A Study on Multicollinearity using Ridge Regression

A Project Report Submitted to Bharathiar University in Partial Fulfillment
of the Requirements for the Award of the Degree of

Master of Science in Statistics

by

K. SUGANYA

(Reg. No.17STAC32)

Under the Guidance of

Dr. R. VIJAYARAGHAVAN

Department of Statistics
Bharathiar University
Coimbatore 641 046
Tamil Nadu, India

April 2019

Certificate

This is to certify that the project report entitled '**A Study on Multicollinearity using Ridge Regression'** submitted to Bharathiar University, Coimbatore - 641046 in partial fulfillment of the requirements for the Award of the Degree of **Master of Science in Statistics** is a record of the original work done by **Ms. K. SUGANYA** during the period of her study in the Department of Statistics, Bharathiar University under my supervision and guidance. This project work has not formed the basis for the award of any Degree / Diploma / Associateship / Fellowship or similar title to any other candidate of any university / institution.

Dr. R. VIJAYARAGHAVAN
Project Guide

Countersigned by

Dr. R. VIJAYARAGHAVAN
Head of the Department

Coimbatore
April 2019

Declaration

I, **K. SUGANYA**, hereby declare that the project report entitled **'A Study on Multicollinearity using Ridge Regression'** submitted to Bharathiar University, Coimbatore 641046 in partial fulfillment of the requirements for the Award of the Degree of **Master of Science in Statistics** is a record of the original work done by me under the guidance and supervision of **Dr. R. VIJAYARAGHAVAN,** Professor and Head, Department of Statistics, Bharathiar University, and it has not formed the basis for the award of any Degree / Diploma / Associateship / Fellowship or similar title to any candidate of any university / institution.

K. SUGNAYA
Candidate

Coimbatore
April 2019

Acknowledgments

“Every great human achievement is preceded by extended periods of dedicated, concentrated efforts”.

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely fortunate to have got this all along the completion of my project work. Whatever I have done is only due to such guidance and assistance and I would not forget to thank them.

“Develop an attitude of gratitude, and give thanks for everything that happens to you, knowing that every step forward is a step toward achieving something bigger and better that your current situation”.

I own my profound gratitude to my supervisor and guide **Dr. R. VIJAYARAGHAVAN,** Professor and Head, Department of Statistics, Bharathiar University, who took keen interest in my project and guided all along, till the completion of my project work by providing all his valuable guidelines. Words are insufficient to express.

I render my sincere gratitude to **Dr. R. JAISANKAR,** Professor, who constantly supported and encouraged throughout the course of completion of this work. Words are insufficient to express.

I would also like to thank **Dr. R. MUTHUKRISHNAN,** Professor, **Dr. K. M. SAKTHIVEL,** Associate Professor, **Dr. V. KAVIYARASU,** Assistant Professor, **Dr. K. PRADEEPAVEERAKUMARI,** Assistant Professor, **Dr. S. JAYALAKSHMI**, Assistant Professor, **Dr. S. GANDHIYA VENDHAN**, Assistant Professor, Department of Statistics, Bharathiar University, Coimbatore for their valuable suggestions and encouragement.

I wish to express my thanks to non-teaching staffs for their kind assistance for providing the necessary facilities to carry out the work.

With all the blessings of my late uncle Mr. K. Arivazhagan, I owe my sincere gratitude to my father Mr. S. Kannan, my mother Mrs. K. Thenmozhi, my sister Ms. K. Kalaipriya and my brother Mr. T. K. Pranav for their constant encouragement and support till this moment.

I express my profound and deep sense of gratitude to all my friends for their constant support, sustained interest and involvement throughout the study. My sincere thanks to all the research scholars of the Department of Statistics, Bharathiar University for their timely help in completing this project work.

K. SUGANYA

Contents

Chapter 1

Introduction

This chapter provides the concepts of regression analysis in brief. The organization of the contents of the subsequent chapters is presented at the end.

Regression Analysis

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent variable, called target or response variable, and independent variable(s), called predictor variables. This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Regression analysis is a statistical technique used to describe relationships among variables. The simplest case to examine is one in which a variable y, referred to as the dependent or target variable, may be related to one variable x, called an independent or explanatory variable, or simply a regressor. If the relationship between $y$and x is believed to be linear, then the equation for a line may be appropriate: $y = \beta_o + \beta_1 x + \varepsilon$

This is called a linear regression model. Customarily x is called the independent variable and y is called the dependent variable.

There are multiple benefits of using regression analysis. They are,

- It indicates the significant relationships between dependent variable and independent variable.
- It indicates the strength of impact of multiple independent variables on a dependent variable.
- Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help us to eliminate and evaluate the best set of variables to be used for building predictive models.

Types of Regression

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line). Some of the widely used modeling techniques are given below:

1. Linear Regression
2. Logistic Regression
3. Polynomial Regression
4. Stepwise Regression
5. Ridge Regression
6. Lasso Regression
7. Elastic Net Regression

Multicollinearity and Its Effects

One of the first steps in a regression analysis is to determine if multicollinearity exits among regressors or predictors.

If there is no linear relationship between the regressors, they are said to be orthogonal. When the regressors are orthogonal, inferences such as those illustrated above can be made relatively easily. Unfortunately, in most applications of regression, the regressors are not orthogonal. Sometimes the lack of orthogonality is not serious. However, in some situations the regressors are nearly perfectly linearly related, and in such cases the inferences based on the regression model can be misleading or erroneous. When there are near - linear dependencies among the regressors, the problem of multicollinearity is said to exist. Specifically, we will examine the causes of multicollinearity, some of its specific effects on inference, methods of detecting the presence of multicollinearity, and some techniques for dealing with the problem.

Sources of Multicollinearity

There are four primary **sources of multicollinearity.** They are,

1. The data collection method employed

2. Constraints on the model or in the population

3. Model specification

4. An over-defined model

It is important to understand the differences among these sources of multicollinearity, as the recommendations for analysis of the data and interpretation of the resulting model depend to some extent on the cause of the problem.

Effects of Multicollinearity

The presence of multicollinearity has several potentially serious effects on the least - squares estimates of the regression coefficients. Some of these effects may be easily demonstrated. Suppose that there are only two regressor variables, x1 and x2. The model, assuming that X_1, X_2 and Y are scaled to unit lengths,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

and the least - squares normal equations are

$$(\mathrm{X}'\mathrm{X})\hat{\beta} = \mathrm{X}'\mathrm{y}$$

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

where r_{12} is the simple correlation between x_1 and x_2. r_{jy} is the simple correlation between x_j and y, j = 1, 2. Now the inverse of $(\mathrm{X}'\mathrm{X})$ is

$$\mathrm{C} = (\mathrm{X}'\mathrm{X})^{-1}$$

and the estimates of the regression coefficients are

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}{}^2}$$

and $$\hat{\beta}_2 = \frac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}{}^2}.$$

If there is strong multicollinearity between X_1 and X_2, then the correlation coefficient r_{12} will be large. From the above equation, one may observe that as $|r_{12}| \to 1$, $\mathrm{Var}(\hat{\beta}_j) \to \infty$ and $\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \mathrm{C}_{12}\sigma^2 \to \pm\infty$ depending on whether $r_{12} \to +1$ or $r_{12} \to -1$.

Therefore, strong multicollinearity between X_1 and X_2 results in large variances and covariances for the least - squares estimators of the regression coefficients. This implies that different samples taken at the same x levels could lead to widely different estimates of the model parameters. When there are more than two regressor variables, multicollinearity produces similar effects. It can be shown that the diagonal elements of the $\mathrm{C} = (\mathrm{X'X})^{-1}$ matrix are

$$\mathrm{C_{jj}} = \frac{1}{1 - R_j^{\,2}}, j = 1,2,..,p$$

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

Methods for Detecting Multicollinearity

Desirable characteristics of a diagnostic procedure are that it directly reflects the degree of the multicollinearity problem and provide information helpful in determining which regressors are involved. Several techniques have been proposed for detecting multicollinearity. They are:

1. Examination of the Correlation Matrix
2. Variance Inflation Factors (VIF)
3. Eigensystem Analysis of X′X

Examination of the Correlation Matrix

A very simple measure of multicollinearity is inspection of the off-diagonal elements r_{ij} in X′X. If regressors X_i and X_j are nearly linearly dependent, then $|r_{ij}|$ will be near unity. The matrix X′X reveals the correlation between any two regressors. Thus, inspection of the correlation matrix indicates whether there is any near–linear dependencies in the data. Examining the simple correlations r_{ij} between the regressors is helpful in detecting near - linear dependence between pairs of regressors only. Unfortunately, when more than two regressors are involved in a near - linear dependence, there is no assurance that any of the pairwise correlations r_{ij} will be large. Generally, inspection of the r_{ij} is not sufficient for detecting anything more complex than pair-wise multicollinearity. In this case, we check the variation inflation factors for all the regressors.

Variance Inflation Factors (VIF)

We know that the diagonal elements of the $\mathrm{C} = (\mathrm{X'X})^{-1}$ matrix are very useful in detecting multicollinearity. The j^{th} diagonal element of C can be written as

$$C_{jj} = (1 - {R_j}^2)^{-1},$$

where ${R_j}^2$ is the coefficient of determination obtained when x_j is regressed on the remaining $p - 1$ regressors.

If x_j is nearly orthogonal to the remaining regressors, ${R_j}^2$ is small and C_{jj} is close to unity, while if x_j is nearly linearly dependent on some subset of the remaining regressors, ${R_j}^2$ is near unity and is large. Since the variance of the j^{th} regression coefficients is $C_{jj}\sigma^2$, we can view C_{jj} as the factor by which the variance of $\hat{\beta}_j$ is increased due to near-linear dependences among the regressors. We called

$$\mathbf{VIF_j} = \boldsymbol{C_{jj}} = (\mathbf{1} - \boldsymbol{R_j}^{\mathbf{2}})^{-\mathbf{1}},$$

the variance inflation factor.

The VIF for each term in the model measures the combined effect of the dependences among the regressors on the variance of that term. One or more large VIFs indicate multicollinearity. Practical experience indicates that if any of the VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

Eigensystem Analysis of $X'X$

The characteristic roots or eigenvalues of $X'X$, say $\lambda_1, \lambda_2, \dots, \lambda_p$, can be used to measure the extent of multicollinearity in the data. † If there are one or more near-linear dependences in the data, then one or more of the characteristic roots will be small. One or more small eigenvalues imply that there are near - linear dependences among the columns of X. Some analysts prefer to examine the condition number of $X'X$, defined as

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

This is just a measure of the spread in the eigenvalue spectrum of $X'X$.

Generally, if the condition number is less than 100, there is no serious problem with multicollinearity. Condition numbers between 100 and 1000 imply moderate to strong multicollinearity, and if κ exceeds 1000, severe multicollinearity is indicated. The condition indices of the $X'X$ matrix is

$$\kappa_j = \frac{\lambda_{max}}{\lambda_j}, j = 1,2,\dots,p$$

Clearly, the largest condition index is the condition number defined as κ. The number of condition indices that are large (say ≥ 1000) is a useful measure of the number of near - linear dependences in $X'X$.

Methods for Removing Multicollinearity

- Collecting Additional Data
- Model Respecification
- Ridge Regression
- Principal - Component Regression

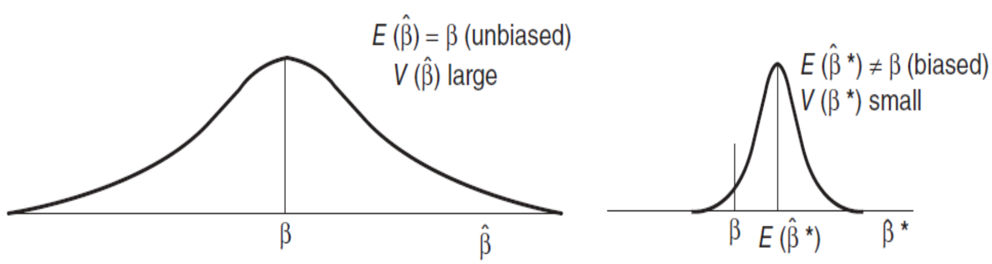
Chapter 2

Concept of Ridge Regression

This chapter provides the concept and the significance of Ridge Regression.

Ridge Regression

When the method of least squares is applied to non-orthogonal data, very poor estimates of the regression coefficients can be obtained. The variance of the least- squares estimates of the regression coefficients maybe considerably inflated, and the length of the vector of least-squares parameter estimates is too long on the average. This implies that the absolute value of the least-squares estimates are too large and that they are very unstable, that is, their magnitudes and signs may change considerably given a different sample.



(a) (b)

The problem with the method of least squares is the requirement that $\hat{\beta}$ be an unbiased estimator of β. The Gauss - Markov assures us that the least - squares estimator has minimum variance in the class of unbiased linear estimators, but there is no guarantee that this variance will be small. The situation is illustrated in Figure (a) and (b), where the sampling distribution of $\hat{\beta}$, the unbiased estimator of β, is shown. The variance of $\hat{\beta}$ is large, implying that confidence intervals on β would be wide and the point estimate $\hat{\beta}$ is very unstable. One way to alleviate this problem is to drop the requirement that the estimator of β be unbiased. Suppose that we can find a biased estimator of β, say $\hat{\beta}^*$, that has a smaller variance than the unbiased estimator $\hat{\beta}$. The mean square error of the estimator $\hat{\beta}^*$ is defined as

$$MSE\left(\hat{\beta}^*\right) = Var\left(\hat{\beta}^*\right) + (Bias\ in\ \hat{\beta}^*)^2$$

Note that the MSE is just the expected squared distance from $\hat{\beta}^*$ to β. By allowing a small amount of bias in $\hat{\beta}^*$, the variance of $\hat{\beta}^*$ can be made small such that the MSE of $\hat{\beta}^*$is less than the variance of the unbiased estimator β. Figure (b) illustrates a situation where the variance of the biased estimator is considerably smaller than the variance of the unbiased estimator (Figure (a)). Consequently, confidence intervals on β would be much narrower using the biased estimator. The small variance for the biased estimator also implies that $\hat{\beta}^*$ is a more stable estimator of β than is the unbiased estimator $\hat{\beta}$.

A number of procedures have been developed for obtaining biased estimators of regression coefficients. One of these procedures is ridge regression. The ridge estimator is found by solving a slightly modified version of the normal equations. Specifically, we define the ridge estimator $\hat{\beta}^*$ as the solution to

$$(\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} = \mathbf{X}'\mathbf{y}$$

or

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} = (\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

where $k \geq 0$ is a constant selected by the analyst. The procedure is called ridge regression because the underlying mathematics are similar to the method of ridge analysis used for describing the behavior of second – order response surfaces. Note that when k = 0, the ridge estimator is the least – squares estimator. The ridge estimator is a linear transformation of the least - squares estimator since

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{R}} = (\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{Z}_{\mathbf{k}}\hat{\boldsymbol{\beta}}$$

Therefore, since $E\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}\right) = E\left(\mathbf{Z}_{\mathbf{k}}\hat{\boldsymbol{\beta}}\right) = \mathbf{Z}_{\mathbf{k}}\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ is a biased estimator of β.

We usually refer to the constant k as the biasing parameter. The covariance matrix of $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ is,

$$V\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}\right) = \sigma^2(\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})^{-1}\mathrm{X}'\mathrm{X}(\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})^{-1}$$

The mean square error of the ridge estimator is

$$MSE\ \left(\hat{\beta}^*\right) = Var\left(\hat{\beta}^*\right) + \left(Bias\ in\ \hat{\beta}^*\right)^2$$

$$= \sigma^2\mathrm{Tr}[(\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})^{-1}\mathrm{X}'\mathrm{X}(\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})^{-1}] + \mathrm{k}^2\beta'(\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})^{-2}\beta$$

where $\lambda_1, \lambda_2, \dots, \lambda_p$ are the eigenvalues of $\mathrm{X}'\mathrm{X}$.

The first term on the right-hand side of this equation is the sum of variances of the parameters in $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ and the second term is the square of the bias. If k > 0, note that the bias in $\hat{\beta}_R$ increases with k. However, the variance decreases ask increases. In using ridge regression, we would like to choose a value of k such that the reduction in the variance term is greater than the increase in the squared bias. If this can be done, the mean square error of the ridge estimator $\hat{\beta}_R$ will be less than the variance of the least - squares estimator. It has been proved that there exists a nonzero value of k for which the MSE of $\hat{\beta}_R$ is less than the variance of the least -squares estimator $\hat{\beta}$, provided that $\beta'\beta$ is bounded. The residual sum of squares is

$$SS_{Res} = \left(y - X\hat{\beta}_R\right)^{-1}\left(y - X\hat{\beta}_R\right)$$

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

Since the first term on the right-hand side of the above equation is the residual sum of squares for the least-squares estimates $\hat{\beta}$, we see that as k increases, the residual sum of squares increases. Consequently, because the total sum of squares is fixed, R^2decreases as k increases. Therefore, the ridge estimate will not necessarily provide the best "fit" to the data, but this should not overly concern us, since we are more interested in obtaining a stable set of parameter estimates. The ridge estimates may result in an equation that does a better job of predicting future observations than would least squares (although there is no conclusive proof that this will happen). It has been suggested that an appropriate value of k may be determined by inspection of the ridge trace. The ridge trace is a plot of the elements of$\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}$versus k for values of k usually in the interval 0 – 1. It has been suggested using up to about 25 values of k, spaced approximately logarithmically over the interval [0, 1]. If multicollinearity is severe, the instability in the regression coefficients will be obvious from the ridge trace. As k is increased, some of the ridge estimates will vary dramatically. At some value of k, the ridge estimates $\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ will stabilize. The objective is to select a reasonably small value of k at which the ridge estimates $\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ are stable. Hopefully this will produce a set of estimates with smaller MSE than the least - squares estimates.

Method of Choosing *k*

Ridge Trace

To obtain the ridge solution, we must solve the equations

$$(\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} = \mathbf{X}'\mathbf{y}$$

for several values $0 \leq \mathrm{k} \leq 1$, with $\mathbf{X'X}$ and $\mathbf{X'y}$ in correlation form.

The ridge coefficients for several values of k are to be listed in the table. The table also shows the residual mean square and $R^2$for each ridge model.

Notice that as k increases, MS_{Res} increases and R^2decreases. The ridge trace illustrates the instability of the least - squares solution, as there are large changes in the regression coefficients for small values of k. However, the coefficients stabilize rapidly as k increases. Judgment is required to interpret the ridge trace and select an appropriate value of k. We want to choose k large enough to provide stable coefficients, but not unnecessarily large ones, as this introduces additional bias and increases the residual mean square. From the Ridge Trace, reasonable coefficient stability is achieved in the certain region of$0< k <1$ without a severe increase in the residual mean square. By choosing the value of k, we can obtain the ridge regression model.

Chapter 3

Case Studies

This chapter provides the data analysis on different case studies on multicollinearity in a detailed manner.

Data Analysis on Multicollinearity

The study of multicollinearity is carried out based on three sets of data taken from three different sources. The data considered are industrial data, economic data, health data.

Industrial Data

Montgomery, et al., (2012) considered data (Table 3.1, here) concerning the percentage of conversion of *n*-heptane to acetylene and three explanatory variables, Reactor temperature (T), mole ratio (H) and contact time (C). These are typical chemical process data for which a full quadratic response surface in all three regressors is considered to be an appropriate tentative model.

Table 3.1: Acetylene data

Observation	P	T	H	C
1	49.0	1300	7.5	0.0120
2	50.2	1300	9.0	0.0120
3	50.5	1300	11.0	0.0115
4	48.5	1300	13.5	0.0130
5	47.5	1300	17.0	0.0135
6	44.5	1300	23.0	0.0120
7	28.0	1200	5.3	0.0400
8	31.5	1200	7.5	0.0380
9	34.5	1200	11.0	0.0320
10	35.0	1200	13.5	0.0260
11	38.0	1200	17.0	0.034
12	38.5	1200	23.0	0.0410
13	15.0	1100	5.3	0.0840
14	17.0	1100	7.5	0.0980
15	20.5	1100	11.0	0.0920
16	29.5	1100	17.0	0.0860

0In Table 3.1, P is the dependent variable representing conversion of n - Heptane to Acetylene (%) and T, H and C are the regressors representing reactor temperature (°C), ratio of H_2 ton-Heptane (mole ratio) and contact time (sec), respectively.

Linear Model for the Data

Each of the original regressors has been scaled using the unit normal scaling [subtracting the average (centering) and dividing by the standard deviation]. The squared and cross - product terms are generated from the scaled linear terms. The centering the linear terms is helpful in removing on essential ill - conditioning when fitting polynomials. The least-squares fit is,

$$\hat{P} = 35.897 + 4.019T + 2.781H - 8.031C - 3.768HC - 12.54T^2 - 0.973H^2 - 11.594C^2$$

Summary Statistics for the Model

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
T	16	2.499999E-10	1	-1.395391	1.085304
H	16	6.250002E-11	1	-1.261693	1.864392
C	16	-1.875001E-10	1	-0.9106779	1.823331
TH	16	0.209651	0.8844809	-1.122874	2.023432
TC	16	-0.8983163	0.897006	-2.54426	0.07013789
HC	16	-0.2252166	0.8390673	-1.742187	1.162027
T2	16	0.9375	0.7929049	0.02403846	1.947115
H2	16	0.9375	1.105251	0.03480075	3.475957
C2	16	0.9375	1.027236	9.76E-05	3.324537
P	16	36.10625	11.89877	15	50.5

For each variable, the descriptive statistics for the given data are computed. This report is particularly useful for checking that the correct variables were selected.

Multicollinearity Detection

Plotting each two of the independent variables to check the present of linearity.

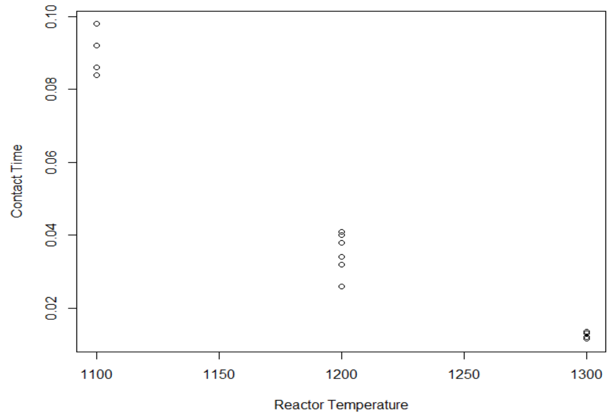


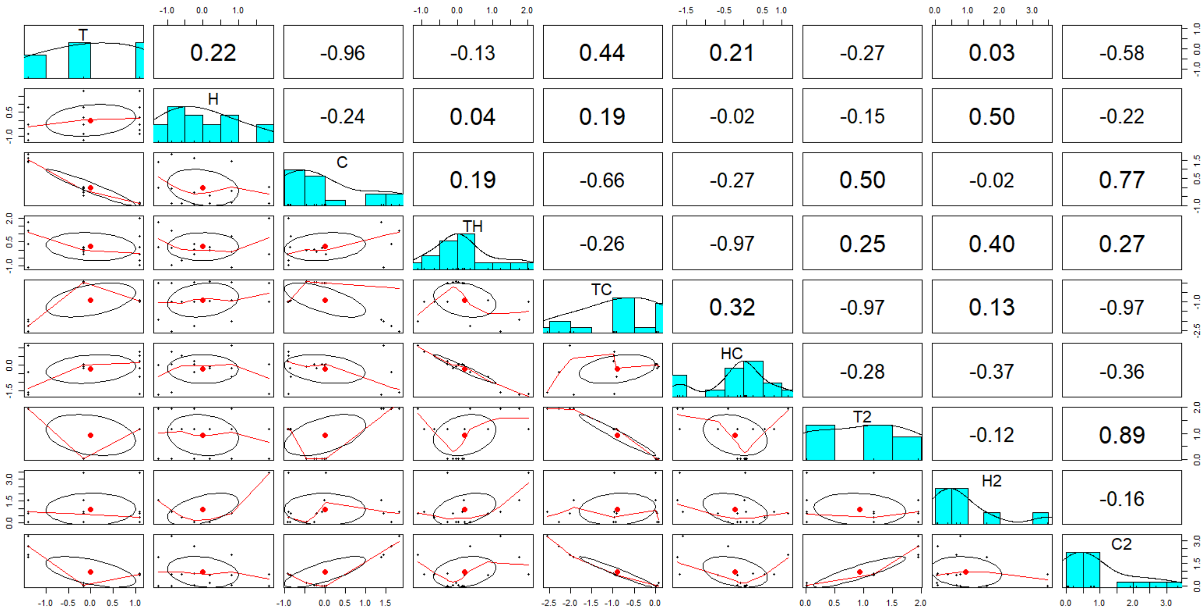
Fig (c)

Here from the Fig(c), as the Reactor Temperature increases, the Contact time decreases. So, there is the linear relation in between these two variables. Then, the correlation between these variables may be high.

Correlation Matrix $X'X$

	T	H	C	TH	TC	HC	T2	H2	C2
T	1.0000000	0.22362776	-0.95820405	-0.13241739	0.4428236	0.20553866	-0.2707456	0.03095990	-0.5767868
H	0.2236278	1.00000000	-0.24023098	0.03868762	0.1922627	-0.02306559	-0.1477108	0.49754636	-0.2239058
C	-0.9582041	-0.24023098	1.00000000	0.19498531	-0.6605265	-0.27411884	0.5009622	-0.01751058	0.7651532
TH	-0.1324174	0.03868762	0.19498531	1.00000000	-0.2648504	-0.97448134	0.2463122	0.39789760	0.2747680
TC	0.4428236	0.19226267	-0.66052652	-0.26485039	1.0000000	0.32351596	-0.9722442	0.12583104	-0.9721670
HC	0.2055387	-0.02306559	-0.27411884	-0.97448134	0.3235160	1.00000000	-0.2792725	-0.37460454	-0.3585293
T2	-0.2707456	-0.14771083	0.50096224	0.24631222	-0.9722442	-0.27927248	1.0000000	-0.12359068	0.8936644
H2	0.0309599	0.49754636	-0.01751058	0.39789760	0.1258310	-0.37460454	-0.1235907	1.00000000	-0.1579789
C2	-0.5767868	-0.22390584	0.76515318	0.27476802	-0.9721670	-0.35852931	0.8936644	-0.15797895	1.0000000

The $X'X$ matrix reveals the high correlation between reactor temperature (T) and contact time (C) suspected earlier from inspection of Fig (C), since $r_{13} = -0.958$. Thus, inspection of the correlation matrix indicates that there are several near-dependencies in the acetylene data.



From the above correlation plot, it is clear that T and C are correlated. Generally, inspection of the r_{ij} is not sufficient for detecting anything more complex than pairwise multicollinearity. So, we find Variation Inflation Factor for the regressors.

Variation Inflation Factor

Independent Variable	Variance Inflation	R-Squared Vs Other X's	Tolerance
T	375.2477	0.9973	0.0027
H	1.7406	0.4255	0.5745
C	680.2800	0.9985	0.0015
TH	31.0371	0.9678	0.0322
TC	6563.3445	0.9998	0.0002
HC	35.6113	0.9719	0.0281
T2	1762.5752	0.9994	0.0006
H2	3.1643	0.6840	0.3160
C2	1156.7662	0.9991	0.0009

Maximum of VIF = 6563.345

Hence, we conclude that the multicollinearity problem exists since VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

Eigensystem Analysis of $\mathbf{X}'\mathbf{X}$

Extent of Multicollinearity

No.	Eigenvalue	Incremental Percent	Cumulative Percent	Condition Index
1	4.205230	46.72	46.72	1.00
2	2.161999	24.02	70.75	1.95
3	1.138677	12.65	83.40	3.69
4	1.040475	11.56	94.96	4.04
5	0.385230	4.28	99.24	10.92
6	0.049538	0.55	99.79	84.89
7	0.013625	0.15	99.94	308.63
8	0.005128	0.06	100.00	820.08
9	0.000097	0.00	100.00	43381.31

The above table gives the eigenvalue analysis of the independent variables after they have been centered and scaled.

Eigen Value

Eigen value is used to determine the extent of multicollinearity in the data. One or more small eigenvalues imply that there are near - linear dependences among the columns of X.

There are four very small eigenvalues, therefore, this is a symptom of seriously ill-conditioned data.

Condition number (k) = 43381.31

Since, k exceeds 1000, **severe** multicollinearity is indicated.

Condition numbers greater than 1000 indicate a severe multicollinearity problem while condition numbers between 100 and 1000 indicate a mild multicollinearity problem.

Condition Index

Useful measure of the number of near-linear dependences in $\mathrm{X}'\mathrm{X}$**.**

Since one of the condition indices **exceeds 1000** (and two others exceed 100), we conclude that there is **at least one strong near - linear dependence** in the acetylene data.

Ridge Estimator

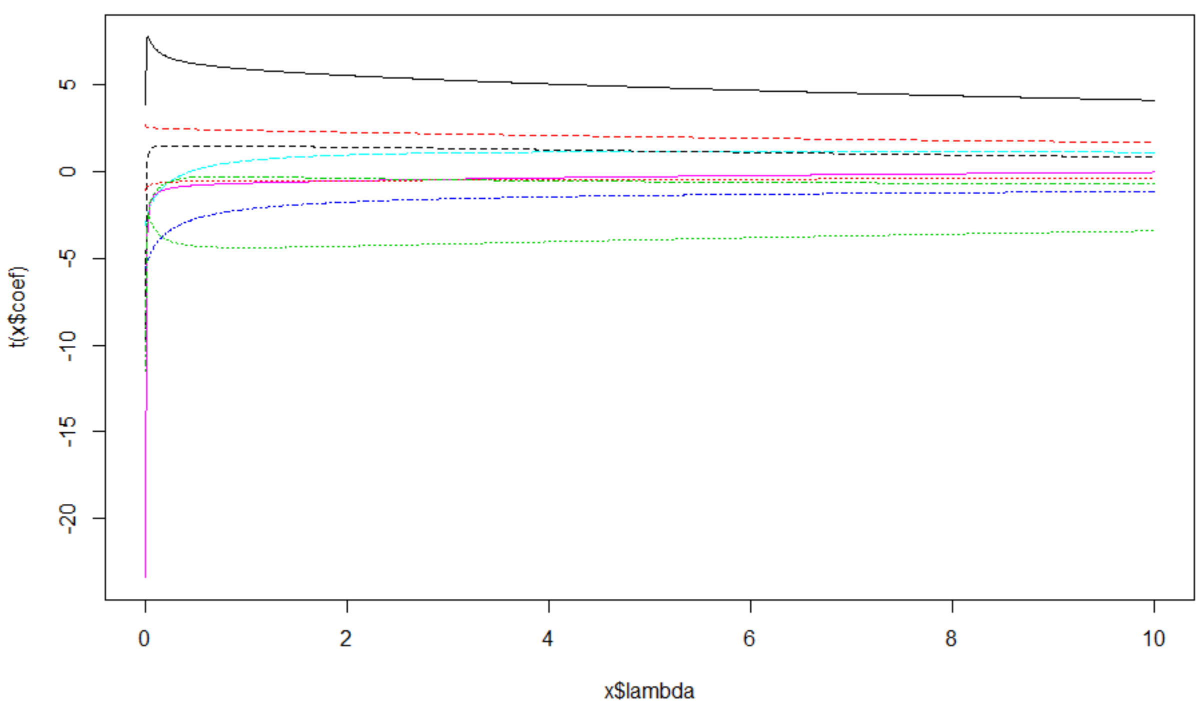
Ridge Trace

> betacR

	k-0	k-0.001	k-0.002	k0.004	k-0.008	k-0.016	k-0.032	k-0.064	k-0.128	k-0.256	k-0.512
r1	0.33648680	0.67653541	0.66493869	0.63594221	0.60008268	0.56702466	0.53902338	0.51210345	0.48052510	0.43785361	0.378353470
r2	0.23349593	0.22397787	0.22204157	0.21972026	0.21723449	0.21472311	0.21165039	0.20655026	0.19712252	0.18066860	0.155417456
r3	-0.67589625	-0.21325337	-0.22863131	-0.26719030	-0.31348412	-0.35148279	-0.37351660	-0.37992192	-0.37234429	-0.35000790	-0.310842419
r12	-0.47995686	-0.44790490	-0.42584712	-0.39131072	-0.34374773	-0.28798585	-0.23291655	-0.18626476	-0.15089099	-0.12501418	-0.104520642
r13	-2.03395608	-0.27757217	-0.18886480	-0.13515462	-0.10176999	-0.08097999	-0.06758417	-0.05703527	-0.04549198	-0.02995039	-0.009331586
r23	-0.26571829	-0.21733993	-0.19202366	-0.15355436	-0.10197365	-0.04331978	0.01227576	0.05622512	0.08484699	0.09844806	0.099127601
r11	-0.83454188	0.06419349	0.10344940	0.12128356	0.12616501	0.12530612	0.12477337	0.12568245	0.12288621	0.10954530	0.082659881
r22	-0.09035418	-0.07312709	-0.06817921	-0.06205042	-0.05580347	-0.05092459	-0.04804334	-0.04640976	-0.04443561	-0.04060448	-0.034157279
r33	-1.00085767	-0.24501050	-0.18528657	-0.13132424	-0.08250316	-0.04553240	-0.02663261	-0.02501237	-0.03383078	-0.04629422	-0.058454547

From the above table, the reasonable coefficient stability is achieved in the region

0.016<k<0.064



From the above figure, we observe that k = 0.032, where the reasonable coefficient is achieved. This is a plot that we have added that shows the impact of k on the variance Inflation factors. Since the major goal of ridge regression is to remove the impact of multicollinearity, it is important to know at what point multicollinearity has been dealt with. This plot shows this.

Since the rule-of-thumb is that multicollinearity is not a problem once all VIFs are less than 10, we inspect the graph for this point. In this example, it appears that all VIFs are small enough once k is greater than 0.03.

Hence, this is the value of k that this plot would indicate we use. Since this plot indicates k = 0.03 and the ridge trace indicates a value near 0.03, we would select 0.032 as our final result. The rest of the reports are generated for this value of k.

Variance Inflation Factor

k	T	H	C	TH	TC	HC
0.000000	375.2477	1.7406	680.2800	31.0371	6563.3445	35.6113
0.010000	9.0461	1.4205	12.3815	10.9811	1.3959	11.6261
0.020000	4.0308	1.3547	4.7245	6.0003	0.5668	6.1776
0.030000	2.4770	1.3022	2.5580	3.9202	0.3686	3.9608
0.040000	1.7705	1.2551	1.6502	2.8303	0.2881	2.8217
0.050000	1.3801	1.2117	1.1840	2.1786	0.2465	2.1511
0.060000	1.1372	1.1713	0.9124	1.7537	0.2214	1.7194
0.070000	0.9735	1.1333	0.7399	1.4590	0.2047	1.4231
0.080000	0.8565	1.0976	0.6230	1.2451	0.1927	1.2098
0.090000	0.7692	1.0639	0.5399	1.0841	0.1836	1.0504
0.109594	0.6500	1.0029	0.4331	0.8643	0.1706	0.8343

Variance Inflation Factor (Continued)

k	T2	H2	C2
0.000000	1762.5752	3.1643	1156.7662
0.010000	7.0873	1.8113	10.9611
0.020000	4.5800	1.6420	6.3887
0.030000	3.3390	1.5431	4.4439
0.040000	2.5924	1.4656	3.3420
0.050000	2.0968	1.3990	2.6323
0.060000	1.7469	1.3396	2.1402
0.070000	1.4888	1.2854	1.7819
0.080000	1.2920	1.2356	1.5115
0.090000	1.1379	1.1893	1.3017
0.109594	0.9182	1.1075	1.0060

This report gives the values that are plotted on the variance inflation factor plot. It is to determine when all three VIFs are less than 10.

K Analysis

k	R2	Sigma	B'B	Ave VIF	Max VIF
0.000000	0.9977	0.9014	6.7689	1178.8630	6563.3445
0.010000	0.9896	1.9167	0.6465	7.4123	12.3815
0.020000	0.9835	2.4157	0.5851	3.9406	6.3887
0.030000	0.9778	2.8021	0.5566	2.6570	4.4439
0.040000	0.9723	3.1285	0.5383	2.0018	3.3420
0.050000	0.9670	3.4157	0.5245	1.6089	2.6323
0.060000	0.9619	3.6746	0.5131	1.3491	2.1402
0.070000	0.9568	3.9116	0.5031	1.1655	1.7819
0.080000	0.9518	4.1311	0.4942	1.0293	1.5115
0.090000	0.9469	4.3359	0.4859	0.9245	1.3017
0.109594	0.9375	4.7030	0.4712	0.7763	1.1075

This report provides a quick summary of the various statistics that might go into the choice of k.

k

This is the actual value of k. Note that the value found by the analytic search (0.03) sticks out as you glance down this column because it does not end in zeros.

R2

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

Sigma

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

B’B

This is the sum of the squared standardized regression coefficients. Ridge regression assumes that this value is too large and so the method tries to reduce this. We want to find a value for k at which this value has stabilized.

Ave VIF

This is the average of the variance inflation factors.

Max VIF

This is the maximum variance inflation factor. Since we are looking for that value of k which results in all VIFs being less than 5 or 10, this value is very helpful in your selection of k.

Ridge versus Least Squares Comparison

Independent Variable	Regular Ridge Coeff's	Regular L.S. Coeff's	Stand'zed Ridge Coeff's	Stand'zed L.S. Coeff's	Ridge Standard Error	L.S. Standard Error
Intercept	35.01574	35.89579				
T	6.413713	4.003777	0.5390	0.3365	1.123036	4.508698
H	2.518378	2.778313	0.2117	0.2335	0.8427978	0.3070756
C	-4.444386	-8.042332	-0.3735	-0.6759	1.128595	6.07066
TH	-3.133385	-6.456775	-0.2329	-0.4800	1.60116	1.466033
TC	-0.8965026	-26.98039	-0.0676	-2.0340	0.4870013	21.02129
HC	0.1740819	-3.768136	0.0123	-0.2657	1.693956	1.655347
T2	1.872418	-12.52359	0.1248	-0.8345	1.662389	12.3238
H2	-0.5172187	-0.9727232	-0.0480	-0.0904	0.8286957	0.3746029
C2	-0.3084931	-11.59322	-0.0266	-1.0009	1.475083	7.706276

This report provides a detailed comparison between the ridge regression solution and the ordinary least squares solution to the estimation of the regression coefficients. From the above report, it is clear that the standard error of the Ridge is much smaller than that of the Least Square standard error. Therefore, the optimum standardized ridge coefficients are obtained at k = 0.032.

Ridge Regression Coefficient

Independent Variable	Regression Coefficient	Standard Error	Standardized Regression Coefficient	VIF
Intercept	35.01574			
T	6.413713	1.123036	0.5390	2.2948
H	2.518378	0.8427978	0.2117	1.2924
C	-4.444386	1.128595	-0.3735	2.3176
TH	-3.133385	1.60116	-0.2329	3.6492
TC	-0.8965026	0.4870013	-0.0676	0.3472
HC	0.1740819	1.693956	0.0123	3.6758
T2	1.872418	1.662389	0.1248	3.1613
H2	-0.5172187	0.8286957	-0.0480	1.5264
C2	-0.3084931	1.475083	-0.0266	4.1776

This report provides the details of the ridge regression solution. Since the VIF is lesser than 5, the multicollinearity is vanished.

Therefore, the Ridge regression model with finite standard error is,

$$\hat{y} = 0.5390\, T + 0.2117H - 0.3735\, C - 0.2329\, TH - 0.0676\, TC + 0.0123\, HC + 0.1248\, T^2 - 0.0480\, H^2 - 0.0266C^2$$

Ridge Parameter

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} = (\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

```
> ridgeParameter
                P
 [1,]  0.53902338
 [2,]  0.21165039
 [3,] -0.37351660
 [4,] -0.23291655
 [5,]  0.01227576
 [6,] -0.06758417
 [7,]  0.12477337
 [8,] -0.04804334
 [9,] -0.02663261
```

Economic Data: (Longley Data)

Gujarati, et al, (2017) considered a data set collected by Longley (1067) to illustrate the problem of multicollinearity. The data set is presented here in Table 3.2 where the data are time series for the years 1947–1962 and pertain to the following variables:

Y = number of people employed (in thousands);

X_1 = gross national product (GNP) implicit price deflator;

X_2 = GNP, millions of dollars;

X_3 = number of people unemployed in thousands;

X_4 = number of people in the armed forces;

X_5 = noninstitutionalized population over 14 years of age

X_6 = year

Data Model

Each of the original regressors has been scaled using the unit normal scaling [subtracting the average (centering) and dividing by the standard deviation]. The squared and cross - product terms are generated from the scaled linear terms. The centering the linear terms is helpful in removing nonessential ill - conditioning when fitting polynomials. The least-squares fit is,

$$\hat{y} = 65317 - 523X_1 + 7156.8X_2 - 377.4X_3 - 390.1X_4 - 2806.8X_5$$

Table 3.2: Longley' Time Series

y	x1	x2	x3	x4	x5
60323	830	234289	2356	1590	107608
61122	885	259426	2325	1456	108632
60171	882	258054	3682	1616	109773
61187	895	284599	3351	1650	110929
63221	962	328975	2099	3099	112075
63639	981	346999	1932	3594	113270
64989	990	365385	1870	3547	115094
63761	1000	363112	3578	3350	116219
66019	1012	397469	2904	3048	117388
67857	1046	419180	2822	2857	118734
68169	1084	442769	2936	2798	120445
66513	1108	444546	4681	2637	121950
68655	1126	482704	3813	2552	123366
69564	1142	502601	3931	2514	125368
69331	1157	518173	4806	2572	127852
70551	1169	554894	4007	2827	130081

Summary Statistics for the Model

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
x1	16	-6.249996E-11	1.032796	-1.787872	1.456496
x2	16	-2.775558E-17	1.032796	-1.594051	1.7373
x3	16	-6.938894E-17	1.032796	-1.462561	1.782387
x4	16	-6.250001E-11	1.032796	-1.707704	1.465244
x5	16	-6.250001E-11	1.032796	-1.457414	1.879227
y	16	65317	3511.968	60171	70551

For each variable, the descriptive statistics for the given data are computed. This report is particularly useful for checking that the correct variables were selected.

Multicollinearity Detection

Plotting each two of the independent variables to check the present of linearity:

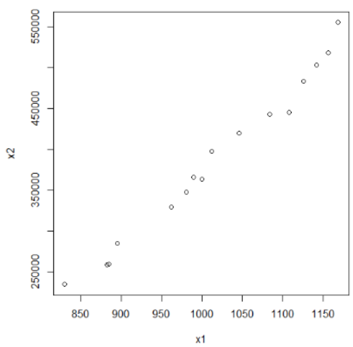


Fig (d)

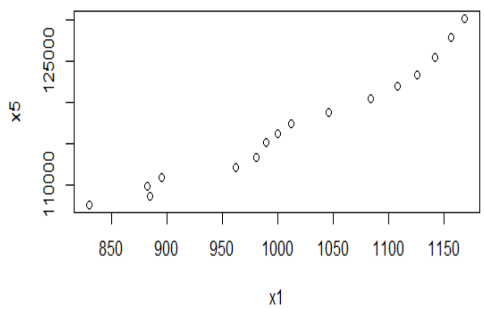


Fig (e)

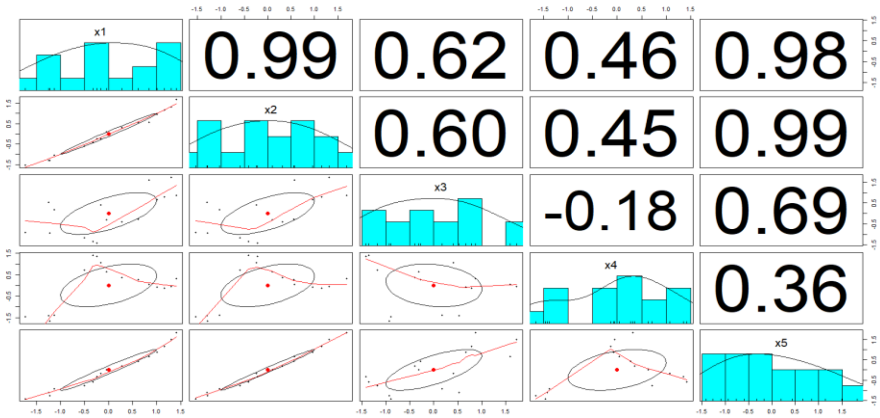
Here, in Fig (d), as the x1 increases, the x2 increases and also in Fig (e), as x1 increases, x5 also increases. So, there is the linear relation in between these two pairs of variables. Then, the correlation between these variables may be high.

Correlation Matrix $X'X$

	x1	**x2**	**x3**	**x4**	**x5**	**y**
x1	1.000000	0.991589	0.620633	0.464744	0.979163	0.970899
x2	0.991589	1.000000	0.604261	0.446437	0.991090	0.983552
x3	0.620633	0.604261	1.000000	-0.177421	0.686552	0.502498
x4	0.464744	0.446437	-0.177421	1.000000	0.364416	0.457307
x5	0.979163	0.991090	0.686552	0.364416	1.000000	0.960391
y	0.970899	0.983552	0.502498	0.457307	0.960391	1.000000

The $\mathrm{X}'\mathrm{X}$ matrix reveals the high correlation between gross national product (GNP) implicit price deflator (x1) and GNP, millions of dollars (x2), since $r_{12} = 0.991589$. Then the correlation between x2 and x5 is high since $r_{25} = 0.991090$, suspected earlier from inspection of Fig (d) and

Fig (e). Thus, inspection of the correlation matrix indicates that there are several near-dependencies in the Longley data.



From the above correlation plot, it is clear that x1 and x2, x2 and x5 are correlated. Generally, inspection of the r_{ij} is not sufficient for detecting anything more complex than pairwise multicollinearity. So, we find Variation Inflation Factor for the regressors.

Variation Inflation Factor

Independent Variable	Variance Inflation	R-Squared Vs Other X's	Tolerance
x1	130.8292	0.9924	0.0076
x2	639.0498	0.9984	0.0016
x3	10.7869	0.9073	0.0927
x4	2.5058	0.6009	0.3991
x5	339.0117	0.9971	0.0029

Since some VIF's are greater than 10, multicollinearity is a problem.

Maximum of VIF = 639.0498

Hence, we conclude that the multicollinearity problem exists since VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

Eigensystem Analysis of $\mathbf{X'X}$

Extent of Multicollinearity

No.	Eigenvalue	Incremental Percent	Cumulative Percent	Condition Index
1	3.609669	72.19	72.19	1.00
2	1.175340	23.51	95.70	3.07
3	0.199155	3.98	99.68	18.12
4	0.014882	0.30	99.98	242.55
5	0.000953	0.02	100.00	3785.97

The above table gives the eigenvalue analysis of the independent variables after they have been centered and scaled.

Eigen Value:

Eigen value is used to determine the extent of multicollinearity in the data. One or more small eigenvalues imply that there are near - linear dependences among the columns of X.

There are two very small eigenvalues, therefore, this is a symptom of seriously ill-conditioned data.

Condition number (k) = 3785.97

Since, k exceeds 1000, **severe** multicollinearity is indicated.

Condition numbers greater than 1000 indicate a severe multicollinearity problem while condition numbers between 100 and 1000 indicate a mild multicollinearity problem.

Condition Index

Useful measure of the number of near-linear dependences in $\mathrm{X'X}$.

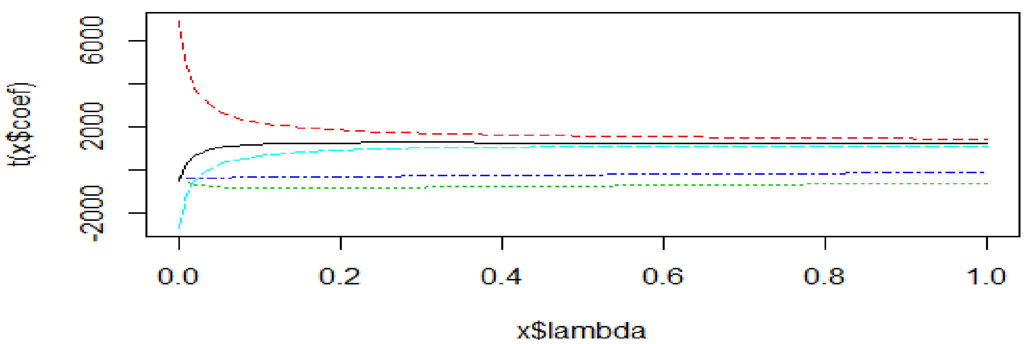
Since one of the condition indices **exceeds 1000** (and two others exceed 100), we conclude that there is **at least one strong near - linear dependence** in the Longley data.

Ridge Estimator

Ridge Trace

```
> betacR
          k-0      k-0.01      k-0.02      k0.025      k-0.03     k-0.035      k-0.04     k-0.045
r1 -0.1489163  0.36389531  0.36959472  0.36983923  0.36830008  0.36609965  0.36365866  0.36115445
r2  2.0378367  0.56853289  0.52174558  0.49564165  0.47828027  0.46545934  0.45532200  0.44692025
r3 -0.1074620 -0.25164421 -0.24757951 -0.24158270 -0.23497776 -0.22821887 -0.22149624 -0.21489815
r4 -0.1110659 -0.09932649 -0.09224938 -0.08564087 -0.07943719 -0.07358643 -0.06804896 -0.06279376
r5 -0.7992234  0.24710216  0.28078160  0.29813717  0.30831055  0.31466714  0.31875453  0.32138928
```

From the above table, the reasonable coefficient stability is achieved in the region 0.01<k<0.025



From the above figure, we observe that at the point of k = 0.02, the reasonable coefficient is achieved. This is a plot that we have added that shows the impact of k on the variance Inflation factors. Since the major goal of ridge regression is to remove the impact of multicollinearity, it is important to know at what point multicollinearity has been dealt with. This plot shows this.

Since the rule-of-thumb is that multicollinearity is not a problem once all VIFs are less than 10, we inspect the graph for this point. In this example, it appears that all VIFs are small enough once k is greater than 0.02.

Hence, this is the value of k that this plot would indicate we use. Since this plot indicates k = 0.02 and the ridge trace indicates a value near 0.02, we would select 0.02 as our final result. The rest of the reports are generated for this value of k.

Variance Inflation Factor

k	x1	x2	x3	x4	x5
0.000000	130.8292	639.0498	10.7869	2.5058	339.0117
0.010000	15.4326	5.8081	2.6893	2.0791	11.6397
0.020000	7.7949	2.0398	2.4356	1.8739	5.5016
0.020000	7.7949	2.0398	2.4356	1.8739	5.5016
0.030000	4.7572	1.1994	2.2431	1.7277	3.3296
0.040000	3.2319	0.8632	2.0783	1.6097	2.2749
0.050000	2.3559	0.6869	1.9339	1.5095	1.6767
0.060000	1.8054	0.5785	1.8060	1.4222	1.3022
0.070000	1.4362	0.5046	1.6918	1.3449	1.0509
0.080000	1.1761	0.4504	1.5895	1.2759	0.8734
0.090000	0.9857	0.4085	1.4973	1.2137	0.7427
0.375120	0.1446	0.1296	0.5082	0.5076	0.1347

This report gives the values that are plotted on the variance inflation factor plot. It is to determine when all three VIFs are less than 10.

K Analysis

k	R2	Sigma	B'B	Ave VIF	Max VIF
0.000000	0.9977	0.9014	6.7689	1178.8630	6563.3445
0.010000	0.9896	1.9167	0.6465	7.4123	12.3815
0.020000	0.9835	2.4157	0.5851	3.9406	6.3887
0.030000	0.9778	2.8021	0.5566	2.6570	4.4439
0.040000	0.9723	3.1285	0.5383	2.0018	3.3420
0.050000	0.9670	3.4157	0.5245	1.6089	2.6323
0.060000	0.9619	3.6746	0.5131	1.3491	2.1402
0.070000	0.9568	3.9116	0.5031	1.1655	1.7819
0.080000	0.9518	4.1311	0.4942	1.0293	1.5115
0.090000	0.9469	4.3359	0.4859	0.9245	1.3017
0.109594	0.9375	4.7030	0.4712	0.7763	1.1075

This report provides a quick summary of the various statistics that might go into the choice of k.

k

This is the actual value of k. Note that the value found by the analytic search (0.02) sticks out as you glance down this column because it does not end in zeros.

R2

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

Sigma

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

B'B

This is the sum of the squared standardized regression coefficients. Ridge regression assumes that this value is too large and so the method tries to reduce this. We want to find a value for k at which this value has stabilized.

Ave VIF

This is the average of the variance inflation factors.

Max VIF

This is the maximum variance inflation factor. Since we are looking for that value of k which results in all VIFs being less than 5 or 10, this value is very helpful in your selection of k.

Ridge versus Least Squares Comparison

Independent Variable	Regular Ridge Coeff's	Regular L.S. Coeff's	Stand'zed Ridge Coeff's	Stand'zed L.S. Coeff's	Ridge Standard Error	L.S. Standard Error
Intercept	65317	65317				
x1	1257.619	-506.3821	0.3698	-0.1489	499.3552	1381.84
x2	1685.404	6929.56	0.4956	2.0378	255.4464	3054.027
x3	-821.4896	-365.4189	-0.2416	-0.1075	279.1294	396.783
x4	-291.2174	-377.6738	-0.0856	-0.1111	244.8347	191.2391
x5	1013.8	-2717.718	0.2981	-0.7992	419.515	2224.4
R-Squared	0.9723	0.9874				
Sigma	715.4263	483.2430				

This report provides a detailed comparison between the ridge regression solution and the ordinary least squares solution to the estimation of the regression coefficients.

From the above report, it is clear that the standard error of the Ridge is much smaller than that of the Least Square standard error.

Therefore, the optimum standardized ridge coefficients are obtained at k = 0.02

Ridge Regression Coefficient

Independent Variable	Regression Coefficient	Standard Error	Stand'zed Regression Coefficient	VIF
Intercept	65317			
x1	1257.619	499.3552	0.3698	7.7949
x2	1685.404	255.4464	0.4956	2.0398
x3	-821.4896	279.1294	-0.2416	2.4356
x4	-291.2174	244.8347	-0.0856	1.8739
x5	1013.8	419.515	0.2981	5.5016

This report provides the details of the ridge regression solution. Since the VIF is lesser than 10, the multicollinearity is vanished.

Therefore, the Ridge regression model with finite standard error is,

$$\hat{y} = 0.3698X_1 + 0.4956X_2 - 0.2416X_3 - 0.0856X_4 + 0.2981X_5$$

Ridge Parameter

$$\widehat{\boldsymbol{\beta}}_R = (\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

```
> ridgeParameter
               y
[1,]  0.36983923
[2,]  0.49564165
[3,] -0.24158270
[4,] -0.08564087
[5,]  0.29813717
```

Health Data

Daniel, et al., (2012) considered a study of obesity and metabolic syndrome using data collected from 15 students on systolic blood pressure (SBP), weight, and BMI. The data set is presented in Table 3.3.

Table 3.3 Data on SBP, Weight and BMI

SBP	Weight (lbs)	BMI
126	125	24.41
129	130	23.77
126	132	20.07
123	200	27.12
124	321	39.07
125	100	20.9
127	138	22.96
125	138	24.44
123	149	23.33
119	180	25.82
127	184	26.4
126	251	31.37
122	197	26.72
126	107	20.22
125	125	23.62

Model for the Data

Each of the original regressors has been scaled using the unit normal scaling [subtracting the average (centering) and dividing by the standard deviation]. The squared and cross - product terms are generated from the scaled linear terms. The centering the linear terms is helpful in removing nonessential ill - conditioning when fitting polynomials. The least-squares fit is,

$$\hat{y} = 124.8667 - 2.7454X_1 + 2.1276X_2$$

Summary Statistics for the Model

Variable	**Count**	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**
Weight	15	6.666662E-11	1	-1.096668	2.624372
BMI	15	6.666668E-11	1	-1.100875	2.862109
SBP	15	124.8667	2.416215	119	129

For each variable, the descriptive statistics for the given data are computed. This report is particularly useful for checking that the correct variables were selected.

Multicollinearity Detection

Plotting each two of the independent variables to check the present of linearity:

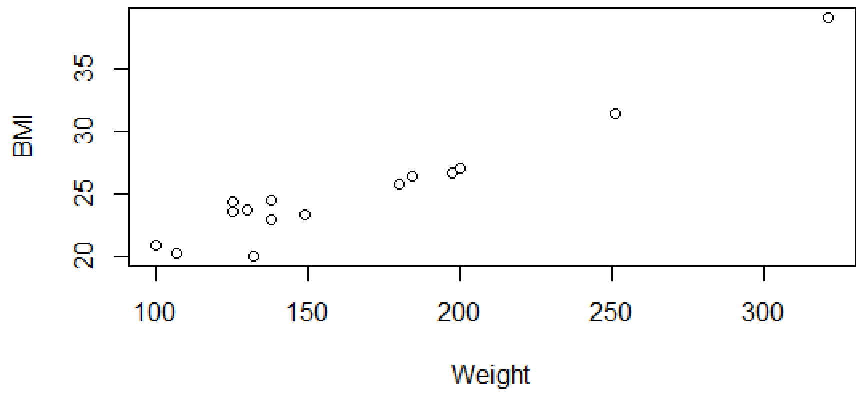


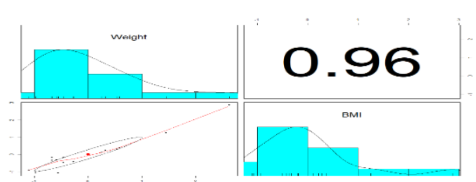
Fig (f)

Here, from the Fig(f), as the Weight increases, the BMI rate also increases. So, there is the linear relation in between these two variables. Then, the correlation between these variables may be high.

Correlation Matrix $X'X$

	Weight	BMI	SBP
Weight	1.000000	0.962103	-0.289058
BMI	0.962103	1.000000	-0.212629
SBP	-0.289058	-0.212629	1.000000

The $X'X$ matrix reveals the high correlation between Weight (X1) and BMI (X2) suspected earlier from inspection of Fig (f), since $r_{12} = 0.962103$. Thus, the inspection of the correlation matrix indicates that there is a near-dependencies in the Health data.



From the above correlation plot, it is clear that Weight and BMI are correlated.

Variation Inflation Factor

Independent Variable	Variance Inflation	R-Squared Vs Other X's	Tolerance
Weight	13.4486	0.9256	0.0744
BMI	13.4486	0.9256	0.0744

Since some VIF's are greater than 10, multicollinearity is a problem.

Maximum of VIF = 13.4486

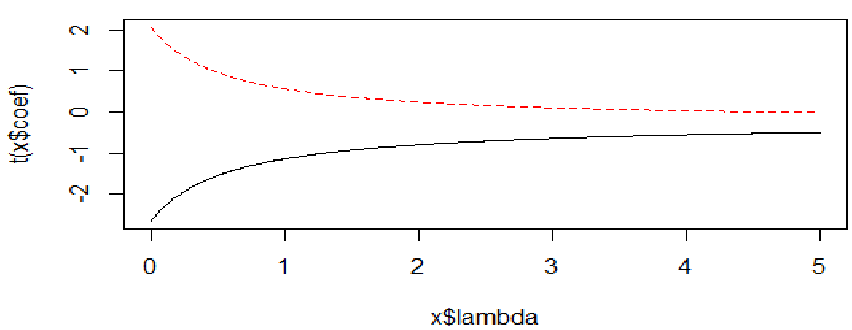
Hence, we conclude that the multicollinearity problem exists since VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

Ridge Estimator

Ridge Trace

```
> betacR
          k-0       k-0.01      k-0.02      k-0.04      k-0.08       k-0.16
r1 -1.1362288 -0.14099234 -0.7865998 -0.6158688 -0.4469711 -0.31130836
r2  0.8805403 -0.04398855  0.5334913  0.3652887  0.2012993  0.07489795
         k-0.32       k-0.64       k-0.75      k-0.88       k-0.95          k-1
r1 -0.216692916 -0.15277247 -0.14099234 -0.1298925 -0.12482098 -0.12150346
r2 -0.003142577 -0.04002821 -0.04398855 -0.0466272 -0.04745564 -0.04786514
```

From the above table, the reasonable coefficient stability is achieved in the region $0.95 < k \leq 1$



From the above figure, we observe that k = 1, where the reasonable coefficient is achieved.

This is a plot that we have added that shows the impact of k on the variance Inflation factors. Since the major goal of ridge regression is to remove the impact of multicollinearity, it is important to know at what point multicollinearity has been dealt with.

Since the rule-of-thumb is that multicollinearity is not a problem once all VIFs are less than 10, we inspect the graph for this point. In this example, it appears that all VIFs are small enough once k is greater than 1.

Hence, this is the value of k that this plot would indicate we use. Since this plot indicates k = 1 and the ridge trace indicates a value near 1, we would select 1 as our final result. The rest of the reports are generated for this value of k.

Ridge versus Least Squares Comparison

Independent Variable	Regular Ridge Coeff's	Regular L.S. Coeff's	Stand'zed Ridge Coeff's	Stand'zed L.S. Coeff's	Ridge Standard Error	L.S. Standard Error
Intercept	124.8667	124.8667				
Weight	-0.2935785	-2.745373	-0.1215	-1.1362	0.2451604	2.37043
BMI	-0.1156525	2.127575	-0.0479	0.8805	0.2451604	2.37043
R-Squared	0.0453	0.1412				
Sigma	2.5500	2.4185				

This report provides a detailed comparison between the ridge regression solution and the ordinary least squares solution to the estimation of the regression coefficients.

From the above report, it is clear that the standard error of the Ridge is much smaller than that of the Least Square standard error.

Therefore, the optimum standardized ridge coefficients are obtained at k = 1

Ridge Regression Coefficient

Independent Variable	Regression Coefficient	Standard Error	Stand'zed Regression Coefficient	VIF
Intercept	124.8667			
Weight	-0.2935785	0.2451604	-0.1215	0.1294
BMI	-0.1156525	0.2451604	-0.0479	0.1294

This report provides the details of the ridge regression solution. Since the VIF is lesser than 5, the multicollinearity is vanished.

Therefore, the Ridge regression model with finite standard error is,

$$\hat{y} = -0.1215\,X_1 - 0.0479\,X_2$$

Ridge Parameter

$$\widehat{\boldsymbol{\beta}}_R = (\mathbf{X}'\mathbf{X} + \mathrm{k}\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

```
> ridgeParameter
              SBP
[1,] -0.12150346
[2,] -0.04786514
```

Chapter 4

Conclusion

Multicollinearity, if left untouched, can have a detrimental impact on the generalizability and accuracy of your model. If multicollinearity exists, the traditional ordinary least squares estimators are imprecisely estimated, which leads to the inaccuracy in the judgment as to how each predictor variable impacts the target outcome variable. Given the information it is essential to detect and solve the issue of multicollinearity before estimating the parameters based on a fitted regression model.

Detecting multicollinearity is a fairly simple procedure involving the employment of VIF tool. The correlation procedure is also useful in multicollinearity detection. After discovering the existence of multicollinearity, the ridge regression technique is implemented to control the effect of multicollinearity.

A number of Monte Carlo simulation studies have been conducted to examine the effectiveness of biased estimators and to attempt to determine which procedures perform best. The Dempster et al. [1977] study compared 57 different estimators for 160 different model configurations. While no single procedure emerges from these studies as best overall, there is considerable evidence indicating the superiority of biased estimation to least squares if multicollinearity is present.

Our own preference in practice is for ordinary ridge regression with k selected by inspection of the ridge trace. The procedure is straightforward and easy to implement on a standard least - squares computer program, and one can learn to interpret the ridge trace very quickly. It is also occasionally useful to find the "optimum" value of k and the iteratively estimated "optimum" k and compare the resulting models with the one obtained via the ridge trace.

Several authors have noted that while one can prove that there exists a k , such that the mean square error of the ridge estimator is always less than the mean square error of the least - squares estimator, there is no assurance that the ridge trace (or any other method that selects the biasing parameter stochastically by analysis of the data) produces the optimal k.

The regressors and the response should be centered and scaled so that $\mathbf{X'X}$ and $\mathbf{X'y}$ are in correlation form. This results in an artificial removal of the intercept from the model. Effectively the intercept in the ridge model is estimated by $\bar{y}$. Centering tends to minimize any nonessential ill – conditioning in the data. Centering and scaling allow us to think of the parameter estimates as standardized regression coefficients, which is often intuitively appealing. Furthermore, centering the regressors can remove nonessential ill - conditioning, thereby reducing variance inflation in the parameter estimates. Consequently, both centering and scaling is recommended in the data.

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

In effect, variable elimination often shrinks the vector of parameter estimates, as ridge regression does. Properly used biased estimation methods are a valuable tool in the data analyst's kit. Through the steps outlined in this study, one should not only able to detect any issue of multicollinearity, but also resolve it in only a few short steps.

Bibliography

1. Daniel, W. W., and Cross, C. L. (2012), Biostatistics: A Foundation for Analysis in the Health Sciences, Tenth Edition, John Wiley & Sons, New York, US, **2012.**

2. Guajarati, D. N., Porter, D., and Gunasekar, S. Basic Econometrics, Fifth Edition, McGraw Hill Education, New York, US, **2017**.

3. Longley, J. An Appraisal of Least-Squares Programs from the Point of the User, *Journal of the American Statistical Association*, *V*ol. 62, pp. 819–841, **1967**.

4. Montgomery, D. C., Peck, E. A., and Vining, G. G. Introduction to Linear Regression Analysis, Fifth Edition, John Wiley & Sons, New York, US, **2012**