CONTENTS

LIST OF TABLES

ABSTRACT

Multiple linear regression also simply known as multiple regression is a statistical technique that uses several explanatory or independent variables to predict the outcome of a response or dependent variable.

The use and interpretation of a multiple regression model often depends explicitly or implicitly on the estimates of the individual regression coefficients.

Some examples of inferences that are frequently made include the following:
1. Identifying the relative effects of the regressor variables
2. Prediction and/or estimation
3. Selection of an appropriate set of variables for the model

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.
In this project we will discuss methods of detecting the presence of multicollinearity, and some techniques for dealing with the problem.

We shall use the data on world's gender inequality index (a composite metric of gender inequality using three dimensions: reproductive health, empowerment and the labour market , whose low value indicates a low inequality between women and men) for our project .

1.INTRODUCTION

A linear model is usually fit to a dataset where the response (y) and the regressor (x_i) variables have a clear linear relationship .

Let there be p independent non-stochastic variables (x_1 , x_2 ,……..,x_p) and a dependent or a response variable (y).Suppose the dataset consists of n observations of (p+1) variables . Let $x_{1i},x_{2i},…..,x_{pi}$ are fixed p values of p independent variables and $y_1,y_2,….,y_n$ are corresponding values of dependent variables.

The assumed model is:

$y_i = \beta_0 x_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots + \beta_p x_{pi} + \varepsilon_i$

where the β_i's are the regression coefficients and ε_i's are the random error.

The model is valid only under following assumptions:-

1) x_1 , x_2 ,……..,x_p are non-stochastic variables
2) $y_1,y_2,….,y_n$ as well as $\varepsilon_1,\varepsilon_2,….,\varepsilon_i$ are the independently and identically distributed random variables with mean 0 and constant variance σ^2 .

Regression model often depends on the estimates of the individual regression coefficients . The estimation process also depends on the existence of near linear dependency of the regression variables or existence of multicollinearity.

1.1 SOURCES OF MULTICOLLINEARITY

We write the multiple regression model as :-

$\mathbf{y} = \mathbf{X}\beta + \mathcal{E}$

where , $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ matrix of the regressor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and ε is an $n \times 1$ vector of random errors, with $\varepsilon \sim NID(0, \sigma^2)$.

It will be convenient to assume that the regressor variables and the response have been centered and scaled to unit length.

Consequently, $\mathbf{X'X}$ is a $p \times p$ matrix of correlations between the regressors and $\mathbf{X'y}$ is a $p \times 1$ vector of correlations between the regressors and the response.

We may formally define multicollinearity in terms of the linear dependence of the columns of $\mathbf{X}$. The vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$ are linearly dependent if there is a set of constants $t_1, t_2, \ldots, t_p$, not all zero, such that

$$\sum_{j=1}^{p} t_j \mathbf{X}_j = 0$$

If the above equation holds exactly for a subset of the columns of $\mathbf{X}$, then the rank of the $\mathbf{X'X}$ matrix is less than p and $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist. However, suppose that the above equation is approximately true for some subset of the columns of $\mathbf{X}$ then there will be a near -linear dependency in $\mathbf{X}'\mathbf{X}$ and the problem of **multicollinearity** is said to exist.

As we shall see, the presence of multicollinearity can make the usual least -squares analysis of the regression model dramatically inadequate.

There are four primary **sources of multicollinearity** :

1. The data collection method employed
2. Constraints on the model or in the population
3. Model specification
4. An over-defined model

1.2 EFFECTS OF MULTICOLLINEARITY

The presence of multicollinearity has a number of 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, x_1 and x_2 .

The model, assuming that x_1, x_2 , and y are scaled to unit length, is

$$y=\beta_1 x_1+\beta_2 x_2+\mathcal{E}$$

and the least - square normal equations are :--

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

$$\begin{pmatrix} 1 & r12 \\ r21 & 1 \end{pmatrix} \begin{matrix} \widehat{\beta 1} \\ \widehat{\beta 2} \end{matrix} = \begin{matrix} r1y \\ r2y \end{matrix}$$

where r_{12} is the simple correlation between x_1 and x_2 and r_{jy} is the simple correlation between x_j and y, j=1,2.

If there is strong multicollinearity between x_1 and x_2 , then the correlation coefficient r_{12} will be large. 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 $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ matrix are—

$$C_{jj} = 1/(1 - R_j^2) \text{ , } j=1,2,\ldots,p$$

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 $\text{Var}(\hat{\beta}_j) = C_{jj}\,\sigma^2 = (1 - R_j^2)^{-1}\,\sigma^2$, strong multicollinearity implies that the variance of the least - squares estimate of the regression coefficient β_j is very large.

Multicollinearity also tends to produce least - squares estimates $\hat{\beta}_j$ that are **too large** in absolute value .

METHODOLOGY

2.1 The Data

In this project the Gender Inequality Index (GII) Dataset has been used. GII is a composite metric of gender inequality using three dimensions: reproductive health, empowerment and the labor market. A low GII value indicates low inequality between women and men, and vice-versa.
It is a collection of records of GII and related factors of 170 countries. It contains attributes like country name, human development category, gender inequality index ,maternal mortality ratio (deaths per 10,0000 live births) , adolescent birth rate (births per 1,000 women aged 15-19) ,share of seats in parliament (in percentage) , females with at least some secondary education (in percentage ,aged 25 and older) ,males with at least some secondary education (in percentage ,aged 25 and older) , female labor force (in percentage aged 15 and older).

SOURCE : This dataset has been taken from kaggle.

VARIABLES IN THE DATASET: The following variables have been used from the dataset for this project :--

y = Gender Inequality Index
x_1 = maternal mortality ratio (deaths per 10,0000 live births)
x_2 = adolescent birth rate (births per 1,000 women aged 15-19)
x_3= share of seats in parliament (in percentage)
x_4= females with atleast some secondary education (in percentage ,aged 25 and older)
x_5= males with atleast some secondary education (in percentage ,aged 25 and older)
x_6= female labour force (in percentage aged 15 and older)

Here y is the response variable or the dependent variable and x ,x_2,x_3,x_4,x_5,x_6 are the regressors or the independent variables. The total number of observations here is denoted by n and the number of independent variables is denoted by p.
The dataset is given in Table 1 in Appendix.

2.2 METHODOLOGY FOR DETECTING MULTICOLLINEARITY

2.2.1 Examination of the Correlation Matrix

A very simple measure of multicollinearity is inspection of the off-diagonal elements r_{ij} in $\mathbf{X'X}$. If regressors x_i and x_j are nearly linearly dependent, then $|r_{ij}|$ will be near unity . 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. Therefore, inspection of the r_{ij} is not sufficient for detecting anything more complex than pair-wise multicollinearity.

2.2.2 Variance Inflation Factors

The diagonal elements of the matrix $\mathbf{C} = (X'X)^{-1}$ are very useful in detecting multicollinearity . the *j*th diagonal element of **C** , can be written as $C_{jj} = (1-R_j^2)^{-1}$ is the coefficientof 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 C_{jj} 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.

The variance inflation factor is given by, $\mathbf{VIF_j = C_{jj} = (1-R_j^2)^{-1}}$. 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.

2.2.3 Eigensystem Analysis of X´X

The characteristic roots or **eigen values** of $\mathbf{X'X}$, say $\lambda_1,\lambda_2,\ldots..,\lambda_P$ 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 eigen values imply that there are near - linear dependencies among the columns of $\mathbf{X}$. Some analysts prefer to examine the **condition number** of $\mathbf{X'X}$, defined as,

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

This is just a measure of the spread in the eigen value spectrum of $\mathbf{X'X}$. 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 $\mathbf{X'X}$ matrix are----

$$\kappa_j = \frac{\lambda \max}{\lambda j} \quad , j=1,2,\ldots,p$$

Eigensystem analysis can also be used to identify the nature of the near – linear dependences in data. The $\mathbf{X'X}$ matrix may be decomposed as,

$$\mathbf{X'X=T\Lambda T'}$$

where $\mathbf{\Lambda}$ is a $p \times p$ diagonal matrix whose main diagonal elements are the **eigenvalues** λ_j of $\mathbf{X'X}$ and $\mathbf{T}$ is a $p \times p$ orthogonal matrix whose columns are the eigenvectors of $\mathbf{X'X}$.

Let the columns of $\mathbf{T}$ be denoted by $\mathbf{t_1.t_2,\ldots.,t_p}$. If the Eigenvalue λ_j is close to zero, indicating a near-linear dependence in the data, the elements of the associated eigenvector $\mathbf{t_j}$ describes the nature of this linear dependence.

There are various methods for dealing with **multicollinearity** collecting additional data, model re-specification , ridge regression. However, we shall stick to the method of ridge regression for dealing with multicollinearity in this project.

2.3 METHOD FOR DEALING WITH MULTICOLLINEARITY

2.3.1 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-square estimates of the regression coefficients may
be 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.

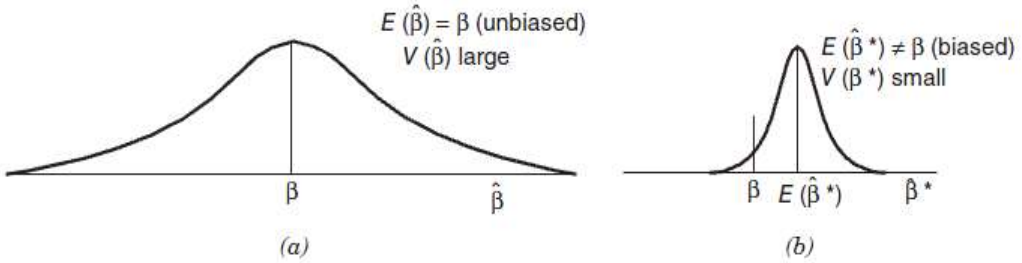

Figure depicting Sampling distribution of (a) unbiased and (b) biased estimator of β

The problem with the method of least squares is the requirement that $\hat{\boldsymbol{\beta}}$ should be an unbiased estimator of **β** .The Gauss - Markov property 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. One way to alleviate this problem is to drop the requirement that the estimator of $\boldsymbol{\beta}$ be unbiased. Suppose that we can find a biased estimator of β, say $\hat{\beta}^*$ that has a smaller variance than the unbiased estimator $\widehat{\boldsymbol{\beta}}$.

A number of procedures have been developed for obtaining biased estimators of regression coefficients. One of these procedures is ridge regression, originally proposed by Hoed and Kennard. The ridge estimator is found by solving a slightly modified version of the normal equations. Specifically we define the ridge
estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{R}}$ as the solution to,

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

which is equal to,

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

where k ≥ 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 earlier by Hoed [1959] 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,

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

Therefore $\beta \mathcal{R}$ is a biased estimator of $\boldsymbol{\beta}$. We usually refer to the constant k as the biasing parameter. 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 $\hat{\beta}_{\mathrm{R}}$ will be less than the variance of the least-squares estimator $\hat{\beta}$. Hoed and Kennard have 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 $\widehat{\boldsymbol{\beta}}_{\mathbf{R}}$ versus k for values of k usually in the interval 0 – 1.

Plot 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 $\widehat{\boldsymbol{\beta}}_{\mathbf{R}}$ will vary dramatically. At some value of k , the ridge estimates will stabilize. The objective is to select a reasonably small value of k at which the ridge estimates $\widehat{\boldsymbol{\beta}}_{\mathbf{R}}$ are stable. Hopefully this will produce a set of estimates with smaller MSE than the least - squares estimates

3.RESULT AND ANALYSIS

3.1 Calculation of correlation matrix

In order to get an idea about the near-linear dependencies among the regressor variables, we calculate the correlation matrix based on the regressor variables. If the absolute value of off diagonal elements of the correlation matrix is more than or equal to 0.5, it indicates the presence of multicollinearity.

The correlation matrix based on the regressor variables is given below in Table 3.1.1.

Table 3.1.1 :- CORRELATION MATRIX BASED ON THE REGRESSOR VARIABLES

	MATERNAL MORTALITY RATIO	ADOLESCENT BIRTH RATE	SHARE OF SEATS IN PARLIAMENT	FEMALES WITH ATLEAST SOME SECONDARY EDUCATION	MALES WITH ATLEAST SOME SECONDARY EDUCATION	FEMALE LABOUR FORCE
MATERNAL MORTALITY RATIO	1.00000000	0.75276893	-0.16222076	-0.69801427	-0.64257249	0.23051072
ADOLESCENT BIRTH RATE	0.7527689	1.00000000	-0.09473649	-0.72872473	-0.69175990	0.26047064
SHARE OF SEATS IN PARLIAMENT	-0.1622208	-0.09473649	1.00000000	0.16948270	0.16882344	0.27901459
FEMALES WITH ATLEAST SOME SECONDARY EDUCATION	-0.6980143	-0.72872473	0.16948270	1.00000000	0.97292084	-0.09871135
MALES WITH ATLEAST SOME SECONDARY EDUCATION	-0.6425725	-0.69175990	0.16882344	0.97292084	1.00000000	-0.08317605
FEMALE LABOUR FORCE	0.2305107	0.26047064	0.27901459	-0.09871135	-0.08317605	1.00000000

From the correlation matrix given in TABLE 3.1.1 , we observe that there is some near-linear dependencies among the regressor variables.

For instance, we observe that there is high correlation between the regressor variable **MALES WITH ATLEAST SOME SECONDARY EDUCATION (x_4)** and **FEMALES WITH ATLEAST SOME SECONDARY EDUCATION (x_5)** since $r_{45} = 0.97292084$.
Furthermore, there are other large correlation coefficients between x_4 and x_2 , x_1 and x_2 and so on . Thus, inspection of the correlation matrix indicates that there are several near–linear dependencies in the gender inequality index 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 pair-wise correlations r_{ij} will be large.
Therefore we conclude that the inspection of the correlation matrix is not sufficient for detecting anything more complex than pair-wise multicollinearity. Hence we calculate variance inflation factors next.

3.2 Calculation of variance inflation factors

The diagonal elements of the matrix $\mathbf{C} = (X'X)^{-1}$ are the variance inflation factors for the given dataset.

The inverse of the correlation matrix is given as follows---

TABLE 3.2.1- INVERSE OF THE CORRELATION MATRIX

2.7761312	-1.30279166	0.24935496	2.31563753	-1.43408009	-0.26086409
-1.3027917	2.94727773	-0.04410002	1.18889249	0.02449739	-0.33567256
0.2493550	-0.04410002	1.15724045	0.05552888	-0.15094277	-0.37595269
2.3156375	1.18889249	0.05552888	23.02256900	-20.1197307	-0.25983500
-1.4340801	0.02449739	-0.15094277	-20.1197307	19.69738191	0.01860977
-0.2608641	-0.33567256	-0.37595269	-0.25983500	0.01860977	1.22836033

From table 3.2.1, we observe that the principal diagonal elements of the inverse of the correlation matrix or the variance inflation factors are given as follows :--

$VIF_1 = C_{11} = 2.7761312$, $VIF_2 = C_{22} = 2.94727773$

$VIF_3 = C_{33} = 1.15724045$, $VIF_4 = C_{44} = 23.02256900$

$VIF_5 = C_{55} = 19.69738191$, $VIF_6 = C_{66} = 1.22836033$

As discussed in section 3.2. 2, if any of the VIFs exceeds 5 or 10,
it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity. Since VIF_4 and VIF_5 are greater than 10, we conclude that the problem of multicollinearity exists in the data .

3.3 Calculation of Eigen Values , Condition number , Condition indices

The characteristic roots or the Eigen values of the correlation matrix are given as follows----

$\lambda_1 = 3.32465878$, $\lambda_2 = 1.30425512$, $\lambda_3 = 0.68812962$
$\lambda_4 = 0.41950403$, $\lambda_5 = 0.23949984$, $\lambda_6 = 0.02395261$

Condition number is given as follows :----

$$\kappa = \frac{3.32465878}{0.02395261} = 138.8015$$

Condition indices are :---

$\kappa_1 = \frac{3.32465878}{3.32465878} = 1.0000$, $\kappa_2 = \frac{3.32465878}{1.30425512} = 2.549086$, $\kappa_3 = \frac{3.32465878}{0.68812962} = 4.831443$

$\kappa_4 = \frac{3.32465878}{0.41950403} = 7.925213$, $\kappa_5 = \frac{3.32465878}{0.23949984} = 13.881674$, $\kappa_6 = \frac{3.32465878}{0.02395261} = 138.801534$

We draw the following conclusion from the above given eigen values, condition numbers and condition indices----

1) The eigenvalues λ_5=0.23949984 and λ_6= 0.02395261 imply that there are near - linear dependencies among the regressor variables .

2) The condition number κ=138.8015 which is greater than 100 , also indicates the presence of multicollinearity in the data.

3) Since one of the condition index exceed 100, we conclude that there is atleast one near linear dependence in the gender inequality index data.

The smallest eigen value is λ_6= 0.02395261 , so the elements of eigenvector corresponding to λ_6 are the coefficients of the regressors in the equation $\sum_{j=1}^{p} t_j X_j = 0$.

The eigen vector corresponding to λ_6 is given by ,

t_6 =(0.0678774 0.0198492 0.0039617 0.7346525 -0.6747122 -0.0056589)

This implies that,

$$\mathbf{0.0678774x_1+0.0198492x_2+0.0039617x_3+0.7346525x_4-0.6747122x_5-0.0056589x_6=0}$$

Assuming that 0.0039617 and -0.0056589 are approximately zero and re-arranging terms gives,

$$x_1=-0.292427x_2-10.823227x_4-9.940159x_5$$

that is the regressors x_1,x_2,x_4,x_5 approximately add up to a constant. Thus, the elements of $\mathbf{t_6}$ directly reflect the relationship used to generate x_1,x_2,x_4,x_5 .

3.4 Calculation of best value of k using ridge regression and to find the coefficient estimates for the model

We perform ridge regression on the given data of gender inequality index and compare the estimates for various values of k, some of which are given below in table 3.4.1

Table3.4.1:Estimate of regression coefficients for different values of k

	k=0.02	**k=0.67**	**k=0.34**	**k=0.5**
β_1	0.0001809621	0.0001809721	0.0001809670	0.0001809695
β_2	0.0044688013	0.0044687900	0.0044687958	0.0044687930
β_3	-0.0024143106	-0.0024142160	-0.0024142640	-0.0024142407
β_4	-0.0019256396	-0.0019250885	-0.0019253682	-0.0019252326
β_5	0.0038768776	0.0038763093	0.0038765978	0.0038764579
β_6	0.0003753530	0.0003753532	0.0003753531	0.0003753532

we choose k= 0.67 as the best value of k. The Table 3.4.2 shows the ridge coefficients for best value of k.

Table 3.4.2: The ridge coefficients for k=0.67

β_1	0.0001809721
β_2	0.0044687900
β_3	-0.0024142160
β_4	-0.0019250885
β_5	0.0038763093

total sum of squares (sst)= 6.565734

sum of squares due to error (sse)= 2.731308

R-squared =1 - $\frac{sse}{sst}$ =1 – $\frac{2.731308}{6.565734}$ = 0.5840057

The R-squared turns out to be 0.5840057**. That is, ridge regression was able to explain** 58.4% **of the variation in the response values of the data.**

4. CONCLUSION

From the above results, it is clear that there is multicollinearity present in the data. The correlation matrix , the variance inflation factors, the condition number and the condition indices, all indicate near linear-dependence among the regressors. Through Eigensystem analysis (in section 3.3) we have also established the linear relation between the regressors. Section 3.4 gives the ridge coefficients for the best value of k. After performing the ridge regression on the data R-squared turns out to be 0.5840057. That is, the ridge regression was able to explain **58.40 %** of the variation in the response values of the data.

Therefore,
1) There exists multicollinearity among the regressor variables of Gender Inequality Dataset.
2) Ridge Regression is the technique used to overcome multicollinearity in the data.

5.REFERENCE

1)Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining (fifth edition):Introduction To Linear Regression Analysis

2)Alvin C. Rencher and G. Bruce Schaalje (second edition) : Linear Models In Statistics , John Wiley

3)Gareth James, Daniela Witten , Trevor Hastie, Robert Tibshirani : An Introduction To Statistical Learning

4)John Fox(Second Edition):Regression Diagnostics—An Introduction

6.ACKNOWLEDGEMENT

I would like to express my sincere gratitude and appreciation to all those who have contributed to the completion of this project. Without their support, guidance, and assistance, this project would not have been possible.

First and foremost, I would like to thank my project supervisor Professor Tanusree Banerjee, for her invaluable guidance, expertise, and constant support throughout the project. Her deep knowledge and insightful feedback have been instrumental in shaping the direction and quality of my work.

I would also like to extend my gratitude to the faculty of Department of Statistics, Bethune College, who provided me with a conducive learning environment.

I would like to thank my family and friends for their unwavering support, encouragement, and understanding throughout the project. Their belief in my ability and constant motivation have been the driving force behind my perseverance and dedication.

In conclusion, I extend my deepest gratitude to everyone who has contributed to the successful completion of this project. Their support, guidance, and encouragement have been invaluable, and I am truly grateful for the opportunity to work on this project and for the knowledge and experience gained along the way.

Thank You

7.APPENDIX

The data used is as follows:

1. Title –Gender-Inequality Index Data

2. Sources
a) Origin—This dataset was taken from StatLib Library which is maintained at Carnegie Mellon University.

3.Relevant Information—
The Gender Inequality Index (GII) dataset provides a comprehensive measure of gender inequality across countries, capturing gender disparities in health, education, and economic opportunities. Developed by the United Nations Development Program (UNDP), the GII measures gender inequality by analyzing health, empowerment, and labor market participation indicators. This dataset includes GII scores, as well as component scores for each indicator, for over **190 countries, in 2021**.

4. Number of instances---170

5. Number of attributes ---- 6

6. Attribute information:
a) Name of the country

b) Gender Inequality Index : continuous

c) Maternal Mortality Ratio (Deaths per 100000 live births) : continuous

d) Adolescent Birth Rate(Births per 1000 women age 15-19) : continuous

e) Share of seats in parliament (in percentage) : continuous

f) Females with at least some secondary education (percent age 25 and older) : continuous

g) Males with at least some secondary education (percent age 25 and older): continuous

h) Female labor force(percent age 15 and older):continuous

Dataset: Kaggle.com

Data Link: https://www.kaggle.com/datasets/gianinamariapetrascu/gender-inequality-index?resource=download

The first 20 rows of the Gender Inequality Index data used in the project for the analysis is given in the table below :---

TABLE 7.1:-First 20 rows of the Gender Inequality Index Data

Country name	GII	Maternal Mortality Ratio	Adolescent Birth Rate	Share Of Seats In Parliament	Females With Secondary Education	Males With Secondary Education	Female Labour Force
Switzerland	0.018	5	2.2	39.8	96.9	97.5	61.7
Norway	0.016	2	2.3	45	99.1	99.3	60.3
Iceland	0.043	4	5.4	47.6	99.8	99.7	61.7
Australia	0.073	6	8.1	37.9	94.6	94.4	61.1
Denmark	0.013	4	1.9	39.7	95.1	95.2	57.7
Sweden	0.023	4	3.3	47	91.8	92.2	61.7
Ireland	0.074	5	5.9	27.3	88.1	86	56.5
Germany	0.073	7	7.5	34.8	96.1	96.5	56.8
Netherlands	0.025	5	2.8	39.1	89.8	92.7	62.4
Finland	0.033	3	4.2	46	99	98.5	56.5
Singapore	0.04	8	2.6	29.8	80.5	85.9	59.4
Belgium	0.048	5	5.3	42.9	87.2	89.7	49.8
New Zealand	0.088	9	12.6	49.2	82	81.8	65.1
Canada	0.069	10	7	34.4	100	100	60.8
Luxembourg	0.044	5	4.3	35	100	100	58.5
United Kingdom	0.098	7	10.5	31.1	99.8	99.8	58
Japan	0.083	5	2.9	14.2	95.9	92.7	53.3
South Korea	0.067	11	2.2	19	83.1	93.1	53.4
United States	0.179	19	16	27	96.5	96.4	55.2
Israel	0.083	3	7.6	28.3	91.6	93.7	58.5