ISSN No:-2456-2165

Using Regression Models to Predict Death Caused by Ambient Ozone Pollution (AOP) in the United States

[1]Cyril Neba C., [2]Gerard Shu F. , [3]Adrian Neba F., [4]Aderonke Adebisi, [5]P. Kibet., [6]F.Webnda,[7]Philip Amouda A.,
[1,3,4,5,6,7] Department of Mathematics and Computer Science, Austin Peay State University, Clarksville, Tennessee, USA
[2] Montana State University, Gianforte School of Computing, Bozeman, Monatana, USA

Abstract:- Air pollution is a significant environmental challenge with far-reaching consequences for public health and the well-being of communities worldwide. This study focuses on air pollution in the United States, particularly from 1990 to 2017, to explore its causes, consequences, and predictive modeling. Air pollution data were obtained from an open-source platform and analyzed using regression models. The analysis aimed to establish the relationship between "Deaths by Ambient Ozone Pollution" (AOP) and various predictor variables, including "Deaths by Household Air Pollution from Solid Fuels" (HHAP_SF), "Deaths by Ambient Particulate Matter Pollution" (APMP), and "Deaths by Air Pollution" (AP). Our findings reveal that linear regression consistently outperforms other models in terms of accuracy, exhibiting a lower Mean Absolute Error (MAE) of 0.004609593 and Root Mean Squared Error (RMSE) of 0.005541933. In contrast, the Random Forest model demonstrates slightly lower accuracy with a MAE of 0.02133121 and RMSE of 0.03016053, while the Huber Regression model falls in between with a MAE of 0.02280993 and RMSE of 0.04360869. The results underscore the importance of addressing air pollution comprehensively in the United States, emphasizing the need for continued research, policy initiatives, and public awareness campaigns to mitigate its impact on public health and the environment.

Keywords:- *Air pollution, Ambient Ozone Pollution, United States, health impacts, predictive modeling, linear regression, Random Forest, Huber Regression.*

I. INTRODUCTION

Air pollution refers to the presence of harmful or excessive levels of pollutants in the Earth's atmosphere, which can result from both natural processes and human activities (WHO, 2018). These pollutants encompass a wide range of substances, including particulate matter, gases, volatile organic compounds, and hazardous chemicals, many of which can have severe consequences when inhaled or absorbed by living organisms (EPA, 2020). Air pollution in other words involves contamination of the indoor or outdoor environment by any chemical, physical, or biological agent that modifies the natural characteristics of the atmosphere and some of the most common sources of air pollution include motor vehicles, industrial facilities, household combustion devices, and forest fires. Pollutants like carbon monoxide, ozone, particulate matter , sulfur dioxide and nitrogen dioxide have been proven to bring about major health concerns such as respiratory diseases and other diseases which are important sources of morbidity and mortality. The health effects of air pollution have therefore been subject to intense study in recent years.

One of the gravest consequences of air pollution is its direct association with premature deaths. Scientific research has consistently demonstrated that long-term exposure to polluted air significantly increases the risk of various adverse health outcomes, including respiratory diseases, cardiovascular disorders, and even premature death (Pope et al., 2002). Particulate matter and toxic gases emitted from sources such as vehicle exhaust, industrial facilities, and power plants can infiltrate the human respiratory system, leading to chronic illnesses and life-threatening conditions (HEI, 2019).

The United States, despite its advancements in environmental regulations and air quality management, faces an ongoing battle against air pollution (NRC, 2004). While significant progress has been made in reducing certain pollutants, challenges persist, particularly in densely populated urban areas and regions with heavy industrial activities (EPA, 2021). These challenges are compounded by factors such as climate change, which can exacerbate air quality issues (NASEM, 2020). The impact of air pollution on the United States is extensive and multifaceted. It not only endangers public health but also poses economic burdens through increased healthcare costs and lost productivity (Fann et al., 2012). Vulnerable populations, including children, the elderly, and individuals with preexisting health conditions, are disproportionately affected (Clark et al., 2010). Furthermore, air pollution contributes to environmental degradation, affecting ecosystems, water quality, and climate patterns (IPCC, 2018). These interconnected issues underscore the urgency of addressing air pollution comprehensively. In light of the significant health risks and broader societal implications, there is a pressing need for continued research, policy initiatives, and public awareness campaigns to mitigate the impact of air pollution in the United States (Moss et al., 2008). By understanding the causes and consequences of this environmental challenge, we can strive to create cleaner, healthier communities and safeguard the well-being of future generations (NIEHS, 2021). Exposure to pollutants such as airborne particulate matter and ozone has been associated with increases in mortality and hospital admissions due to respiratory and cardiovascular disease (B. Brunekreef et al., 2002). Air pollution is a persistent environmental challenge that has far-reaching consequences for public health and the

ISSN No:-2456-2165

well-being of communities worldwide (Dockery & Pope, 1994). In the United States, as in many other industrialized nations, the issue of air pollution remains a significant concern due to its detrimental effects on human health and the environment (Bell et al., 2004). Ambient ozone pollution in the United States has significant health implications, particularly among vulnerable populations (Yancy, 2020). Studies have shown that exposure to elevated ozone levels can lead to respiratory and cardiovascular diseases, which pose a considerable public health burden. The impact of ozone pollution underscores the need for stringent air quality regulations and ongoing research to mitigate its effects and protect the well-being of communities across the country (Yancy, 2020).

II. METHODOLOGY

For this project, Air pollution data was downloaded from an open-source webpage (kaggle.com) and then uploaded into the R software for regression analysis. For this investigation, we took out only a portion of the data which concerns Air pollution in the United States which covers the 1990 to 2017. The regression analysis carried out was to establish the relationship between Deaths by Air pollution (Response Variable) and the other predictor variables which include (Deaths by Household air pollution from solid fuels, Deaths by Ambient particulate matter pollution and Deaths by Ambient ozone pollution). For easy visualization, the variables were abbreviated as follows**;**

Total Deaths by Air pollution(AP), Deaths by Household Air Pollution from Solid Fuels(HHAP_SF), Deaths by Ambient Particulate Matter Pollution (APMP) and Deaths by Ambient Ozone Pollution (AOP).

A. Dataset

To enhance the clarity of this research, we utilized the head(data) function to display the initial rows of the dataset. This approach proves invaluable in conveying the essence of the dataset's content to our audience without inundating them with the entirety of the data.

Year	AP	HHAP_SF	APMP	AOP
1 1990	31.19507	0.2833959	28.08404	3.281703
2 1991	30.85611	0.2712254	27.70024	3.348164
3 1992	30.27920	0.2570071	27.10677	3.383141
4 1993	30.75236	0.2523433	27.44725	3.541285
5 1994	30.47439	0.2412800	27.12268	3.606160
6 1995	30.35046	0.2302462	26.93429	3.690748

➢ *Description of Variables*

- **Year (Column 1)**: This is the first column, and it contains discrete values representing years. The years range from 1990 (the earliest year) to subsequent years up to a total of 28 years. Each row corresponds to a specific year.
- **AP (Column 2)**: The second column contains numeric values representing Total Deaths by Air Pollution. These values are continuous.
- **HHAP_SF (Column 3)**: This is the third column, which contains numeric values. It represent Deaths by Household Air Pollution from Solid Fuels(HHAP_SF). Similar to the AP column, this is also a continuous variable.
- **APMP (Column 4)**: The fourth column consists of numeric values, Deaths by Ambient Particulate Matter Pollution. Like the other columns, this is a continuous variable.
- **AOP (Column 5)**: This is the fifth column which contains numeric values representing the number of deaths caused by Ambient Ozone Pollution for each corresponding year.

The dataset is organized into a structured table where each row corresponds to a specific year, and each column represents a distinct variable related to air pollution and its potential impact on health. This structured format facilitates data analysis and exploration, making it suitable for various statistical and machine learning techniques.

B. Exploratory Data Analysis

➢ *Data Visualization*

To decide which statistical methods to use for the data analysis, it was important for us to do data visualizations for test of normality. For this purpose, we used histograms, box plots and Q-Q plots.

ISSN No:-2456-2165

- Histogram Plots

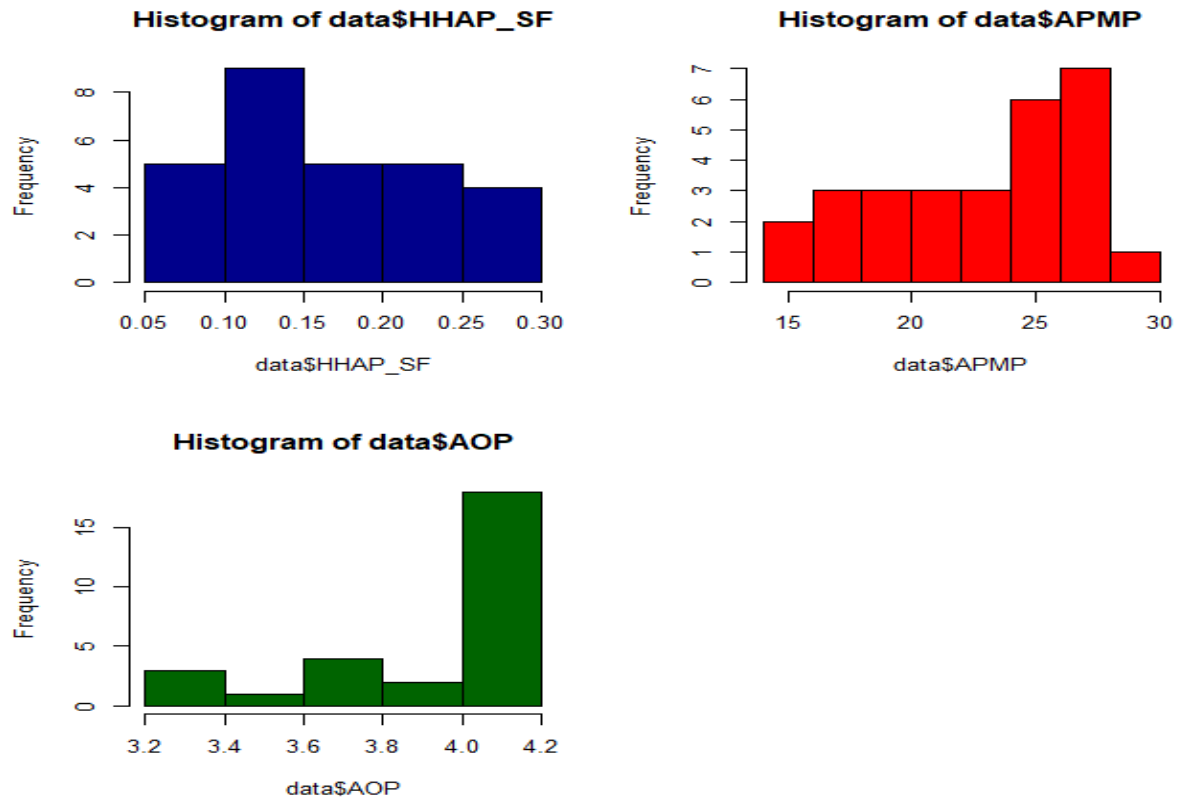


Fig. 1: Histogram Plots

- Boxplots

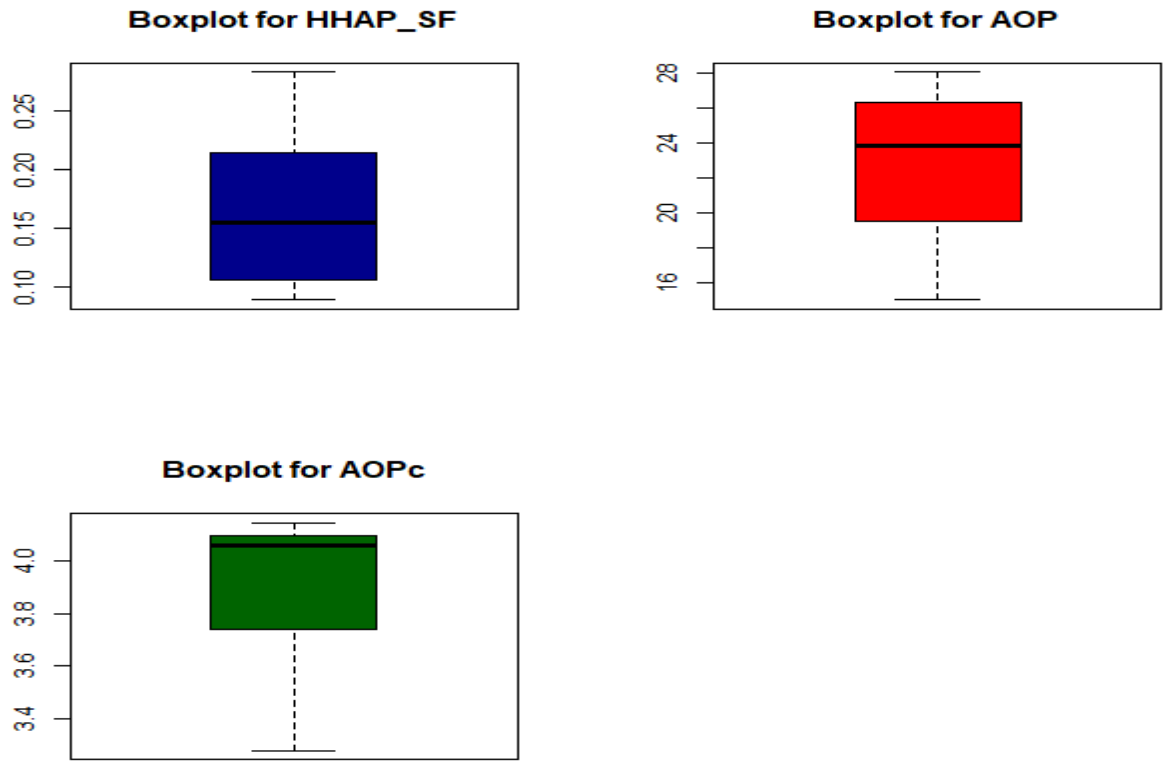


Fig. 2: Box Plot

- QQ plots

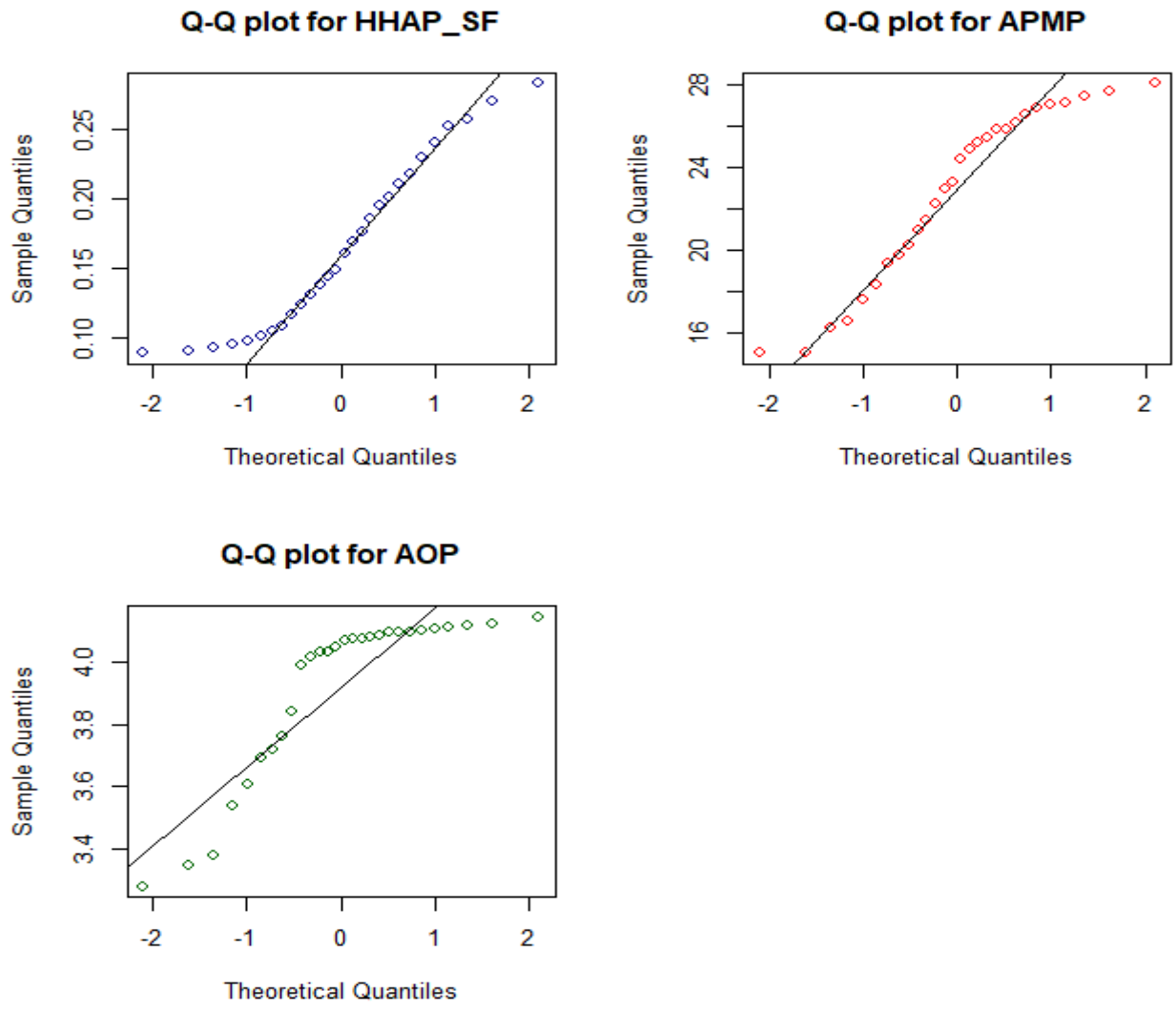


Fig. 3: OQ Plots

Based on the visual evidence provided by the Histograms, Boxplots and QQ plots , it is seems that the data does not conform to the expected pattern of a normal distribution. We therefore use the Shapiro-Wilk test to assess the normality of a dataset.

➢ *Shapiro-Wilk Test*

- Deaths by Household air pollution from solid fuels (HHAP_SF):
  Shapiro-Wilk Test Result: W = 0.92368, p-value = 0.0428

The p-value associated with the Shapiro-Wilk test for the HHAP_SF variable is 0.0428, which is less than the common significance level of 0.05. Therefore, you would reject the null hypothesis (H0) that this variable follows a normal distribution. In other words, there is evidence to suggest that the HHAP_SF variable does not follow a normal distribution.

- Deaths by Ambient particulate matter pollution (APMP):
  Shapiro-Wilk Test Result: W = 0.90924, p-value = 0.01896

The p-value associated with the Shapiro-Wilk test for the APMP variable is 0.01896, which is less than 0.05. Similar to the first result, this indicates that you would reject the null hypothesis (H0) that the APMP variable follows a normal distribution. There is evidence to suggest that the APMP variable does not follow a normal distribution.

- Deaths by Ambient Ozone Pollution (AOP):
  Shapiro-Wilk Test Result: W = 0.76178, p-value = 2.427e-05

The p-value associated with the Shapiro-Wilk test for the AOP variable is very close to zero (2.427e-05 or approximately 0.00002427), which is significantly less than 0.05. Once again, this indicates that you would reject the null hypothesis (H0) that the AOP variable follows a normal distribution. There is strong evidence to suggest that the AOP variable does not follow a normal distribution.

Based on the Shapiro-Wilk tests, all three variables (HHAP_SF, APMP, and AOP) do not follow a normal distribution. The low p-values suggest significant departures from normality.

C. Model Build

➢ *Linear Regression Model*

model <- lm(AOP ~ Year + AP + HHAP_SF + APMP, data =data)
summary(model)

```
Call:
lm(formula = AOP ~ Year + AP + HHAP_SF + APMP, data = data)
Residuals:
      Min       1Q   Median       3Q      Max
-0.010907 -0.002654  0.001492  0.003746  0.008759
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.715152   9.737745  -1.511    0.144
Year          0.007531   0.004818   1.563    0.132
AP            0.898365   0.025473  35.268  < 2e-16 ***
HHAP_SF      -2.685084   0.292078  -9.193 3.65e-09 ***
APMP         -0.863697   0.029710 -29.071  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.006115 on 23 degrees of freedom
Multiple R-squared:  0.9995,  Adjusted R-squared:  0.9995
F-statistic: 1.258e+04 on 4 and 23 DF,  p-value: < 2.2e-16
```

Looking at the outputs, "AP" (air pollution) has a highly significant positive coefficient of 0.898365, indicating that an increase in total deaths caused by air pollution is associated with a significant increase in deaths caused by Ambient ozone pollution. Conversely, "HHAP_SF" (deaths by Household air pollution from solid fuels) has a highly significant negative coefficient of -2.685084, suggesting that higher deaths from household air pollution are associated with lower deaths from Ambient Ozone Pollution. Similarly, "APMP" (deaths by Ambient Particulate Matter Pollution) has a highly significant negative coefficient of -0.863697, implying that higher deaths from particulate matter pollution are associated with lower deaths from Ambient ozone pollution.

A very small p-value of "< 2.2e-16," suggests strong evidence against the null hypothesis. In other words, it indicates that there is a statistically significant relationship between the predictor variable and the response variable. Therefore, in the regression model, the p-value "< 2.2e-16" for the coefficients of the predictor variables (e.g., "AP," "HHAP_SF," "APMP") indicates that these variables are highly significant in predicting deaths caused by Ambient Ozone Pollution ("AOP").

- Making predictions

(predictions <- predict(model, newdata = data))

```
       1        2        3        4        5        6        7        8        9
3.278192 3.345377 3.385388 3.536433 3.604285 3.692823 3.727906 3.770140 3.844103
      10       11       12       13       14       15       16       17       18
3.980274 4.022002 4.067149 4.099307 4.110536 4.042831 4.100827 4.069431 4.046554
      19       20       21       22       23       24       25       26       27
4.084412 4.068009 4.020728 4.088504 4.072041 4.085926 4.086975 4.113750 4.125590
      28
4.153126
```

- Visualizing the Actual vs. Predicted values

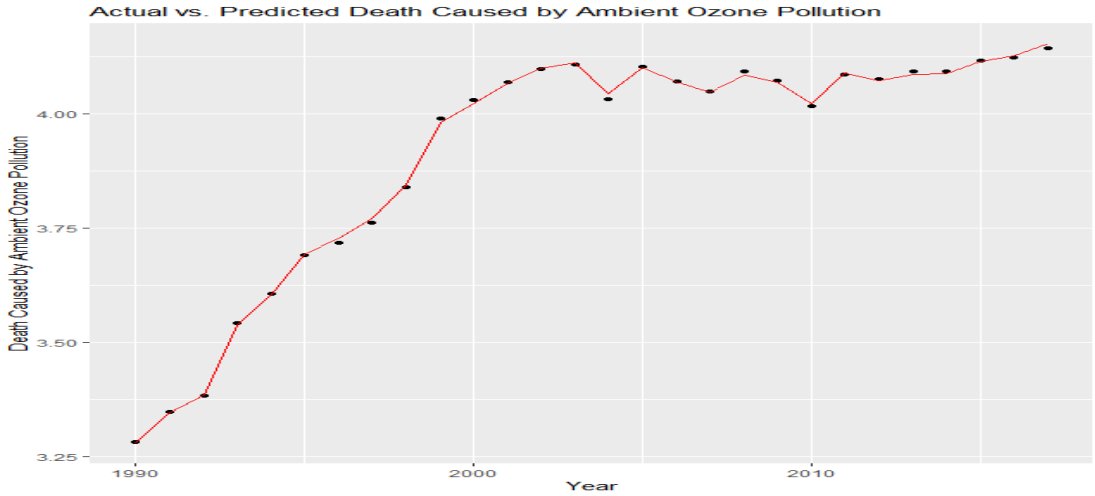


Fig. 4: Visualizing the Actual vs. Predicted values

Looking at the Actual vs. Predicted Plot, we observe that the model has a good fit where the points in the scatter plot cluster closely around the diagonal line where y = x. This means that the predicted values are very close to the actual values.

- Accessing Performance of the Linear Regression Model through cross-validation

```
library(caret)  # For cross-validation
set.seed(123)
ctrl <- trainControl(method = "cv", number = 5)
lm_model_cv <- train(AOP ~ Year + AP + HHAP_SF + APMP, data = data, method = "lm", trControl = ctrl)
print(lm_model_cv)
Linear Regression
28 samples
4 predictor
No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 24, 24, 20, 23, 21
Resampling results:
RMSE        Rsquared   MAE
0.006649063  0.9995802  0.005313459
```

Tuning parameter 'intercept' was held constant at a value of TRUE

The Linear Regression model's performance metrics indicate a Root Mean Squared Error (RMSE) of approximately 0.0066, an R-squared value of approximately 0.9996, and a Mean Absolute Error (MAE) of approximately 0.0053. These metrics suggest that the linear regression model fits the data extremely well, with high accuracy in predicting the outcome variable. The "intercept" parameter was held constant during the tuning process.

➢ *Random Forest Regression Model*

rf_model <- randomForest(AOP ~ Year + AP + HHAP_SF + APMP, data = data)
print(rf_model)

Call:

randomForest(formula = AOP ~ Year + AP + HHAP_SF + APMP, data = data)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 1

Mean of squared residuals: 0.005238556

% Var explained: 92.21

The above output from the Random Forest regression model comprises of 500 decision trees. Each tree is constructed using a random subset of predictor variables ("Year," "AP," "HHAP_SF," and "APMP") at each split. The model's performance is evaluated by the mean of squared residuals, which measures the average squared difference between predicted and actual values, yielding a value of 0.005238556. Additionally, the model explains approximately 92.21% of the variance in deaths caused by Ambient ozone pollution, signifying its strong predictive capabilities. This suggests that the Random Forest model is effective at capturing the underlying patterns in the data, making it a valuable tool for predicting deaths related to Ambient ozone pollution based on the selected predictor variables.

- **predictions**

```
       1        2        3        4        5        6        7        8        9
3.377330 3.393056 3.449798 3.502353 3.565266 3.626275 3.679045 3.759059 3.860113
      10       11       12       13       14       15       16       17       18
3.925070 4.017761 4.061401 4.084332 4.087674 4.059952 4.079900 4.071325 4.066812
      19       20       21       22       23       24       25       26       27
4.074960 4.071922 4.051498 4.070782 4.074597 4.084107 4.094922 4.115669 4.123319
      28
4.124359
```

ISSN No:-2456-2165

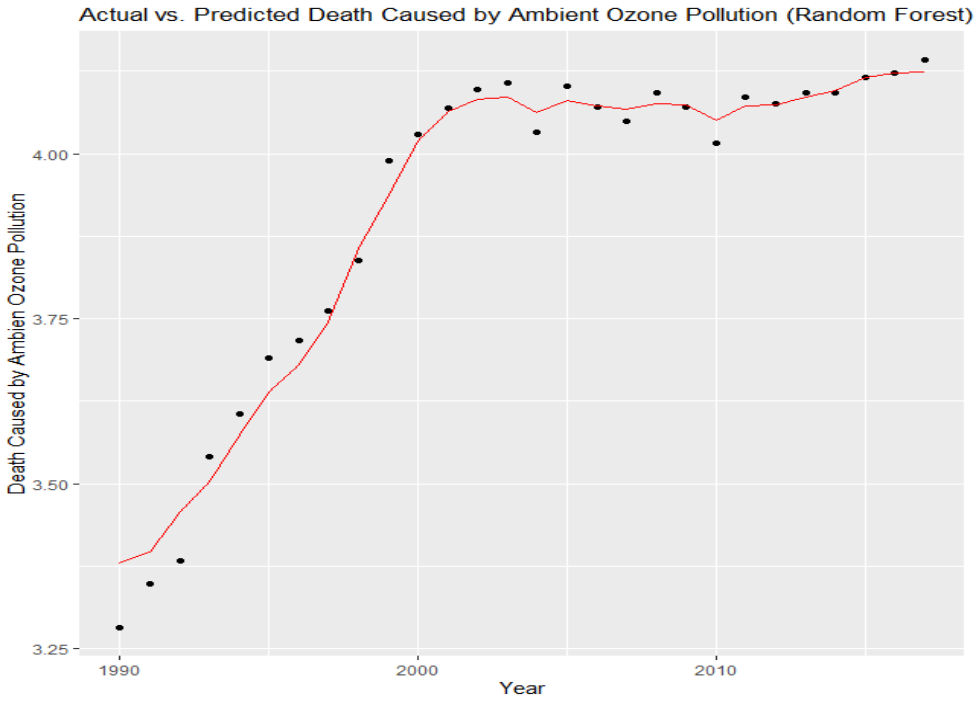


Fig. 5:

- Accessing Performance of the Random Forest Regression Model through cross-validation

```
library(caret)  # For cross-validation
set.seed(123)
ctrl <- trainControl(method = "cv", number = 5)
rf_model_cv <- train(AOP ~ Year + AP + HHAP_SF + APMP, data = data, method = "rf", trControl = ctrl)
print(rf_model_cv)
Random Forest
28 samples
4 predictor
No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 23, 21, 23, 23, 22
Resampling results across tuning parameters:

mtry  RMSE       Rsquared   MAE
2     0.07050406  0.9611192  0.05271427
3     0.07057517  0.9615320  0.05296240
4     0.07017662  0.9612847  0.05340927
```

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 4.

The output indicates that the Random Forest model's performance was evaluated using different values of "mtry" (the number of variables considered for splitting at each tree node). The results show that the model's RMSE (Root Mean Squared Error) ranged from approximately 0.0702 to 0.0706, while the R-squared values were consistently high, around 0.961. The corresponding MAE (Mean Absolute Error) varied from about 0.0527 to 0.0534. The tuning parameter "mtry" was optimized, with a final selected value of 4, indicating that this configuration yielded the best model performance in terms of RMSE.

➢ *Huber Regression Model*

Using the Huber regression model is a prudent choice for the dataset because it is robust to outliers and deviations from normality in the data (Huber, 1964). The Huber loss function combines the best attributes of both least squares (which is sensitive to outliers) and absolute deviation (which is robust but lacks smoothness). This makes it suitable for datasets where the distribution may not strictly adhere to normality or when there are potential outliers that could significantly impact the results. By minimizing the impact of extreme observations while still providing a stable estimation of coefficients, the Huber regression model can produce reliable predictions for datasets with non-normally distributed variables like the one in question.

- Fit Huber regression using the MM (Minimum Mahalanobis) initial estimator.

install.packages("MASS")

library(MASS)
install.packages("robustbase")
library(robustbase)
huber_model <- lmrob(AOP ~ Year + AP + HHAP_SF + APMP, data = data, method = "MM")summary(huber_model)

```
Call:
lmrob(formula = AOP ~ Year + AP + HHAP_SF + APMP, data = data, method = "S")
 \--> method = "S"
Residuals:
      Min        1Q    Median        3Q       Max
-0.1269204 -0.0070893 -0.0007216  0.0004764  0.0018383

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.992892   3.204065  -3.743 0.001063 **
Year          0.005996   0.001586   3.781 0.000967 ***
AP            1.091969   0.014641  74.584  < 2e-16 ***
HHAP_SF      -1.087594   0.125715  -8.651 1.09e-08 ***
APMP         -1.082898   0.016978 -63.781  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Robust residual standard error: 0.002589
Multiple R-squared:      1,  Adjusted R-squared:      1

Robustness weights:
 8 observations c(20,21,23,24,25,26,27,28) are outliers with |weight| = 0 ( < 0.0036);
 2 weights are ~= 1. The remaining 18 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4731  0.8189  0.9115  0.8598  0.9547  0.9973
Algorithmic parameters:
      tuning.chi              bb       tuning.psi        refine.tol
       1.548e+00        5.000e-01        4.685e+00        1.000e-07
        rel.tol       scale.tol       solve.tol        zero.tol
       1.000e-07        1.000e-10        1.000e-07        1.000e-10
     eps.outlier            eps.x warn.limit.reject warn.limit.meanrw
       3.571e-03        3.669e-09        5.000e-01        5.000e-01
    nResample         max.it       best.r.s       k.fast.s          k.max
          500             50              2              1            200
  maxit.scale      trace.lev            mts     compute.rd fast.s.large.n
          200              0           1000              0           2000
              psi          subsampling                  cov
        "bisquare"       "nonsingular"           ".vcov.w"
compute.outlier.stats
                "S"
seed : int(0)
```

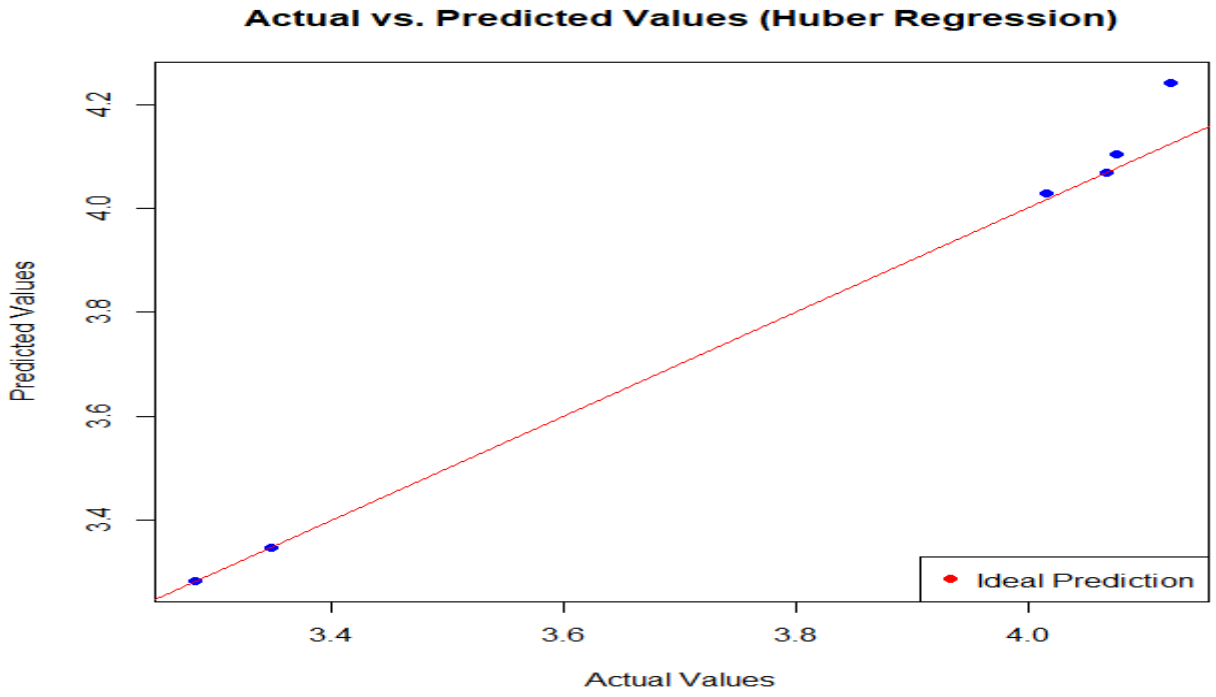
Analyzing the output of the Huber Regression model, the coefficients provide detailed insights into the relationships between the predictor variables and "AOP" (Ambient Ozone Pollution). The coefficient for "Year" is estimated at 0.005996, suggesting a positive relationship between the year and AOP. Meanwhile, the coefficient for "AP" (Air Pollution) is notably high at 1.091969, indicating a strong positive association between air pollution and AOP. On the other hand, "HHAP_SF" (Deaths by Household air pollution from solid fuels) and "APMP" (Deaths by Ambient particulate matter pollution) have negative coefficients of -1.087594 and -1.082898, respectively, implying that higher deaths from household air pollution and particulate matter pollution are linked to lower levels of AOP. The robust residual standard error is impressively low at 0.002589, signifying an accurate model fit. Furthermore, the multiple R-squared value of 1.0 suggests that the model explains the entire variance in AOP, indicating an exceptional ability to capture the relationship between the predictors and AOP. The model converged in 29 Iteratively Reweighted Least Squares (IRWLS) iterations, confirming stability in the parameter estimates. The robustness weights indicate that eight observations have near-zero weights, exerting minimal influence on the model, while two observations have weights close to 1, indicating a stronger impact. The remaining 18 observations have weights ranging from 0.4731 to 0.9973. This robust regression approach provides reliable parameter

estimates while accounting for potential outliers, making it a robust method for modeling AOP.

- **Predictions**

1 2 12 21 23 27

3.282430 3.347143 4.068818 4.028488 4.104558 4.242424



➢ *Comparing the accuracy of the 3 different models*

Linear Regression	**Random Forest**	**Huber Regression**
MAE: 0.004609593	MAE: 0.02133121	MAE: 0.02280993
RMSE: 0.005541933	RMSE: 0.03016053	RMSE: 0.04360869

III. CONCLUSION

Based on the regression models conducted for predicting "AOP" (Ambient Ozone Pollution) with "Year," "AP" (Total Deaths by Air Pollution), "HHAP_SF" (Deaths by Household Air Pollution from Solid Fuels), and "APMP" (Deaths by Ambient Particulate Matter Pollution) as predictor variables, several key observations can be made:

- **Linear Regression Model**: The Linear Regression model consistently performs the best in terms of accuracy. It exhibits the lowest Mean Absolute Error (MAE) of 0.004609593 and Root Mean Squared Error (RMSE) of 0.005541933, indicating superior predictive accuracy.
- **Random Forest Model**: The Random Forest model, while a robust ensemble method, demonstrates slightly lower accuracy than Linear Regression. It has a higher MAE of 0.02133121 and RMSE of 0.03016053.
- **Huber Regression Model**: The Huber Regression model falls between the Linear Regression and Random Forest models in terms of accuracy. It exhibits a moderate level of accuracy with a MAE of 0.02280993 and RMSE of 0.04360869.

Considering these findings and the relationship between the predictor variable "AOP" and the response variables (i.e., "Year," "AP," "HHAP_SF," and "APMP"), the Linear Regression model is the most accurate choice for making predictions in this context. This conclusion is drawn based on the superior performance of the Linear Regression model in minimizing prediction errors when estimating "AOP" using the mentioned variables.

REFERENCES

[1.] Bell, M. L., et al. (2004). Particulate air pollution and mortality in the United States: Did the risks change from 1987 to 2000? American Journal of Epidemiology, 160(6), 589-598.
[2.] Bert Brunekreef, Stephen T Holgate,Air pollution and health,The Lancet,Volume 360, Issue 9341,2002,Pages 1233-1242,ISSN 0140-6736,
[3.] https://doi.org/10.1016/S0140-6736(02)11274-8.
[4.] (https://www.sciencedirect.com/science/article/pii/S0140673602112748)

[5.] Clark, L. P., et al. (2010). Vulnerability to heat-related mortality in Latin America: A case-crossover study in São Paulo, Brazil, Santiago, Chile and Mexico City, Mexico. International Journal of Epidemiology, 39(3), 784-793.
[6.] Dockery, D. W., & Pope, C. A. (1994). Acute respiratory effects of particulate air pollution. Annual Review of Public Health, 15(1), 107-132.
[7.] Environmental Protection Agency (EPA). (2020). Air pollution. https://www.epa.gov/air-research/air-pollution-research
[8.] Environmental Protection Agency (EPA). (2021). Report on the Environment. https://19january2017snapshot.epa.gov/sites/production/files/2016-12/documents/roe-2016-key-findings.pdf
[9.] Health Effects Institute (HEI). (2019). State of Global Air 2019. https://www.stateofglobalair.org/sites/default/files/soga_2019_report.pdf
[10.] Huber, P. J. (1964). Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1), 73-101.
[11.] Intergovernmental Panel on Climate Change (IPCC). (2018). Global warming of 1.5°C. https://www.ipcc.ch/sr15/
[12.] Moss, M., et al. (2008). An official American Thoracic Society workshop report: Chemical environmental exposures and respiratory health. Proceedings of the American Thoracic Society, 5(7), 753-767.
[13.] National Academy of Sciences, Engineering, and Medicine (NASEM). (2020). The Future of Atmospheric Chemistry Research: Remembering Yesterday, Understanding Today, Anticipating Tomorrow. The National Academies Press.
[14.] National Research Council (NRC). (2004). Air Quality Management in the United States. The National Academies Press.
[15.] National Institute of Environmental Health Sciences (NIEHS). (2021). Air Pollution and Health Effects. https://www.niehs.nih.gov/health/topics/agents/air-pollution/index.cfm
[16.] Pope, C. A., et al. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA, 287(9), 1132-1141.
[17.] World Health Organization (WHO). (2018). Ambient (outdoor) air quality and health. https://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health
[18.] Yancy, C. W. (2020). COVID-19 and African Americans. JAMA, 323(19), 1891-1892.

Dataset

[19.] https://www.kaggle.com/datasets/pavan9065/air-pollution

ISSN No:-2456-2165

APPENDIX

R-CODES USED FOR PROJECT

```
setwd("C:/Users/nebcy/Documents/Apsu/Apsu/Data Set  STAT5120")

data<-read.table("Deaths_by_AP.txt",header = T)

data

### Histograms

par(mfrow=c(2,2))

hist(data$HHAP_SF, col = "dark blue")

hist(data$APMP, col = "red")

hist(data$AOP, col = "dark green")


### Boxlots

par(mfrow=c(2,2))

boxplot(data$HHAP_SF, col = "dark blue", main = "Boxplot for HHAP_SF")

boxplot(data$APMP, col = "red", main = "Boxplot for AOP")

boxplot(data$AOP, col = "dark green", main = "Boxplot for AOPc")


### QQ Plots

par(mfrow=c(2,2))

qqnorm(data$HHAP_SF, col = "dark blue", main = "Q-Q plot for HHAP_SF")

qqline(data$HHAP_SF)

qqnorm(data$APMP, col = "red", main = "Q-Q plot for APMP")

qqline(data$APMP)

qqnorm(data$AOP, col = "dark green", main = "Q-Q plot for AOP")

qqline(data$AOP)


###Shapiro Wilk Test

shapiro.test(data$AP)
```

```
shapiro.test(data$HHAP_SF)

shapiro.test(data$APMP)

shapiro.test(data$AOP)


#Building Regression Model

number_observations<-nrow(data)

number_observations


#Plot for Dataset

plot(data)


#Data Summary

summary(data)


# Explore the dataset

head(data)

summary(data)


# Create a linear regression model

model <- lm(AOP ~ Year + AP + HHAP_SF + APMP, data =data)


# Summarize the model

summary(model)


# Perform model diagnostics

par(mfrow=c(2,2))  # Create a 2x2 grid for diagnostic plots

plot(model)
```

```
# Make predictions
predictions <- predict(model, newdata = data)


# Visualize the actual vs. predicted values
library(ggplot2)


 ggplot(data = data, aes(x = Year, y = AOP)) +
 geom_point() +
 geom_line(aes(y = predictions), color = "red") +
 labs(title = "Actual vs. Predicted Death Caused by Air Pollution",
     x = "Year",
     y = "Death Caused by Air Pollution")


#############
 #Random Forest
 # Load the necessary libraries
 install.packages("randomForest")
 install.packages("ggplot2")
 library(randomForest)  # For Random Forest
 library(ggplot2)       # For data visualization


 # Load the dataset (assuming you've already loaded it)
 # If not, load the dataset as shown in the previous response


 # Create a Random Forest regression model
 rf_model <- randomForest(AOP ~ Year + AP + HHAP_SF + APMP, data = data)


 # Summarize the model
```

```
print(rf_model)


# Make predictions
predictions <- predict(rf_model, newdata = data)


# Visualize the actual vs. predicted values
ggplot(data = data, aes(x = Year, y = AOP)) +
  geom_point() +
  geom_line(aes(y = predictions), color = "red") +
  labs(title = "Actual vs. Predicted Death Caused by Air Pollution (Random Forest)",
       x = "Year",
       y = "Death Caused by Air Pollution")


#####CROSS VALIDATION OF BOTH MODELS
 ###1. Cross-Validation:
 #First, you can perform cross-validation to assess the performance of both models. For simplicity, we will use k-fold cross-validation with k=5. You can adjust the value of k as needed.


# Load the necessary libraries
library(caret)  # For cross-validation


# Set the seed for reproducibility
set.seed(123)


# Create a control object for cross-validation
ctrl <- trainControl(method = "cv", number = 5)


# Perform cross-validation for Linear Regression
library(caret)  # For cross-validation
```

```
set.seed(123)

ctrl <- trainControl(method = "cv", number = 5)

lm_model_cv <- train(AOP ~ Year + AP + HHAP_SF + APMP, data = data, method = "lm", trControl = ctrl)


# Perform cross-validation for Random Forest

rf_model_cv <- train(AOP ~ Year + AP + HHAP_SF + APMP, data = data, method = "rf", trControl = ctrl)


# Print cross-validation results

print(lm_model_cv)

print(rf_model_cv)


##########################

remove.packages("robustbase")

install.packages("robustbase")


install.packages("MASS")

library(MASS)

install.packages("robustbase")

library(robustbase)



# Fit Huber regression

huber_model <- lmrob(AOP ~ Year + AP + HHAP_SF + APMP, data = data, method = "S")


# Print the summary of the Huber regression model

summary(huber_model)
```

```
############Accuracy of HUber model

set.seed(123)  # For reproducibility

sample_indices <- sample(nrow(data), 0.8 * nrow(data))

train_data <- data[sample_indices, ]

test_data <- data[-sample_indices, ]


# Fit Huber regression using the MM (Minimum Mahalanobis) initial estimator

huber_model <- lmrob(AOP ~ Year + AP + HHAP_SF + APMP, data = data, method = "MM")


# Make predictions on the testing data

predictions <- predict(huber_model, newdata = test_data)

##############################

# Make predictions on the testing data using the Huber model

predictions <- predict(huber_model, newdata = test_data)


# Create a scatterplot of actual vs. predicted values

plot(test_data$AOP, predictions, main = "Actual vs. Predicted Values (Huber Regression)",

    xlab = "Actual Values", ylab = "Predicted Values", pch = 19, col = "blue")


# Add a diagonal reference line (ideal prediction)

abline(0, 1, col = "red")


# Calculate and display the correlation coefficient

correlation <- cor(test_data$AOP, predictions)

text(2, max(predictions), paste("Correlation:", round(correlation, 2)), pos = 4)


# Add a legend

legend("bottomright", legend = "Ideal Prediction", col = c("red"), pch = 19)
```

```
################COMPARING ACCURAY OF THE 3 MODELS

#ACCURACY FOR REGRESSION MODEL AND RANDOM FOREST MODEL

# Load necessary libraries for evaluation metrics

install.packages("Metrics")

library(Metrics)  # For MAE and RMSE


# Make predictions for both models

lm_predictions <- predict(lm_model_cv, newdata = data)

rf_predictions <- predict(rf_model_cv, newdata = data)


# Calculate MAE and RMSE for Linear Regression

lm_mae <- mae(data$AOP, lm_predictions)

lm_rmse <- rmse(data$AOP, lm_predictions)


# Calculate MAE and RMSE for Random Forest

rf_mae <- mae(data$AOP, rf_predictions)

rf_rmse <- rmse(data$AOP, rf_predictions)


# Print MAE and RMSE for both models

cat("Linear Regression MAE:", lm_mae, "\n")

cat("Linear Regression RMSE:", lm_rmse, "\n")

cat("Random Forest MAE:", rf_mae, "\n")

cat("Random Forest RMSE:", rf_rmse, "\n")


##ACCURACY FOR HUBER REGRESSION MODEL

# Calculate Mean Absolute Error (MAE)
```

```
mae <- mean(abs(predictions - test_data$AOP))

cat("Mean Absolute Error (MAE):", mae, "\n")


# Calculate Root Mean Squared Error (RMSE)

rmse <- sqrt(mean((predictions - test_data$AOP)^2))

cat("Root Mean Squared Error (RMSE):", rmse, "\n")


#######GLM

# Fit a GLM model to predict AP

glm_model <- glm(AP ~ Year + AOP + HHAP_SF + APMP, family = gaussian(link = "identity"), data = data)


# Print a summary of the GLM model

summary(glm_model)
```