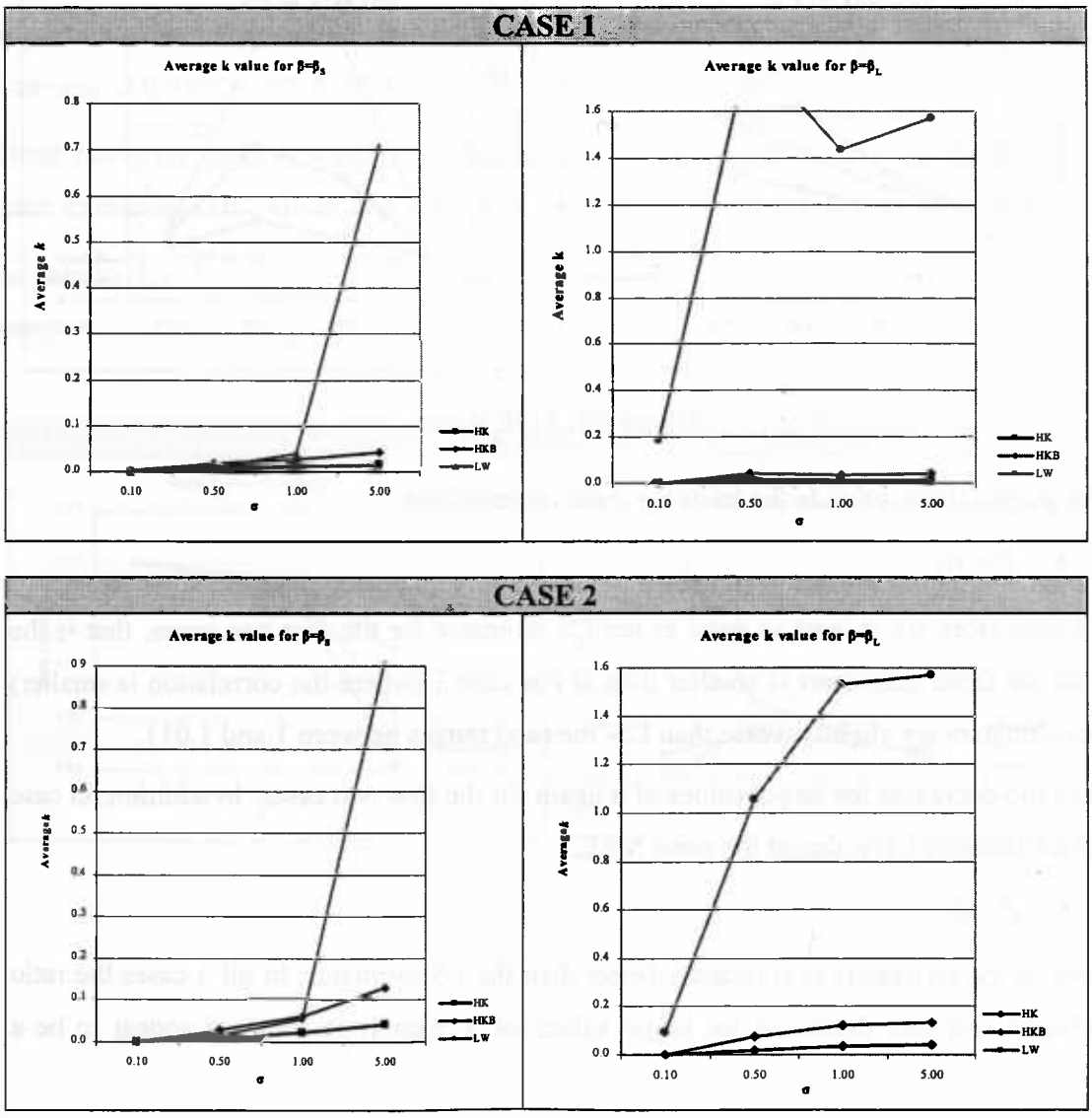
https://doi.org/10.26219/heal.aueb.6167

monotone function of σ. Moreover, HK and HKB rules appear to perform better than the rule of LW.

As it was expected the ridge estimators perform better for higher degree of multicollinearity, that is in case 1.

5.2.2 Average *k*

The average *k* and standard deviation of the *k* values observed in 100 samples are recorded for the three rules described above (HK-HKB-LW). The next figure presents the results for the three cases.



https://doi.org/10.26219/heal.aueb.6167

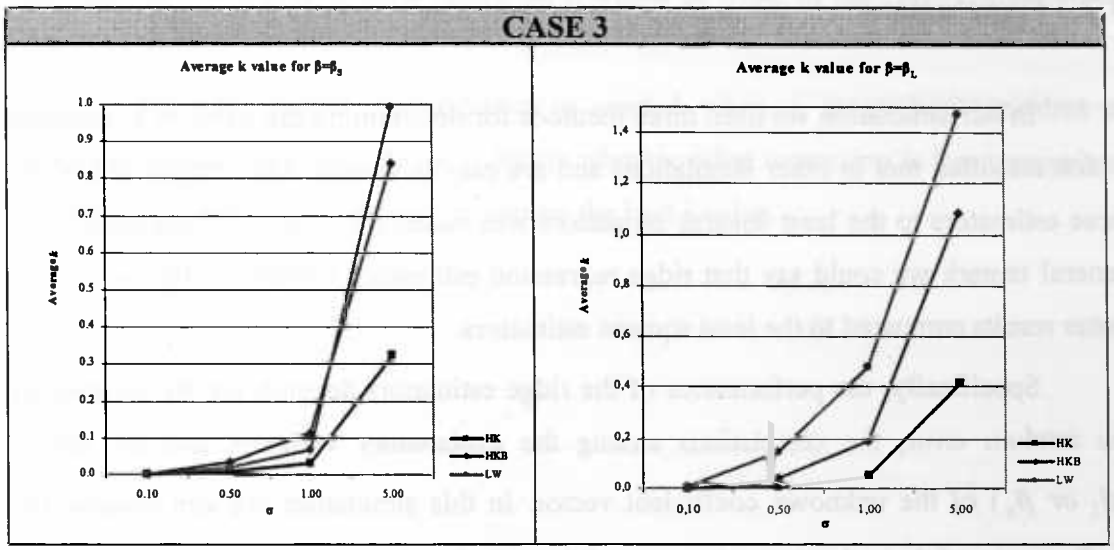


Figure 5.2: Average *k*

Note that the values for the ISRM estimator are 0.3, 0.15, and 0.7 for each of the three cases and are not presented in the figure. This rule tends to give much larger values than the other rules. The following observations can be made for each case:

Case 1: For this extreme case of multicollinearity we notice that HK and HKB rule give values for k between 0.0004 to 0.044 irrespective of the value of σ and β. The LW k increases rapidly to 0.7 for $\beta = \beta_S$ and $\sigma > 1$. For $\beta = \beta_L$ the same rule gives values of k larger to one except for σ=0.1 (k = 0.18).

Case 2: As in case 1 HK and HKB rule give small values for k, specifically, between 0.0003 and 0.1332. The values of k for LW rule exceed 1 for $\beta = \beta_L$ and $\sigma > 0.1$

Case 3: All rules give larger values for k than for cases 1 and 2. These values increase as σ increases and for $\sigma > 1$ become close or larger to 1.

In general:

- The average *k* is smaller for all estimators when $\beta = \beta_S$.
- Tke *k* values associated with estimators LW are not restricted to values less than one. This estimator assumes relatively large average *k* values as σ approaches one.

https://doi.org/10.26219/heal.aueb.6167

5.2.3 Conclusions

In our simulation we used three methods for determining the value of k, methods which are often met in other simulations and are easy to handle. The comparison of the three estimators to the least squares estimators was made using the MSE criterion. As a general remark we could say that ridge regression estimators yielded similar or slightly better results compared to the least squares estimators.

Specifically, the performance of the ridge estimators depends on the variance of the random error, the correlations among the explanatory variables and the choice (β_L *or* β_S) of the unknown coefficient vector. In this simulation one can observe the performance of the estimators when one of these factors is changed while the remaining two are fixed. As one could expect, no ridge estimator is shown to be better than the LS in all cases. While in some cases the HK or the HKB rule achieves a reduction in MSE in other cases the MSE increases. However, in case of high multicollinearity the HK rule appears to be a good alternative to the classical LS estimator.

Overall, in this study ridge regression reduces the mean squared error of the estimated coefficients under conditions of multicollinearity, low signal to noise ratios and as long as β is equal to the eigenvector of the smallest eigenvalue. Yet, when β is equal to the eigenvector of the largest eigenvalue, ridge in general performs poorly. However, our conclusions must be viewed with reservation since the size of the regression problem presented was fixed (a five parameter model with n=30) and the number of replication samples taken 100. Moreover, the size of the ratio of the number of predictors to the sample (n) is relatively small (5/30). It has been found that when the ratio is too small, there really is no difference between the least squares and ridge regression. But when the ratio is large, i.e., many predictors with a small sample, ridge regression has been demonstrated to be more accurate than least squares. (Dempster et al. , 1977). In order to check this conclusion we also simulated data for a 5 parameter model with n=15 and thus a large ratio (1/3). For high multicollinearity of regressors (case 1 in section 5.1) the MSE of all three ridge estimators were smaller than least squares (for both β_L and β_S),

https://doi.org/10.26219/heal.aueb.6167

especially for low signal to noise ratios. So it would be wiser to use ridge regression in respective cases.

In practice, careful investigation is needed when a researcher considers a particular regression problem so as to decide whether ridge regression is the appropriate alternative to least squares and how to choose the best k value.

https://doi.org/10.26219/heal.aueb.6167

APPENDIX

PART 1

A. Generalized Inverse

Definition 1: Let A be an $m \times n$ matrix. Then a matrix A^{-}: $n \times m$ is said to be a generalized inverse of A if

$$AA^{-}A = A$$

holds (see Rao and Toutenburg (1999), p.372).

A generalized inverse always exists although it is not unique in general.

Definition 2: (Moore-Penrose) A matrix A^{+} satisfying the following conditions is called the Moore-Penrose inverse of A:

(i) $AA^{+}A = A$,

(ii) $A^{+}AA^{+} = A^{+}$,

(iii) $\left(A^{+}A\right)' = A^{+}A$,

(iv) $\left(AA^{+}\right)' = AA^{+}$.

A is unique.

https://doi.org/10.26219/heal.aueb.6167

B. The Augmented Model

Let the X-matrix and the observation vector Y be augmented by $\sqrt{k}I_p$ and Y_A respectively (subscript "A" denoting augmentation). The model will then take the form,

$$\begin{bmatrix} Y_X \\ \cdots \\ Y_A \end{bmatrix} = \begin{bmatrix} X \\ \cdots \\ k^{1/2} I_p \end{bmatrix} \beta + \varepsilon, \tag{A.1}$$

where Y_x is the same as original Y, Y_A is a $p \times 1$ observation vector corresponding to the augmented part, I_p is a $p \times p$ identity matrix, and ε is $(n+p)\times 1$ error vector. In this augmented model, we have $E(Y_X) = X\beta$ and $E(Y_A) = \sqrt{k}\beta$. The (unbiased) least squares estimates of β in the augmented model are given by

$$\begin{aligned} \hat{\beta}_A &= (X'X + kI)^{-1}\left(X'Y + \sqrt{k}Y_A\right) \\ &= \hat{\beta}^* + \sqrt{k}(X'X + kI)^{-1} Y_A . \end{aligned}$$

One might say that we use, in fact, the biased estimator $\hat{\beta}^*$ in place of the unbiased estimator $\hat{\beta}_A$, and not in place of $\hat{\beta}$, and that in using $\hat{\beta}^*$, the part, $\Delta = \sqrt{k}(X'X + kI)^{-1} Y_A$, is omitted from the estimation procedure. The bias in estimation will therefore come from this omitted part. Thus, if an unbiased estimator was to be used at all, it would be $\hat{\beta}_A$ and not $\hat{\beta}$. So if $\hat{\beta}_A$ is adopted as the unbiased estimator, the mean squared error of the biased estimator shall be compared with the variance of $\hat{\beta}_A$.

Hoerl and Kennard have shown that the wquared bias of $\hat{\beta}^*$ is given by

$$k^2 \beta'(X'X + kI)^{-2} \beta . \tag{A.2}$$

On the other hand we have

$$\begin{aligned} E(\Delta) &= E\left[\sqrt{k}(X'X + kI)^{-1} Y_A\right] \\ &= k(X'X + kI)^{-1} \beta . \end{aligned} \tag{A.3}$$

Squaring (A.3), we have $\{E(\Delta)\}^2 = k^2 \beta'(X'X + kI)^{-2} \beta$, which is the same as (A.2).

A more general model than (A.1) could also be considered. Thus, rather than considering the additional data (Y_A, kI_p), we might consider the data (Y_A, V), where $V'V = K$ is a

https://doi.org/10.26219/heal.aueb.6167

diagonal matrix with diagonal elements k_i. However, we are considering model (A.1) in view of the following reasons:

(i) Hoerl and Kennard ultimately thought in terms of one k, and not in terms of k_i.

(ii) Model (A.1) will show how little of the observed Y (if observable) is being discarded to obtain the biased estimator.

https://doi.org/10.26219/heal.aueb.6167

C. Influence Analysis

The usual multiple regression model can be defined as

$y = 1\beta_0 + X\beta_1 + \varepsilon$, where y is an n vector of observable random variables, X is an $n \times r$ centred and standardized matrix of known constants, β_0 is an unknown parameter, β_1 is an r vector of unknown parameters and ε is an n vector of unobservable disturbances.

If $Z = (1, X)$ then the LS estimator is $b = (Z'Z)^{-1}Z'y$ and the vector of fitted responses $\hat{y} = Zb$. The estimator of σ^2 is $s^2 = e'e/(n-p)$, where e is the vector of residuals.

A particularly appealing perturbation scheme is case deletion. The influence of a case can be viewed as the product of two factors, the first a function of the residual and the second a function of the position of the point in the Z space. The position or leverage of the ith point is measured by h_i, the ith diagonal element of the "hat" matrix $H = Z(Z'Z)^{-1}Z'$.

Among the most popular single-case influence measures is the difference in fit standardized (*DFFITS*), which evaluated at the ith case is given by

$$DFFITS(i) = z_i\,(b - b(i))/SE(z_i b), \tag{A.4}$$

where $b(i)$ is the LS estimator of β without the ith case and $SE(z_i b)$ is an estimator of the standard error (SE) of the fitted value.

DFFITS is the standardized change in the fitted value of a case when it is deleted. Thus it can be considered a measure of influence on individual fitted values. *DFFITS* can be written as the product of two factors, one depending on the residual and the other depending on leverage,

$$DFFITS(i) = \left[\frac{e_i}{s(i)}\right]\left[\frac{h_i^{1/2}}{(1-h_i)}\right], \tag{A.5}$$

where $s(i)$ is the LS estimator of σ when the ith case has been deleted, e_i is the ith residual, and h_i is the leverage of the point.

Another useful measure of influence is Cook's D, which evaluated at the ith case is given by

https://doi.org/10.26219/heal.aueb.6167

$$D_i = \frac{(b-b(i))' Z'Z(b-b(i))}{ps^2}. \tag{A.6}$$

D_i is a measure of the change in all of the fitted values when a case is deleted. It can also be written as

$$D_i = \frac{e_i^2}{ps^2} \frac{h_i}{(1-h_i^2)}. \tag{A.7}$$

To determine influential cases, Cook and Weisberg suggested that D_i to be compared with an $F(p, n-p)$ distribution.

These measures are useful for detecting single cases having an unduly high influence. However, they suffer from the problem of masking- that is, the presence of cases that can disguise or mask the potential influence of other cases (Walker and Birch, 1988).

https://doi.org/10.26219/heal.aueb.6167

PART 2

Table 1 Longley data

PEOPLE EMPLOYED	GNP DEFLATOR	GNP	UNEMPLOYED	ARMED FORCES	POPULATION	YEAR
60,323	83.0	234,289	2,356	1,590	107,608	1947
61,122	88.5	259,426	2,325	1,456	108,632	1948
60,171	88.2	258,054	3,682	1,616	109,773	1949
61,187	89.5	284,599	3,351	1,650	110,929	1950
63,221	96.2	328,975	2,099	3,099	112,075	1951
63,639	98.1	346,999	1,932	3,594	113,270	1952
64,989	99.0	365,385	1,870	3,547	115,094	1953
63,761	100.0	363,112	3,578	3,350	116,219	1954
66,019	101.2	397,469	2,904	3,048	117,388	1955
67,857	104.6	419,180	2,822	2,857	118,734	1956
68,169	108.4	442,769	2,936	2,798	120,445	1957
66,513	110.8	444,546	4,681	2,637	121,950	1958
68,655	112.6	482,704	3,813	2,552	123,366	1959
69,564	114.2	502,601	3,931	2,514	125,368	1960
69,331	115.7	518,173	4,806	2,572	127,852	1961
70,551	116.9	554,894	4,007	2,827	130,081	1962

Source: J. Longley (1967) "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User", *Journal of the American Statistical Association*, vol. 62. September, pp. 819-841

https://doi.org/10.26219/heal.aueb.6167

TABLE 2: RESULTS OF THE SIMULATION

CASE 1 α and α_* equal to 0.99

	$\beta = \beta_S$				$\beta = \beta_L$			
Signal-to-noise ratio, ρ	100	4	1	0.04	100	4	1	0.04
HK								
Ratio of total mean square errors	*0.999*	*0.991*	*0.973*	*0.916*	*0.866*	*1.069*	*0.976*	*0.916*
k values	0.0004	0.0070	0.0108	0.0146	0.0004	0.0160	0.0132	0.0146
St.deviation of k	(0.0001)	(0.0033)	(0.0106)	(0.0212)	(0.0002)	(0.0533)	(0.0287)	(0.0212)
HKB								
Ratio of total mean square errors	*0.999*	*0.987*	*0.958*	*0.859*	*0.599*	*1.163*	*0.990*	*0.860*
k values	0.0016	0.0180	0.0270	0.0423	0.0019	0.0410	0.0366	0.0438
St.deviation of k	(0.0005)	(0.0095)	(0.0277)	(0.0556)	(0.0012)	(0.1140)	(0.0735)	(0.0593)
LW								
Ratio of total mean square errors	*0.999*	*0.987*	*0.946*	*0.804*	*3.472*	*1.444*	*1.110*	*0.796*
k values	0.0004	0.0098	0.0368	0.7048	0.1816	1.869	1.4367	1.5727
St.deviation of k	(0.0011)	(0.0027)	(0.0142)	(0.6996)	(0.0811)	(5.7995)	(2.3857)	(1.9047)

CASE 2 α equal to 0.99, α_* equal to 0.10

	$\beta = \beta_S$				$\beta = \beta_L$			
Signal-to-noise ratio, ρ	100	4	1	0.04	100	4	1	0.04
HK								
Ratio of total mean square errors	*1.000*	*0.997*	*0.986*	*0.953*	*0.939*	*1.202*	*1.053*	*1.002*
k values	0.0003	0.0074	0.0191	0.0408	0.0003	0.0199	0.0358	0.0416
St.deviation of k	(0.0001)	(0.0027)	(0.0134)	(0.0656)	(0.0001)	(0.0495)	(0.0721)	(0.0799)

https://doi.org/10.26219/heal.aueb.6167

HKB								
Ratio of total mean square errors	*1.000*	*0.996*	*0.979*	*0.923*	*0.941*	*1.481*	*1.126*	*1.013*
k values	0.0015	0.0263	0.0584	0.1278	0.0017	0.0731	0.1224	0.1332
St.deviation of k	(0.0004)	(0.0116)	(0.0446)	(0.1731)	(0.0007)	(0.1758)	(0.2127)	(0.1909)
LW								
Ratio of total mean square errors	*1.000*	*0.996*	*0.9741*	*0.908*	*22.316*	*2.078*	*1.282*	*1.025*
k values	0.0005	0.0127	0.0527	0.9082	0.0851	1.0525	1.5282	1.5738
St.deviation of k	(0.0001)	(0.0041)	(0.0194)	(0.7026)	(0.0363)	(0.9687)	(2.0644)	(1.5073)

CASE 3 α equal to 0.70, α_s equal to 0.30

	$\beta = \beta_S$				$\beta = \beta_L$			
Signal-to-noise ratio, ρ	100	4	1	0.04	100	4	1	0.04
HK								
Ratio of total mean square errors	*1.000*	*1.000*	*1.002*	*1.004*	*1.000*	*1.020*	*1.040*	*1.013*
k values	0.0003	0.0073	0.0278	0.3219	0.0003	0.0081	0.0503	0.4112
St.deviation of k	(0.0001)	(0.0022)	(0.0098)	(0.2878)	(0.0001)	(0.0033)	(0.1256)	(0.5522)
HKB								
Ratio of total mean square errors	*1.000*	*1.004*	*1.011*	*1.010*	*1.000*	*1.088*	*1.125*	*1.026*
k values	0.0014	0.0348	0.1149	0.8411	0.0014	0.0388	0.1892	1.0878
St.deviation of k	(0.0004)	(0.0104)	(0.0412)	(0.7317)	(0.0004)	(0.0158)	(0.3746)	(1.3966)
LW								
Ratio of total mean square errors	*1.000*	*1.002*	*1.006*	*1.010*	*1.022*	*1.242*	*1.236*	*1.035*
k values	0.0007	0.0180	0.0655	0.9934	0.0057	0.1420	0.4779	1.4743
St.deviation of k	(0.0002)	(0.0053)	(0.0229)	(1.2711)	(0.0016)	(0.0555)	(0.5317)	(1.5955)

REFERENCES

Banerjee, K. S. and R. N. Carr. (1971). A comment on ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* 13, 895-898.

Brown, K. G. (1978). On ridge estimation in rank deficient models. *Comm. in Statistics, Part A Theory and Methods* A7(2), 187-192.

Carmer, S., G. and Hsieh, W. T. (1978). A simulation study of five biased estimators for straight line regression. *Communications in Statistics, Simulation and Computation*, B7(6), 529-548.

Coniffe, D. and Stone J. (1973). A critical view of ridge regression. *The Statistician* 22,181-187.

Dempster, A. P., Schatzoff M. and Wermuth N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association* 72, 77-91.

Dillon, W. R. and Goldstein, M. (1984). *Multivariate analysis: Methods and Applications*. Wiley.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. Wiley.

Draper, N. R. and Van Nostrand R. C. (1979). Ridge regression and James-Stein estimation: review and comments. *Technometrics* 21, 451-466.

Dwividi, T.D. and Srivastava, V., K. (1978). On the minimum mean square error estimators in a regression model. *Communications in Statistics, Part A Theory and Methods* A7(5), 487-494.

Dwividi, T.D. and Srivastava, V. K. Hall, R. L. (1980). Finite sample properties of ridge estimators. *Technometrics* 22, 205-212.

Erickson, G. M. (1981). Using ridge regression to estimate directly lagged effects in marketing. *Journal of the American Statistical Association* 76, 766-773.

Faraway, Julian J. (2000). Practical regression and Anova using *R*.

Farrar, D. and Glauber, R. (1976). Multicollinearity in regression analysis: the problem revisited. *Review of Economics and Statistics* 49, 92-107.

Firinguetti, L. (1989). A simulation study of ridge regression estimators with auto-correlated errors. *Communications in Statistics, Simulation and Computation* 18, 673-702.

Fox, John (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. SAGE Publictions.

Gibbons, D. G. (1981). A simulation study of some ridge estimators. *Journal of the American Statistical Association* 76, 131-139.

Golstein, M. and Smith, A. F. M. (1974). Ridge-type estimators for regression analysis. *Journal of the Royal Statistical Society* B-36, 284-291.

Greene, W. H. (1993). *Econometric analysis*. Macmillan.

Guilkey, D. K. and Murphy, J. L. (1975). Directed ridge regression techniques in cases of multicollinearity. *Journal of the American Statistical Association* 70, 769-775.

Gunst, R. T. and Mason, R. L. (1977). Biased estimation in regression: An evaluation using MSE. *Journal of the American Statistical Association* 72, 616-628.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics* 12, 55-67.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications to non-orthogonal problems. *Technometrics* 12, 69-82, correction 12, 723.

Hoerl, A. E. and Kennard, R. W. and Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics* 4, 105-123.

Hoerl, R. W. Schuenemeyer, J. H. and Hoerl, A. E. (1986). A simulation of biased estimation and subset selection regression techniques. *Technometrics* 28, 369-380.

Huang, D. (1970). *Regression and Econometric Methods*. Wiley.

Jackson, E. (1991). *A User's Guide to principal components*, Wiley.

Judge, G., Griffiths,W., Hill, R., Lutkepohl, H., Lee, T. (1985). *The theory and practice of econometrics.* Wiley.

Koutsoyiannis, A. (1977). *Theory of Econometrics*. Macmillan.

Lawless, J. F. (1978). Ridge and related estimation procedures: Theory and practice. *Communications in Statistics* 7, 139-163.

Lee, A. H. Silvapulle, M. J. (1988). Ridge estimation in logistic regression. *Communications in Statistics, Simulation and Computation* 17, 1231-1257.

Lin, K. and Kmenta, J. (1982). Ridge regression under alternative loss criteria. *Review of Economics and Statistics* 64,488-494.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society* B, 34, 1-41.

Liski, E., P. (1982). A test for the mean square error criterion for shrinkage estimators. *Communications in Statistics, Simulation and Computation* 11(5), 543-562.

Longley, J. W. (1967). An appraisal of least-squares programs from the point of view of the user. *Journal of the American Statistical Association*, 62, 819-841.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics* 15, 591-612.

Marquardt, D. W. and Snee, R. D. (1975). Ridge regression in practice. *American Statistician* 29, 3-20.

Mason, R. L. and Gunst, R. F. (1985). Outlier-Induced Collinearities. *Technometrics* 27, 401-407.

McDonald, Gary C. (1980). Some algebraic properties of ridge coefficients. *Journal of the Royal Statistical Society Series B* 42, 31-34.

McDonald, Gary C. and Galarneau D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association,* 70, 407-416.

Nordberg, L. (1982). A procedure for determination of a good ridge parameter in linear regression *Communications in Statistics, Simulation and Computation* 11(3), 285-309.

Pagel, M. D. (1981). Comment on "Hoerl and Kennard's Simulation Methodology". *Comm. in Statistics, Part A Theory and Methods* 10,2361-2367.

Qannari E. M., Vigneau E. and Semenou M. (1997). New approach in biased regression. *Journal of Applied Statistics,* 24, 647-657.

Seber, G., A., F., (1977). *Linear regression analysis*. Wiley.

Rao, C. R. and Toutenburg H. (1999). *Linear Models: Least squares and alternatives*. Springer.

Shi L. and Wang X. (1999). Local influence in ridge regression. *Computational statistics and Data analysis,* 31, 341-353.

Singh, B., Chaubey, Y.P. and Dwivedi, T.D. (1986). An almost unbiased ridge estimator.*Sankhya: The Indian Journal of Statistics*, B48, 342-346.

Smith, G. and Campbell, F. (1980). A critique of some ridge regression methods. *Journal of the American Statistical Association* 75, 74-81, discussion 81-103.

Swamy, P. A. V. B. Menta, J. S. and Rappoport, P. N. (1978). Two methods of evaluating Hoerl and Kennard's ridge regression. *Communications in Statistics, Part A, Theory and Practice* 12, 1133-1155.

Tikhonov, A.N. and Arsenin, V.A. (1977). *Solutions of ill-posed problems*. W.H. Winston, Washington D.C.

Tobias, R. D. (1999). *An introduction to partial least squares regression*. Cary, NC: SAS Institute.

Vinod, H. D. (1978). Equivariance of ridge estimators through standardization, A note. *Communications in Statistics, Part A, Theory and Methods*, A7(12), 1157-1161.

Vinod, H. D. (1976). Application of ridge regression methods to a study of Bell system scale economies. *Journal of the American Statistical Association* 71, 835-841.

Vinod, H. D. and Ullah, A. (1981). *Recent Advances in Regression Methods*. Dekker, NY.

Walker, E., Birch, J. B., (1988). Influence measures in ridge regression. *Technometrics* 30, 221-227.

Wichern, D. W. and Churchill, G. A. (1978). A comparison of ridge estimators. *Technometrics* 20, 301-311.

Willan, A. R. and Watts, D. G. (1978). Meaningful multicollinearity measures. *Technometrics* 20, 407-412.

https://doi.org/10.26219/heal.aueb.6167

Δωρεά.