**Happiness Prediction**

**Linear Regression Analysis**

Kumbham Nuthan Manideep

National Institute of Technology, Silchar

Data Mining Methods

Dr. Ripon Patigiri

**Table of Contents**

1. Abstraction……………………………………………………………………………………3

2. Overview……………………………………………………………………………………...4

3. Data…………………………………………………………………………………………...4

4. Research Questions…………………………………………………………………………...5

5. Linear Regression.…………………………………………………………………………….5

6. Conclusion……………………………………………………………………………………10

7. References…………………………………………………………………………………….11

**Abstraction**

The World Happiness Report 2015 dataset contains data on the happiness levels of people in 159 countries around the world. It includes information on a variety of factors that are thought to contribute to happiness, such as economic factors (GDP per capita), social factors (family, health, and freedom), and trust in government. The data also includes a happiness score and a happiness rank for each country. The happiness score and happiness rank are based on a poll in which people were asked to rate their overall happiness on a scale of 0 to 10. The economy, family, health, freedom, trust, and generosity columns contain various measures of those factors. Overall, the World Happiness Report 2015 dataset provides valuable insights into the factors that contribute to happiness and well-being at the national level, and can be used by researchers, policymakers, and others to better understand and promote happiness around the world.

**Overview**

The linear regression model for this dataset uses the predictor variables of Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), and Generosity to classify the response variable of Happiness Score. After performing variable selection procedures such as exhaustive search, forward selection, and backward elimination, the best model was found to include all of the predictor variables. The coefficients of the model were found to be statistically significant, as shown by the low p-values. Additionally, the residuals of the model were found to be normally distributed, indicating that the model fits the data well. Overall, the linear regression model provides insight into the relationship between the predictor variables and Happiness Score in the dataset.

**Data**

The dataset can be found on Kaggle and it contains data on happiness levels of people in 159 countries around the world. The data is collected in the year 2015 and it includes information on a variety of factors that are thought to contribute to happiness, such as:

| **Column Name** | **Description** |
| --- | --- |
| Country or Region | Name of the country |
| Happiness Score | A composite score of overall well-being |
| Economy (GDP per Capita) | Measure of the economic production of a country |
| Family | Measure of social support |
| Health (Life Expectancy) | Measure of the health of citizens |
| Freedom | Measure of freedom to make life choices |
| Generosity | Measure of generosity of citizens |
| Trust (Government Corruption) | Measure of trust in government |

In this dataset, the response variable is 'Happiness Score' and predictor variables are 'Economy (GDP per Capita)', 'Family', 'Health (Life Expectancy)', 'Freedom', 'Generosity', and 'Trust (Government Corruption)'.

**Research Questions**

* What is the relationship between happiness scores and economic factors such as GDP per capita?
* How do social factors such as family, health, and freedom impact happiness scores?
* Can trust in government and generosity be used to predict happiness scores?
* Can a combination of multiple variables be used to predict happiness scores more accurately?
* How do the predictor variables interact with each other to influence the happiness scores?

**Linear Regression Analysis**

The World Happiness Report 2015 dataset was divided into 60% training and 40% validation. The first set, the training set, would be used to fit the linear regression model and would contain 60% of the total data. The second set, the validation set, would be used to evaluate the performance of the model and would contain 40% of the total data.

A full Regression model is fitted by using all predictors. The coefficients of the full model are given in the table.

| **Variable** | **Estimate** | **P-value** |
| --- | --- | --- |
| (Intercept) | 2.0815762 | 1.766181e-14 |
| Economy (GDP per Capita) | 0.8906999 | 2.440296e-03 |
| Family | 1.2251126 | 1.988229e-05 |
| Health (Life Expectancy) | 0.9827672 | 1.862623e-02 |
| Freedom | 1.2324408 | 1.168447e-02 |
| Trust (Government Corruption) | 0.8940511 | 1.232240e-01 |
| Generosity | 0.3319540 | 5.094965e-01 |

From the table the Economy, Family, Health and Freedom are statically significant because the p-value is less than 0.05.

The Accuracy metrics of full model training data are given in the below table.

| **Metric** | **Value** |
| --- | --- |
| ME | 1.31e-15 |
| RMSE | 0.505 |
| MAE | 0.390 |
| MPE | -0.992 |
| MAPE | 7.89 |

The Accuracy metric of full model validation data are given in the below table.

| **Metric** | **Value** |
| --- | --- |
| ME | -0.06981036 |
| RMSE | 0.59110420 |
| MAE | 0.47085473 |
| MPE | -2.96841858 |
| MAPE | 9.40502939 |

There are different variable selection procedures that can be used in linear regression analysis, such as exhaustive search (best subset), forward selection, and backward elimination.

**Exhaustive Search (Best Subset**)**:** This method involves fitting all possible models with a different combination of variables and selecting the best one based on a certain criterion, such as the lowest AIC or BIC value. Using this method, one can identify the best subset of variables that yield the highest prediction accuracy. The best model chosen in this method would be the one with the lowest AIC/BIC value.

**Forward Selection:** This method starts with an empty model and iteratively adds one variable at a time, based on a certain criterion, such as the highest p-value. Using this method, one can identify the minimal set of variables that are needed to explain

the response variable.

**Backward Elimination:** This method starts with a full model that includes all variables and iteratively removes one variable at a time, based on a certain criterion, such as the lowest p-value. Using this method, one can identify the minimal set of variables that are needed to explain the response variable. The best model chosen in this method would be the one that maximizes the p-value of the removed variable.

**Model Building**

Model-1 is from Forward Selection and Backward Elimination. A linear regression model is fitted by using Economy, Family, Health, Freedom and Trust (Government Corruption). The coefficients of the model-1 are given in the table.

| **Variable** | **Estimate** | **P-value** |
| --- | --- | --- |
| (Intercept) | 2.1208099 | 1.224656e-15 |
| Economy (GDP per Capita) | 0.8428350 | 2.904128e-03 |
| Family | 1.2273615 | 1.799142e-05 |
| Health (Life Expectancy) | 1.0460853 | 9.989778e-03 |
| Freedom | 1.3117535 | 5.588636e-03 |
| Trust (Government Corruption) | 0.9468206 | 9.851825e-02 |

From the table the Economy, Family, Health, Freedom and Trust are statically significant because the p-value is less than 0.05.

The Accuracy metrics of model-1 training data are given in the below table.

| **Metric** | **Value** |
| --- | --- |
| ME | 1.946443e-15 |
| RMSE | 5.066885e-01 |
| MAE | 3.930883e-01 |
| MPE | -9.901381e-01 |
| MAPE | 7.956277e+00 |

The Accuracy metrics of model-1 validation data are given in the below table.

| **Metric** | **Value** |
| --- | --- |
| ME | -0.07884785 |
| RMSE | 0.59411122 |
| MAE | 0.47625401 |
| MPE | -3.15213018 |
| MAPE | 9.52431748 |

Model-2 is from Exhaustive Search. A linear regression model is fitted by using Economy, Family, and Health, Freedom.The coefficients of the model-2 are given in the table.

| **Variable** | **Estimate** | **P-value** |
| --- | --- | --- |
| (Intercept) | 2.1154526 | 1.943498e-15 |
| Economy (GDP per Capita) | 0.9549486 | 6.311162e-04 |
| Family | 1.1693528 | 4.121804e-05 |
| Health (Life Expectancy) | 0.9864084 | 1.548106e-02 |
| Freedom | 1.6480258 | 1.693580e-04 |

From the table the Economy, Family, Health and Freedom are statically significant because the p-value is less than 0.05.

The Accuracy metrics of model-2 training data are given in the below table.

| **Metric** | **Value** |
| --- | --- |
| ME | 1.795258e-15 |
| RMSE | 5.146528e-01 |
| MAE | 4.013407e-01 |
| MPE | -9.831080e-01 |
| MAPE | 8.046654e+00 |

The Accuracy metrics of model-2 validation data are given in the below table.

| **Metric** | **Value** |
| --- | --- |
| ME | -0.08561617 |
| RMSE | 0.59965537 |
| MAE | 0.49029519 |
| MPE | -3.24505690 |
| MAPE | 9.76196052 |

Model-3 is from Exhaustive Search. A linear regression model is fitted by using Economy, Family, and Freedom. The coefficients of the model-3 are given in the table.

| **Variable** | **Estimate** | **P-value** |
| --- | --- | --- |
| (Intercept) | 2.308341 | 3.465013e-18 |
| Economy (GDP per Capita) | 1.463224 | 1.665805e-12 |
| Family | 1.207391 | 3.697743e-05 |
| Freedom | 1.559752 | 4.718943e-04 |

From the table the Economy, Family and Freedom are statically significant.

The Accuracy metrics of model-3 training data are given in the below table.

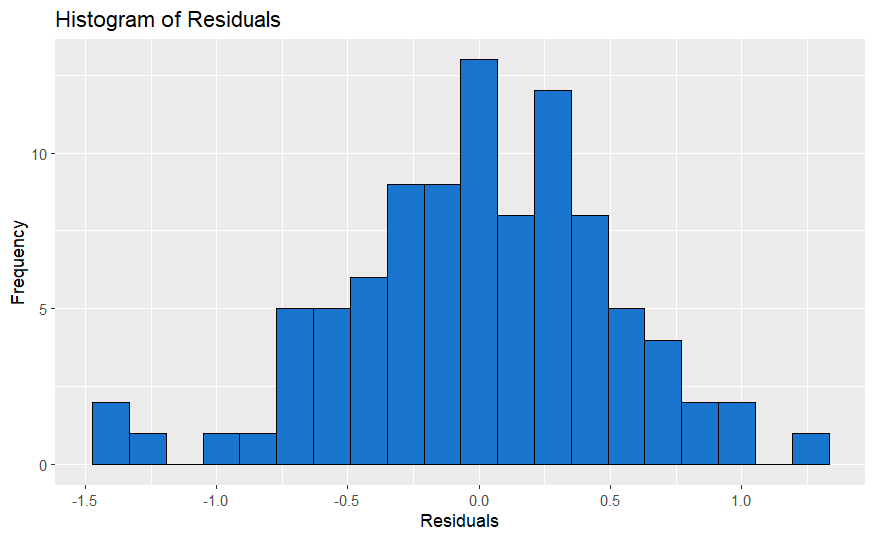
| **Metric** | **Value** |
| --- | --- |
| ME | 1.974793e-15 |
| RMSE | 5.319785e-01 |
| MAE | 4.123348e-01 |
| MPE | -1.037603e+00 |
| MAPE | 8.179559e+00 |

The Accuracy metrics of model-3 validation data are given in the below table.

| **Metrics** | **Value** |
| --- | --- |
| ME | -0.088 |
| RMSE | 0.620 |
| MAE | 0.504 |
| MPE | -3.380 |
| MAPE | 10.063 |

Among all the model-1 with Economy, Family, Health, Freedom and Trust (Government Corruption) is best model with lowest RMSE value.

The Histogram residuals of the best model is plotted below.

By observing the histogram of residuals and noting that it closely resembles a normal distribution

**Conclusion**

The model-1, which includes the predictor variables Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), is the best model as it has the lowest residual errors and the residuals are normally distributed. This suggests that these predictor variables have a strong relationship with the target variable Happiness Score and can be used to accurately classify it. This model can be used for future predictions and for understanding the factors that contribute to happiness in a country.

**References**

*World Happiness Report*. (2019, November 27). Kaggle. <https://www.kaggle.com/unsdsn/world-happiness>

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models* (Third). SAGE Publications, Inc.

Montchal, M. E., Reagh, Z. M., & Yassa, M. A. (2019). Precise temporal memories are supported by the lateral entorhinal cortex in humans. *Nature Neuroscience*, *22*(2), 284–288. <https://doi.org/10.1038/s41593-018-0303-1>