The Happiness Scorecard

MSDA3055 : Linear Regression and Time Series Analysis

Session: Jan 2024 – Apr 2024

Professor – Aghil Alaee Khangha

Submitted by

Kunal Malhan

Neha Manjrekar

Nithin Rachakonda

Prasanna Rushi

Ritika Kapoor

School of Professional Studies
Clark University
Worcester, MA

Abstract

This investigates the relationship between various national-level factors and a country's Happiness Score using linear regression. We explore how factors like Human Development Index (HDI), healthcare, population density, pollution, literacy, demographics, and more correlate with happiness scores.

The methodology utilizes multiple linear regression analysis. We will employ various diagnostic tests to identify and address potential issues like outliers, multicollinearity, and violations of normality and homoscedasticity. Additionally, we will explore model accuracy through adjusted R-squared and potentially build alternative models with different variable combinations to achieve the best fit. Finally, we will assess the model's adherence to linear regression assumptions through appropriate hypothesis tests.

This study provides valuable insights into the complex interplay of factors contributing to national happiness. The findings can inform future research and potentially guide policy decisions aimed at enhancing national well-being.

List of Figures

List of Tables

Table of Contents

1. Introduction

Happiness is something every living being is looking for and human is not an exception. According to Plato in 400 BC:

"*The man who makes everything that leads to happiness depends upon himself, and not upon other men, has adopted the very best plan for living happily*."

However, most intelligent species on earth human try to control happiness based on the factors and environment surrounding humans. This is why the definition of Happiness changes over the period and many definitions of happiness came up during different times with one of the few recently given by Henry David Thoreau, during the 19th century:

" *Happiness is like a butterfly; the more you chase it, the more you chase it, the more it will elude you, but if you turn your attention to other things, it will come and sit softly on your shoulder*."

The objectives and motivation of performing this study is around factors that are correlated to Happiness.

1.1 Happiness Index

The Happiness Index is one tool that can be utilized by researchers, community organizers and policymakers to understand more clearly and promote issues surrounding social justice; promote economic equity both nationally and internationally; and encourage environmental sustainability among other benefits relating to happiness including communal wellness. The survey instrument and data have been made openly available to community organizers, educators, researchers, students, organizations, government, and others to foster societal transformation. No other index is as unique as this one in terms of being available online at no cost for anyone who responds to surveys all over the world. In addition, users can include their questions in the survey tool or tailor it to suit specific groups thereby obtaining information quickly from their samples.

1.2 Motivation

The motivation for the study is to prepare a model that can help in finding the happiness index/score of any country/region based on the various other independent factors/parameters. This will help government bodies to pay attention to other things (i.e. factors correlated with happiness) so the Happiness Butterfly sits on the shoulders of the people of the country, and increases the happiness index/score.

2. Dataset

2.1 Independent and Dependent Variables

S.No	Data Used	Source	Type of Variable
0	Happiness Index	World Population Review	Target Variable (Continuous)
1	Population Density	World Population Review	Continuous
2	Migrants (Net)	World Population Review	Continuous
3	Fertility rate	World Population Review	Continuous
4	Median Age	World Population Review	Continuous
5	Urban Population	World Population Review	Continuous
6	Developed/Developing Status	World Population Review	Categorical
7	Human Development Index (HDI)	World Population Review	Continuous
8	Healthcare Index	World Population Review	Continuous
9	Constitutional form	Wikipedia	Categorical
10	Literacy Rate	Data Pandas	Continuous
11	Country wise mean Latitude	Git Hub	Continuous
12	Country wise mean Longitude	Git Hub	Continuous
13	Water to land ratio in %	Nation Master	Continuous
14	IQ Air 2022 World Air Quality	Wikipedia	Continuous

Table 1: Dataset IQ

2.2 Categorical Variables

- Status of the country – Developed/Developing
- Constitutional form – Provisional/Monarchy/Constitutional Monarchy/Absolute Monarchy

3. Methodology

3.1 Objectives and Research Questions

The objective for the study is to understand the Happiness Index behaviour for a particular country in correlation with the human development index (HDI), health care index, population density, pollution index, literacy rate, median age, urban population, and various other parameters. This leads to the following important research questions:

- Do the various parameters like human development index (HDI), health care index, population density, pollution index, literacy rate, median age, and urban population have any correlation with Happiness Score/Index?
- Do individual parameters have any outliers, if so, do we need to handle the outliers?
- Are all individual parameters independent to each other's? In case of any correlation among any two independent variables, should one of the independent parameters be dropped from the study?
- Does each parameter have any correlation with the Happiness Score/Index?

- What is the factor by which each parameter is correlated with the happiness score/index in the final determinant model?
- What is the accuracy of the model generated? If required multiple models be created by trying different combinations of independent parameters.

- Is the model prepared to follow all the assumptions related to the model generated?

3.2 Hypotheses

Major hypotheses that will be used to answer study questions during the project study are as follows:

- Do the various parameters like human development index (HDI), health care index, population density, pollution index, literacy rate, median age, and urban population have any correlation with Happiness Score/Index?

  Null Hypothesis (H0): None of the independent variables is correlated with the target variable i.e. $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_n = 0$
  Alternate Hypothesis (HA): At least one of the independent variables is correlated with the target variable i.e. $\beta_i \neq 0$

- Are individual parameters have any outliers, if so, do we need to handle the outliers?

  Various plots like box plots, time plots, DFFITS, DFBETAS and Cook's D bar plots will be used to study any possibility of outliers.

- Are all individual parameters independent of each other's? In case of any correlation among any two independent variables, should one of the independent parameters be dropped from the study?

  The correlation matrix provides any insights that will help in finding any existing correlation between independent variables and target variables. Under basis hypothesis for any correlation can be defined as:

  Null Hypothesis(H0):Correlation coefficient(ρ) between any 2 independent variables = 0
  Alternate Hypothesis (HA): $\rho \neq 0$ for the respective pair of independent variables

- Does each parameter have any correlation with the Happiness Score/Index?

  Null Hypothesis (H0): $\beta_i = \mathbf{0}$ (for each parameter i)

  Alternate Hypothesis (H): $\beta_i \neq \mathbf{0}$ (for respective individual parameter i)

- What is the factor by which each parameter is correlated with the happiness score/index in the final determinant model?

Estimated regression coefficients (β_i) from the final model will provide the factors by which each parameter is correlated with the target variable. The magnitude and sign of β_i indicate the strength and direction of the relationship between the corresponding independent variable (i) and the Happiness Score.

- What is the accuracy of the model generated?

  If required multiple models be created by trying different combinations of independent parameters. Adjusted R-squared assesses how well the model explains the variation in Happiness Score (dependent variable, Y) based on the independent variables (X_i). Comparison of multiple models using adjusted R-squared will be done to choose the model that best explains the data with the fewest parameters.

- Is the model prepared to follow all the assumptions related to the model generated?

  Below are the hypothesis tests for various assumptions included for all the selected predictor variables

  - Linearity Test:
    H_0: $E(Y) = \beta_0 + \beta_1 X_i$ (states relation is linear)
    H_A: $E(Y) \neq \beta_0 + \beta_1 X_i$ (states no linear relation)

  - Homoscedasticity Test:
    H_0: $\sigma^2(\varepsilon_i) = \sigma^2$ (states constant variance)
    H_A: $\sigma^2(\varepsilon_i) \neq \sigma^2$ (states unequal variance)

  - Independence Assumption:
    H_0: $\rho = 0$ (states independent error ε_t)
    H_A: $\rho > 0$ (states positively correlated)

  - Normality Assumption:
    H_0: $r_{eE} > r_c$ (normality assumption holds true)
    H_A: $r_{eE} <= r_c$ (normality assumption is inconsistent)

3.3 Potential Models

Multiple Linear Regression will be a potential model for this project study. We might be exploring the transformation of target or independent variables or other models if a case may arise.

4. Regression Analysis

4.1 Global Variables

We will adopt a significance level of alpha(α) = 0.05 for this analysis and n = 0. Note that here n is not number of records in our dataset. It is just a global variable that will be used many times in our analysis.

4.2 Box plots for Dependent and Independent Variables

Let us look at the box plots of both the predictor and target variables.

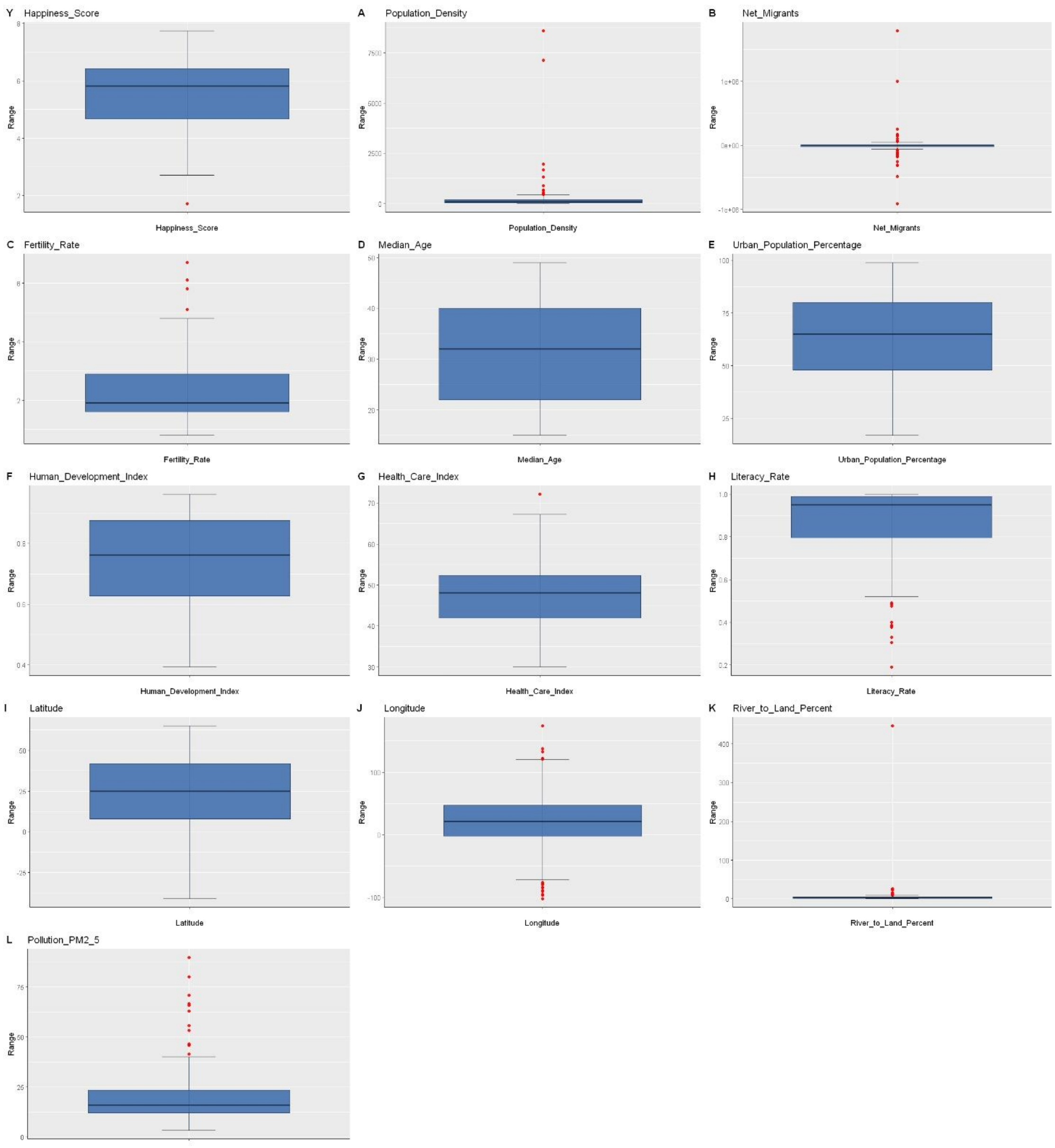

Figure 4.2-1: Box plots of Continuous Independent Variables

Attributes like population density, net migrants, and river to land percentage exhibited a presence of very wide outliers. Few outliers were identified for fertility rate, literacy rate, longitude, and pollution PM 2.5. The remaining attributes showed no outliers.

It's important to note that the ranges of these attributes also differed significantly. For instance, literacy rate ranged between 0 and 1, while population density and net migrants were measured in thousands. This substantial variation in ranges highlights the need for a data standardization technique before proceeding with the regression analysis.

By standardizing the data, we ensure all variables are placed on a common scale, mitigating the influence of outliers and allowing for a more accurate assessment of the relationships between variables in the linear regression model.

4.3 Exploratory Analysis of Data

4.3.1 Diagnostics for relationships and strong interactions

A scatter variable plot and correlation matrix was plotted for the dataset excluding the Target variable – Price.

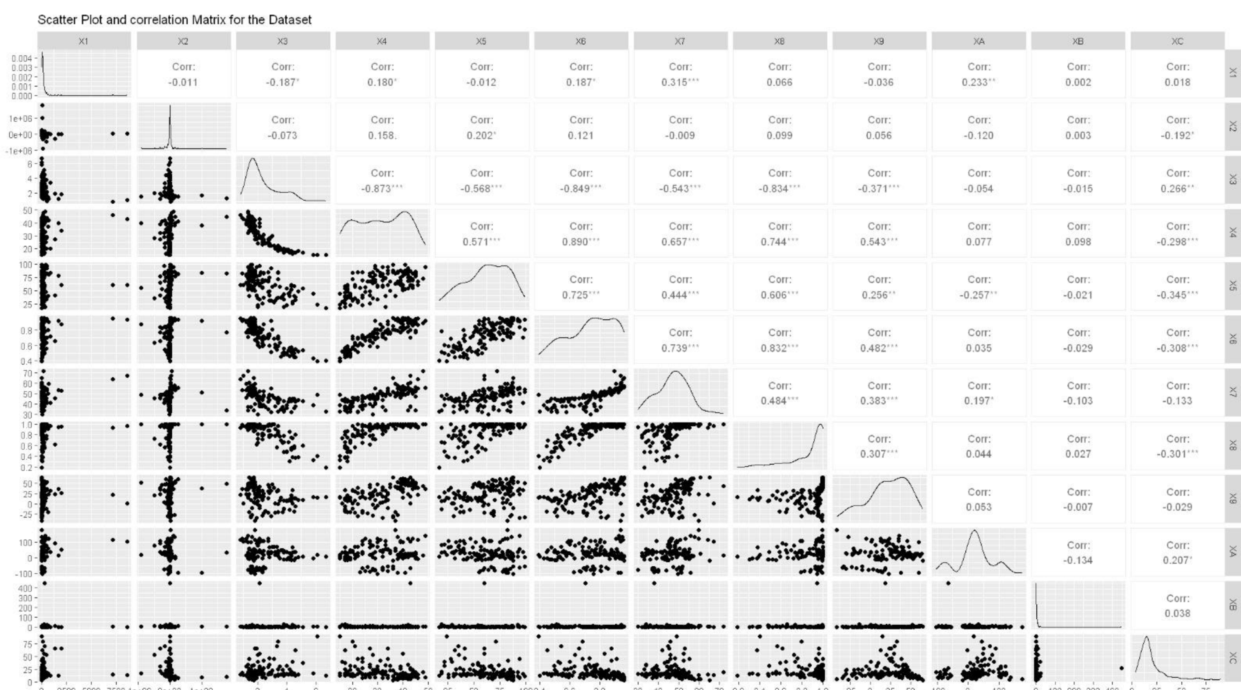

Figure 4.3-1: Scatter Plot and Correlation Matrix of Independent Variables

A heat table has been created based on the correlation coefficient value between the Independent variables. Given below is the criteria between absolute value of the correlation coefficient and the colour :

- 0 < 0.18 – Green
- 0.18 < 0.5 – Yellow
- 0.5 < 1.0 - Red

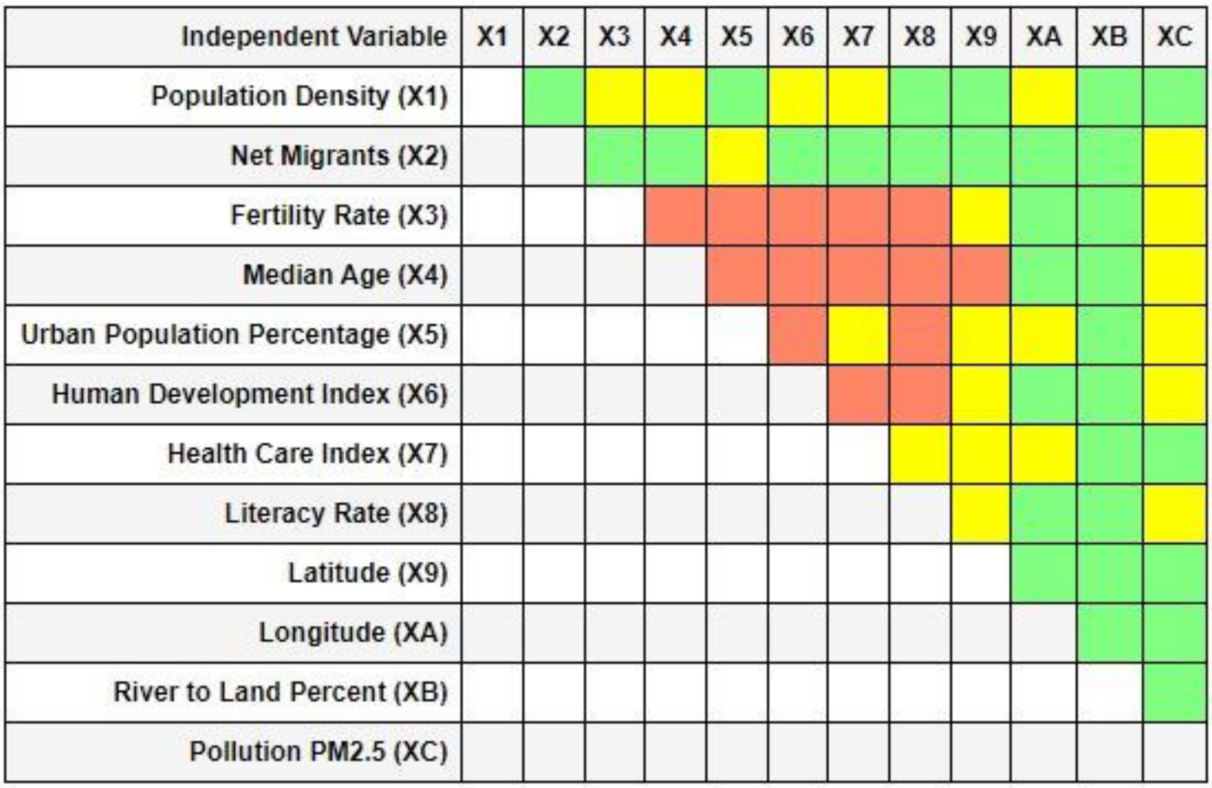

Figure 4.3-2: Heat Table of Variables

There exists high collinearity between independent variables X4, X5, X6, X7 and X8. Including all these variables in the model affects the predicting power of the same. We will filter these variables in the model selection section of the project.

4.3.2 Determine several potentially useful subsets of explanatory variables

Let us have a look at how the independent variables are corelated with the target variable.

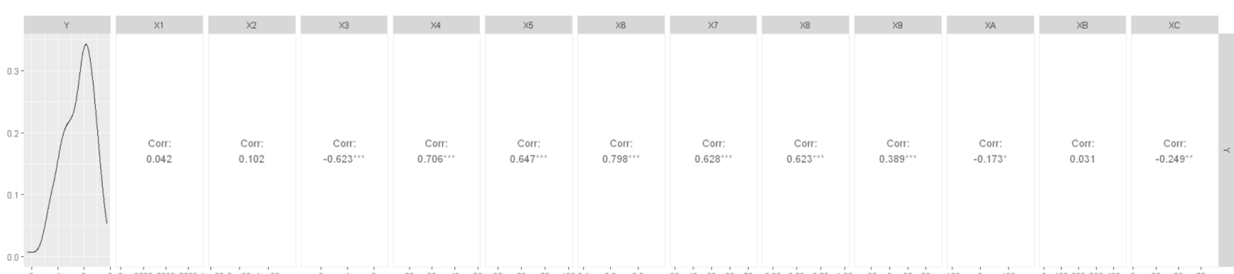

Figure 4.3-3: Correlation values between predictor variables and target variable

Based on the correlation values of independent variables with Price, the variables have been divided into 3 groups:

- Essential variables with high correlation with Price (more than 60%)
  - Fertility Rate (X3)
  - Median Age (X4)
  - Urban Population Percentage (X5)
  - Human Development Index (X6)
  - Health Care Index (X7)
  - Literacy rate (X8)
- Variables with moderate correlation with Price (10% to 60%)
  - Net Migrants (X2)
  - Latitude (X9)
  - Longitude (XA)
  - Pollution PM 2 5 (XC)
- Variables with no or very little correlation with Price (less than 10%)
  - Population Density (X1)
  - River to Land Percentage (XB)

As the correlation values are near zero, Population Density and River to Land Percentage have been excluded to be part of further study

4.4 Model Refinement

4.4.1 Converting categorical variables to dummy variables

Two categorical variables were present in the dataset: " Country status" and "Form of government." To facilitate their inclusion in the linear regression model, these categorical variables were converted into dummy variables. This process involved creating n-1 dummy variables for each categorical variable, where n represents the number of categories.

In our case, "Country status" was converted into one dummy variable, named "IsDeveloped" ("Not Developed" as reference category). The "Form of government" variable was converted into three dummy variables, named "IsRepublic," "IsMonarchyConstitutional," and "IsAbsoluteMonarchy" ("Provisional" as reference category).

This approach ensures that the categorical variables are effectively incorporated into the model while maintaining the original information about the data.

4.4.2 Correlation Transformation of variables

Given the significant variation observed in the ranges of our attributes, a data standardization technique was employed to ensure all variables were placed on a common scale. We opted for correlation transformation, a method that transforms each variable to have a mean of 0 and a standard deviation of 1. This process effectively eliminates the influence of differing measurement scales and mitigates the impact of outliers on the regression analysis.

By standardizing the data, we create a more suitable environment for linear regression, where the coefficients can be interpreted directly in relation to the relative importance of each predictor variable. Both the Dependent and Independent variables have been transformed.

4.4.3 Making all Interaction and Quadratic terms for Polynomial regression

Centering of the continuous variables has been done, such that, there won't exist any collinearity between them and their quadratic terms when added to the model. Quadratic terms have been created seperately. Then all the possible second order interaction terms have been created for all the independent variables. After this, the quadratic terms have been added to the dataframe containing interaction terms.

The final dataframe consistes of the following terms :

```
'X2' · 'X3' · 'X4' · 'X5' · 'X6' · 'X7' · 'X8' · 'X9' · 'XA' · 'XC' · 'X2X3' · 'X2X4' · 'X2X5' · 'X2X6' · 'X2X7' · 'X2X8' · 'X2X9' · 'X2XA' · 'X2XC' · 'X3X4' · 'X3X5' · 'X3X6' · 'X3X7' · 'X3X8' · 'X3X9' · 'X3XA' · 'X3XC' · 'X4X5' · 'X4X6' · 'X4X7' · 'X4X8' · 'X4X9' · 'X4XA' · 'X4XC' · 'X5X6' · 'X5X7' · 'X5X8' · 'X5X9' · 'X5XA' · 'X5XC' · 'X6X7' · 'X6X8' · 'X6X9' · 'X6XA' · 'X6XC' · 'X7X8' · 'X7X9' · 'X7XA' · 'X7XC' · 'X8X9' · 'X8XA' · 'X8XC' · 'X9XA' · 'X9XC' · 'XAXC' · 'Y' · 'X2X2' · 'X3X3' · 'X4X4' · 'X5X5' · 'X6X6' · 'X7X7' · 'X8X8' · 'X9X9' · 'XAXA' · 'XCXC' · 'D1' · 'D2' · 'D3' · 'D4' · 'D1X2' · 'D1X3' · 'D1X4' · 'D1X5' · 'D1X6' · 'D1X7' · 'D1X8' · 'D1X9' · 'D1XA' · 'D1XC' · 'D2X2' · 'D2X3' · 'D2X4' · 'D2X5' · 'D2X6' · 'D2X7' · 'D2X8' · 'D2X9' · 'D2XA' · 'D2XC' · 'D3X2' · 'D3X3' · 'D3X4' · 'D3X5' · 'D3X6' · 'D3X7' · 'D3X8' · 'D3X9' · 'D3XA' · 'D3XC' · 'D4X2' · 'D4X3' · 'D4X4' · 'D4X5' · 'D4X6' · 'D4X7' · 'D4X8' · 'D4X9' · 'D4XA' · 'D4XC'
```

4.5 Model Selection

Automatic Search Procedure along with the Best Subset Selection method has been used to build the models. The automatic search procedure efficiently evaluated a vast number of variable combinations and shortlisted a promising set of initial predictors. The best subset selection method then, refined the model selection by evaluating all possible models containing only the variables identified from the output of automatic search method.

This two-step approach ensured a balance between efficiency and precision. The automatic search procedure provided a good starting point by quickly exploring a large number of variables, while the best subset selection method further refined the model by focusing on the most impactful predictors. This ultimately resulted in a more concise and interpretable model, reducing the risk of overfitting and potentially improving its overall accuracy.

The automatic search procedures used are:

- Stepwise Regression Method
- Forward Selection Method
- Backward Elimination Method

To implement the above mentioned methods, 3 preliminary models were created:

- Full model with all variables
- Base model with all essential variables discussed in 2.3.2
    - Fertility Rate (X3)
    - Median Age (X4)
    - Urban Population Percentage (X5)
    - Human Development Index (X6)
    - Health Care Index (X7)
    - Literacy rate (X8)
- Null model with no variables

With the help of these models and different automatic search methods 3 models have been built. Note that the build models are not the final models. Refinement of these models are done in further stages of the report but we will use the same names for the models after refinement at each stage.

4.5.1 Model 1 - Stepwise method + Best Subset Model Selection

Parameters given and the model selected after automatic search procedure is given below:

- Base model – 'Base model with all essential variables'
- Upper limit – 'Full model with all variables'
- Lower limit – 'Base model with all essential variables'
- Direction – 'Both'

Model 1 – (Y ~ X3 + X4 + X5 + X6 + X7 + X8 + XA + XAXA + X8X9 + X4X4 + X2XA + XAXC + X8XA + X7XC + D1XC + X5X9 + X5X6 + X3X5 + X3X3 + X5X5 + X3X4 + D2X5 + X5XA + D1 + D2X8 + X3XC + X5X7 + X4X7 + D1X5 + X7XA)

Model selected after Best subset selection method with 8 variables is given below:

Model 1 – (Y ~ X6 + XA + X^25 + X^2A + X3XC + X5X7 + X5X9 + X8X9)

4.5.2 Model 2 - Forward method + Best Subset Model Selection

Parameters given and the model selected after automatic search procedure is given below:

- Base model – 'Base model with all essential variables'
- Upper limit – 'Full model with all variables'
- Lower limit – 'Base model with all essential variables'

- Direction – ‘Forward’

Model 2 – (Y ~ X3 + X4 + X5 + X6 + X7 + X8 + XA + XAXA + X8X9 + X4X4 + X2XA + XAXC + D1X6 + X8XA + X7XC + D1XC + X5X9 + D3X3 + X5X6 + X3X5 + X3X3 + X5X5 + X3X4 + D2X5 + X5XA + D1 + D2X8 + D1XA + X4X7 + X5X7 + X3XC + D1X5 + X7XA + XC + D1X3)

Model selected after Best subset selection method with 8 variables is given below:

Model 2 – (Y ~ X6 + XA + XC + $X^2$5 + X^2A + X5X6 + X5XA + X8XA)

4.5.3 Model 3 - Backward method + Best Subset Model Selection

Parameters given and the model selected after automatic search procedure is given below:

- Base model – ‘Full model with essential variables’
- Direction – ‘Backward’

Model 3 – (Y ~ X2 + X4 + X5 + X6 + X2X5 + X2X7 + X2X8 + X3X4 + X3X6 + X3X7 + X4X8 + X4XC + X5X6 + X5X8 + X5XC + X6X8 + X6X9 + X7X9 + X7XC + X9XA + XAXC + X2X2 + X3X3 + X5X5 + X8X8 + XAXA + XCXC + D2 + D1X3 + D1X4 + D1X5 + D1X8 + D2X3 + D2X4 + D2X5 + D2X8 + D3X2)

Model selected after Best subset selection method with 8 variables is given below:

Model 3 – (Y ~ X2 + X6 + $X^2$5 + X^2A + X2X8 + X5X6 + X5XC + XAXC)

4.6 Investigate Curvature and interaction effects more fully

Our analysis included models containing interaction terms. It's important to note that for an interaction term (Ex: X5X6) to be statistically meaningful, both of its corresponding main effects (X5 and X6) should be included in the model as well. This ensures a proper interpretation of the interaction term.

The significance of each interaction term was carefully evaluated through a ggplot with the target variable along with confidence band. If an interaction term was found to significantly impact the model, we ensured the presence of its corresponding main effects. Conversely, non-significant interaction terms were removed from the final model to avoid potential issues of multicollinearity. This approach ensures a statistically robust model that focuses on the most relevant relationships between variables.

4.6.1 Model 1 Study

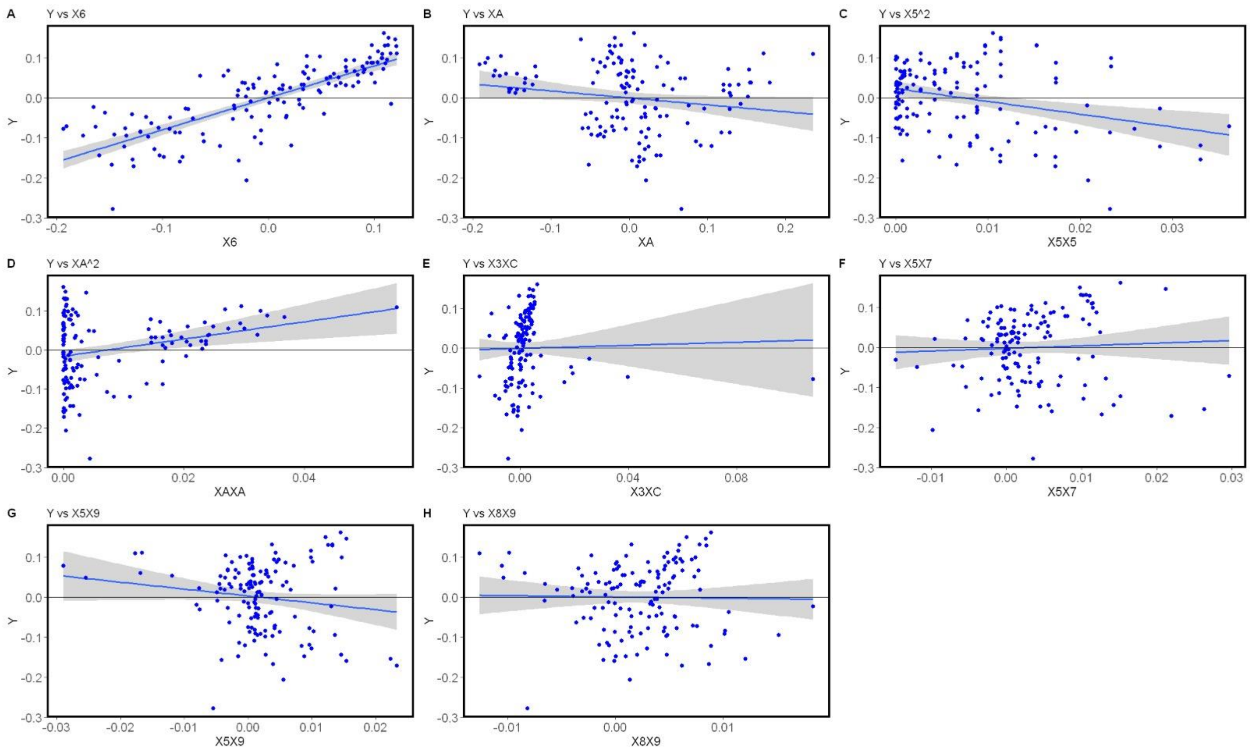

Figure 4.6-1: Model 1 - Curvature and Interaction effects on Y

- X3XC: The slope of the regression line is very low and the confidence band is also wide
- X5X7: Slope of regression line is very low the confidence band is a bit narrower
- **X5X9**: There is a significant negative slope for the regression line
- X8X9: No presence of slope for the regression line

Conclusion – Add X5, X9 and Remove X3XC, X5X7, X8X9 from the model

Model 1 – (Y ~ X5 + X6 + X9 + XA + X^{2}5 + X^{2}A + X5X9)

4.6.2 Model 2 Study

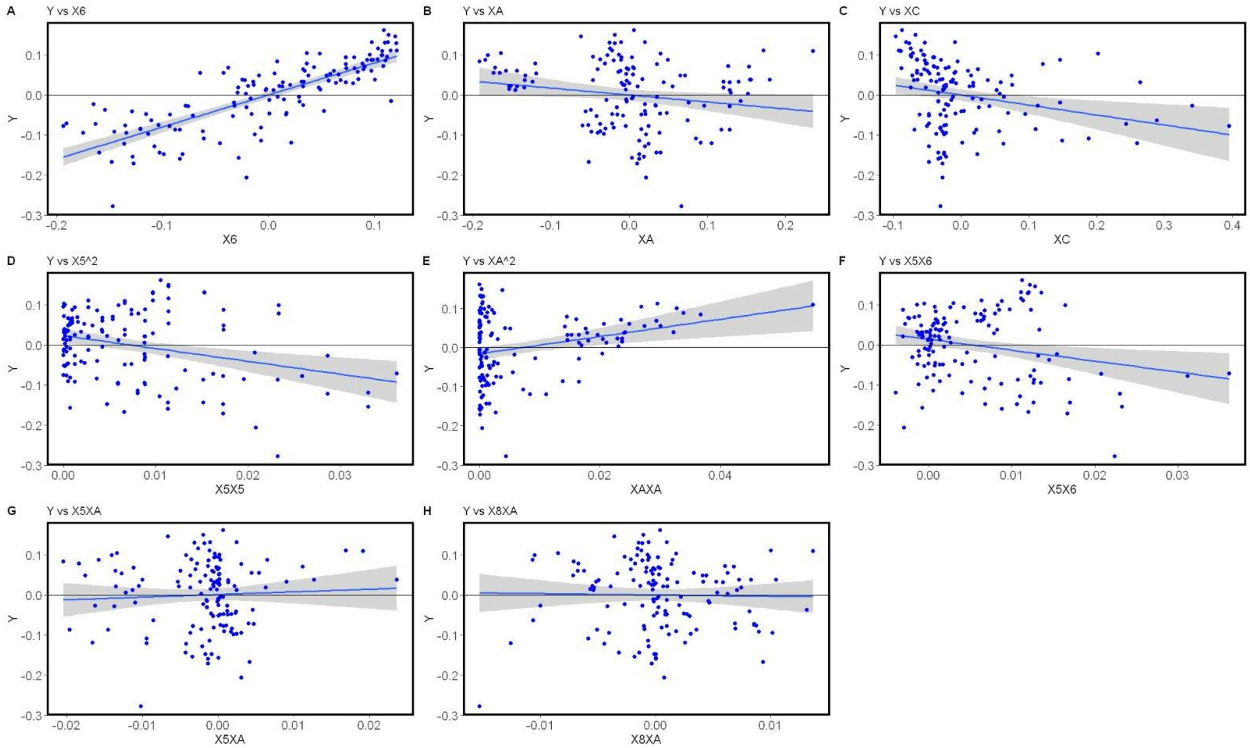

Figure 4.6-2: Model 2 - Curvature and Interaction effects on Y

- **X5X6**: There is a significant negative slope for the regression line
- X5XA: Slope of regression line is very low
- X8XA: No presence of slope for the regression line

Conclusion – Add X5 and Remove X5XA, X8XA from the model

Model 2 – (Y ~ X5 + X6 + XA + XC + X^25 + X^2A + X5X6)

4.6.3 Model 3 Study

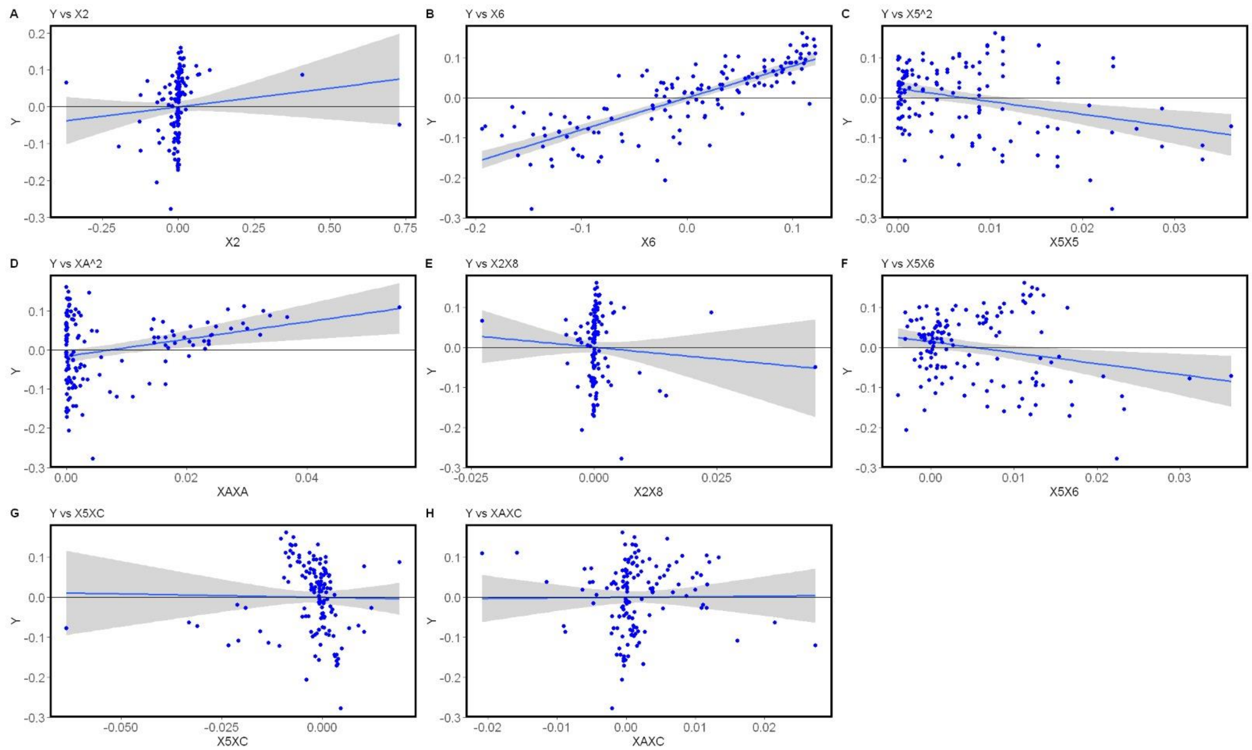

Figure 4.6-3: Model 3 - Curvature and Interaction effects on Y

- **XAXA**: There is a significant positive slope for the regression line
- **X2X8**: There is a significant negative slope for the regression line
- **X5X6**: There is a significant negative slope for the regression line
- X5XC: No presence of slope for the regression line
- XAXC: No presence of slope for the regression line

Conclusion – Add X5, X8, XA and Remove X5XC, XAXC from the model

Model 3 – (Y ~ X2 + X5 + X6 + X8 + XA + $X^2$5 + X^2A + X2X8 + X5X6)

4.7 Train-Test Split

Splitting the data into training and testing sets is a crucial step. This practice ensures our model generalizes well beyond the data used to train it. The training set allows the model to learn the underlying relationships, while the unseen testing set provides an honest assessment of how well the model performs on new data. This fosters a statistically rigorous evaluation of the model's generalizability, ensuring its ability to make accurate predictions beyond the training data.

To ensure the effectiveness of the train-test split and mitigate potential biases, we evaluated the similarity between the training and testing sets. This assessment involved comparing key metrics like the Mean Squared Prediction Error (MSPE) or Mean Squared Error (MSE) between the two sets. Ideally, the ratio of these values should be close to 1. Values close to 1 indicate that the training and testing sets share similar characteristics, suggesting a successful random split.

We performed the Train-Test split based on the model including all the linear continuous and dummy variables

Train/Test Model – (Y ~ X2+X3+X4+X5+X6+X7+X8+X9+XA+XC+D1+D2+D3+D4)

The ratio of MSPE and MSE turned out to be **1.00544** confirming a successful random split. Carrying on from now, we will be using train dataset for further model refinement.

4.8 Multicollinearity Check

The Variance Inflation Factor (VIF) plays a critical role in diagnosing multicollinearity. Multicollinearity can inflate variances of estimated coefficients, making them appear less reliable and hindering the interpretation of individual variable effects. VIF is calculated for each variable:

- VIF < 5 indicates minimal multicollinearity
- Mean of VIF < 3 indicates minimal multicollinearity across independent variables as a whole

By analyzing VIF values, we can identify variables contributing significantly to multicollinearity and take corrective measures by removing redundant variables. This ensures the model's coefficients are statistically sound and interpretations are accurate.

4.8.1 VIF test for Model 1

VIF1						
X5	X6	X9	XA	X^25	X^2A	X5X9
2.681041	3.416722	1.789780	1.192934	1.259176	1.390519	1.284611

Table 2: VIF test - Model 1

VIF for each individual term is less than 5 (highest is 3.41). Mean VIF of all terms is 1.86, which is not considerably high. So, we can conclude that Model 1 is free from any form of high multicollinearity between terms. No change to Model 1.

Model 1 – (Y ~ X5 + X6 + X9 + XA + X^25 + X^2A + X5X9)

4.8.2 VIF test for Model 2

VIF2						
X5	X6	XA	XC	X^25	X^2A	X5X6
2.686252	2.698322	1.217935	1.159395	2.946631	1.067990	3.118618

Table 3: VIF test - Model 2

For model 2, VIF for each individual term is less than 5 (highest is 3.12). Mean VIF of all terms is 2.13, which is not considerably high. So, we can conclude that Model 1 is free from any form of high multicollinearity between terms. No change to Model 2.

Model 2 – (Y ~ X5 + X6 + XA + XC + X^25 + X^2A + X5X6)

4.8.3 VIF test for Model 3

VIF3_i								
X2	X5	X6	X8	XA	X^25	X^2A	X2X8	X5X6
11.17372	2.69354	5.62912	5.33146	1.24125	3.51621	1.13276	10.82162	4.80331

Table 4: VIF test - Model 3

VIF for few individual terms are more than 10. Also, mean VIF of all terms is 5.15 which is more than to 3. To handle this situation, term with highest VIF i.e. X2 should be removed from model. Also, as linear term X2 has been removed from model, so interaction term containing X2, i.e. X2X8, should also be removed. As model 3 has been modified, VIF test should be performed again to validate, if there is any serious multicollinearity exists between terms, in modified model 3.

VIF3_i						
X5	X6	X8	XA	X^25	X^2A	X5X6
2.685372	5.610383	5.246337	1.205252	3.342227	1.054928	4.462977

Table 5: VIF test 2 - Model 3

Even after modification for model 3, VIF for few individual terms are more than 5. Also, mean VIF of all terms is 3.37 which is more than 3. To handle this situation, term with highest VIF i.e. X6 should be removed from model. Also, as linear term has been removed from model, so interaction term containing X6, i.e. X5X6, should also be removed. As model 3 has been modified futher, VIF test should be performed again to validate, if there is any serious multicollinearity exists between terms, in modified model 3.

VIF3_i				
X5	X8	XA	X^25	X^2A
1.851003	1.902165	1.171483	1.209580	1.051992

Table 6: VIF test 3 - Model 3

With modified model 3, VIF for each individual term is less than 5 (highest is 1.90). Mean VIF of all terms is 1.44 (<3), which is not considerably high. So, we can conclude that current Model 3 is free from any form of high multicollinearity between terms. After modification final model 3 is as follows:

Model 3 – (Y ~ X5 + X8 + XA + $X^2$5 + X^2A)

4.9 Outlier Study and Influential Cases

4.9.1 Methodology for Identifying the outliers and Influential cases

4.9.1.1 Identifying Outlying Y observations

Test for outlying Y observations was done by calculating the studentized deleted residuals (t_i). Firstly, the residuals were calculated for each model (e_i). Here, n is the number of Observations and p is equal to number of predictor variables + 1. The Deleted Residuals were then computed using the below equation:

$$d_i = \frac{e_i}{1 - h_{ii}}$$

where, h_{ii} = hat matrix for the model

The Studentized Deleted Residuals were then calculated:

$$t_i = \frac{d_i}{s(d_i)}$$

Where, $s^2(d_i) = \frac{n-p}{n-p-1} MSE(1 - h_{ii}) - \frac{e_i^2}{n-p-1}$

n = number of observations
p = number of prdictor variables + 1

- The ith case is an outlying Y observation if $\boldsymbol{|t_i| \geq t\left(1 - \frac{\alpha}{2n}; n - p - 1\right)}$

4.9.1.2 Identifying Outlying X observations

The ith case is an outlying X observation if $\boldsymbol{h_{ii} > \frac{2p}{n}}$

4.9.1.3 Identifying Influential Cases

We identified the influential cases based on DFFITS (Studentized Deleted Fit), DFBETAS (Change in Beta) and Cook's Distance.

DFFITS considers a point's leverage (how much it pulls on the regression line) and the change in the predicted value for that point when excluded from the model fit. Large DFFITS values (positive or negative) indicate influential points with high leverage that cause a substantial, unexpected change in the fit when removed, suggesting they might be outliers or have an undue influence on the model.

DFBETAS calculates the difference in a coefficient estimate when a specific point is excluded from the model fit. Large DFBETAS values (positive or negative) for a specific coefficient highlight points that significantly alter the estimate of that particular coefficient when removed. These points might be outliers or have an unexpected relationship with the variable represented by that coefficient.

Cook's distance considers both a point's leverage and the magnitude of the change in fitted values when the point is excluded. Large Cook's distance values indicate influential points with high leverage that cause a substantial change in the overall model fit when removed, suggesting they might be outliers or have unexpected relationships with the dependent variable. Analyzing Cook's distance alongside DFFITS and DFBETAS provides a more comprehensive understanding of influential points in your model.

Model plots, Cooks D Bar plot and DFBETAS plots were computed and the influential cases were identified for each model.

4.9.2 Model 1

4.9.2.1 Outliers

Outlying Y observations are:

- Afghanistan, Lebanon

Outlying X observations are:

- Japan, Canada, Lesotho, Niger, Chile, Mauritius, Malawi, Argentina, New Zealand

4.9.2.2 Influential Cases

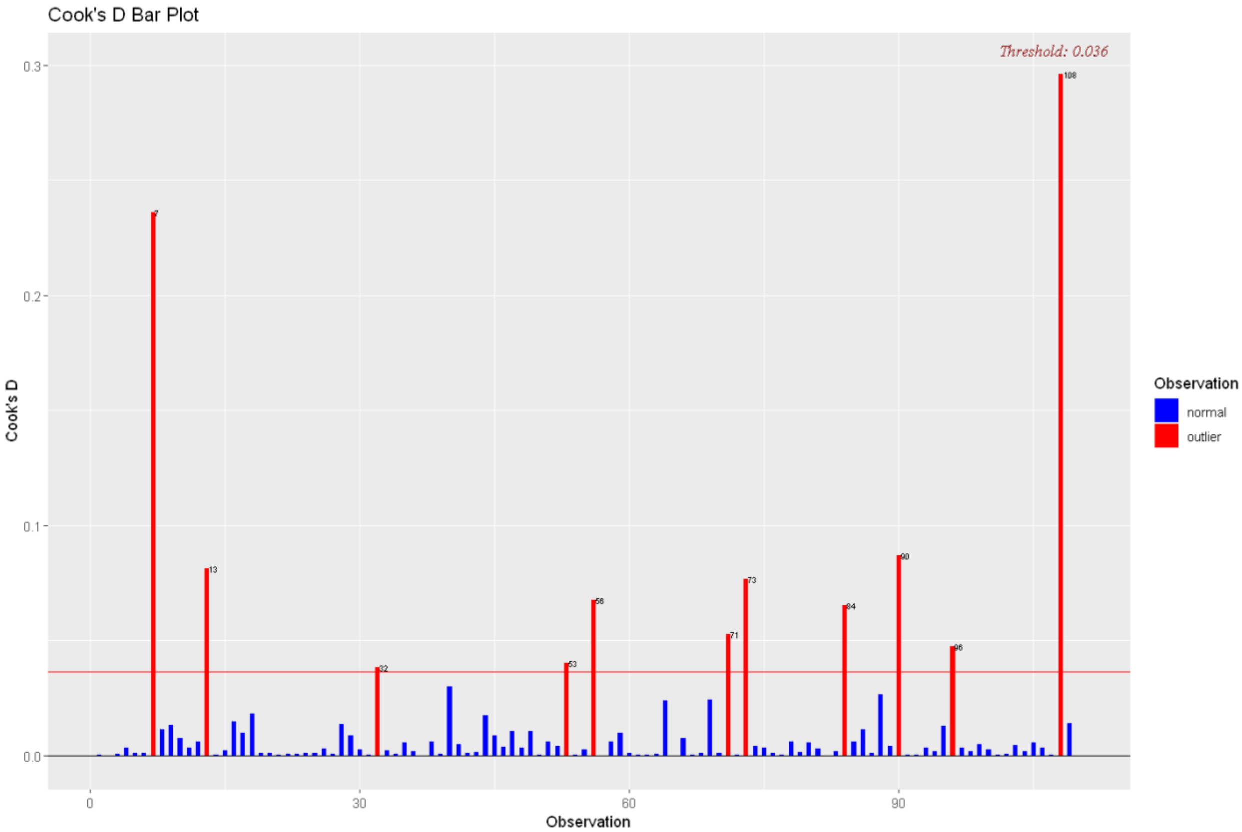

Figure 4.9-1: Model 1 - Cook's D Bar plot

Note: DFFITS and DFBETAS plots for model 1 are attached in appendix

4.9.2.3Modification

Following are the observations that are Outlying as well as Influential:

- Afghanistan, Japan, Chile, Mauritius, Malawi, Argentina, New Zealand and Lebanon

All the above mentioned observations are removed from dataset for Model 1

4.9.3 Model 2

4.9.3.1 Outliers

Outlying Y observations are:

- Afghanistan

Outlying X observations are:

- Japan, Nepal, Bahrain, Chad, Niger, Tajikistan, New Zealand, Iraq, Lebanon

4.9.3.2 Influential Cases

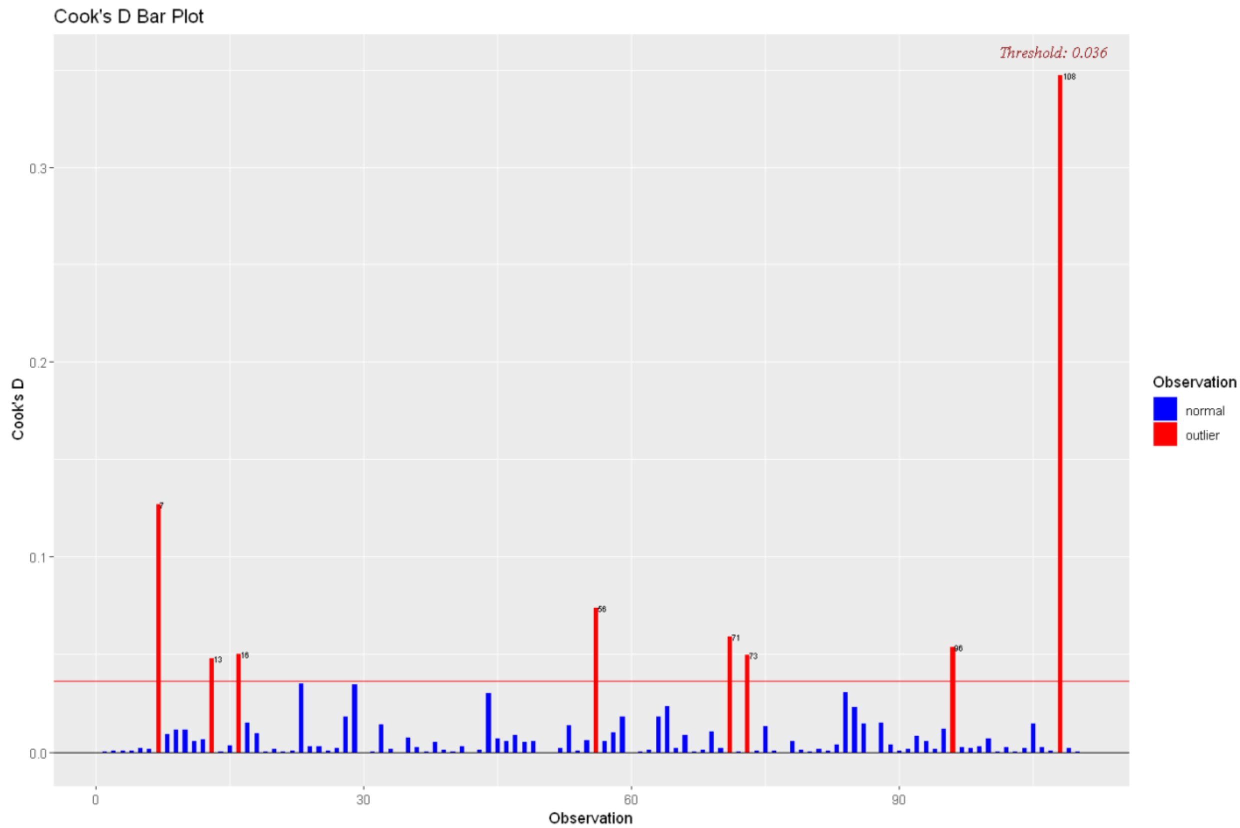

Figure 4.9-2: Model 2 - Cook's D bar plot

Note: DFFITS and DFBETAS plots for model 2 are attached in appendix

4.9.3.3 Modification

Following are the observations that are Outlying as well as Influential:

- Afghanistan, Japan, Chad, New Zealand, Iraq and Lebanon

All the above mentioned observations are removed from dataset for Model 2

4.9.4 Model 3

4.9.4.1 Outliers

Outlying Y observations are:

- Lebanon

Outlying X observations are:

- Japan, Belgium, Niger, Malawi, New Zealand, United States

4.9.4.2 Influential Cases

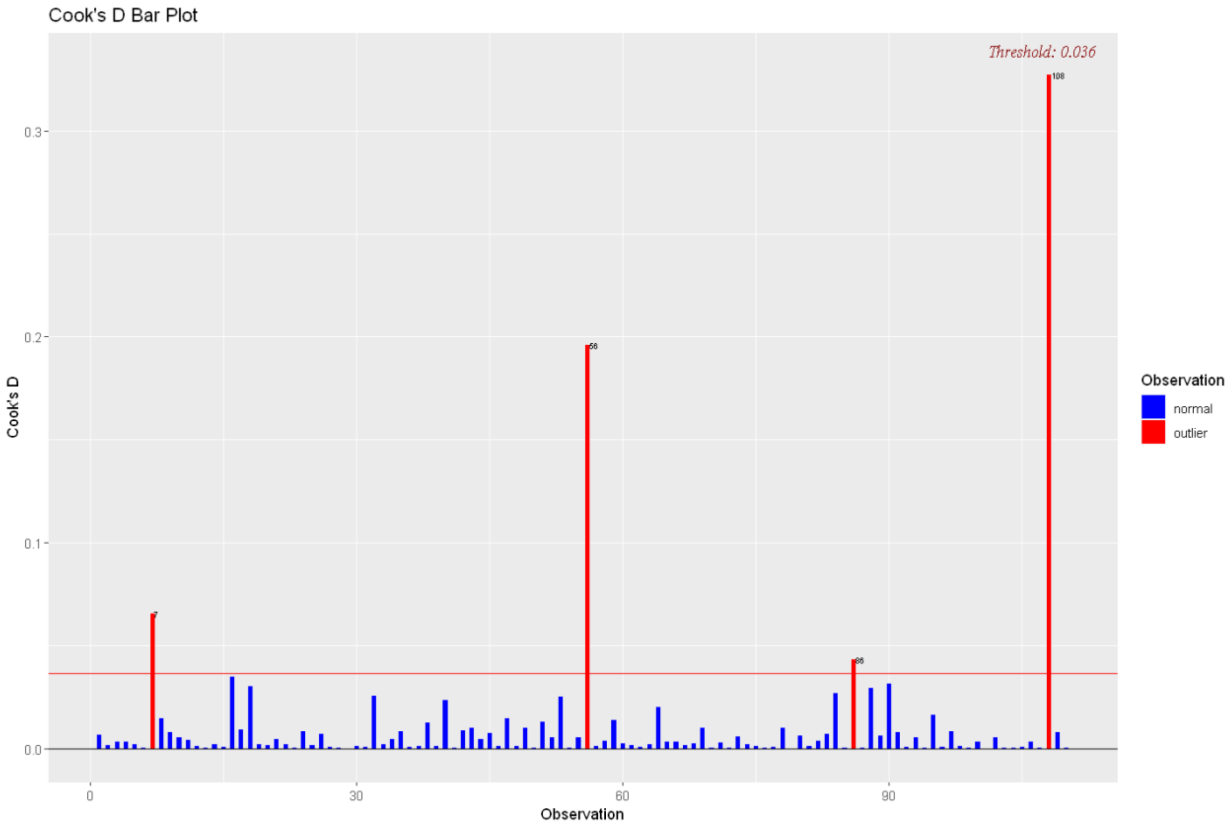

Figure 4.9-3: Model 3 - Cook's D bar Plot

Note: DFFITS and DFBETAS plots for model 3 are attached in appendix

4.9.4.3 Modification

Following are the observations that are Outlying as well as Influential:

- Lebanon, Japan, Belgium, Niger, Malawi, New Zealand and United States

All the above mentioned observations are removed from dataset for Model 3

4.10 Model Preparation and Analysis

4.10.1 Model 1

From summary table following important points are observed:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.009833	0.006621	-1.485	0.14086	
X5	0.036313	0.086783	0.418	0.67658	
X6	0.578070	0.091886	6.291	9.9e-09	***
X9	0.261670	0.078338	3.340	0.00120	**
XA	-0.170466	0.058158	-2.931	0.00424	**
X5X5	0.717447	0.648554	1.106	0.27145	
XAXA	1.414978	0.482706	2.931	0.00424	**
X5X9	-0.196610	0.813128	-0.242	0.80947	

Figure 4.10-1: Summary of model 1

- Overall model fitment is high with F-statistic 44.21, and p-value near 0
- Attributes X6, X9, XA, X^2A have p-value < α
- Confidence for estimation of coefficient for terms X5,X^25 and X5X9 is very low

4.10.1.1 Test for existence of regression relation in model

Hypothesis:

- Null Hypothesis (H_0): $\beta_5, \beta_6, \beta_9, \beta_A, \beta_{55}, \beta_{AA}, \beta_{59} = 0$
- Alternate Hypothesis (H_A): Not all β_i are zero

Test method: F statistic, α = 0.05

Analyse sample data: F-value = 2.19, F-statistic = 44.21, p-value ≈ 0

Decision Rule: If F-statistic <= F-value, conclude H_0, otherwise conclude H_A.

Result: As F-statistic > F-value, conclude H_A

Conclusion: Not all β_i are zero and therefore there exist a relation between dependent and independent variables

4.10.1.2 "Extra Sum of Square" test for attributes with low confidence

Hypothesis:

- Null Hypothesis (H_0): $\beta_5, \beta_{55}, \beta_{59} = 0$
- Alternate Hypothesis (H_A): Not all β_i are zero

Test method: F statistic, α = 0.05

Analyse sample data: F-value = 2.70, F-statistic = 0.45, p-value = 0.72

Decision Rule: If F-statistic <= F-value, conclude H_0, otherwise conclude H_A.

Result: As F-statistic < F-value, conclude H_0

Conclusion: All $\beta_5, \beta_{55}, \beta_{59}$ are zero and therefore all three terms should be dropped from the model

4.10.1.3 Modified Model 1

Model 1 – (Y ~ X6 + X9 + XA + X^2A)

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.006092   0.005225  -1.166 0.246490
X6           0.578168   0.063992   9.035 1.63e-14 ***
X9           0.269953   0.076409   3.533 0.000631 ***
XA          -0.170348   0.051965  -3.278 0.001451 **
XAXA         1.440433   0.461253   3.123 0.002361 **
```

Figure 4.10-2: Summary of Modified model 1

4.10.2 Model 2

From summary table following important points are observed:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.005832   0.006583  -0.886  0.37788
X5           0.008026   0.085766   0.094  0.92564
X6           0.812250   0.076783  10.579  < 2e-16 ***
XA          -0.175628   0.060570  -2.900  0.00464 **
XC           0.112056   0.068410   1.638  0.10473
X5X5        -1.467241   0.998060  -1.470  0.14484
XAXA         1.245692   0.465503   2.676  0.00878 **
X5X6         2.564762   1.146013   2.238  0.02756 *
```

Figure 4.10-3: Summary of Model 2

- Overall model fitment is high with F-statistic 42.78, and p-value near 0
- Attributes X6, XA, X^2A and X5X6 have p-value < α
- Confidence for estimation of coefficient for terms X5, XC and $X^2$5 is very low

4.10.2.1 Test for existence of regression relation in model

- Null Hypothesis (H_0): $\beta_5, \beta_6, \beta_A, \beta_C, \beta_{55}, \beta_{AA}, \beta_{56} = 0$
- Alternate Hypothesis (H_A): Not all β_i are zero

Test method: F statistic, $\alpha = 0.05$
Analyse sample data: F-value = 2.11, F-statistic = 42.78, p-value ≈ 0
Decision Rule: If F-statistic <= F-value, conclude H_0, otherwise conclude H_A.
Result: As F-statistic > F-value, conclude H_A
Conclusion: Not all β_i are zero and therefore there exist a relation between dependent and independent variables

4.10.2.2 "Extra Sum of Square" test for attributes with low confidence

- Null Hypothesis (H_0): $\beta_5, \beta_C, \beta_{55} = 0$
- Alternate Hypothesis (H_A): Not all β_i are zero

Test method: F statistic, $\alpha = 0.05$
Analyse sample data: F-value = 2.70, F-statistic = 1.49, p-value = 0.23

Decision Rule: If F-statistic <= F-value, conclude H_0, otherwise conclude H_A.
Result: As F-statistic < F-value, conclude H_0
Conclusion: All $\beta_5, \beta_C, \beta_{55}$ are zero and therefore all three terms should be dropped from the model

Model 2 – (Y ~ X6 + XA + X^2A + X5X6)

4.10.2.3 "Extra Sum of Square" test for extra interaction term

- Null Hypothesis (H_0): $\beta_{56} = 0$
- Alternate Hypothesis (H_A): $\beta_6 \neq 0$

Test method: F statistic, $\alpha = 0.05$
Analyse sample data: F-value = 3.94, F-statistic = 2.64, p-value = 0.11
Decision Rule: If F-statistic <= F-value, conclude H_0, otherwise conclude H_A.
Result: As F-statistic < F-value, conclude H_0
Conclusion: β_{56} is zero and it should be dropped from the model

4.10.2.4 Modified Model 2

Model 2 – (Y ~ X6 + XA + X^2A)

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.001117   0.005122  -0.218  0.82788
X6           0.761909   0.047296  16.110  < 2e-16 ***
XA          -0.169318   0.053958  -3.138  0.00224 **
XAXA         0.955372   0.458267   2.085  0.03967 *
```

Figure 4.10-4: Summary of Modified model 2

4.10.3 Model 3

From summary table following important points are observed:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.002003   0.008871   0.226    0.822
X5           0.407978   0.098297   4.150 7.13e-05 ***
X8           0.418867   0.087689   4.777 6.32e-06 ***
XA          -0.087889   0.081496  -1.078    0.284
X5X5        -0.126847   0.899085  -0.141    0.888
XAXA         0.677597   0.643198   1.053    0.295
```

Figure 4.10-5: Summary of Model 3

- Overall model fitment is high with F-statistic 25.47, and p-value near 0
- R^2 for model is 0.54, which is very less than other two models
- Attributes X5, X8 have p-value < α
- Confidence for estimation of coefficient for terms XA,$X^2$5 and X^2A is very low

4.10.3.1 Test for existence of regression relation in model

- Null Hypothesis (H_0): $\beta_5, \beta_8, \beta_A, \beta_{55}, \beta_{AA} = 0$
- Alternate Hypothesis (H_A): Not all β_i are zero

Test method: F statistic, $\alpha = 0.05$
Analyse sample data: F-value = 2.11, F-statistic = 17.82, p-value ≈ 0
Decision Rule: If F-statistic <= F-value, conclude H_0, otherwise conclude H_A.
Result: As F-statistic > F-value, conclude H_A
Conclusion: Not all β_i are zero and therefore there exist a relation between dependent and independent variables

4.10.3.2 "Extra Sum of Square" test for attributes with low confidence

- Null Hypothesis (H_0): $\beta_A, \beta_{55}, \beta_{AA} = 0$
- Alternate Hypothesis (H_A): Not all β_i are zero

Test method: F statistic, $\alpha = 0.05$
Analyse sample data: F-value = 2.70, F-statistic = 0.97, p-value = 0.41
Decision Rule: If F-statistic <= F-value, conclude H_0, otherwise conclude H_A.
Result: As F-statistic < F-value, conclude H_0
Conclusion: All $\beta_A, \beta_{55}, \beta_{AA}$ are zero and therefore all three terms should be dropped from the model

4.10.3.3 Modified Model 3

Model 3 – (Y ~ X5 + X8)

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.005986   0.005686   1.053    0.295
X5          0.460278   0.087727   5.247 8.69e-07 ***
X8          0.399451   0.080451   4.965 2.83e-06 ***
```

Figure 4.10-6: Summary of Modified model 3

4.11 Residual Analysis and other diagnostic studies

4.11.1 Model 1

4.11.1.1 Linearity assumption validation

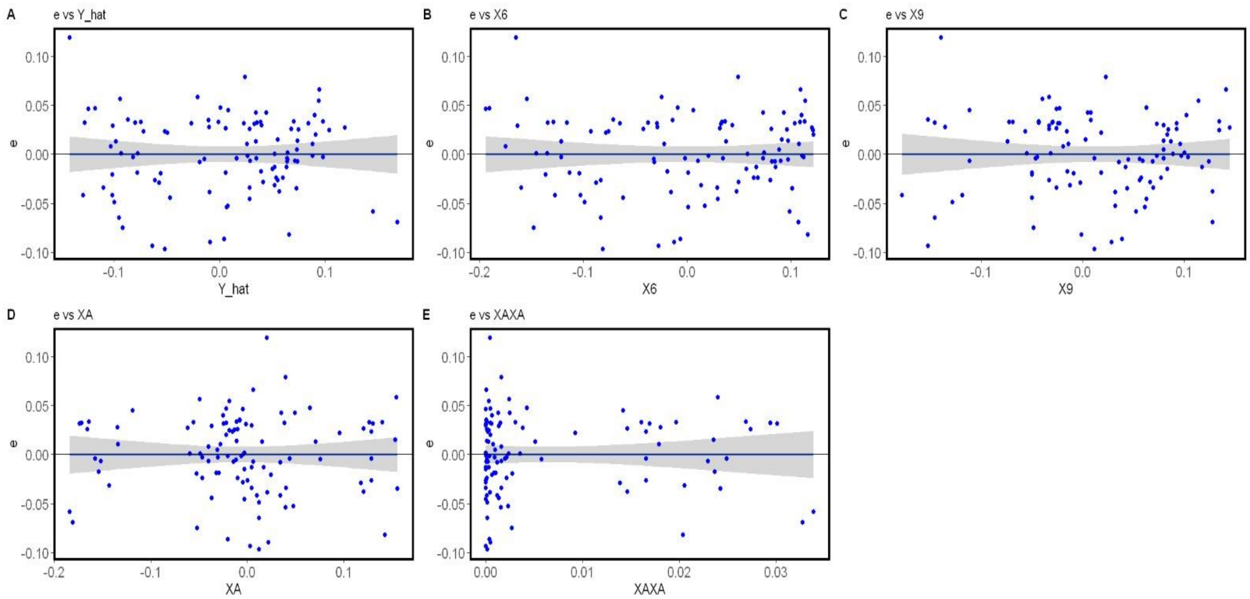

Figure 4.11-1: Residual plots - Model 1

The residuals are randomly scatterd around the horizontal line at zero and there is no pattern. This indicates that there is no consistent pattern in errors.

4.11.1.2 Normality Test

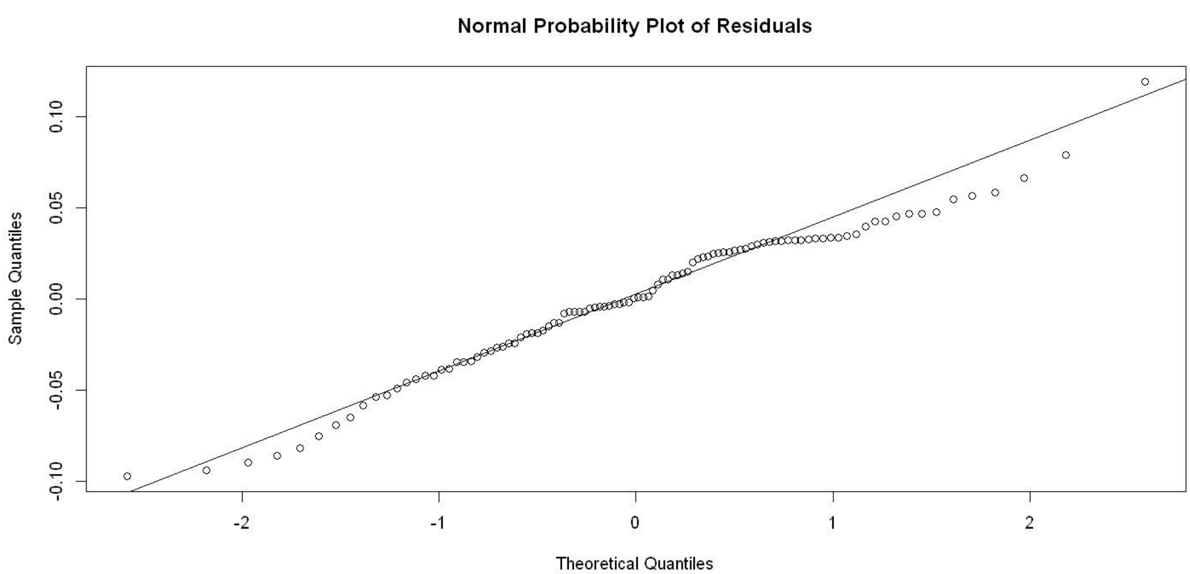

Figure 4.11-2: QQ Plot - Model 1

No heavy tails present, which indicates that the error terms are normally distributed. This has been verified using the correlation test.

Hypothesis:

- Null Hypothesis (H_0): The null hypothesis states that residuals follow normal probability distribution
- Alternate Hypothesis (H_A): The alternative hypothesis states that residuals does not follow normal probability distribution

Test method: Coeffiecient of Correlation
Analyse sample data: r-critical = 0.987, Coefficient Correlation = 0.988
Decision Rule: If |Coefficient Correlation| >= r-critical, conclude H_0, or else conclude H_A
Result: As |Coefficient Correlation| > r-critical, conclude H_0
Conclusion: Residuals follow a normal probability distribution

4.11.1.3 Homoscedasticity Validation (Breusch-Pagan Test)

Since, the normality assumption is correct, we can use the Breusch-Pagan Test to verify if the error terms have constant variance
Hypothesis:

- Null Hypothesis (H_0): $\sigma^2(\varepsilon_i) = \sigma^2$ (Equal variance)
- Alternate Hypothesis (H_A): $\sigma^2(\varepsilon_i) \neq \sigma^2$ (Unequal variance)

Analyse sample data: p-value = 0.17, $\alpha = 0.05$
Decision Rule: If p-value >= α, conclude H_0, or else conclude H_A
Result: As p-value >= α, conclude H_0
Conclusion: Residuals have constant variance

4.11.1.4 Independence Assumption

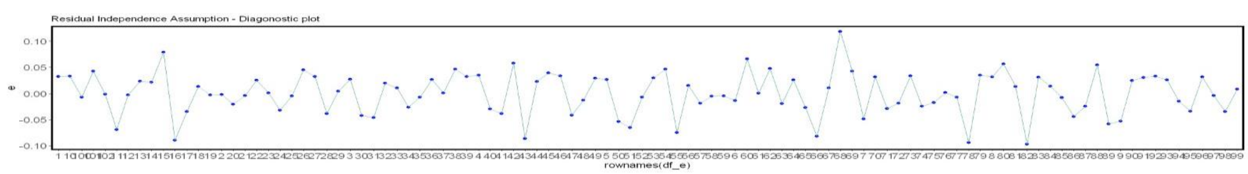

Figure 4.11-3: Sequence plot of residuals - Model 1

The residuals are fluctuating in a more or less random pattern around zero. This validates the Independence assumption.

4.11.2 Model 2

4.11.2.1 Linearity assumption validation

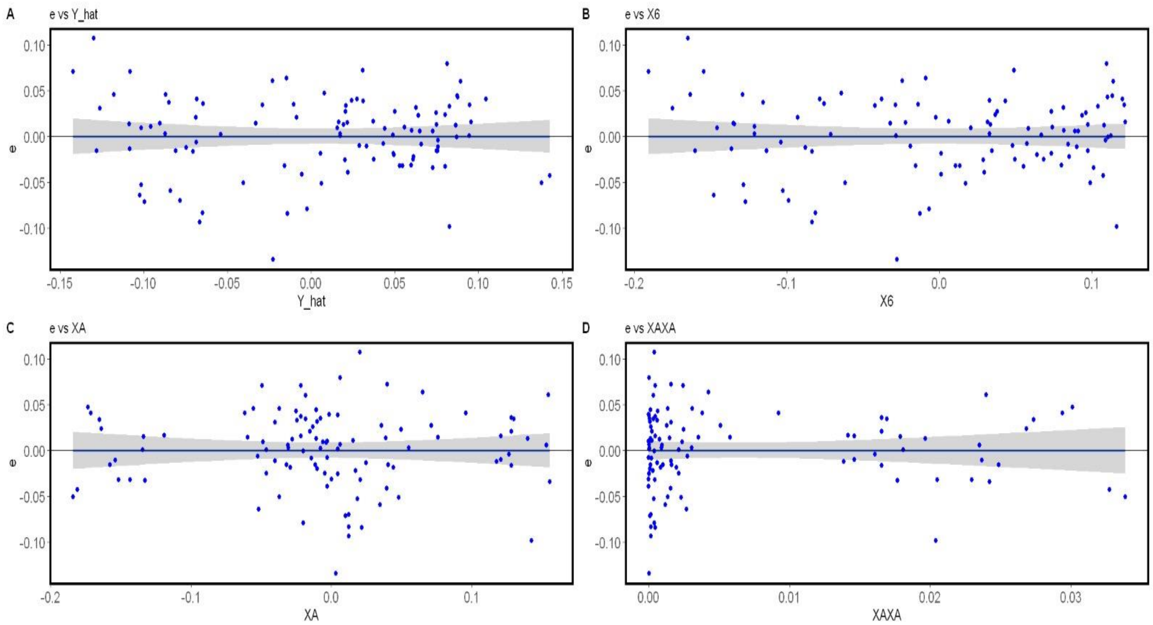

Figure 4.11-4: Residual Plots - Model 2

The residuals are randomly scatterd around the horizontal line at zero and there is no pattern. This indicates that there is no consistent pattern in errors.

4.11.2.2 Normality Test

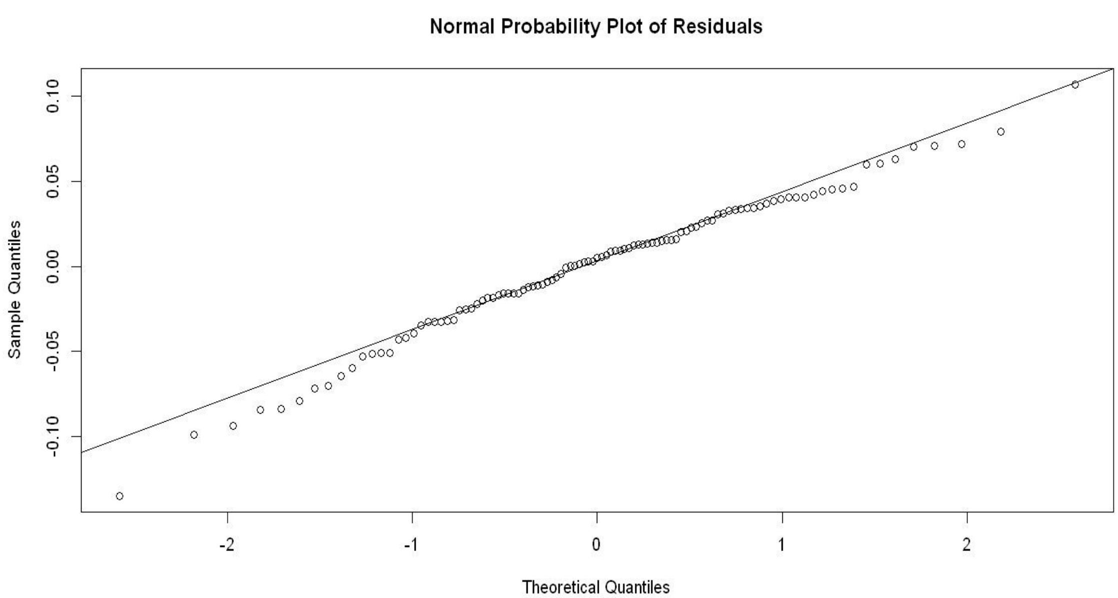

Figure 4.11-5: QQ Plot - Model 2

No heavy tails present, which indicates that the error terms are normally distributed. This has been verified using the correlation test.

Hypothesis:
- Null Hypothesis (H_0): The null hypothesis states that residuals follow normal probability distribution
- Alternate Hypothesis (H_A): The alternative hypothesis states that residuals does not follow normal probability distribution

Test method: Coeffiecient of Correlation
Analyse sample data: r-critical = 0.987, Coefficient Correlation = 0.991
Decision Rule: If |Coefficient Correlation| >= r-critical, conclude H_0, or else conclude H_A
Result: As |Coefficient Correlation| > r-critical, conclude H_0
Conclusion: Residuals follow a normal probability distribution

4.11.2.3 Homoscedasticity Validation (Breusch-Pagan Test)

Since, the normality assumption is correct, we can use the Breusch-Pagan Test to verify if the error terms have constant variance
Hypothesis:
- Null Hypothesis (H_0): $\sigma^2(\varepsilon_i) = \sigma^2$ (Equal variance)
- Alternate Hypothesis (H_A): $\sigma^2(\varepsilon_i) \neq \sigma^2$ (Unequal variance)

Analyse sample data: p-value = 0.085, α = 0.05
Decision Rule: If p-value >= α, conclude H_0, or else conclude H_A
Result: As p-value >= α, conclude H_0
Conclusion: Residuals have constant variance

4.11.2.4 Independence Assumption

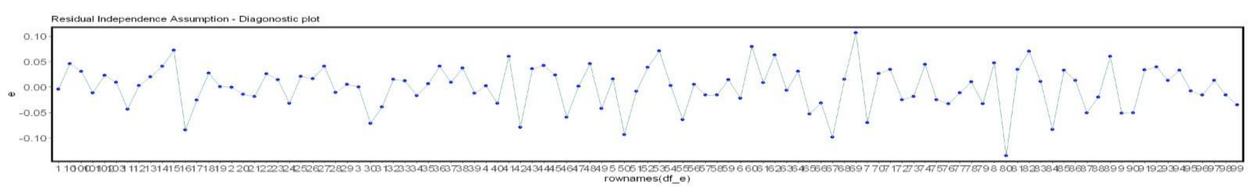

Figure 4.11-6: Sequence plot of residuals - Model 2

The residuals are fluctuating in a more or less random pattern around zero. This validates the Independence assumption.

4.11.3 Model 3

4.11.3.1 Linearity assumption validation

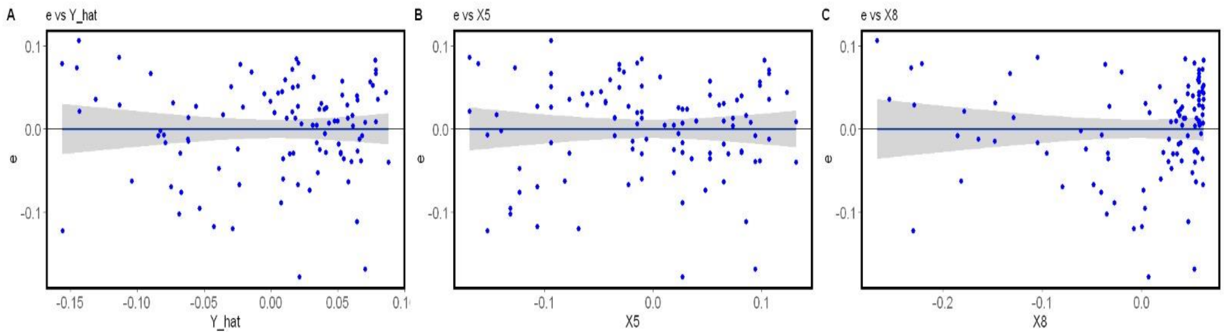

Figure 4.11-7: Residual Plots - Model 3

The residuals are randomly scatterd around the horizontal line at zero and there is no pattern. This indicates that there is no consistent pattern in errors.

4.11.3.2 Normality Test

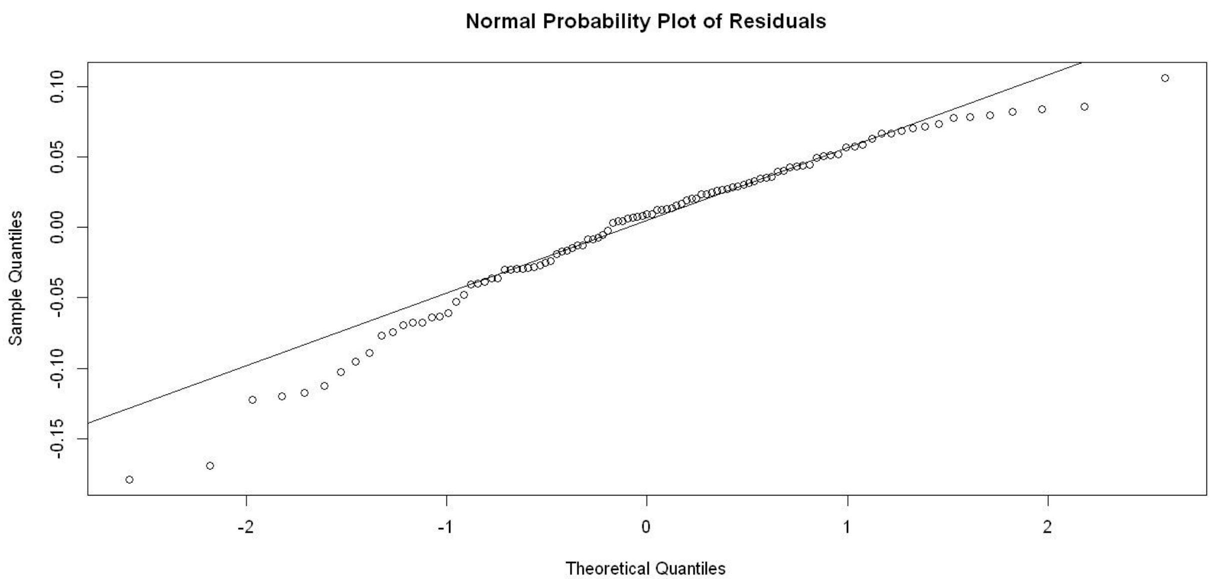

Figure 4.11-8: QQ plot - Model 3

Mild tails present, which raises a doubt whether the residuals are normally distributed or not. This has been verified using the correlation test.

<u>Hypothesis</u>:

- Null Hypothesis (H_0): The null hypothesis states that residuals follow normal probability distribution
- Alternate Hypothesis (H_A): The alternative hypothesis states that residuals does not follow normal probability distribution

Test method: Coeffiecient of Correlation
Analyse sample data: r-critical = 0.987, Coefficient Correlation = 0.980
Decision Rule: If |Coefficient Correlation| >= r-critical, conclude H_0, or else conclude H_A
Result: As |Coefficient Correlation| < r-critical, conclude H_0
Conclusion: Residuals follow a normal probability distribution

4.11.3.3 Homoscedasticity Validation (Breusch-Pagan Test)

Since, the normality assumption is correct, we can use the Breusch-Pagan Test to verify if the error terms have constant variance

Hypothesis:

- Null Hypothesis (H_0): $\sigma^2(\varepsilon_i) = \sigma^2$ (Equal variance)
- Alternate Hypothesis (H_A): $\sigma^2(\varepsilon_i) \neq \sigma^2$ (Unequal variance)

Analyse sample data: p-value = 0.33, $\alpha = 0.05$
Decision Rule: If p-value >= α, conclude H_0, or else conclude H_A
Result: As p-value >= α, conclude H_0
Conclusion: Residuals have constant variance

4.11.3.4 Independence Assumption

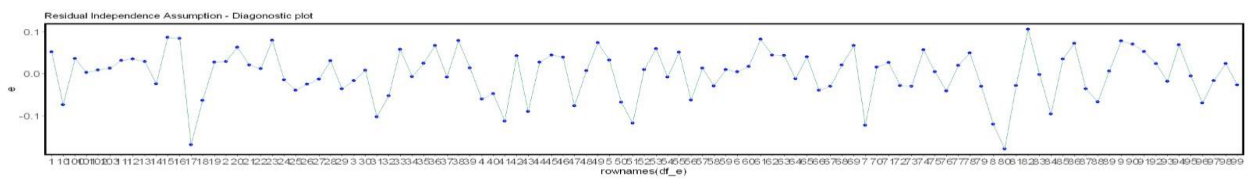

Figure 4.11-9: Sequence plot of residuals - Model 3

The residuals are fluctuating in a more or less random pattern around zero. This validates the Independence assumption.

4.11.4 Conclusion

Model 3 fails for Normality Assumption. Also, R^2 of Model 2 is only 0.54. On the other hand, both Model 1 and Model 2 follows all the assumptions for residuals. Model 1 and Model 2 will be used for model validation and Model 3 has been dropped.

4.12 Model Validation

Mean square Prediction error (MSPR) of the test data and Mean square error (MSE) of the train data were calculated for both the models to check the predictability of the models with unseen data.

4.12.1 Model 1

```
Residual standard error: 0.04747 on 22 degrees of freedom
Multiple R-squared:  0.6657,	Adjusted R-squared:  0.6049
F-statistic: 10.95 on 4 and 22 DF,  p-value: 4.85e-05
```

```
Residual standard error: 0.04076 on 97 degrees of freedom
Multiple R-squared:  0.7637,	Adjusted R-squared:  0.7539
F-statistic: 78.37 on 4 and 97 DF,  p-value: < 2.2e-16
```

Figure 4.12-1: Summary of Test data (left) and Train data (right)

$$\frac{MSPR}{MSE} = \frac{0.04747^2}{0.04076^2} = 1.356$$

4.12.2 Model 2

```
Residual standard error: 0.04653 on 23 degrees of freedom
Multiple R-squared:  0.6643,	Adjusted R-squared:  0.6205
F-statistic: 15.17 on 3 and 23 DF,  p-value: 1.162e-05
```

```
Residual standard error: 0.04281 on 99 degrees of freedom
Multiple R-squared:  0.7413,	Adjusted R-squared:  0.7334
F-statistic: 94.55 on 3 and 99 DF,  p-value: < 2.2e-16
```

Figure 4.12-2: Summary of Test data (left) and Train data (right)

$$\frac{MSPR}{MSE} = \frac{0.04653^2}{0.04281^2} = 1.181$$

4.13 Final Model

Although, Model 1 performed best on train dataset, however it shows more variation when validated against Test dataset. So, **Model 2** is choosen as best possible model.

Final Model = Model 2 – (Y ~ X6 + XA + X^2A)

4.13.1 Final model confidence intervals

The following are the confidence intervals for the final model i.e, Model 2.

	b_L	b_U
(Intercept)	-0.01128085	0.009047433
X6	0.66806402	0.855753507
XA	-0.27638224	-0.062253851
XAXA	0.04607072	1.864672957

Table 7: Confidence intervals of Final Model

4.13.2 Final model in form of Original variables

We have standardized our independent variables before fitting the model. Now, the model has been transformed back into the form of original variables (reversing the transformation) as below:

- $b_6' = \left(\frac{s_Y}{s_6}\right) b_6$
- $b_A' = \left(\frac{s_Y}{s_A}\right) b_A - \frac{2\bar{X}_A}{s_{A\sqrt{n-1}}} b_{AA}$
- $b_{AA}' = \left(\frac{s_Y}{s^2_{A\sqrt{n-1}}}\right) b_{AA}$
- $b_0' = \bar{Y} + b_0 - \left(\frac{s_Y}{s_6}\right) \bar{X}_6 b_6 - \left(\frac{s_Y}{s_A}\right) \bar{X}_A b_A + \left(\frac{s_Y}{s_A^2\sqrt{n-1}}\right) \bar{X}^2_{AA} b_{AA}$

The calculated final regression coefficients are:

$\mathbf{b_6}$ = 5.8048, $\mathbf{b_A}$ = -0.00518, $\mathbf{b_{AA}}$ = 3.6209, $\mathbf{b_0}$ = 1.32134

4.13.3 Final Estimated Regression Function

$$\hat{Y} = 1.32134 + 5.8048X_6 - 0.00518X_A + 3.6209X_A^2$$

Where, $\mathbf{X_6}$ = Human Development Index,

$\mathbf{X_A}$ = Longitude

5. Conclusion and Discussion

Our project delved into the intricate relationship between socio-economic, environmental factors, and happiness, blending ancient wisdom with modern scientific inquiry to understand national well-being better. Starting with clear objectives, we gathered data from reputable sources like World Population Review and Wikipedia, assembling a comprehensive dataset spanning human development, health, demographics, and governance.

We employed rigorous analytical techniques, including exploratory data analysis and regression modeling, aiming for robust insights into the complex dynamics of happiness determinants. Through iterative refinement, we navigated challenges like multicollinearity and outliers, striving for robust and interpretable models.

Our analysis culminated in the development of a final estimated regression function:

$$\hat{Y} = 1.32134 + 5.8048X_6 - 0.00518X_A + 3.6209X_A^2$$

This equation reveals the significant impact of variables like the Human Development Index and geographic longitude on national happiness levels.While the positive association with HDI aligns with expectations, the negative relationship with longitude presents a surprising finding.

It's important to acknowledge the limitations of interpreting correlation as causation. While HDI likely contributes to happiness, other unmeasured factors might influence both HDI and happiness scores. Furthermore, this analysis is based on the available data and might not capture the full complexity of factors influencing happiness across nations. Cultural nuances, historical events, and social structures likely play a significant role that this model cannot currently account for.

Despite these limitations, this study provides a starting point for further exploration. Future research could involve:

- Including additional relevant variables to capture a more holistic picture of national happiness.
- Expanding the data collection to include more countries. This would allow for a more comprehensive understanding of how various factors interact to influence happiness on a global scale.

In conclusion, this analysis highlights the complex interplay of factors contributing to national happiness. While the results offer intriguing insights, further investigation is necessary to fully understand the drivers of happiness across countries. This study serves as a stepping stone for future research to delve deeper into this critical topic

6. References

[1] Albert, "Average Latitute & Longitude of Countries," 2013. [Online]. https://github.com/albertyw/avenews/blob/master/old/data/averagelatitude-longitude-countries.csv.

[2] "Water - Countries Compared," [Online]. https://www.nationmaster.com/country-info/stats/Geography/Area/Water.

[3] "List of Countries by System of Government," [Online]. https://en.wikipedia.org/wiki/List_of_countries_by_system_of_government.

[4] "Literacy rate by Country," [Online]. https://www.datapandas.org/ranking/literacy-rate-by-country.

[5] "IQAir 2022 World Air Quality," [Online]. https://en.wikipedia.org/wiki/List_of_countries_by_air_pollution.

[6] Laura M, "Happiness Index Methodology," *Journal of Social Change,* vol. 9, pp. 4-31, 2017.

[7] "World Population Data," [Online]. https://worldpopulationreview.com/.

7. Appendix

7.1 Dataset

- No of records – 137
- Data Dictionary

```
$ Country                     <chr> "Afghanistan", "Albania", "Algeria", "Arge~
$ Happiness_Score             <dbl> 1.72, 5.30, 5.36, 6.19, 5.46, 7.06, 6.91, ~
$ Population                  <int> 42239854, 2832439, 45606480, 45773884, 277~
$ Land.Area..KM2.             <int> 652860, 27400, 2381740, 2736690, 28470, 76~
$ Population_Density          <int> 65, 103, 19, 17, 98, 3, 109, 126, 1955, 13~
$ Net_Migrants                <int> -65846, -8000, -9999, 3718, -5000, 139991,~
$ Fertility_Rate              <dbl> 4.4, 1.4, 2.8, 1.9, 1.6, 1.6, 1.5, 1.7, 1.~
$ Median_Age                  <int> 17, 38, 28, 32, 35, 38, 43, 32, 34, 27, 41~
$ Urban_Population_Percentage <int> 26, 67, 75, 94, 67, 86, 59, 57, 60, 41, 99~
$ Developed_Developing        <chr> "Developing", "Developing", "Developing", ~
$ Human_Development_Index     <dbl> 0.478, 0.796, 0.745, 0.842, 0.759, 0.951, ~
$ Health_Care_Index           <dbl> 45.00, 48.86, 49.00, 50.04, 48.71, 57.77, ~
$ Constitutional_Form         <chr> "Provisional", "Republic", "Republic", "Re~
$ Literacy_Rate               <dbl> 0.3817, 0.9755, 0.7961, 0.9809, 0.9977, 0.~
$ Latitude                    <dbl> 33.00, 41.00, 28.00, -34.00, 40.00, -27.00~
$ Longitude                   <dbl> 65.00, 20.00, 3.00, -64.00, 45.00, 133.00,~
$ River..Sq.KM.               <int> 0, 1350, 0, 43710, 1540, 58920, 1426, 3971~
$ River_to_Land_Percent       <dbl> 0.00, 4.93, 0.00, 1.60, 5.41, 0.77, 1.73, ~
$ Pollution_PM2_5             <dbl> 15.0, 14.5, 17.8, 7.7, 31.4, 4.2, 10.6, 18~
```

Figure 7.1-1: Data Dictionary

- Data Sample

	Y			X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12		X13	X14
Country	Happiness_Score	Population	Land Area (KM2)	Population_Density	Net_Migrants	Fertility_Rate	Median_Age	Urban_Population Perc	Developed_Developing	Human_Development_Index	Health_Care_Index	Constitutional_Form	Literacy_Rate	Latitude	Longitude	River (Sq KM)	River_to_Land_Percent	Pollution_PM2_5
Afghanistan	1.72	42239854	652860	65	-65846	4.40	17	26	Developing	0.478	45.00	Provisional	0.3817	33.00	65.00	0	0.00	15.00
Albania	5.3	2832439	27400	103	-8000	1.40	38	67	Developing	0.796	48.86	Republic	0.9755	41.00	20.00	1350	4.93	14.50
Algeria	5.36	45606480	2381740	19	-9999	2.80	28	75	Developing	0.745	49.00	Republic	0.7961	28.00	3.00	0	0.00	17.80
Argentina	6.19	45773884	2736690	17	3718	1.90	32	94	Developed	0.842	50.04	Republic	0.9809	-34.00	-64.00	43710	1.60	7.70
Armenia	5.46	2777970	28470	98	-5000	1.60	35	67	Developing	0.759	48.71	Republic	0.9977	40.00	45.00	1540	5.41	31.40
Australia	7.06	26439111	7682300	3	139991	1.60	38	86	Developed	0.951	57.77	Constitutional monarchy	0.99	-27.00	133.00	58920	0.77	4.20
Austria	6.91	8958960	82409	109	19999	1.50	43	59	Developed	0.916	54.69	Republic	0.98	47.33	13.33	1426	1.73	10.60
Azerbaijan	4.89	10412651	82658	126	0	1.70	32	57	Developing	0.745	48.66	Republic	0.9981	40.50	47.50	3971	4.80	18.90
Bahrain	5.96	1485509	760	1955	0	1.80	34	60	Developed	0.875	52.83	Constitutional monarchy	0.9572	26.00	50.55	0	0.00	66.60
Bangladesh	3.89	172954319	130170	1329	-309977	1.90	27	41	Developing	0.661	45.39	Republic	0.6149	24.00	90.00	13830	10.62	65.80
Belgium	6.89	11686140	30280	386	23999	1.60	41	99	Developed	0.937	53.99	Constitutional monarchy	0.99	50.83	4.00	250	0.83	10.80
Benin	4.38	13712828	112760	122	-200	4.80	18	48	Developing	0.525	44.00	Republic	0.3845	9.50	2.25	2000	1.77	15.00
Bolivia	5.78	12388571	1083300	11	-3000	2.50	24	69	Developing	0.692	37.00	Republic	0.9514	-17.00	-65.00	15280	1.41	7.30
Bosnia and He	5.88	3210847	51000	63	-500	1.30	42	54	Developing	0.780	34.00	Republic	0.9849	44.00	18.00	10	0.02	33.60

Figure 7.1-2: Data Sample

7.2 Code and Output