

A study on Crime rates in India
socio-economic regional factors affecting them

*Thesis to be submitted in partial fulfilment of the requirement
of the degree*

of

BSc. Statistics Honours.

by

Biswarup Majumdar

Roll No: 21-300-4-07-0439

Registration No: A01-1112-0852-21

Session-2021-2024

Under the supervision of

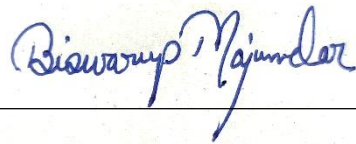
Prof. Mausumi Bose



DEPARTMENT OF STATISTICS
ST. XAVIER'S COLLEGE (AUTONOMOUS), KOLKATA

DECLARATION

I affirm that I identify all my sources and that no part of my dissertation paper uses unacknowledged materials.



Biswarup Majumdar

Department of Statistics

St Xavier's College (Autonomous), Kolkata

Date: April, 2024

Index

| | |
|---|----|
| 1. Introduction..... | 02 |
| 2. Research Objectives..... | 02 |
| 3. Research Importance..... | 02 |
| 4. Source of data..... | 02 |
| 5. Description of the data..... | 03 |
| 6. Graphical representation of data (scatter plot)..... | 04 |
| 7. Regression Analysis..... | 07 |
| 8. Regression Diagnostics..... | 08 |
| 8.1.Detection of influential points..... | 08 |
| 8.2.Multicollinearity..... | 10 |
| 8.3.Heteroscedasticity | 12 |
| 9. Test for significance of predictors..... | 14 |
| 10. Regression fit..... | 15 |
| 10.1. Multiple correlation coefficients..... | 15 |
| 10.2. Partial correlation coefficients..... | 16 |
| 11. Goodness of fit..... | 16 |
| 12. Conclusion..... | 17 |
| 13. Appendix of codes..... | 17 |
| 14. References..... | 17 |
| 15. Acknowledgement..... | 18 |

1. Introduction:

Crime, a social and legal construct, is a phenomenon that has been present in societies across the world and throughout history. It is an issue of significant concern due to its impact on victims, communities, and the overall society. The frequency and nature of criminal offences not only the mirror of the evolving socio-economic environment but also underscore the efficiency of governance and law enforcement measures. It is a complex task to understanding the patterns and causes of crime in India, a diverse country with population over a billion. This complexity arises from the myriad socio-economic factors and regional variations that influence crime rates.

The backbone of criminal justice system in India, The Indian Penal Code (IPC) and Special and Local Laws (SLL), serves as both a reflexion and a catalyst for the nation's socio-economic development. While IPC establishes a uniform code of criminal laws applicable throughout the nation, SLL address specific legal provisions enacted by central or state governments for specialized or regional concerns, despite of the diverse tapestry of India's legal landscape.

2. Research Objectives:

This research endeavours to dissect the intricate relationship between crime rates, socio-economic factors throughout the whole nation, with the following objectives: -

1. Exploration of socio-Economic Determinants:

Study the impact of crucial socio-economic measures, including unemployment rates, GDP, inflation, and the economic performance at the state level, on the occurrence of IPC & SLL crimes in various states and territories.

2. Regional Disparities:

Assess variations in crime rates and police efficacy across diverse geographical regions of India, elucidating the underlying factors contributing to regional disparities in law enforcement and crime prevention strategies.

3. Identification of potential factors:

Employ rigorous statistical analyses to identify casual factors driving divergent trends in crime rates, thereby informing evidence-based policy interventions and proactive crime prevention measures at both regional and national level.

3. Research Importance:

The connection between crime rates and socio-economic factors can be better understood through certain theoretical frameworks. These are crucial in understanding the complicated aspects of criminal behaviours. Strain theory, Social disorganization theory, Routine activity theory represent seminal perspectives in criminology that offer insights into how socio-economic conditions shape patterns of criminal activity.

This part of study provides a critical analysis of these theories, outlining their significance in understanding the levels of crime and their regional differences.

4. Source of data:

- Crimes in India,2021 (from national Crime Record Bureau)
- Handbook of statistics on Indian states,2021-22 (from Reserve Bank of India)

5. Description of the data used:

Here we have data on 36 states and territories of India for the year 2021 on variables namely proportion of IPC crimes, proportion of SLL crimes, Unemployment rate for rural areas, Unemployment rate for urban areas, GDP, state-wise average general inflation, state-wise average inflation of food & beverages and Net State Domestic Product (NSDP).

We restrict ourselves in some cases. For reference we don't include some territories of India due to unavailability of the data required. Hence the total number of observation boils down to 33.

All the variables are renamed for further use and detailed description of them are given below:

➤ **Response variable:**

- Rate of IPC crimes (Y1):
the data is given in the ratio scale where,
IPC crime rate is given due to crime per lakh population.
- Rate of SLL crimes(Y2):
The data is given in the ration scale where,
IPC crime rate is given due to crime per lakh population

➤ **Predictors:**

- Unemployment rate for rural areas(X1):
It denotes the rate of unemployment in the rural areas throughout the different states and territories of India for per thousand people.
- Unemployment rate for urban areas(X2):
It denotes the rate of unemployment in the urban areas throughout the different states and territories of India for per thousand people.
- GDP(X3):
Logarithmic value of GDP or Gross Domestic Product is given for different states and territories of India with respect to the year 2021-2022 taking base year as 2011-2012 in per lakh Indian rupees with respect to current prices.
- NSDP(X4):
Logarithmic value of NSDP or Net State Domestic Product is given for different states and territories of India with respect to the year 2021-2022 taking base year as 2011-2012 in per lakh Indian rupees with respect to current prices.
- State-wise average general inflation(X5):
Wages or State-wise average inflation for general i.e., Consumer Price Index (CPI) for general is given in percentage.
- State-wise average inflation of food and beverages(X6):
Wages or State-wise average inflation for food and beverages i.e., Consumer Price Index (CPI) for food and beverages is given in percentage for India.

States and territories are divided into 3 categories according to their population Such that if population less than 500 lakh we assign 1, more than 1000 lakh we assign 3 and 2 for rest in-between.

- Dummy variable1 (X7):
Assigned 1 to those where population lies below 500 lakhs and 0 to rest.
- Dummy variable2 (X8):
Assigned 1 to those where population lies between 500 lakhs and 1000 lakhs and 0 to rest.

➤ **Cases:**

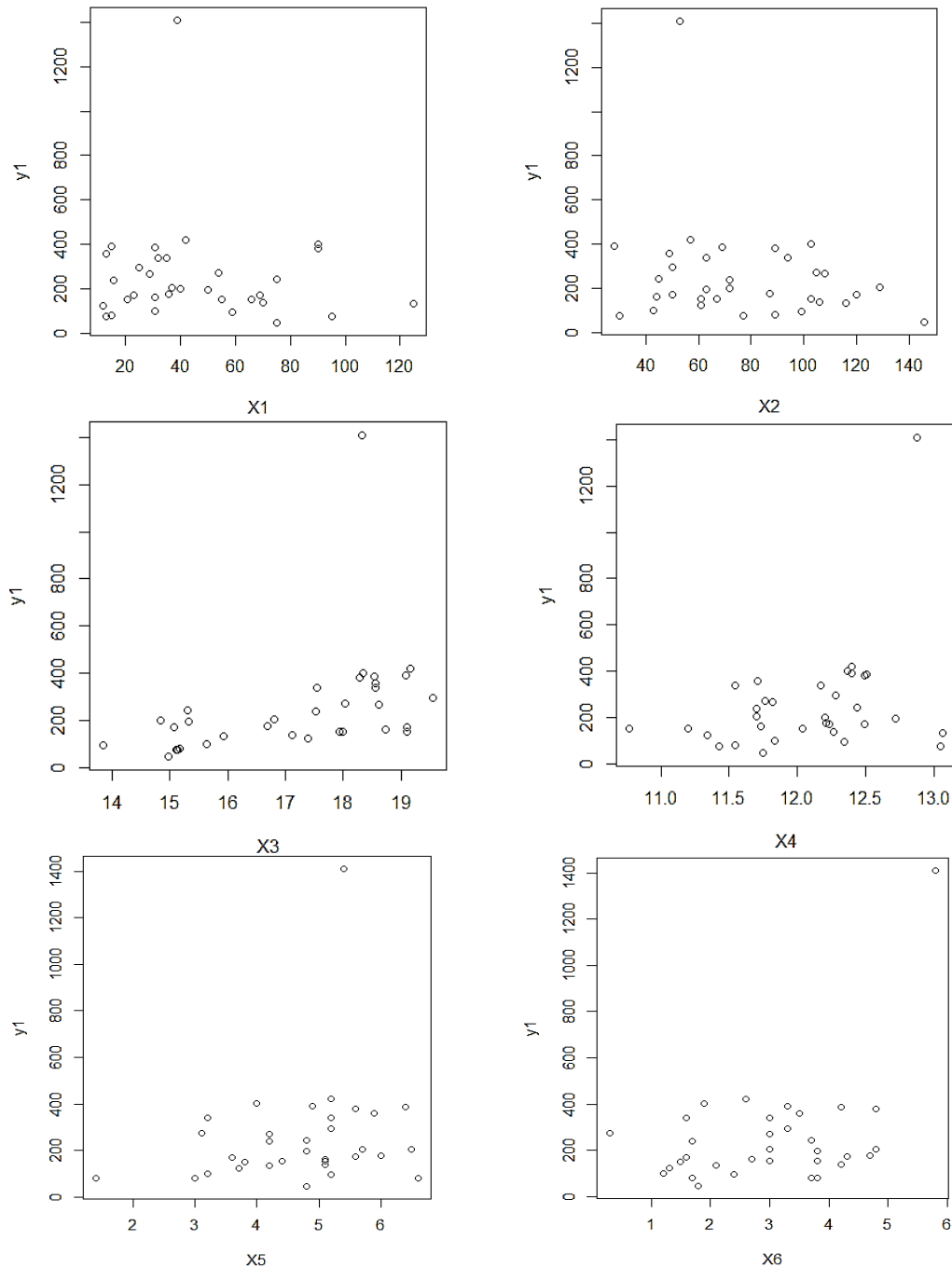
- Case 1: we will regress Y1 jointly on X1,X2,X3,X4,X5,X6,X7,X8

$$Y1 \sim X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8$$
, for all observations
- Case 2: we will regress Y2 jointly on X1,X2,X3,X4,X5,X6,X7,X8

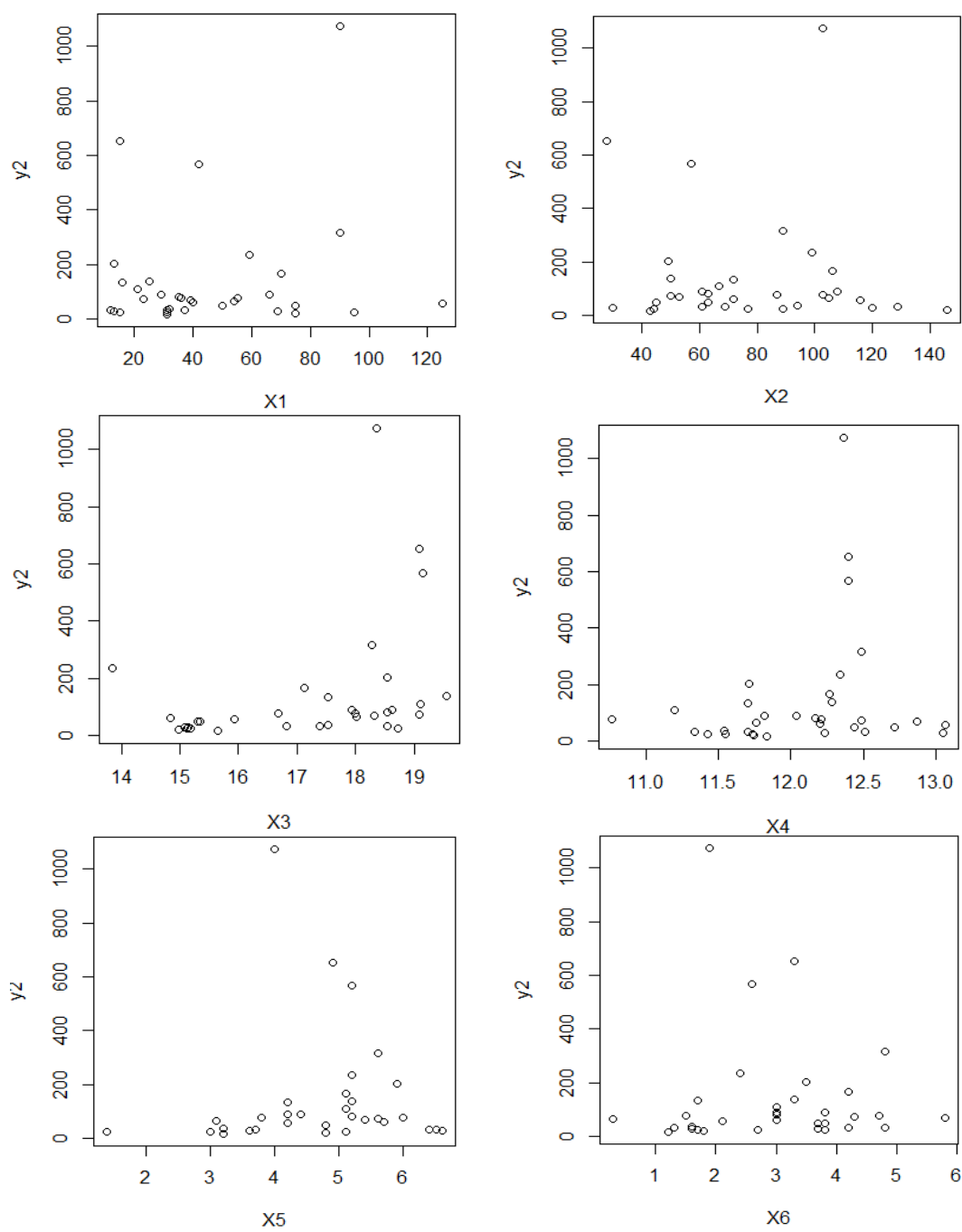
$$Y2 \sim X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8$$
, for all observations

6. graphical representation of data(scatter plot):

Case1:



Case 2:



7. Regression Analysis:

The term “Regression” was proposed by Sir Francis Galton, which literally means “ backward movement, return to an earlier stage of development.”.

The idea is to check how the predictor variables jointly influences the response variable though the linear relationship i.e., to describe and evaluate the relationship between the Y (explained/dependent/response variable) with other X_k 's (explanatory/independent/predictor variables). When $k=1$, we called it simple regression and otherwise ($k > 0$) multiple regression.

Objective of the regression analysis:

- Check if X is significant for Y or not.
- Effects on Y for changing value of X.
- Predict or forecast the value of Y for a given set of X.

Model:

Let Y be the response variable and $X_1, X_2, X_3, \dots, X_p$ be p independent predictors and they can be modelled as follows,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i \\ = E(Y | x_{1i}, x_{2i}, \dots, x_{pi}) + \varepsilon_i$$

Where, ε_i are the residuals(error terms associated with) of Y_i for all values of i.

β_0 is the intercept parameter and β_j 's is the partial regression coefficient associated with X_j which can be interpreted as the change in the value of Y, for unit change in X_j , keeping the other predictors fixed. $j= 1(1)p$

Assumptions:

Under classical linear model there are some assumptions as follows:

- The regression model must be linear in parameter. However it may or may not be linear in variables.
- X_i 's are non-stochastic i.e., data matrix X is non stochastic or non-probabilistic in nature.
- Errors are normally distributed with zero mean and constant variance. i.e., $\varepsilon_i \sim N(0, \sigma^2) \forall i$.
- The errors are uncorrelated. i.e., $\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j$
- The number of observations must be greater than the number of parameters to be estimated.
- Variance of X must be positive and should not include any outliers.

Estimation of coefficients:

To estimate the regression coefficients we have to minimize the sum of the errors of predictors, in the method of ordinary least squares. i.e., we have to minimize the following,

$$\Delta = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \dots - \beta_p x_{pi})^2$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \dots - \beta_p \bar{x}_p$$

$$\hat{\beta}_j = - \frac{R_{yj}}{R_{yy}} \frac{\sigma_y}{\sigma_j}, j=1(1)p$$

Where R_{yy} is the cofactor of $(1,1)^{\text{th}}$ element of R and R_{yj} is the cofactor of $(1,j+1)^{\text{th}}$ element of R.

Fitted regression model:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_p X_{pi} = E(Y | X_{1i}, X_{2i}, \dots, X_{pi})$$

Where, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \dots - \hat{\beta}_p \bar{x}_p$ and $\hat{\beta}_j = -\frac{R_{yj}}{R_{yy}} \frac{\sigma_y}{\sigma_j}, j=1(1)p$

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value |
|----------|-------|---------|---------|---------|
| Constant | -3228 | 1334 | -2.42 | 0.023 |
| X1 | -1.29 | 1.94 | -0.66 | 0.513 |
| X2 | 0.39 | 1.72 | 0.23 | 0.820 |
| X3 | 86.3 | 30.2 | 2.86 | 0.009 |
| X4 | 172 | 112 | 1.53 | 0.139 |
| X5 | -67.8 | 51.9 | -1.31 | 0.204 |
| X6 | 65.8 | 40.8 | 1.61 | 0.120 |
| X7 | 117 | 164 | 0.71 | 0.482 |
| X8 | -22 | 153 | -0.14 | 0.889 |

Case1

Residual standard error: 202.4 on 24 degrees of freedom

Multiple R-squared: 0.4418, Adjusted R-squared: 0.2558

F-statistic: 2.375 on 8 and 24 DF, p-value: 0.04841

Comment: as per adjusted r square, only 25.58 % of the total variation can be explained by the linear fit of Y1 jointly on X1,X2,X3,X4,X5,X6,X7 and X8. The fit is not seeming to be good. So, we will go for some test to remove unwanted values and refit the regression.

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value |
|----------|-------|---------|---------|---------|
| Constant | -2260 | 1423 | -1.59 | 0.125 |
| X1 | 2.00 | 2.07 | 0.96 | 0.344 |
| X2 | -0.18 | 1.83 | -0.10 | 0.921 |
| X3 | 65.4 | 32.3 | 2.03 | 0.054 |
| X4 | 100 | 120 | 0.83 | 0.413 |
| X5 | 13.4 | 55.4 | 0.24 | 0.811 |
| X6 | -45.7 | 43.6 | -1.05 | 0.305 |
| X7 | 76 | 175 | 0.43 | 0.668 |
| X8 | 100 | 163 | 0.62 | 0.544 |

Case2

Residual standard error: 216 on 24 degrees of freedom

Multiple R-squared: 0.2872, Adjusted R-squared: 0.04966

F-statistic: 1.209 on 8 and 24 DF, p-value: 0.3354

Comment: as per adjusted r square, only 4.97 % of the total variation can be explained by the linear fit of Y2 jointly on X1,X2,X3,X4,X5,X6,X7 and X8. The fit seems to be pretty bad. So, we will go for some test to remove unwanted values and refit the regression.

8. Regression diagnostic:

Often regression based on different subsets of the data sometimes produces various results which can lead to question the model stability. Reasons of that can be unusual circumstances like unknown error in data collection and collinearity can be also a potential cause of it.

❖ 8.1 Detection of influential points:

In general, an outlier is any unusual data point which is discordant with other points.

There are two kinds of outliers: -

- Outlier in y direction (called error outlier or outlier)
- Outlier in x direction (called high leverage point)

Residuals:

The residual is defined as the difference between the fitted and observed value of the study variable. It can also be considered as the deviation between the data and the fit.

It can also be considered as the model error. So, it is expected that if there is any departure from the model it will be reflected by the residual. Thus, model inadequacies can be found by residual analysis.

We define residuals as u_i and estimated residuals as \hat{u}_i

Then, **Standardized residuals** $t_i = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ where, \hat{u}_i = estimated residuals

$\hat{\sigma}$ = estimated standard errors

h_{ii} = leverage point for i^{th} observation

Where leverage is calculated $h_i = \frac{1}{(n-1)} \left(\frac{x_i - \bar{x}}{s_x} \right)^2 + \frac{1}{n}$ as,

Rule for detection of outlier: the i^{th} observation is an $|t_i|$ outlier if > 2

Rule for detection of high leverage points: the i^{th} observation is a high leverage point if

$h_{ii} > l_0 = 3*(p+1)/n$; where p is the total number of prediction variables and n is the total number of observations.

for **case1** , we get,

Fits and Diagnostics for Unusual Observations

| Obs | y1 | Fit | Resid | Std Resid | R |
|-----|------|-----|-------|-----------|---|
| 31 | 1410 | 666 | 744 | 4.47 | R |

R Large residual

for **case2** , we get,

Fits and Diagnostics for Unusual Observations

| Obs | y2 | Fit | Resid | Std Resid | R |
|-----|------|-----|-------|-----------|---|
| 7 | 655 | 261 | 394 | 2.04 | R |
| 12 | 1076 | 374 | 702 | 3.73 | R |

R Large residual

In both cases, $l_0 = 0.81818$

For case1, we have point 31 as a potential outlier and for case2, we have points 7 and 12 as potential outliers but there is no potential leverage point found.

If a dataset is small, then deletion of values greatly affects the fit and statistical conclusions. So, in measurement of that if a point is an influential point or not, we should consider location of both in x and y space.

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{ps^2}$$

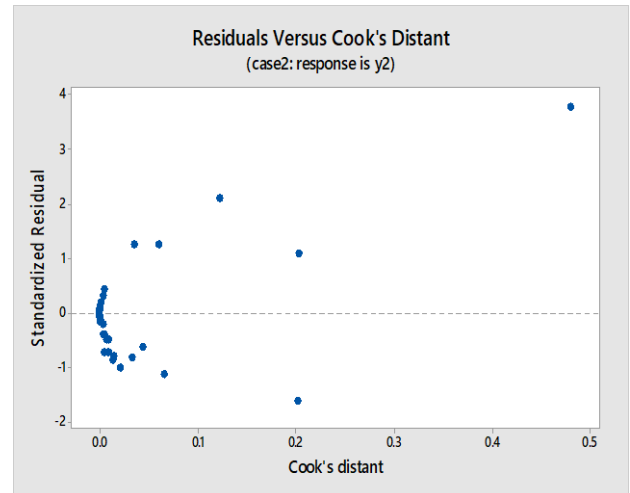
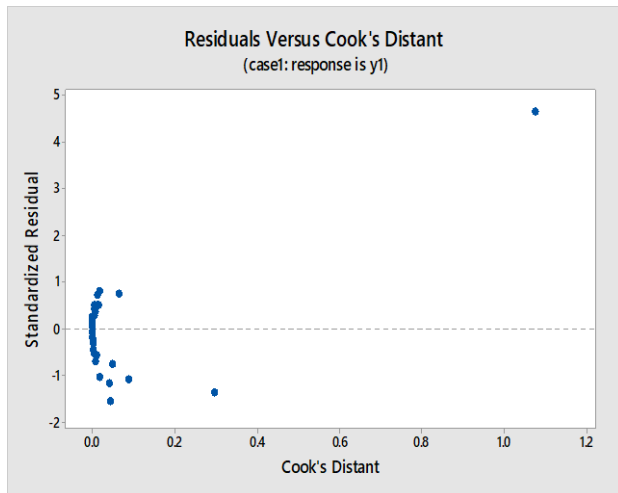
Cook's Distance:

where $\hat{y}_{j(i)}$ is the fitted response value obtained when excluding i,

$$s^2 = \frac{\sum \hat{u}_i^2}{(n-p)} = \text{MSE of regression model, where } \hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i$$

And p: number of parameters in the model , n: number of observations

Rule for detection of influential point: a value more than 1 indicates that the point should be studied further and consider dropping.



Standard solution: assessing the effect of the data point has on the regression by deleting the point and refitting the regression. If the quantities (i.e., estimated coefficients, fitted values, standard errors and so on.) changes markedly, the point is highly influential.

Comment:

In case1, the value of Cook's distance for the 31th point is $1.07514 > 1$, hence we can take this potentially outlier point as high influential point; thus, we delete that from our dataset before further analysis.

In case2, the value of Cook's distance for both the potential outlier point 7th and 12th is less than 1, hence we can conclude that those two points are not high influential points, and we proceed without deleting them in further analysis.

❖ 8.2 Multicollinearity:

Before diving into modelling, a crucial initial inquiry arises: do we need all predictors for our study? It's possible to exclude certain predictors if a subset proves adequate for predicting diabetes, or conversely, retain all predictors if they hold equal importance. So, we examine the concept of **Multicollinearity**.

The term "Multicollinearity" was initially introduced by Ragnar Frisch in statistics, referring to the presence of exact or "perfect" linear relationships among some or all predictors in a multiple regression model.

In such cases, the coefficient estimates of the regression model may exhibit irregular fluctuations in response to minor changes in the model or data. While multicollinearity might not diminish the overall reliability or predictive capacity of the model, it does impact the accuracy of calculations pertaining to individual predictors. Consequently, a multiple regression model with correlated predictors can accurately assess the collective predictive capability of the predictors for the outcome variable. However, it may not yield reliable insights into the significance of any individual predictor or identify which predictors are redundant in relation to others.

We have perfect multicollinearity if the correlation between explanatory variables is ± 1 which in practice is rarely observed. The issue arises when there is an approximate linear relationship among two or more predictors.

Mathematically, this relationship represented as:

$$b_0 + b_1X_{1i} + b_2X_{2i} + \dots + b_kX_{ki} = 0 \quad \forall i = 1(1)n$$

Where, b_j 's are constant and X_{ji} is the i^{th} observation on k^{th} covariate. $\forall j = 1(1)k$

Even after eliminating redundancies, nearly multicollinear variables may persist due to inherent correlations within the studied system. In such instances, rather than adhering strictly to the previously mentioned equation, we adapt it into a modified form, incorporating an error term represented by ϵ_i . We denote variables as nearly perfectly multicollinear when the variance of ϵ_i 's is minimal for certain values of the b_j coefficients.

i. Visualization Method - using Correlation Plot:

Correlation plot is the graph plot of correlation matrix also called heatmap. In this plot, correlation coefficients are coloured according to the value. Higher the intensity of colour, higher the correlation.

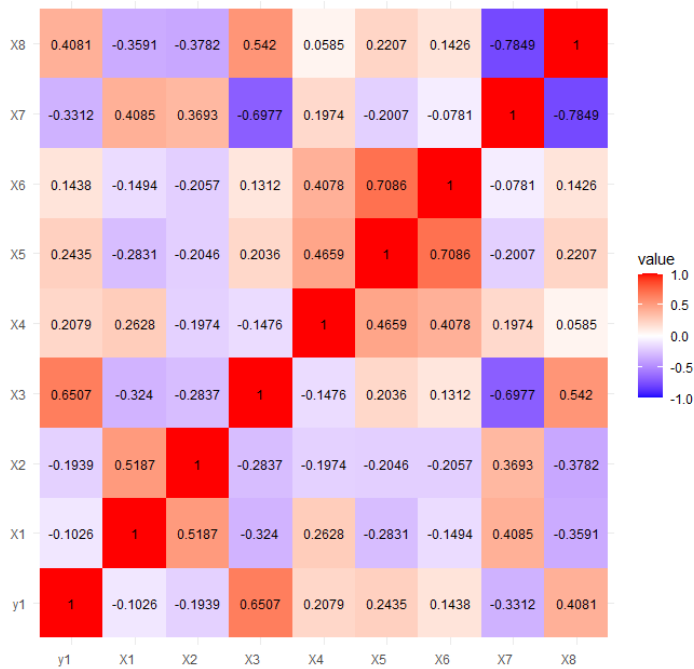
ii. Using Variance Inflation Factors(VIFs):

Existence of collinearity in data set can be checked by **Tolerance** which is $1 - R_k^2$ where R_k^2 is the multiple correlation coefficient between x_k and other independent variables. Also, one can consider **variance inflation factor (VIF)** which is nothing but the reciprocal of tolerance. VIF for the k^{th} predictor is given as:

$$\text{VIF}_k = 1/(1 - R_k^2)$$

Range of tolerance is from 0 to 1. A high value of tolerance & low value of VIF is acceptable. Tolerance value < 0.2 i.e., $\text{VIF} > 5$ is really concerning.

Comments:

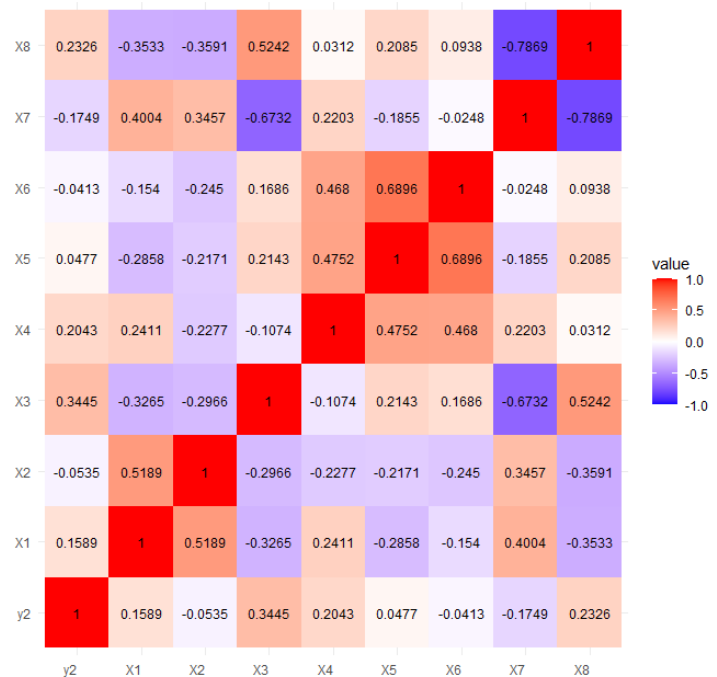


Correlation heatmap for case 1

There is high negative correlation between the variables X3, X7 and X7, X8. Again, high positive correlation is noticed among Y1, X3 and X5, X6. Rest of the variables seems to be more or less correlated among themselves.

Correlation heatmap for case 2

There is high negative correlation between X3, X7 and X7, X8 where high positive correlation is seen between X5, X6. rest of the variables are more or less correlated among themselves and some moderately correlated



Values of Variance Inflation Factors:

```
> vif(model_1)# collinearity for casel
      X1      X2      X3      X4      X5      X6      X7      X8
2.396081 1.981236 1.988227 2.704271 3.071645 2.101823 4.501144 3.180293
```

From the above snapshot we can observe that for all the predictor variables, the values of VIF are less than 5. Hence, all the independent variables can be considered as uncorrelated and should be included in our further study.

```
> vif(model_2)# collinearity for case2
      X1      X2      X3      X4      X5      X6      X7      X8
2.338242 2.021236 1.923238 2.760027 2.811491 2.119333 4.560045 3.143896
```

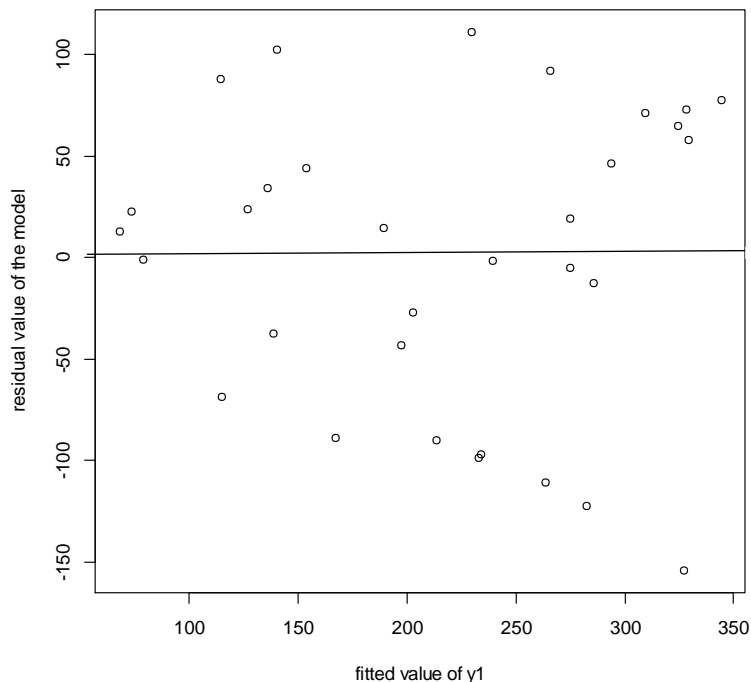
We can conclude same for case2 also as we said in case1.

❖ 8.3 Heteroscedasticity:

Heteroscedasticity is a phenomenon in regression analysis where the variability of the residuals changes across different levels of the predictor variables. It violates the assumption of homoscedasticity, potentially leading to biased estimates and incorrect inferences.

We plot residuals values against the fitted y values and check if it shows any systematic pattern or not. If the data shows any systematic pattern then we can assume that the data may be potentially heteroscedastic, and some tests will be done to confirm it.

We can make variable transformation to reduce the variability of the data i.e., to make the data homoscedastic.

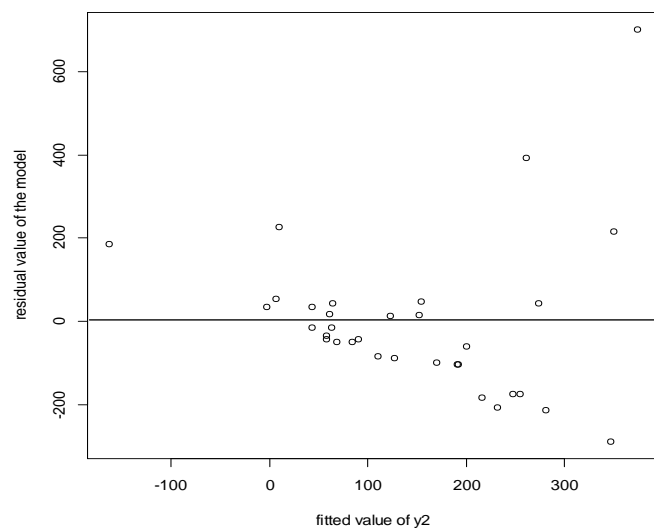


Case1:

Residual values do not seem to have any pattern with the fitted value of Y1 variable. So it can be safe to say that the data may not have any presence of heteroscedasticity.

Case 2:

Residual do not seem to have any Pattern with the fitted variable y2. As it shows no pattern, it is safe to say that the data may not have any presence of heteroscedasticity.



Hence we can perform two consecutive test's for confirming the presence of heteroscedasticity as follows:

i. Glejser's Test:-

This test is based on the assumption that σ^2 is influenced by only one variable which is influencing the heteroscedasticity.

for this the model chosen is, $|\hat{u}_i| = \delta_0 + \delta_1 x_i^{\delta_2} + u_i^* \quad \forall i = 1(1)n$
where u_i^* is associated disturbance term.

We are to test for,

$$h_0: \delta_1 = 0 \text{ vs } h_1: \delta_1 \neq 0$$

We perform the test for, $\delta_2 = -1, -1/2, 1/2, 1$

$$\text{Test statistic, under } h_0, \quad T = \frac{\hat{\delta}_1}{se(\hat{\delta}_1)} \sim t_{n-2}$$

The value is said to be significant under 5% level iff $|T_{obs}| > t_{\frac{\alpha}{2}, n-2}$

Alternatively, the value is said to be significant iff the p-value is less than α where $\alpha = 0.05$

Note: among significant values of $|T_{obs}|$ the highest one is taken and further that model is used in Goldfeldt-Quant test.

Comment: with Glejser's test, we find that for each variable $\delta_2 = 1$ giving the highest significant value for both of the cases. To reconfirm our finding we will proceed for Goldfeldt-Quant test.

ii. Goldfeldt-Quant test:-

The test is applicable if one assumes that the heteroscedastic variance σ^2 is positively related to one of the explanatory variables in the regression model.

For simplicity we assume the two variable model as,

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \text{ where suppose that } \sigma_i^2 = \sigma^2 X_i^2, \text{ keeping in mind } \sigma^2 \text{ is constant.}$$

Here we divide the model into 2 equal group and find out the fitted regression for group one and two, while leaving c observations in the middle such that each part contains $(n-c)/2$ observations. In general c is preferred to be taken as $n/3$.

We have to test for,

$$h_0: \sigma_i^2 = \sigma^2 \text{ vs } h_1: \sigma_i^2 \propto X_i^2 \quad \forall i = 1(1)n$$

$$\text{Test statistic under } h_0 \text{ is, } F = \frac{RSS_{ii}}{RSS_i} \sim F_{\left(\frac{n-c}{2}-k, \frac{n-c}{2}-k\right)}$$

Where k is the number of variables.

The assumption of homoscedasticity is rejected under 5% level of significance iff

$$|F_{obs}| > F_{\alpha; \left(\frac{n-c}{2}-k, \frac{n-c}{2}-k\right)}$$

Comment: for both cases, this test confirms that there is not present any heteroscedasticity for the variables we are testing for.

9. Test for significance of predictors:

Our objective is to check which predictors are significant in determining the effect of crime.

Testing problem is given by:

$$H_0: \beta_0 = 0 \text{ vs } H_{1j} : \beta_0 \neq 0 \forall j$$

The test statistic under H_0 is then given by:

$$T_j = \frac{\hat{\beta}_i}{s.e(\hat{\beta}_i)}, \text{ where } \hat{\beta}_i \text{ is the estimated coefficient of the } i^{\text{th}} \text{ predictor}$$

Under 10% level of significance, we reject H_0 iff $|T_{j \text{ obs}}| > \tau_{\alpha/2}$ and $\tau_{\alpha/2} = 1.64$

As per the data shown in regression fit, decision table can be shown as follows:

Decision table :

| Case 1 | | Case 2 | |
|------------|----------|------------|----------|
| parameters | Decision | parameters | Decision |
| X1 | Rejected | X1 | Rejected |
| X2 | Rejected | X2 | Rejected |
| X3 | Accepted | X3 | Accepted |
| X4 | Rejected | X4 | Rejected |
| X5 | Rejected | X5 | Rejected |
| X6 | Rejected | X6 | Rejected |
| X7 | Rejected | X7 | Rejected |
| X8 | Rejected | X8 | Rejected |

Interpretation:

In the light of the given data,

We can see at 10% level of significance in both cases only variable X3 found to be significant.

i.e., we can state that IPC crimes or SLL crimes significantly depends on GDP of that state or territory.

10. Regression Fit:

After all the corrections made in the model like determining the influential points, multicollinearity check and the checking for heteroscedasticity and fixing them as per needed we get out fitted model with coefficients as,

Case1 :

$$Y_1 = -532.168 + 43.695 X_3$$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|--------------|
| (Intercept) | -532.168 | 160.418 | -3.317 | 0.00239 ** |
| X3 | 43.695 | 9.309 | 4.694 | 5.52e-05 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 85.78 on 30 degrees of freedom
Multiple R-squared: 0.4234, Adjusted R-squared: 0.4042
F-statistic: 22.03 on 1 and 30 DF, p-value: 5.522e-05

Comment: according to adjusted R square value, only 40.42% of the total observation can be explained by the fit of Y1 on X3.

Case2:

$$Y_2 = -655.88 + 46.5X_3$$

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|----------|
| (Intercept) | -655.88 | 392.95 | -1.669 | 0.1052 |
| X3 | 46.50 | 22.76 | 2.043 | 0.0496 * |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 211.3 on 31 degrees of freedom
Multiple R-squared: 0.1187, Adjusted R-squared: 0.09027
F-statistic: 4.175 on 1 and 31 DF, p-value: 0.04959

Comment: according to adjusted R square value, only 9.03% of the total observation can be explained by the fit of Y2 on X3.

❖ 10.1 Multiple Correlation coefficient:

Here we consider the problem of measuring the degree to which one of the random variables may be said to be dependent on the other variables jointly. i.e., let X_1, X_2, \dots, X_p be p random variables then the joint influence of X_2, X_3, \dots, X_p on X_1 called to be multiple correlation coefficients and it is denoted as $\rho_{1.23\dots p}$.

$$\rho_{1.23\dots p} = \left(1 - \frac{|R|}{R_{11}}\right)^{1/2}$$

where R is the correlation matrix and R_{11} is the co-factor of it's $(1,1)^{\text{th}}$ element.

Comment: in case 1, the value of multiple correlation is 0.6507 between Y1 and X3 i.e., they have moderate positive correlation among themselves.

In case 2, the value of multiple correlation is -0.0535 between Y1 and X3 i.e., they have moderate negative correlation among themselves.

❖ 10.2 Partial correlation coefficient:

Here we consider the problem of measuring the degree to which two random variables said to be related when the influences of the other random variables are eliminated from each of them i.e., the marginal impact of X_j on X_i keeping the other predictors constant and it is denoted as $\rho_{1j.23...p}$.

$$\rho_{1j.23...p} = - \frac{R_{1j}}{\sqrt{R_{11} R_{jj}}}$$

Where R_{ij} is the co-factor of $(i,j)^{th}$ element of correlation matrix R .

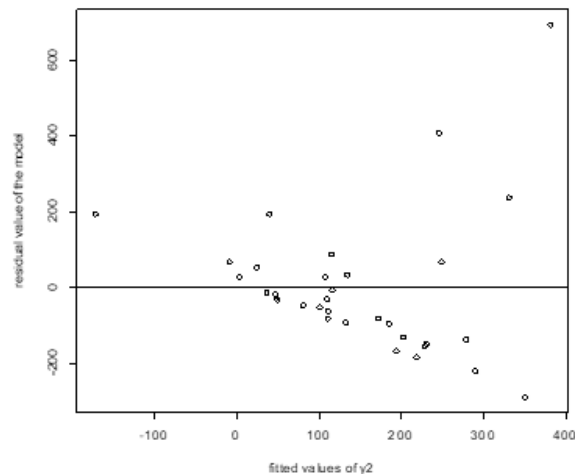
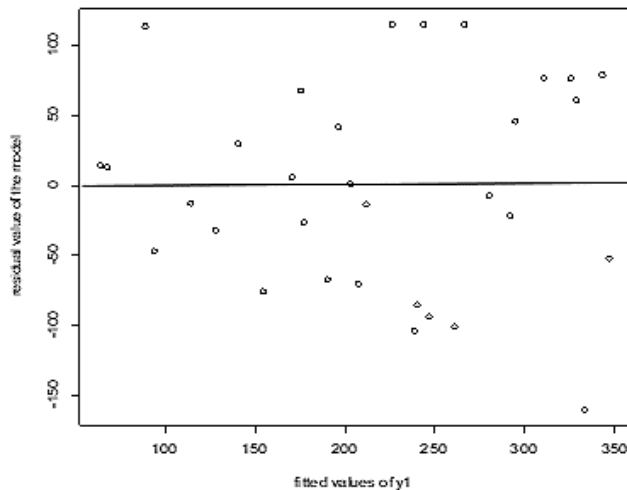
In similar fashion when the true regression of X_1 as well as X_j on X_2, X_3, \dots, X_p is linear then,

$$\rho_{1j.23...p} = \frac{E[cov(X_1, X_j | X_2, X_3, \dots, X_p)]}{\sqrt{E[v(X_1 | X_2, X_3, \dots, X_p)]} \sqrt{E[v(X_j | X_2, X_3, \dots, X_p)]}}$$

Comment: as there are only two variables, so partial correlation would be same as multiple correlation.

11. Goodness of the fit:

To understand the goodness of the fit graphically, we plot residual versus predicted response (in a linear combination of all predictors).



Residual standard error(RSE) = $\frac{RSS}{n-p}$, where RSS is the residual sum of squares

$$R \text{ squared} = R^2 = 1 - \frac{RSS}{TSS} = \rho_{1.23..p}^2 = \left(1 - \frac{|R|}{R_{11}}\right)$$

To get even more accurate result we use adjusted R squared, which is defined as follows,

$$\text{Adjusted } R^2 = 1 - \frac{n-1}{n-p} (1 - R^2) = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$$

Comment: for case 1, $R^2 = 0.4234$ and adjusted $R^2 = 0.4042$

For case 2, $R^2 = 0.1187$ and adjusted $R^2 = 0.09027$

We can see that the fit is not good . i.e., the model doesn't able to explain most of the variation of the data given.

12. Conclusion:

we try to find that how used factors influences the crime of India (IPC and SLL). We at the end come to the conclusion that these factors are not sufficient to explain the crime rate. Only GDP of a place found be significant although not highly effective to explain crimes. So, the is further scope for work with this project if we have significant amount of predicting variables and sufficient amount of data.

13. Appendix of codes:

Codes are uploaded in their original format at Github. The works done with help of Minitab software are not given.

Link: <https://github.com/Rup21/Dissertation-project.git>

14. References:

- ❖ Fundamental of statistics, volume 1: [A.M. Goon, M.K.Gupta, B. Dasgupta](#)
- ❖ Fundamental of statistics, volume 2: [A.M. Goon, M.K.Gupta, B. Dasgupta](#)
- ❖ Crimes in India 2021, statistics volume 1: [docs](#)
- ❖ Crimes in India 2021, statistics volume 2: [docs](#)
- ❖ Crimes in India 2021, statistics volume 3: [docs](#)
- ❖ Handbook on statistics on Indian states: [docs](#)

Reference folder : [references](#)

15. Acknowledgement:

I extend my heartfelt gratitude to St. Xavier's College (Autonomous), Kolkata, for providing me with the opportunity to explore the subject of my dissertation. I am deeply thankful to my supervisor, Prof. Mausumi Bose , for her unwavering guidance, supervision, and encouragement throughout this journey. I am immensely grateful to my parents and friends for their invaluable help and support during the course of this project.

---- Thank you ----