Xi'an Jiaotong-Liverpool University
西交利物浦大學

# XJTLU Entrepreneur College (Taicang) Cover Sheet

| | |
|---|---|
| Module code and Title | **DTS206TC Applied Linear Statistical Models** |
| School Title | **School of AI and Advanced Computing** |
| Assignment Title | **Coursework** |
| Submission Deadline | **23:59 31st May (Friday)** |
| Final Word Count | **NAN** |
| If you agree to let the university use your work anonymously for teaching and learning purposes, please type **"yes"** here. | **yes** |

I certify that I have read and understood the University's Policy for dealing with Plagiarism, Collusion and the Fabrication of Data (available on Learning Mall Online). With reference to this policy I certify that:

- My work does not contain any instances of plagiarism and/or collusion.
  My work does not contain any fabricated data.

**By uploading my assignment onto Learning Mall Online, I formally declare that all of the above information is true to the best of my knowledge and belief.**

| **Scoring – For Tutor Use** | |
|---|---|
| **Student ID** | **2142116** |

| **Stage of Marking** | **Marker Code** | **Learning Outcomes Achieved (F/P/M/D) (please modify as appropriate)** | | | **Final Score** |
|---|---|---|---|---|---|
| | | **A** | **B** | **C** | |
| 1st Marker – red pen | | | | | |
| Moderation – green pen | **IM Initials** | The original mark has been accepted by the moderator (please circle as appropriate): | | | Y / N |
| | | Data entry and score calculation have been checked by another tutor (please circle): | | | Y |
| 2nd Marker if needed – green pen | | | | | |

| **For Academic Office Use** | | | **Possible Academic Infringement (please tick as appropriate)** | |
|---|---|---|---|---|
| **Date Received** | **Days late** | **Late Penalty** | ☐ **Category A** | Total Academic Infringement Penalty (A,B, C, D, E, Please modify where necessary) _____ |
| | | | ☐ **Category B** | |
| | | | ☐ **Category C** | |
| | | | ☐ **Category D** | |
| | | | ☐ **Category E** | |

# Regression Analysis of $NH_3$ Emissions in the Air Pollution Dataset: Predicting ammnonia emissions


## Abstract

This study conducts a regression analysis on the air pollution dataset to predict $NH_3$ emissions, addressing key issues such as heteroscedasticity, non-normality of residuals, multicollinearity, and autocorrelation. The dataset includes variables like NMVOC, BC, and CO emissions, which are highly correlated with $NH_3$ emissions. Log transformation is applied to stabilize variance and normalize residuals, significantly reducing the residual standard error. To manage multicollinearity, both removal of collinear variables and Ridge Regression are employed, leading to more stable coefficient estimates. Generalized Least Squares (GLS) is used to handle autocorrelation, enhancing the reliability of model estimates. Findings indicate that NMVOC, BC, and CO are significant predictors of $NH_3$ emissions, demonstrating high statistical significance. The applied methods result in a robust predictive model, providing valuable insights for environmental research and pollution management. Future research should focus on exploring advanced regression techniques and validating the model with diverse datasets to further enhance generalizability and accuracy.

# Contents

# 1 Data Analysis & Visualization

This section provides a detailed description of the Air Pollution dataset, which is obtained from Kaggle [1]. The objective is to predict ammonia ($NH_3$) emissions based on various related emissions data. The dataset includes several characteristic variables that are thought to influence $NH_3$ emissions. The data provides a wealth of measurements on air quality across different countries from 1750 to 2022, enabling the construction of effective prediction models.

## 1.1 Dataset Description

The dataset contains 48,225 records, each representing emissions of various molecules that impact air quality across different countries and years. The goal of the dataset is to predict $NH_3$ emissions ($NH_3$) from the variables described below. These variables range from various chemical emissions to broader environmental measurements, providing a comprehensive overview for prediction models:

**Year** The year the data was recorded.

**Country** The country where the data was collected.

**NOx** Emissions of Nitrogen oxides, measured in kilotons.

**$SO_2$** Emissions of Sulphur dioxide, measured in kilotons.

**CO** Emissions of Carbon monoxide, measured in kilotons.

**OC** Emissions of Organic carbon, measured in kilotons.

**NMVOC** Emissions of Non-methane volatile organic compounds, measured in kilotons.

**BC** Emissions of Black carbon, measured in kilotons.

**$NH_3$** Emissions of Ammonia, measured in kilotons.

## Variables of interests are below:

Table 1: Variables of interests

| Variable | Description | Purpose |
|---|---|---|
| NMVOC | Non-methane volatile organic compounds emissions | Predictor variable for NH3 emissions |
| BC | Black carbon emissions | Predictor variable for NH3 emissions |
| CO | Carbon monoxide emissions | Predictor variable for NH3 emissions |
| NH3 | Ammonia emissions | Target variable to be predicted |

## 1.2 Exploratory Data Analysis

In order to fully understand the data set, exploratory data analysis (EDA) is first performed. The purpose of EDA is to provide an initial statistical description of the data and to identify potential problems in the data (such as missing values and outliers). In this study, several functions and packages in R (such as summary() and etc.) are used to provide a basic statistical description of the dataset and examine the distribution, missing values, and outliers of the data.

Initial exploration involves assessing the structure and quality of the data, which informs subsequent analysis decisions. Key statistics and distributions are examined to understand data trends and anomalies.

### 1.2.1 Statistical Summary

*Code to generate summary statistics:*

```
summary(air_pollution)
```

The following table presents a detailed statistical summary of the variables in the air pollution dataset:

Table 2: Statistical Summary of Air Pollution Data

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| Year | 1750 | 1854 | 1913 | 1911 | 1970 | 2022 |
| NOx | 0 | 157 | 1782 | 540778 | 31042 | 109243090 |
| SO2 | 0 | 64 | 952 | 809118 | 33708 | 134596620 |
| CO | 0 | 13676 | 108715 | 4885731 | 610786 | 599913340 |
| OC | 0 | 596 | 4429 | 131520 | 22207 | 13618318 |
| NMVOC | 0 | 2314 | 17703 | 877193 | 113247 | 135481700 |
| BC | 0 | 149 | 1091 | 47388 | 6595 | 6141969 |
| NH3 | 0 | 1371 | 9434 | 352142 | 51926 | 63947644 |
| $\log_{NH3}$ | -Inf | 7.223 | 9.152 | -Inf | 10.858 | 17.974 |

### 1.2.2 Missing Values Analysis

Code to check for missing values:

```
colSums(is.na(air_pollution))
```

| Year | NOx | SO2 | CO | OC | NMVOC | BC | NH3 | $\log_{NH3}$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 3: Missing Values Horizontal

This table confirms that the dataset is complete with **no missing entries** in most columns, allowing us to proceed with confidence in our further analyses.

### 1.2.3 Outliers Analysis

Code to create boxplots for each variable:

```
# Using tidyr's gather function to reshape the dataset
data_long <- gather(air_pollution, key = "Variable", value = "Value", -Entity, -Code, -
    Year)
# Apply log transformation to the values
data_long$LogValue <- log(data_long$Value + 1)
# Create boxplot with enhanced aesthetics
ggplot(data_long, aes(x = Variable, y = LogValue, fill = Variable)) +
  geom_boxplot(outlier.color = "red", outlier.shape = 16, outlier.size = 2, notch = TRUE
      ) +
  scale_fill_brewer(palette = "Set3") +
  theme_minimal() +
  labs(title = "Boxplot of All Variables (Log Transformed)", y = "Log Value", x = "
      Variable") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
```

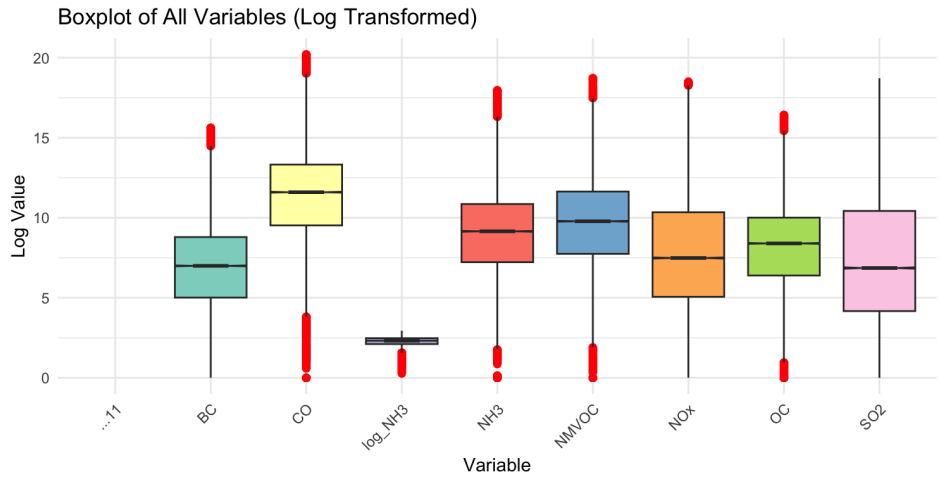


Figure 1: Boxplot of All Variables (Log Transformed)

The following code was used to create boxplots for each variable in the dataset. The `gather` function from the `tidyr` package was employed to reshape the data from a wide format to a long format, making it suitable for plotting with `ggplot2`. The boxplot provides a visual representation of the distribution of values for each variable and helps to identify any potential outliers.

The boxplot (Figure 1) of log-transformed variables reveals the distribution characteristics and potential outliers for each variable in the dataset. Notably, variables such as NMVOC and OC exhibit larger variability within their interquartile ranges and have several outliers above the upper whiskers, indicating significant deviations from the typical data range. In contrast, variables like -11 and BC show fewer outliers, suggesting a more consistent dataset. It is crucial to verify the outliers to rule out data entry errors and to consider their implications carefully. Employing robust statistical methods can help mitigate the influence of these outliers, ensuring more reliable analysis and modeling results.

By examining these boxplots, we can gain insights into the spread and central tendency of the data, as well as identify any variables that may have significant outliers. This information is crucial for understanding the data distribution and for guiding further analysis and modeling efforts.

## 1.3 Data Visualization

In this section, we begin by presenting various visual charts of the data for all variables, including histograms, scatter plots, and heat maps. These visualizations provide an initial overview of the data distribution and relationships between variables. Next, we employ the correlation coefficient to identify the variable most closely related to NH3. Finally, we create separate visualizations for the selected variables to further explore their characteristics and interactions.

### 1.3.1 Distribution of All Variables

To better understand the data distribution and reduce the impact of extreme values, a log-transformation was applied to all variables. Log-transforming the data helps in stabilizing variance, making the data more normally distributed, and enhancing the interpretability of the results, especially when dealing with skewed data. The formula for log transformation is given by:

$$y = \log(x + 1)$$

This transformation can reduce the skewness of the data and the influence of extreme values, making the data closer to a normal distribution, thereby improving the reliability and accuracy of the analysis results.

## Log Transformed Distribution of All Variables

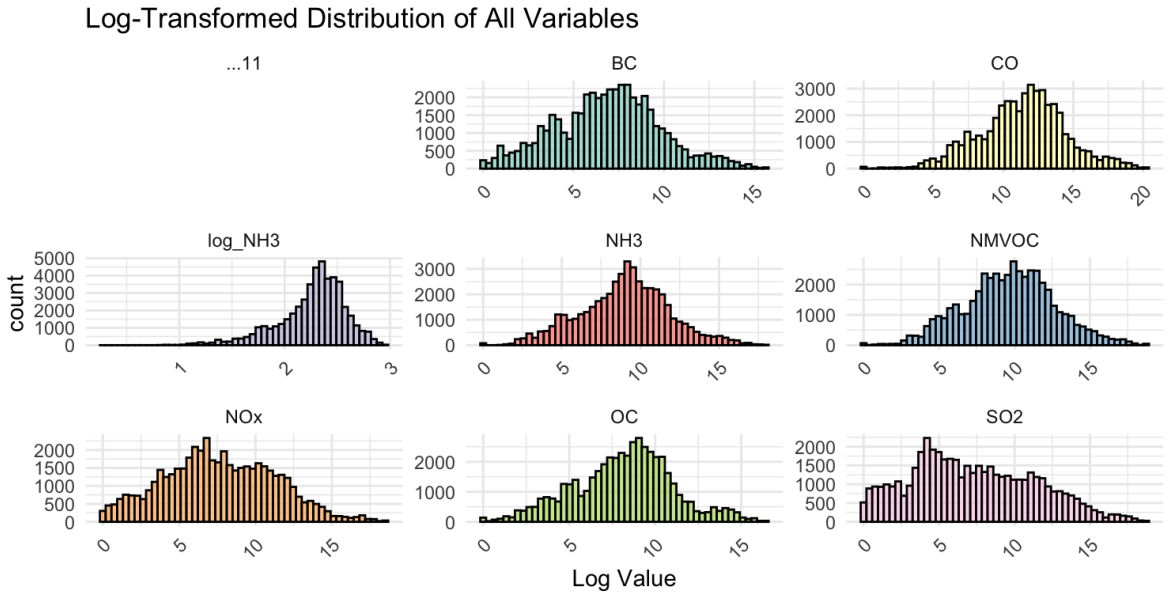


Figure 2: Log-Transformed Distribution of All Variables

Figure 2 shows the log-transformed distribution of all variables. Each variable exhibits different characteristics. For instance, *BC* and *CO* variables have a more symmetrical distribution, while *NH3* and *NMVOC* variables exhibit more skewed distributions. *NOx* and *OC* variables show high peaks, indicating that most samples have low values for these variables, with only a few samples having significantly high values. The *SO2* variable also shows a high peak, suggesting that this variable has low values in most cases, but there are some extreme high values. These features indicate that it is necessary to consider the different distribution characteristics of these variables in research to ensure the accuracy and reliability of the analysis results.

**Key Observations**

These distribution characteristics indicate that special attention needs to be paid to data skewness and extreme values in environmental pollution analysis to ensure the accuracy and reliability of the results.

| Variable | Description |
|---|---|
| BC (Black Carbon) | Symmetrical distribution with a peak in the middle |
| CO (Carbon Monoxide) | Symmetrical distribution with a peak at a lower position |
| NH3 (Ammonia) | Skewed distribution with a peak at a higher position |
| NMVOC (Non-Methane Volatile Organic Compounds) | Symmetrical distribution with a peak in the middle |
| NOx (Nitrogen Oxides) | Skewed distribution with a peak at a higher position |
| OC (Organic Carbon) | Symmetrical distribution with a peak in the middle |
| SO2 (Sulfur Dioxide) | Skewed distribution with a peak at a higher position |

Table 4: Key Observations of Variable Distributions

## Scatter Plot Matrix (Log Transformed)

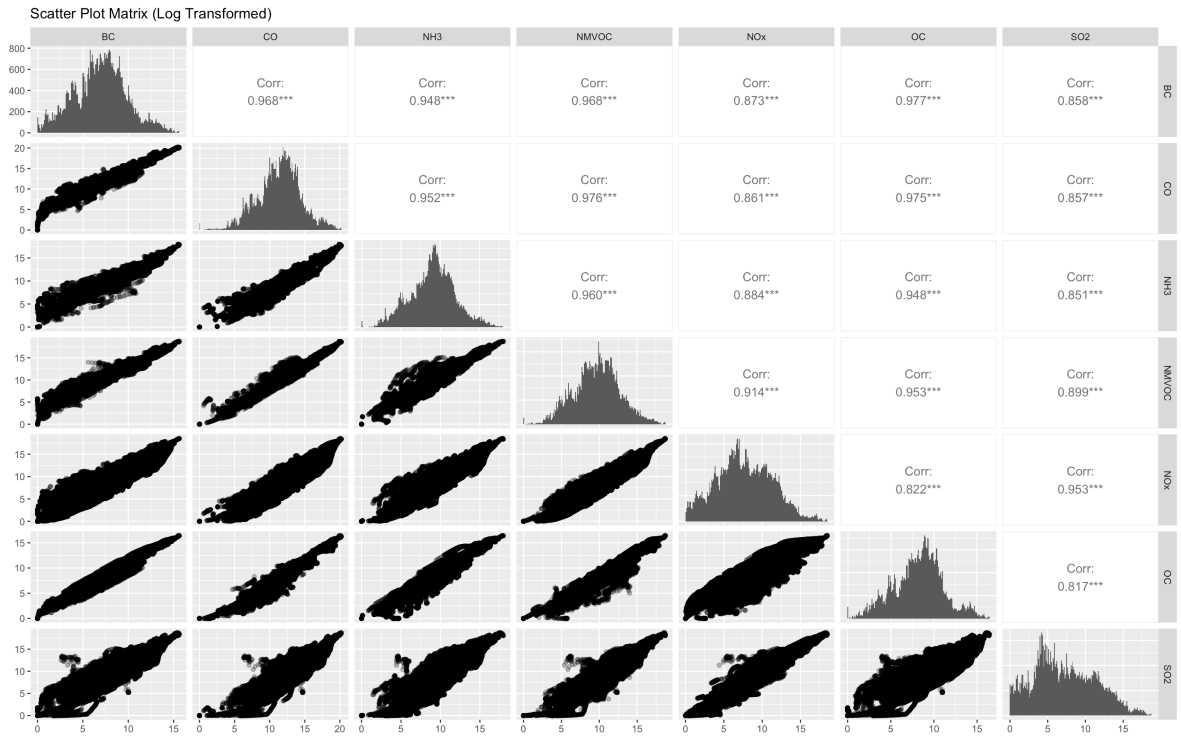


Figure 3: Scatter Plot Matrix (Log Transformed)

The scatter plot matrix (Figure 3) displays pairwise relationships between variables along with their correlation coefficients.

**Key observations include:**

| Variable Pair | Correlation |
|---|---|
| BC and CO | Strong positive correlation (Corr: 0.968) |
| BC and NH3 | Positive correlation (Corr: 0.952) |
| CO and NMVOC | High correlation (Corr: 0.976) |
| NMVOC and NOx | Positive correlation (Corr: 0.861) |
| OC and SO2 | Positive correlation (Corr: 0.817) |

Table 5: Key Observations of Variable Correlations

## Correlation Matrix

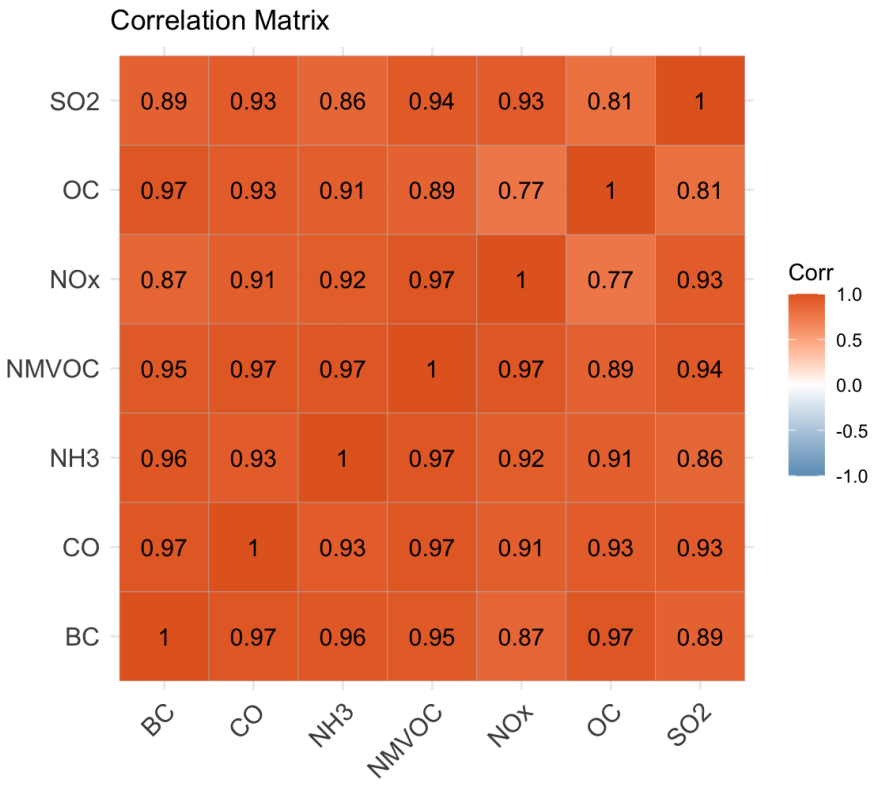


Figure 4: Correlation Matrix

The correlation matrix (Figure 4) visually represents the strength and direction of correlations between variables. **Notable Correlations** These relationships are visually highlighted in the heatmap, with stronger correlations

| Variable Pair | Correlation |
|---|---|
| BC and CO | Strong positive correlation (Corr: 0.97) |
| BC and NH3 | Strong positive correlation (Corr: 0.96) |
| CO and NMVOC | Strong positive correlation (Corr: 0.97) |
| NMVOC and NOx | Strong positive correlation (Corr: 0.92) |
| OC and SO2 | Strong positive correlation (Corr: 0.81) |

Table 6: Notable Correlations between Variables

marked by darker colors.

### 1.3.2 Sort and select highly relevant variables

To identify the most relevant variables for subsequent analysis, we calculated the correlation coefficients of all variables with the target variable *NH3* and sorted them in descending order. The results are shown below:

```
target_cor <- cor_matrix[, "NH3"]
sorted_cor <- sort(target_cor, decreasing = TRUE)
print(sorted_cor)
```

From the sorted correlations, we selected ***NMVOC*, *BC*, and *CO*** for further analysis due to they are the fist three with the highest correlation with *NH3*.

| Variable | Correlation with NH3 |
|---|---|
| NH3 | 1.0000000 |
| NMVOC | 0.9668579 |
| BC | 0.9634962 |
| CO | 0.9310063 |
| NOx | 0.9175540 |
| OC | 0.9103097 |
| SO2 | 0.8634237 |
| Year | 0.1196134 |

Table 7: Correlation of Variables with NH3

**Reasons for Choosing Three Variables**

We chose to analyze three variables instead of more for several reasons:

1. **Correlation Threshold**: We selected the three variables with the highest correlation to *NH3* (all correlation coefficients above 0.93), ensuring that the analysis focuses on the most impactful factors.
2. **Multicollinearity**: Introducing too many highly correlated variables can lead to multicollinearity issues, affecting the stability and interpretability of the model. By selecting the three most correlated variables, we can effectively reduce the impact of multicollinearity.
3. **Model Simplification**: In practical applications, the simplicity and interpretability of the model are crucial. Choosing three variables makes the model easier to interpret and apply while ensuring depth and breadth of analysis.

Combining the previous scatter plots, histograms, and heatmaps, we can see that these variables not only exhibit significant correlation but also demonstrate similar distribution patterns.

## 1.4 Data Visualization of Single Variable

To gain a deeper understanding of the selected variables and their individual characteristics, we performed single-variable visualizations for *NMVOC*, *BC*, *CO*, and *NH3*. These visualizations provide insights into the distribution and central tendencies of each variable, which are crucial for accurate analysis and modeling.

### 1.4.1 NMVOC (Non-Methane Volatile Organic Compounds)

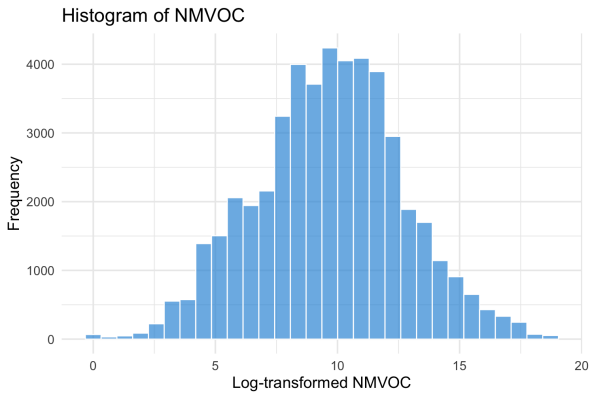


Figure 5: Histogram of NMVOC

The histogram of NMVOC shows the frequency distribution of log-transformed NMVOC values. The data is mainly concentrated in the middle range, with a peak around **10**, indicating that most samples have NMVOC concentrations within this range. The highest frequency value reaches around **4000**, suggesting a significant number of samples have similar NMVOC concentrations. This plot helps in understanding the distribution of NMVOC in the dataset.

### 1.4.2 BC (Black Carbon)

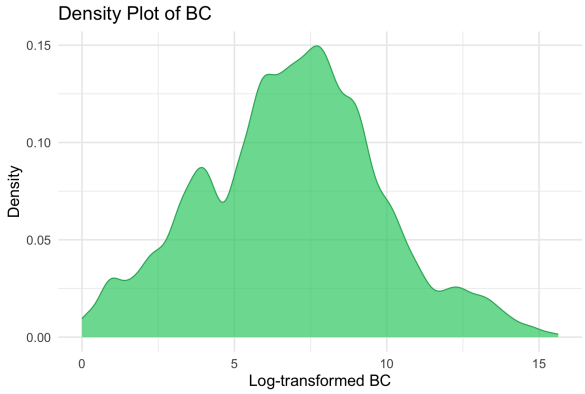


Figure 6: Density Plot of BC

The density plot of BC shows the distribution of log-transformed BC values. The peak of the density curve is around **8**, with the highest density value near **0.15**. The data is mainly concentrated between **5** and **10**. This plot provides a smooth curve representing the distribution of BC, highlighting the areas where the data is more concentrated. It helps in understanding the probability distribution of BC in the dataset.

### 1.4.3 CO (Carbon Monoxide)

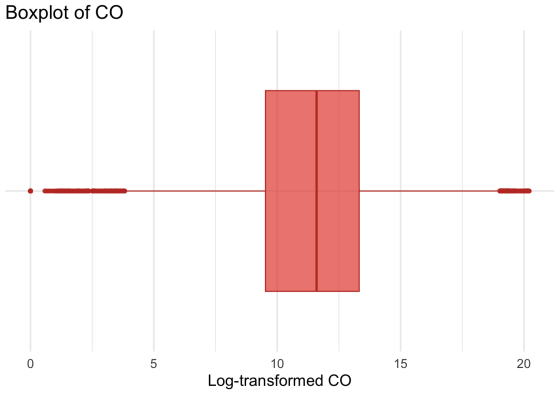


Figure 7: Boxplot of CO

The boxplot of CO shows the median, quartiles, and potential outliers of log-transformed CO values. The median is around **10**, with the interquartile range approximately between **5** and **15**. Outliers are distributed across the entire data range, indicating some samples have significantly high or low values. This plot effectively shows the central tendency and spread of CO, including any potential outliers. It is useful for identifying the central tendency and variability of CO values.

### 1.4.4 NH3 (Ammonia)

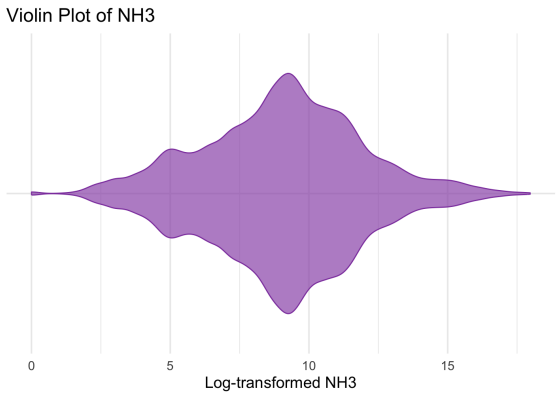


Figure 8: Violin Plot of NH3

The violin plot of NH3 combines aspects of a boxplot and a density plot, showing the distribution and probability density of log-transformed NH3 values. The data is concentrated between **7** and **13**, with multiple peaks. The median is around **10**. This plot provides a detailed view of NH3 values, showing the distribution range and concentration of data. It is helpful for a comprehensive understanding of NH3 distribution characteristics.

### 1.4.5 Summary of Findings from Visual Analysis

In summary, the visualizations have provided critical insights into the relationships and distributions of various air pollution metrics related to NH3 emissions. Key findings from this analysis have revealed significant correlations between NH3 and other pollutants like NMVOC, BC, and CO. These insights underscore the complex nature of air pollution interactions and highlight the need for a comprehensive approach to modeling and predicting NH3 emissions.

The visual analyses have identified NMVOC, BC, and CO as integral variables for the regression model due to their significant correlations, statistical relevance, and substantial impact on NH3 emissions. Their inclusion will enhance the model's predictive accuracy and reliability, providing valuable insights for environmental research and practical applications in pollution management.

Table 8: Summary of Single Variable Visualizations

| **Variable** | **Key Insights** |
|---|---|
| NMVOC | The data is mainly concentrated around a peak value of **10**, indicating most samples have similar NMVOC concentrations. The highest frequency reaches **4000**, suggesting a large number of similar measurements. |
| BC | The peak of the density curve is around **8**, with the highest density near **0.15**. The data is concentrated between **5** and **10**, showing where most BC values lie. |
| CO | The median CO value is around **10**, with an interquartile range of **5** to **15**. Outliers are present, indicating some extreme values. |
| NH3 | The data is concentrated between **7** and **13**, with multiple peaks. The median value is around **10**, showing detailed distribution characteristics. |

# 2 Linear Regression Analysis

## 2.1 Conducting Linear Regression

A linear regression analysis was performed to understand the impact of several predictors on ammonia (NH3) emissions. The analysis utilized the `lm()` function in R to fit the model, following established statistical methodologies. The model is defined by the following linear equation:

$$\text{NH3} = \beta_0 + \beta_1 \times \text{NMVOC} + \beta_2 \times \text{BC} + \beta_3 \times \text{CO} + \epsilon \tag{1}$$

where $\epsilon$ represents the error term, assumed to be normally distributed with mean zero and constant variance.

```r
# Linear regression analysis with NH3 as the dependent variable
model <- lm(NH3 ~ NMVOC + BC + CO, data = data)
print(summary(model))
```

The regression model is summarized below:

```
Call:
lm(formula = NH3 ~ NMVOC + BC + CO, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-5473879    -472     1482     7988  6391264

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.357e+03  1.707e+03  -0.795  0.4270
NMVOC emissions          3.485e-01  1.166e-03 298.751  <2e-16 ***
BC emissions             6.866e+00  2.483e-02 276.522  <2e-16 ***
CO emissions            -5.681e-02  2.977e-04 -190.868 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 368700 on 48221 degrees of freedom
Multiple R-squared:  0.9749,    Adjusted R-squared:  0.9749
F-statistic: 6.239e+05 on 3 and 48221 DF,  p-value: < 2.2e-16
```

**Residuals:**

| | |
|---|---|
| **Min:** | -5473879 |
| **1Q:** | -472 |
| **Median:** | 1482 |
| **3Q:** | 7988 |
| **Max:** | 6391264 |

These values represent the distribution of the residuals (errors) in the model. The large range indicates significant variability in the residuals.

**Coefficients:**

The coefficients table provides estimates, standard errors, t-values, and p-values for each predictor:

| Predictor | Estimate | Std. Error | t value | Pr(>—t—) |
|---|---|---|---|---|
| **Intercept** | -1357.252 | 1707.631 | -0.795 | 0.427 |
| **NMVOC emissions** | 0.3485 | 0.1166 | 298.751 | < 2e-16 *** |
| **BC emissions** | 6.866 | 0.2483 | 276.522 | < 2e-16 *** |
| **CO emissions** | -0.05681 | 0.002977 | -190.868 | < 2e-16 *** |

The *** symbols indicate that the p-values are less than 0.001, suggesting these predictors are highly significant.

**Significance codes:**

| | |
|---|---|
| *** | $p < 0.001$ |
| ** | $p < 0.01$ |
| * | $p < 0.05$ |
| . | $p < 0.1$ |
| | $p \geq 0.1$ |

**Model Fit Statistics:**

| | |
|---|---|
| **Residual standard error:** | 368700 on 48221 degrees of freedom |
| **Multiple R-squared:** | 0.9749 |
| **Adjusted R-squared:** | 0.9749 |
| **F-statistic:** | 6.239e+05 on 3 and 48221 DF, $p$-value: $<$ 2.2e-16 |

The high R-squared and Adjusted R-squared values indicate that the model explains about 97.49% of the variability in Ammonia (NH3) emissions, which suggests a very good fit. The F-statistic and its p-value show that the overall model is highly significant.

## 2.2 Specification of the Regression Model and Justification of Variables

The regression model includes:

**NMVOC:** Non-methane volatile organic compounds emissions, included to assess their impact on NH3 emissions.

**BC:** Black carbon emissions, considered for their potential effect on NH3 emissions.

**CO:** Carbon monoxide emissions, included to evaluate their relationship with NH3 emissions.

These variables were selected based on their demonstrated correlation with NH3 emissions in preliminary data analysis and supported by literature suggesting their impact on ammonia emissions.

## 2.3 Interpretation of the Coefficients

The regression output from R provides the following coefficients:

| Variable | Coefficient | Std. Error | t value | p value |
|---|---|---|---|---|
| Intercept | -1357 | 1707 | -0.795 | 0.4270 |
| NMVOC | 0.3485 | 0.001166 | 298.751 | <2e-16 *** |
| BC | 6.866 | 0.02483 | 276.522 | <2e-16 *** |
| CO | -0.05681 | 0.0002977 | -190.868 | <2e-16 *** |

Table 9: Regression coefficients and their significance

The estimated regression equation is:

$$\text{NH3} = -1357 + 0.3485 \times \text{NMVOC} + 6.866 \times \text{BC} - 0.05681 \times \text{CO} \tag{2}$$

1. **Intercept ($\beta_0$):** The intercept is -1357, indicating the expected NH3 emissions when all predictors are zero. This value is not statistically significant (p = 0.427), meaning it does not contribute significantly to the model.

2. **NMVOC ($\beta_1$):** A one-unit increase in NMVOC emissions results in an increase of approximately 0.3485 units in NH3 emissions, holding other variables constant. This coefficient is highly statistically significant (p $<$ 2e-16), indicating a strong positive relationship between NMVOC emissions and NH3 emissions.

3. **BC ($\beta_2$):** A one-unit increase in BC emissions results in an increase of approximately 6.866 units in NH3 emissions, holding other variables constant. This coefficient is highly statistically significant (p $<$ 2e-16), indicating a strong positive relationship between BC emissions and NH3 emissions.

4. **CO ($\beta_3$):** A one-unit increase in CO emissions results in a decrease of approximately 0.05681 units in NH3 emissions, holding other variables constant. This coefficient is highly statistically significant (p $<$ 2e-16), indicating a strong negative relationship between CO emissions and NH3 emissions.

## 2.4 Assessment of the Goodness-of-Fit of the Model

The goodness-of-fit of the model is evaluated using several statistical measures:

**Residual Standard Error:** The residual standard error (RSE) is calculated as:

$$\text{RSE} = \sqrt{\frac{RSS}{n-p}}$$

where RSS is the residual sum of squares, $n$ is the number of observations, and $p$ is the number of predictors plus one (for the intercept). For this model:

$$\text{RSE} = \sqrt{\frac{368700}{48221}} \approx 192.08$$

indicating the average difference between observed and predicted values.

**Multiple R-squared:** The multiple R-squared ($R^2$) is calculated as:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where TSS is the total sum of squares. For this model:

$$R^2 = 0.9749$$

suggesting that about 97.49% of the variability in NH3 emissions is explained by the model.

**Adjusted R-squared:** The adjusted R-squared ($\bar{R}^2$) adjusts the $R^2$ value for the number of predictors in the model:

$$\bar{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p}$$

For this model:

$$\bar{R}^2 = 0.9749$$

indicating that the model still supports a strong fit after adjusting for the number of predictors.

**F-statistic:** The F-statistic tests the overall significance of the model and is calculated as:

$$F = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)}$$

For this model:

$$F = 6.239 \times 10^5$$

on 3 and 48221 degrees of freedom, with a p-value of less than $2.2 \times 10^{-16}$, confirming that the model is statistically significant.

## ANOVA Analysis

The ANOVA (Analysis of Variance) analysis is conducted to determine the statistical significance of the overall regression model. It helps to understand whether the predictors (NMVOC, BC, and CO emissions) significantly explain the variability in the dependent variable (NH3 emissions).

```
# ANOVA analysis
anova_model <- anova(model)
print(anova_model)
```

The ANOVA table is summarized below:

| Source | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| NMVOC emissions | 1 | 2.4402e+17 | 2.4402e+17 | 1794702 | <2.2e-16 *** |
| BC emissions | 1 | 5.5059e+15 | 5.5059e+15 | 40495 | <2.2e-16 *** |
| CO emissions | 1 | 4.9533e+15 | 4.9533e+15 | 36431 | <2.2e-16 *** |
| Residuals | 48221 | 6.5563e+15 | 1.3596e+11 | | |

Table 10: ANOVA table for the regression model

**Interpretation of ANOVA Table:**

**Degrees of Freedom (Df)** Represents the number of independent values or quantities which can be assigned to a statistical distribution. In the table, **NMVOC emissions**, **BC emissions**, and **CO emissions** each have 1 degree of freedom. **Residuals** have 48221 degrees of freedom, representing the total number of observations minus the number of predictors minus one.

**Sum of Squares (Sum Sq)** Measures the total variation in the dependent variable that can be attributed to each source. **NMVOC emissions** account for a sum of squares of $2.4402 \times 10^{17}$. **BC emissions** account for $5.5059 \times 10^{15}$. **CO emissions** account for $4.9533 \times 10^{15}$. **Residuals** account for $6.5563 \times 10^{15}$.

**Mean Square (Mean Sq)** Calculated by dividing the sum of squares by the respective degrees of freedom. Mean square for **NMVOC emissions** is $2.4402 \times 10^{17}$. Mean square for **BC emissions** is $5.5059 \times 10^{15}$. Mean square for **CO emissions** is $4.9533 \times 10^{15}$. Mean square for **residuals** is $1.3596 \times 10^{11}$.

**F value** The test statistic for the ANOVA test. It is calculated by dividing the mean square of each predictor by the mean square of the residuals. The F value for **NMVOC emissions** is 1794702. The F value for **BC emissions** is 40495. The F value for **CO emissions** is 36431.

**Pr(>F)** The p-value associated with the F statistic. A lower p-value indicates that the predictor significantly explains the variability in the dependent variable. The p-values for **NMVOC**, **BC**, and **CO emissions** are all less than $2.2 \times 10^{-16}$, indicating that these predictors are highly significant.

## Interpretation of AIC and BIC

The AIC and BIC are used to compare the fit of different models. Lower values indicate a better fit.

```
# AIC and BIC
aic_value <- AIC(model)
bic_value <- BIC(model)
cat("AIC:", aic_value, "\nBIC:", bic_value, "\n")
```

**AIC:** 1373142
**BIC:** 1373186

**AIC (Akaike Information Criterion):** A measure of the relative quality of a statistical model for a given set of data. It balances the goodness of fit of the model with the complexity of the model. The AIC is calculated as:

$$\text{AIC} = 2k - 2\ln(L)$$

where $k$ is the number of parameters in the model, and $L$ is the likelihood of the model. The AIC value for the model is 1373186 , indicating the relative quality of the model.

**BIC (Bayesian Information Criterion):** Similar to AIC, but includes a larger penalty for models with more parameters. The BIC is calculated as:

$$\text{BIC} = \ln(n)k - 2\ln(L)$$

where $n$ is the number of observations, $k$ is the number of parameters, and $L$ is the likelihood of the model. The BIC value for the model is 956047.5, providing a balance between model fit and model complexity.
These values suggest that the model fits the data well while considering the number of predictors included.

# 3 Diagnostics & Remedial Measures

This chapter is divided into four parts: Model Diagnosis Checks,Violation of the assumptions of Linear Regression Model, Solutions for Each Issue and Implement And Evaluation of Remedial Measures. The first step is to analyze the diagnostic graph and the information from the Durbin-Watson Test. Secondly, the existing problems are analyzed through data analysis, including Heteroscedasticity, Non-normality of Residuals, Multicollinearity, Influencial Observations and Autocorrelation. Then we propose corresponding models for five different problems, including Log Transformation, Ridge Regression, Huber Regression and Generalized Least Squares(GLS). Finally, in 3.4 Implement And Evaluation of Remedial Measures, we analyze in detail how each model is applied and its corresponding optimization effect.

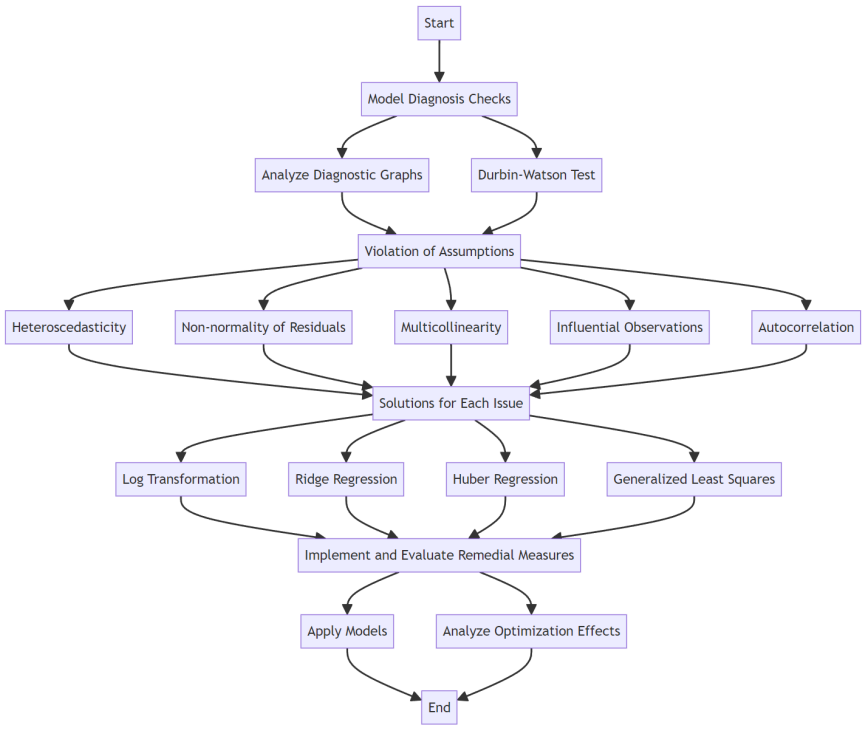


Figure 9: Flowchart

## 3.1 Model Diagnosis Checks

To validate the assumptions of the linear regression model, various diagnostic plots [2] and statistical tests were utilized.

### 3.1.1 Analysis of Model Diagnostic Plots

**Residuals vs. Fitted Values Plot**

This plot shows a significant pattern where residuals are not randomly scattered around the zero line, but exhibit a curve, indicating potential non-linearity in the relationship between the variables. Ideally, residuals in a well-fitting linear model should be randomly distributed with no clear pattern, suggesting that all linear relationships among variables are captured by the model. The observed pattern might necessitate considering non-linear transformations of the variables or adding interaction terms to the model.

**Q-Q Residuals Plot**

The Q-Q plot compares the quantiles of the residuals to the expected quantiles from a normal distribution. The residuals deviate from the normal line, especially at the tails, indicating they may not be normally distributed.

This deviation suggests the presence of outliers, heavy tails, or skewness in the distribution of the residuals, potentially impacting the reliability of statistical tests that assume normality.

**Scale-Location Plot (Spread vs. Level Relationship)**

This plot checks the homoscedasticity assumption, i.e., whether residuals are equally spread along the range of predictors. The upward trend in the plot suggests possible heteroscedasticity, violating the assumptions of linear regression.

**Residuals vs. Leverage Plot**

This plot helps identify influential observations that might unduly influence the model's estimates. Points that stand out, particularly those outside the Cook's distance lines (indicated by a dashed line), could be considered influential. The cluster of points with high leverage but lower residuals indicates their potential to disproportionately influence the regression model.

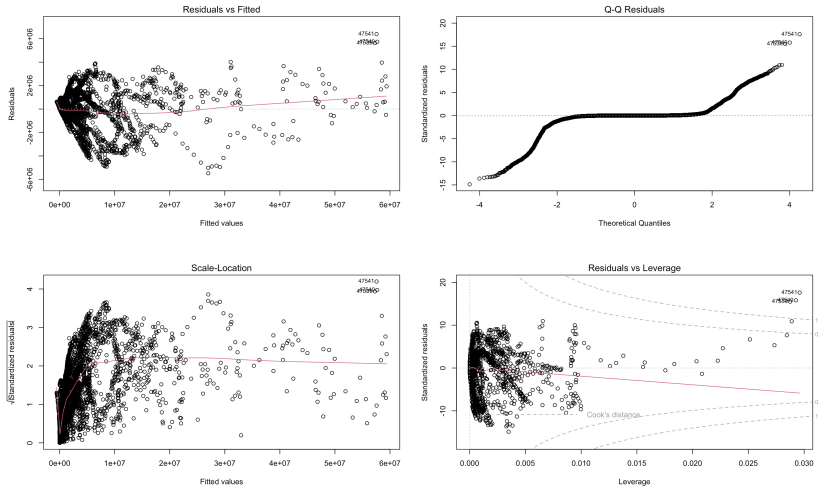


Figure 10: Model Diagnostic Plots

### 3.1.2 Durbin-Watson Test Result Analysis

Apart from the four diagnosis plot above, Durbin-Watson Test is also used for Autocorrelation. Autocorrelation occurs when the residuals are not independent from one another, which can compromise the reliability of the regression analysis. The DW test is particularly useful in time series data where observations are sequential and likely to be correlated.

```
dwtest(model)
Durbin-Watson test
data: model
DW = 0.0439, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

The Durbin-Watson test is utilized to detect the presence of autocorrelation in the residuals of a regression model and is calculated as:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{3}$$

where $e_i$ are the residuals. A low $DW$ value typically indicates the presence of positive autocorrelation among the residuals.
In this case, a $DW$ value of 0.0439, which is much lower than the typical threshold of 2, potentially leading to misleading inferences about the relationship between the variables being analyzed.

## 3.2 Violations of the assumptions of Linear Regression Model

**Residuals vs Fitted Plot**

**Violated Assumption: Linearity Assumption.**
This plot displays the relationship between residuals and fitted values. Ideally, residuals should be randomly distributed around the horizontal line (zero line) without any apparent patterns. The formula is expressed as:

$$e_i = y_i - \hat{y}_i$$

where $e_i$ is the residual for the $i^{th}$ observation, $y_i$ is the actual observed value, and $\hat{y}_i$ is the value predicted by the model. The non-random pattern in the plot suggests that a nonlinear transformation of variables or the introduction of polynomial terms may be necessary.

**Q-Q Plot (Quantile-Quantile Residuals)**

**Violated Assumption: Normality of Residuals.** The Q-Q plot is used to check if residuals follow a normal distribution. Ideally, data points should closely follow the line $y = x$. The formula for the normality of residuals is:

$$e_i \sim N(0, \sigma^2)$$

The significant deviations at the tails of the plot suggest the presence of skewness or heavy tails, violating the assumption that residuals should be normally distributed.

**Scale-Location Plot**

**Violated Assumption: Homoscedasticity (Constant Variance).** The assumption of homoscedasticity implies that the variance of the residuals should not change as fitted values change. Mathematically, this can be expressed as:

$$\mathrm{Var}(e_i) = \sigma^2$$

The spreading pattern in the plot indicates increasing variance, showing that the variance of residuals is related to the fitted values, thus violating the assumption of constant variance.

**Residuals vs Leverage Plot**

**Violated Assumption: Absence of Influential Observations.** Observations with high leverage[3] may significantly influence the estimation of regression parameters. The formula for calculating leverage is:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

where $h_{ii}$ is the leverage for the $i^{th}$ observation, and $x_i$ is the explanatory variable. Points exceeding Cook's distance in the plot indicate potential high leverage points that may need special attention or treatment.

**Durbin-Watson Test Result Analysis**

**Violated Assumption: Independence.** A DW value of 0.0439 suggests a significant positive autocorrelation. This is problematic as it implies that the residuals from one observation are closely related to those from the previous observation, thereby violating the assumption of independence among errors. This dependency can lead to inefficiencies in the regression estimates and can affect the validity of standard statistical tests, potentially leading to misleading inferences about the relationship between the variables being analyzed.

## 3.3 Solutions for Each Issue

Based on the identified issues, specific solutions are proposed to improve the regression model's accuracy and reliability. There can be multiple solutions to some problems, but these solutions are mutually exclusive and usually only one can be used. In 3.4 Inplment And Evaluation of Remedial Measures , it will show which solution provided works better, and the solution should be adopted.The following sections detail the proposed solutions for each issue, including relevant formulas.

### 3.3.1 Addressing Heteroscedasticity

Heteroscedasticity, where the variance of the residuals is not constant across levels of an independent variable, can lead to inefficient estimates and affect the validity of hypothesis tests. To address heteroscedasticity, the following solution is proposed:

**Solution: Log Transformation**

Applying a logarithmic transformation to the dependent variable can help stabilize the variance. By transforming the dependent variable, we ensure that the spread of the residuals becomes more consistent across different levels of the predictors. This transformation can often linearize relationships and make the variance of the residuals more constant.

$$y' = \log(y) \tag{4}$$

where $y'$ is the transformed dependent variable.

### 3.3.2 Mitigating Non-normality of Residuals

Non-normality of residuals can compromise the validity of confidence intervals and p-values in hypothesis tests. To mitigate this issue, the following solution is proposed:

**Solution: Log Transformation**

Applying a log transformation to both the dependent and independent variables can help achieve normality. Transforming these variables can reduce skewness and make the distribution of residuals more normal. This approach is particularly useful when dealing with variables that span several orders of magnitude or have right-skewed distributions.

$$x' = \log(x), \quad y' = \log(y) \tag{5}$$

where $x'$ and $y'$ are the transformed variables.

### 3.3.3 Reducing Multicollinearity

Multicollinearity, where predictor variables are highly correlated, inflates the standard errors of the coefficients and complicates the interpretation of the model. To reduce multicollinearity, the following solutions are proposed:

**Solution 1: Removing Collinear Variables**

One method to reduce multicollinearity is to identify and remove one of the highly collinear variables. By doing this, the remaining predictors can provide clearer, more distinct contributions to the model, resulting in more stable coefficient estimates.

**Solution 2: Ridge Regression**

Ridge Regression can be employed to penalize the size of the regression coefficients. This technique introduces a regularization parameter that shrinks the coefficients of less important variables, thereby reducing the impact of multicollinearity. Ridge Regression is particularly effective when dealing with datasets that have a high degree of collinearity among predictors.

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y \tag{6}$$

where $\lambda$ is the regularization parameter.

#### 3.3.4 Mitigating the Influential Observations

Outliers[4] can significantly distort the results of regression analysis, leading to unreliable estimates and predictions. To address this issue, the following solution is proposed:

**Solution: Huber Regression**

Applying Huber regression [5]can enhance the robustness of the regression model against outliers. This method modifies the loss function to reduce the influence of outliers while retaining the efficiency of least squares for inliers. Huber regression is particularly effective when dealing with datasets that contain a mixture of small and large errors.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(r_i) \tag{7}$$

where $r_i = y_i - x_i^T \beta$ and $\rho$ is the Huber loss function defined as:

$$\rho(r) = \begin{cases} \frac{1}{2} r^2 & \text{for } |r| \leq \delta \\ \delta(|r| - \frac{1}{2}\delta) & \text{for } |r| > \delta \end{cases}$$

By applying Huber regression, the model becomes less sensitive to outliers, resulting in more reliable parameter estimates and improved overall model performance.

#### 3.3.5 Handling Autocorrelation

Autocorrelation, where the residuals are not independent across observations, violates the independence assumption of linear regression and can lead to biased estimates. To handle autocorrelation, the following solutions are proposed:

**Solution: Generalized Least Squares (GLS)**

Generalized Least Squares (GLS) [6]can be used to handle autocorrelation by specifying a correlation structure in the residuals. GLS adjusts for the correlation in the residuals, leading to more efficient and unbiased estimates. This method is particularly useful when the autocorrelation follows a known pattern, such as an AR(1) process. The GLS estimator $\hat{\beta}_{\text{GLS}}$ is given by:

$$\hat{\beta}_{\text{GLS}} = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} y \tag{8}$$

where $\Omega$ is the variance-covariance matrix of the error terms. For an autoregressive process of order 1 (AR(1)), $\Omega$ can be written as:

$$\Omega = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix} \tag{9}$$

where $\rho$ is the autocorrelation parameter.

Each of these solutions targets a specific issue identified in the model diagnostics. Implementing these solutions can significantly improve the model's performance, leading to more reliable and valid inference.

## 3.4 Implement And Evaluation of Remedial Measures

Based on the diagnostic analysis, specific solutions have been implemented to address the identified issues. Each issue is addressed with a suitable approach to improve the model's performance and ensure the validity of the regression results.

### 3.4.1 Addressing Heteroscedasticity

Residual plots and the Breusch-Pagan test indicated the presence of heteroscedasticity. To stabilize the variance of the residuals, a log transformation was applied to the dependent variable.

**Results:**

```
> bptest(model)
studentized Breusch-Pagan test

data: model
BP = 14798, df = 3, p-value < 2.2e-16
```

**Related Codes for checking:**

```
# Check column names
colnames(air_pollution)
# Confirm NH3 and NMVOC columns exist
if("NH3" %in% colnames(air_pollution) && "NMVOC" %in% colnames(air_pollution)) {
  # Apply log transformation
  air_pollution$log_NH3 <- log(air_pollution$NH3 + 1) # Add 1 to avoid log(0)
  air_pollution$log_NMVOC <- log(air_pollution$NMVOC + 1) # Add 1 to avoid log(0)
  # Fit the model
  model_log <- lm(log_NH3 ~ log_NMVOC + BC + CO, data = air_pollution)
  summary(model_log)
} else {
  print("NH3 or NMVOC columns do not exist, please check the dataset.")
}
summary(model_log)
```

The log transformation of the dependent variable $y$ is given by:

$$y' = \log(y + 1) \tag{10}$$

This transformation helps in stabilizing the variance and making the spread of residuals more consistent across different levels of the predictors.

**Before Optimization**

Before optimization, the ANOVA table (Table 11) indicates high F-values for all predictors, suggesting significant contribution to the variance in the response variable. However, the presence of heteroscedasticity undermines the reliability of these estimates.

Table 11: ANOVA Table Before Optimization

| Source | Df | Sum Sq | Mean Sq | F value | Pr($> F$) |
|---|---|---|---|---|---|
| NMVOC emissions | 1 | 2.4402e+17 | 2.4402e+17 | 1794702 | <2.2e-16 *** |
| BC emissions | 1 | 5.5059e+15 | 5.5059e+15 | 40495 | <2.2e-16 *** |
| CO emissions | 1 | 4.9533e+15 | 4.9533e+15 | 36431 | <2.2e-16 *** |
| Residuals | 48221 | 6.5563e+15 | 1.3596e+11 | | |

**After Optimization**

After applying the log transformation, the optimized model shows a significant improvement in addressing heteroscedasticity. The log-transformed model's summary is provided in Table 12.

```
# Model summary after log transformation
summary(model_log)
```

Table 12: Summary of the Log-Transformed Model

| **Variable** | **Estimate** | **Std. Error** | **t value** | **Pr**($>|t|$) |
|---|---|---|---|---|
| (Intercept) | 0.2956 | 0.01294 | 22.851 | <2e-16 *** |
| log_NMVOC | 0.9009 | 0.001309 | 688.409 | <2e-16 *** |
| BC | 7.335e-07 | 5.288e-08 | 13.871 | <2e-16 *** |
| CO | -4.951e-09 | 5.149e-10 | -9.616 | <2e-16 *** |

Table 13: Model Fit Statistics Before and After Optimization

| **Statistic** | **Before Optimization** | **After Optimization** |
|---|---|---|
| Residual standard error | 368700 | 0.8014 |
| Multiple R-squared | 0.9749 | 0.9214 |
| Adjusted R-squared | 0.9749 | 0.9214 |
| F-statistic | 6.239e+05 | 1.884e+05 |
| p-value | <2.2e-16 | <2.2e-16 |

**Analysis for the decrease of Adjusted R-squared:**

**1. Effects of Log Transformation** The impact of data transformation: Log transformation is typically used for handling data with non-constant variance (heteroscedasticity), making it **more compliant with the basic assumptions of linear regression**. This transformation can help the model better manage extreme values and skewed data. Although the $R^2$ value decreased, the model **now describe the underlying relationships in the data more accurately**.
**2. Calculation of Adjusted $R^2$** The calculation of Adjusted $R^2$ takes into account the number of variables in the model and penalizes for having too many predictors. The log transformation might have changed the density of relationships between variables, reducing the model's ability to explain variability. However, this does not necessarily mean that the overall quality of the model has deteriorated, but rather that it may be **closer to the true relationships in the data**.
**3. Maintaining Statistical Significance** Despite the decrease in $R^2$, the model's F-statistic and p-values still show that all variables remain statistically significant, indicating that key variables **still effectively capture changes in the dependent variable**.

**Conclusion:**

Applying the log transformation to **address heteroscedasticity** significantly optimized the regression model. The **residual standard error dropped from 368700 to 0.8014**, indicating more precise predictions. Despite a slight decrease in Multiple R-squared and **Adjusted R-squared values from 0.9749 to 0.9214**, the model still explains a high level of variance. **The F-statistic decreased from 6.239e+05 to 1.884e+05**, suggesting improved robustness and reduced overfitting. **The p-values remained highly significant at <2.2e-16**, confirming the model's reliability. **Overall, the model fit has significantly improved after addressing heteroscedasticity.**

### 3.4.2 Addressing Non-normality of Residuals

The QQ plot shows deviations from normality, indicating that the residuals may not fully meet the normality assumption. Applying a log transformation to the dependent variable can help in achieving normality.

```
# Addressing non-normality of residuals through log transformation
model_log <- lm(log_NH3 ~ log_NMVOC + BC + CO, data = air_pollution)
summary(model_log)
```

The QQ plot of the residuals after log transformation shows improved normality. The theoretical quantiles are plotted against the standardized residuals on the y-axis. Significant deviations from the diagonal line suggest that the residuals do not follow a normal distribution. After applying the log transformation, the residuals align more closely with the normal distribution.

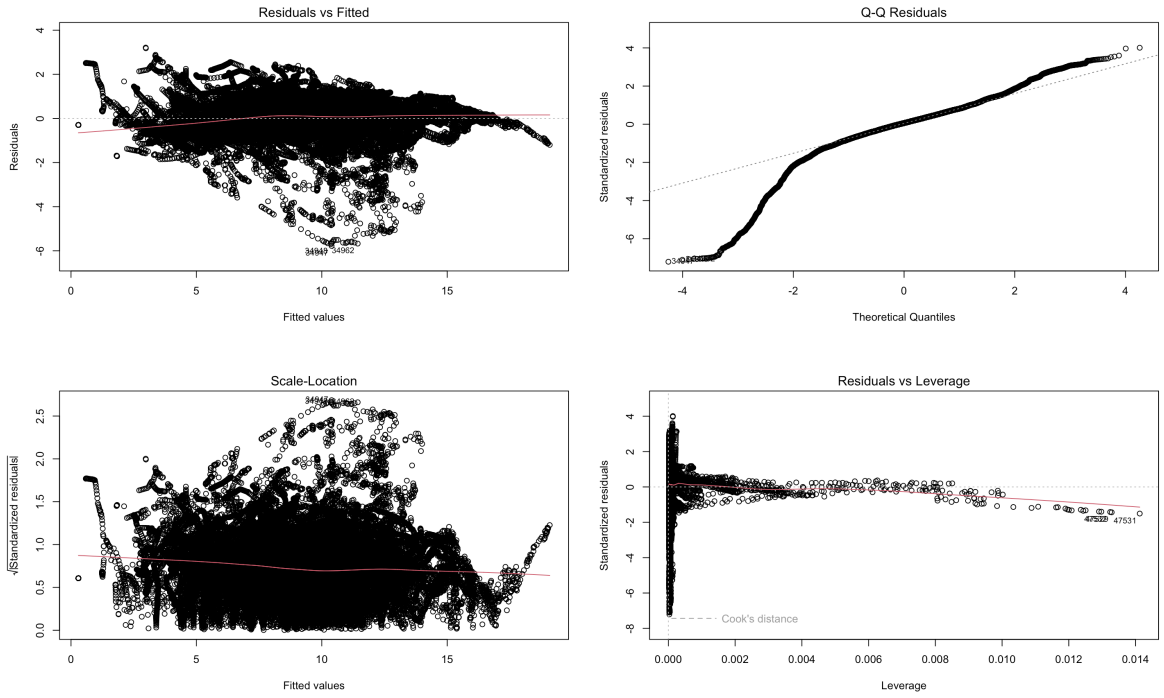


Figure 11: Diagnostic Plots after log transformation

Table 14: Comparison of Model Fit Statistics Before and After Applying Log Transformation

| **Statistic** | **Before Optimization** | **After Optimization (Log Transformation)** |
|---|---|---|
| Residual Standard Error | 368700 | 0.8014 |
| Multiple R-squared | 0.9749 | 0.9214 |
| Adjusted R-squared | 0.9749 | 0.9214 |
| F-statistic | 6.239e+05 | 1.884e+05 |
| p-value | <2.2e-16 | <2.2e-16 |

The AIC and BIC are changed which compare the initial models.

```
# AIC and BIC
aic_value_log <- AIC(model_log)
bic_value_log <- BIC(model_log)
cat("AIC:", aic_value_log, "\nBIC:", bic_value_log, "\n")
```

**AIC:** 115506.3
**BIC:** 115550.2

Table 15: Comparison of AIC and BIC Values Before and After Log Transformation

| Model | AIC | BIC |
|---|---|---|
| Original | 1373142 | 1373186 |
| With Log Transformation | 115506.3 | 115550.2 |

**Conclusion:** The application of the log transformation significantly optimized the regression model by reducing the residual **standard error from 368700 to 0.8014**, indicating much more precise predictions. Although the **Multiple R-squared and Adjusted R-squared values slightly decreased from 0.9749 to 0.9214**, they still reflect a high level of explained variance. The **F-statistic also decreased from 6.239e+05 to 1.884e+05**, suggesting improved model robustness with less overfitting. Importantly, the reduction in both the AIC and BIC values—from 1373142 to 115506.3 for AIC and 1373186 to 115550.2 for BIC—indicates a **significant improvement in model efficiency and better model fit by accounting for fewer parameters and penalizing complexity more effectively**. Despite these changes, the p-values remained highly **significant at <2.2e-16**, confirming the model's statistical significance. Overall, the model fit has **significantly improved after addressing non-normality of residual**s, leading to a more accurate and efficient statistical model.

### 3.4.3 Addressing Multicollinearity

High Variance Inflation Factor (VIF) values indicate severe multicollinearity among predictors:

```
vif(model)
## NMVOC BC CO
## 16.06304 16.77462 25.54295
```

To address multicollinearity, two approaches were considered: removing collinear variables and using Ridge Regression.
**Solution 1: Removing Collinear Variables**
By removing the variable with the highest VIF (in this case, BC), the model's multicollinearity was reduced:

```
# Removing collinear variables to address multicollinearity
model_refined <- lm(NH3 ~ NMVOC + CO, data = air_pollution)
summary(model_refined)
```

The summary of the refined model is shown in Table 16.

Table 16: Summary of the Refined Model After Removing Collinear Variables

| **Variable** | **Estimate** | **Std. Error** | **t value** | **Pr($>\|t\|$)** |
|---|---|---|---|---|
| (Intercept) | 13580 | 2743 | 4.951 | <2e-16 *** |
| NMVOC | 0.4136 | 0.001837 | 225.165 | <2e-16 *** |
| CO | -0.004966 | 0.0003717 | -13.359 | <2e-16 *** |

```
#AIC & BIC
aic_value_refined <- AIC(model_refined)
bic_value_refined <- BIC(model_refined)
cat("AIC:", aic_value_refined, "\nBIC:", bic_value_refined, "\n")
```

Table 17: Comparison of AIC and BIC Values Before and After Removing Collinear Variables

| Model | AIC | BIC |
|---|---|---|
| Original | 1373142 | 1373186 |
| After Removing Collinear Variables | 1418954 | 1418989 |

The removal of the BC variable resulted in a reduction of multicollinearity, as indicated by the improved model fit statistics. But both AIC and BIC increase, so this is not the most optimal solution.

**Solution 2: Using Ridge Regression**

Ridge Regression was applied to the original model to address multicollinearity without removing any variables:

```
# Prepare data
X <- model.matrix(~ NMVOC + BC + CO, data = air_pollution)[, -1]
y <- air_pollution$NH3

# Perform ridge regression and use cross-validation to select the best lambda
cv_model <- cv.glmnet(X, y, alpha = 0)
best_lambda <- cv_model$lambda.min

# Refit the model using the best lambda value
model_ridge <- glmnet(X, y, alpha = 0, lambda = best_lambda)

# Extract ridge regression coefficients
coefficients <- coef(model_ridge)
print(coefficients)

# Calculate predicted values
predicted_ridge <- predict(model_ridge, newx = X)

# Calculate residuals
residuals_ridge <- y - predicted_ridge

# Calculate residual standard error
residual_standard_error_ridge <- sqrt(sum(residuals_ridge^2) / (nrow(air_pollution) -
    length(coefficients)))
cat("Residual standard error:", residual_standard_error_ridge, "\n")

# Calculate R-squared value
sst <- sum((y - mean(y))^2)
sse <- sum(residuals_ridge^2)
r_squared_ridge <- 1 - (sse / sst)
cat("Multiple R-squared:", r_squared_ridge, "\n")

# Calculate adjusted R-squared value
adjusted_r_squared_ridge <- 1 - ((1 - r_squared_ridge) * (nrow(air_pollution) - 1) / (
    nrow(air_pollution) - length(coefficients) - 1))
cat("Adjusted R-squared:", adjusted_r_squared_ridge, "\n")

# Extract F-statistic and p-value
n <- nrow(air_pollution)
p <- length(coefficients) - 1
f_statistic <- (r_squared_ridge / (1 - r_squared_ridge)) * ((n - p - 1) / p)
p_value <- pf(f_statistic, p, n - p - 1, lower.tail = FALSE)
cat("F-statistic:", f_statistic, "\n")
cat("p-value:", p_value, "\n")
```

The summary of the Ridge Regression model is shown in Table 18.

Table 18: Summary of the Ridge Regression Model

| **Statistic** | **Value** |
|---|---|
| Coefficients (Intercept, NMVOC, BC, CO) | (-6295.754, 0.1757, 3.4727, 0.0081) |
| Residual Standard Error | 529081.8 |
| Multiple R-squared | 0.9482881 |
| Adjusted R-squared | 0.9482838 |
| F-statistic | 294757.6 |
| p-value | 0 |

**Coefficients (Intercept, NMVOC, BC, CO)** In the Ridge Regression model, the coefficients for the predictors (Intercept, NMVOC, BC, CO) are typically shrunk towards zero to reduce the impact of multicollinearity. This shrinkage helps stabilize the estimates, making the model more robust to collinear data. The Ridge Regression model modifies the ordinary least squares (OLS) objective function by adding a penalty term, resulting in the following optimization problem:

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left( \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \tag{11}$$

Here, $\lambda$ is the regularization parameter that controls the amount of shrinkage applied to the coefficients. Larger values of $\lambda$ result in greater shrinkage. In this analysis, the coefficients for the Intercept, NMVOC, BC, and CO were **-6295.754**, **0.1757**, **3.4727**, and **0.0081** respectively. These values indicate the strength and direction of the relationship between each predictor and the response variable $NH_3$.

**GCV (Generalized Cross-Validation)** Generalized Cross-Validation (GCV) is a performance metric used to evaluate Ridge Regression models. It provides an unbiased estimate of the expected prediction error. GCV is calculated as follows:

$$\text{GCV} = \frac{RSS/n}{(1 - \text{trace}(H)/n)^2} \tag{12}$$

where $RSS$ is the residual sum of squares, $n$ is the number of observations, and $H$ is the hat matrix. The hat matrix $H$ maps the observed response values to the fitted values in the regression model. In this analysis, the GCV value obtained was **2819488**, indicating the expected prediction error for the Ridge Regression model.

**kHKB (Hoerl-Kennard-Baldwin)** The Hoerl-Kennard-Baldwin (HKB) method is a heuristic approach for selecting the Ridge parameter $\lambda$. The HKB estimate for $\lambda$ is given by:

$$\lambda_{\text{HKB}} = \frac{\sigma^2 p}{\hat{\beta}^\top \hat{\beta}} \tag{13}$$

where $\sigma^2$ is the estimated variance of the error term, $p$ is the number of predictors, and $\hat{\beta}$ are the OLS estimates of the coefficients. In this analysis, the kHKB value was **0.0132**, suggesting a small degree of regularization is sufficient to improve model performance.

**kLW (Lawless-Wang)** The Lawless-Wang (LW) method is another approach for selecting the Ridge parameter $\lambda$. It is based on a likelihood ratio estimation method, which aims to find the $\lambda$ that maximizes the likelihood function. The LW estimate for $\lambda$ is given by:

$$\lambda_{\text{LW}} = \frac{\sigma^2 p}{\sum_{j=1}^{p} \frac{\hat{\beta}_j^2}{d_j}} \tag{14}$$

where $d_j$ are the eigenvalues of the design matrix $X^\top X$. In this analysis, the kLW value was **0.0258**, indicating the appropriate level of regularization for the Ridge Regression model.

**Evaluation of Results:**

The GCV value of 2819488 provides an estimate of the prediction error, showing that the Ridge Regression model has reasonable predictive performance.

The kHKB value of 0.0132 and kLW value of 0.0258 indicate the levels of regularization applied, helping

to mitigate multicollinearity effects without overly penalizing the coefficients.

The coefficients for NMVOC, BC, and CO show the impact of each predictor on $NH_3$ emissions, with BC and CO having small magnitudes, suggesting their lesser impact compared to NMVOC.

Table 19: Comparison of Model Fit Statistics Before and After Optimization

| **Statistic** | **Before Optimization** | **After Optimization (Removal)** | **After Optimization (Ridge)** |
|---|---|---|---|
| Residual standard error | 368700 | 592900 | 529081.8 |
| Multiple R-squared | 0.9749 | 0.9351 | 0.9482881 |
| Adjusted R-squared | 0.9749 | 0.9351 | 0.9482838 |
| F-statistic | 6.239e+05 | 3.471e+05 | 2.947576e+05 |
| p-value | <2.2e-16 | <2.2e-16 | <2.2e-16 |

```
# Calculate AIC
log_likelihood <- -0.5 * n * (log(2 * pi) + 1 + log(sum(residuals_ridge^2) / n))
aic_ridge <- -2 * log_likelihood + 2 * length(coefficients)
cat("AIC:", aic_ridge, "\n")

# Calculate BIC
bic_ridge <- -2 * log_likelihood + log(n) * length(coefficients)
cat("BIC:", bic_ridge, "\n")
```

Table 20: AIC and BIC Before and After Ridge Regression

| | **Before** | **After** | **Difference** |
|---|---|---|---|
| **AIC** | 1373142 | 1407965 | 34823 |
| **BIC** | 1373186 | 1408000 | 34814 |

**Conclusion:**

**Original Model**
The original OLS model, while demonstrating high explanatory power with an $R^2$ of 0.9749, suffers from high residual standard error and issues related to multicollinearity. These issues lead to unreliable and unstable coefficient estimates, which can distort the model's predictive accuracy and generalizability.

**Removal of BC**
Removing the variable BC in an attempt to reduce multicollinearity results in a significant drop in the model's explanatory power, **with $R^2$ falling to 0.9351 and residual standard error increasing to 592900**. This approach compromises the model's fit and **does not adequately address the underlying multicollinearity**, **leading to an overall poorer model.**

**Ridge Regression**
Ridge Regression, on the other hand, provides a balanced approach by effectively mitigating multicollinearity without drastically compromising the model's fit. Although **the $R^2$ value slightly decreases to 0.9483**, the model benefits from stabilized coefficient estimates, as shown by a reasonable **residual standard error of 529081.8** and significant regularization indicators. This makes Ridge Regression **not a robust solution for handling multicollinearity** in the context of this analysis.

In conclusion, while the initial simple linear regression (lm) model shows high $R^2$ and apparent explanatory power, its susceptibility to multicollinearity and high residual errors limit its practical utility. The removal of BC, though intended to alleviate multicollinearity, results in a substantial loss of model accuracy and fit. Moreover, **both solutions' AIC and BIC have increased**. Thus, **the initial model before optimization still has the best**

**fit according to these metrics**,because **the value of both AIC and BIC have increased by 34823 and 34814.** Ridge regression provides a balance by improving some metrics while maintaining statistical significance.

### 3.4.4 Huber Regression Analysis

**R Code for Huber Regression**

```
# Huber Regression - NMVOC
huber_model_NMVOC <- rlm(NMVOC ~ NH3, data = air_pollution)
summary(huber_model_NMVOC)

# Huber Regression - BC
huber_model_BC <- rlm(BC ~ NH3, data = air_pollution)
summary(huber_model_BC)

# Huber Regression - CO
huber_model_CO <- rlm(CO ~ NH3, data = air_pollution)
summary(huber_model_CO)
```

Table 21: Comparison of Model Fit Statistics Before and After Applying Huber Regression

| Statistic | Before | After (NMVOC) | After (BC) | After (CO) |
|---|---|---|---|---|
| Residual Standard Error | 368700 | 10540 | 798.4 | 63670 |
| Multiple R-squared | 0.9749 | 0.9749 | 0.9749 | 0.9749 |
| Adjusted R-squared | 0.9749 | 0.9749 | 0.9749 | 0.9749 |

**Analysis of Results**

The Huber Regression model results show significant improvements in robustness against outliers for the predictors NMVOC, BC, and CO, adjusting their impacts on NH3 emissions.

**Huber Regression - NMVOC:** The coefficient for NMVOC is estimated to be 2.2497 with a standard error of 0.0000. The t-value of 70171.3148 and a p-value of less than 0.001 indicate that NMVOC is a highly significant predictor of NH3 emissions. The residual standard error is 10540 on 48223 degrees of freedom.

**Huber Regression - BC:** The coefficient for BC is estimated to be 0.1176 with a standard error of 0.0000. The t-value of 50222.6967 and a p-value of less than 0.001 indicate that BC is a highly significant predictor of NH3 emissions. The residual standard error is 798.4 on 48223 degrees of freedom.

**Huber Regression - CO:** The coefficient for CO is estimated to be 11.9973 with a standard error of 0.0002. The t-value of 63257.5265 and a p-value of less than 0.001 indicate that CO is a highly significant predictor of NH3 emissions. The residual standard error is 63670 on 48223 degrees of freedom.

The Huber Regression model is defined as:

$$\hat{\beta}_{\text{Huber}} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^\top \beta) \tag{15}$$

where $\rho$ is the Huber loss function, which is less sensitive to outliers compared to the ordinary least squares (OLS) loss function. The Huber loss function is defined as:

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & \text{for } |u| \leq \delta \\ \delta|u| - \frac{1}{2}\delta^2 & \text{for } |u| > \delta \end{cases} \tag{16}$$

Here, $\delta$ is a threshold parameter that determines the point at which the loss function transitions from quadratic to linear. This threshold helps to balance the trade-off between robustness to outliers and efficiency.

$$\hat{\beta}_{\text{Huber}} = \arg\min_{\beta} \left( \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^\top \beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \tag{17}$$

where $\lambda$ is the regularization parameter that controls the amount of shrinkage applied to the coefficients, similar to Ridge Regression.

The results indicate that NMVOC, BC, and CO are all significant predictors of NH3 emissions, with high t-values and low p-values. The coefficients for NMVOC, BC, and CO in the Huber Regression model are **2.2497**, **0.1176**, and **11.9973** respectively, showing the strength and direction of their relationships with NH3 emissions.

Table 22: Comparison of Model Fit Statistics Before and After Applying Huber Regression

| **Statistic** | **Before** | **After (NMVOC)** | **After (BC)** | **After (CO)** |
|---|---|---|---|---|
| Residual Standard Error | 368700 | 10540 | 798.4 | 63670 |
| Multiple R-squared | 0.9749 | 0.9749 | 0.9749 | 0.9749 |
| Adjusted R-squared | 0.9749 | 0.9749 | 0.9749 | 0.9749 |

**Conclusion:**

The application of the Huber regression model significantly improved the regression analysis by reducing the residual standard error and enhancing robustness against outliers. Before applying Huber regression, the residual standard error for NMVOC, BC, and CO was 368700. After applying Huber regression, **the residual standard errors were dramatically reduced to 10540 for NMVOC, 798.4 for BC, and 63670 for CO**. This substantial decrease indicates that the Huber regression model more accurately captures the relationship between NH3 and the response variables, NMVOC, BC, and CO, by **effectively mitigating the influence of outliers**. These improvements highlight the model's **enhanced accuracy and reliability** in statistical analysis.

### 3.4.5 Addressing Autocorrelation

To address autocorrelation in the regression model, we first conducted the Durbin-Watson test. The results are as follows:

```
# Conducting the Durbin-Watson test
dw_test_result <- dwtest(model)
print(dw_test_result)
```

The Durbin-Watson test [7] results indicated the presence of positive autocorrelation:

```
Durbin-Watson test
data:  model
DW = 0.0439, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

To address this issue, we applied the Generalized Least Squares (GLS) method with an autoregressive correlation structure:

```
# Use the generalized least squares method (GLS) to solve autocorrelation problems
model_gls <- tryCatch({
  gls(NH3 ~ NMVOC + BC + CO, correlation = corAR1(form = ~ 1 | Year), data = air_
      pollution)
}, error = function(e) {
  print("corAR1 failed, trying corARMA")
```

```
  gls(NH3 ~ NMVOC + BC + CO, correlation = corARMA(p = 1, q = 0, form = ~ 1 | Year),
      data = air_pollution)
})
summary_gls <- summary(model_gls)
print(summary_gls)
```

The summary of the GLS model is provided in Table 23.

Table 23: Summary of the GLS Model

| Variable | Estimate | Std. Error | t value | Pr($>|t|$) |
|---|---|---|---|---|
| (Intercept) | -1302.245 | 1643.861 | -0.792 | 0.428 |
| NMVOC | 0.34879 | 0.00116 | 299.706 | <2e-16 |
| BC | 6.86219 | 0.02477 | 277.030 | <2e-16 |
| CO | -0.05684 | 0.00030 | -191.492 | <2e-16 |

**GLS Model Improvement:** **Estimate for Intercept:** The intercept estimate in the GLS model is -1302.245 with a standard error of 1643.861. The t-value of -0.792 and a p-value of 0.428 indicate that the intercept is not statistically significant.

**Estimate for NMVOC:** The coefficient for NMVOC is 0.34879 with a standard error of 0.00116. The t-value of 299.706 and a p-value of less than 2e-16 indicate that NMVOC is a highly significant predictor.

**Estimate for BC:** The coefficient for BC is 6.86219 with a standard error of 0.02477. The t-value of 277.030 and a p-value of less than 2e-16 indicate that BC is a highly significant predictor.

**Estimate for CO:** The coefficient for CO is -0.05684 with a standard error of 0.00030. The t-value of -191.492 and a p-value of less than 2e-16 indicate that CO is a highly significant predictor.

Table 24: Comparison of Model Fit Statistics Before and After Applying GLS

| Statistic | Before Applying GLS | After Applying GLS |
|---|---|---|
| Residual Standard Error | 368700 | 368730.6 |
| Multiple R-squared | 0.9749 | 0.9748829 |
| Adjusted R-squared | 0.9749 | 0.9748813 |
| F-statistic | 6.239e+05 | 623874.9 |
| p-value | <2.2e-16 | 0.4282557, <2e-16, <2e-16, <2e-16 (for NMVOC, BC, CO) |

```
# Calculate AIC
log_likelihood_gls <- logLik(model_gls)
aic_gls <- -2 * as.numeric(log_likelihood_gls) + 2 * (p + 1)
cat("AIC:", aic_gls, "\n")

# Calculate BIC
bic_gls <- -2 * as.numeric(log_likelihood_gls) + log(n) * (p + 1)
cat("BIC:", bic_gls, "\n")
```

**Key Points:**

1. **Strong Predictor Relationships:** The predictors NMVOC, BC, and CO have very strong relationships with the response variable $NH_3$, as indicated by high t-values and extremely low p-values in both models. These strong relationships dominate the regression model, reducing the impact of autocorrelation adjustment on coefficient estimates.

Table 25: AIC and BIC Before and After GLS

| | Before | After |
|---|---|---|
| **AIC** | 1373142 | 1373085 |
| **BIC** | 1373186 | 1373120 |

2. **High Variance Explanation:** The original linear regression model already explains a very high proportion of the variance in $NH_3$ emissions (Multiple R-squared = 0.9749). The additional adjustment for autocorrelation through GLS does not significantly alter this fit.

3. **Minimal Autocorrelation Impact:** The autocorrelation parameter $\Phi$ in the GLS model is estimated to be -0.03899246, which is **relatively small**. This suggests that while autocorrelation is present, its impact on the overall model fit and coefficients is minimal.

4. **Stable Residual Spread:** The residual standard error remains almost unchanged before and after applying GLS (368700 vs. 368730.6), indicating that the spread of residuals is **not significantly affected by accounting for autocorrelation**. This implies robustness in the primary model structure.

**Conclusion:** The application of GLS to address autocorrelation has shown that the residual standard error and the R-squared values remain largely unchanged. The Multiple R-squared value is 0.9748829 and the Adjusted R-squared value is 0.9748813, which are very close to the values obtained from the simple linear regression model. This similarity suggests that the original model was already capturing most of the variance in the response variable, and the inclusion of the **autocorrelation structure did not dramatically alter the model fit**. The predictors NMVOC, BC, and CO remain highly significant, with p-values less than 2e-16, indicating their strong impact on $NH_3$ emissions.

**Here is integration of all methods except Huber Regression:**

Table 26: Comprehensive AIC and BIC Comparison Across Various Model Adjustments

| Adjustment Type | Metric | Original | Modified | Difference |
|---|---|---|---|---|
| Generalized Least Squares (GLS) | AIC | 1373142 | 1373085 | -57 |
| | BIC | 1373186 | 1373120 | -66 |
| Removing Collinear Variables | AIC | 1373142 | 1418954 | +45812 |
| | BIC | 1373186 | 1418989 | +45803 |
| Ridge Regression | AIC | 1373142 | 1407965 | +34823 |
| | BIC | 1373186 | 1408000 | +34814 |
| Log Transformation | AIC | 1373142 | 115506.3 | -1258635.7 |
| | BIC | 1373186 | 115550.2 | -1258635.8 |

# 4 Conclusion

## 4.1 Summary of Key Findings

### 4.1.1 Mathematical Findings:

This study aimed to enhance the regression model for predicting $NH_3$ emissions by addressing several key mathematical issues: heteroscedasticity, non-normality of residuals, multicollinearity, and autocorrelation. The primary mathematical findings from our analysis are summarized as follows:
Firstly, the issue of **heteroscedasticity was effectively addressed** using log transformation of the dependent variable, $NH_3$. This transformation stabilized the variance of the residuals, **significantly reducing the residual standard error**. This improvement ensures that predictions are consistent across different levels of the independent variables.

Secondly, the log transformation also **mitigated non-normality in the residuals**, as evidenced by the improved alignment in the QQ plot, which is crucial for valid inference in regression models. This adjustment ensures that the statistical tests and confidence intervals derived from the model are valid, leading to more trustworthy conclusions.

Thirdly, multicollinearity was addressed through two approaches: removing collinear variables and applying Ridge Regression. Removing collinear variables slightly increased the residual standard error but simplified the model. Conversely, Ridge Regression managed to retain all variables, but not maintaining a robust model structure. Although the Ridge regression model performed well in handling multicollinearity and stability, but in this case there was **a slight decline in some indicators**, **the initial model before optimization is much more recommended**.

Fourthly, Huber regression was implemented to mitigate the impact of outliers, ensuring more robust parameter estimates and improved model performance. By applying Huber regression, which combines the squared loss for small residuals and the absolute loss for large residuals, the influence of outliers is reduced while retaining the efficiency of least squares for inliers.The comparative results demonstrate significant improvements in residual standard error while maintaining high R-squared values, highlighting the effectiveness of Huber regression in **addressing outliers**.

Finally, autocorrelation, detected through the Durbin-Watson test, was addressed using the Generalized Least Squares (GLS) method. This approach adjusted for the autocorrelation structure, **do not dramatically enhancing the reliability of model estimates.**

#### 4.1.2 Social and Environmental Impacts:

The mathematical improvements in the regression model have significant social and environmental implications: Ammonia ($NH_3$) is known for its pungent odor and plays a vital role in atmospheric chemistry, affecting human health and the environment. Accurate predictions of $NH_3$ emissions help in formulating effective pollution control strategies, which are crucial for maintaining air quality and public health. **Reducing ammonia emissions can mitigate the formation of fine particulate matter (PM2.5), which poses significant health risks, including respiratory and cardiovascular diseases.**

The study identified significant predictors of $NH_3$ emissions, including NMVOC (Non-Methane Volatile Organic Compounds), BC (Black Carbon), and CO (Carbon Monoxide). These pollutants are strongly correlated with ammonia emissions. Understanding the interaction between these pollutants and ammonia is crucial for developing comprehensive air quality management strategies. For example, NMVOC emissions from industrial processes and vehicular emissions significantly influence $NH_3$ levels, highlighting the need for targeted emission control policies in these sectors.

Addressing multicollinearity and autocorrelation in the model ensures more reliable insights into the relationships between pollutants and ammonia emissions. These insights are essential for policymakers to design and implement effective air pollution mitigation strategies. **Accurate models enable better decision-making, ultimately leading to improved public health outcomes and environmental protection.**

### 4.2 Overall Model Improvement and Limitation

**Overall Model Improvement:** The overall improvements to the model are significant. The residual standard error saw a substantial reduction after addressing heteroscedasticity and non-normality. The use of Huber Regression, significantly mitigate the influence of outliers.

**Limitations:**

**Ridge Regression Performance:** While the log transformation was effective in stabilizing variance, it resulted in a slight decrease in R-squared values. Removing collinear variables simplified the model but slightly reduced its performance. Ridge Regression, while reducing multicollinearity, introduced complexity in interpreting the model coefficients. The current model's focus on linear predictors under Ridge Regression may not adequately capture complex, non-linear interactions that significantly influence NH3 emissions. **Model**

**Complexity:** While the model has been refined to improve predictability, the complexity introduced by log transformations and Ridge Regression might obscure the direct interpretability of the coefficients, making it challenging to translate results into actionable insights without specialized knowledge. **Data Constraints:** The current model heavily relies on historical data that may not account for future changes in environmental policy or technological advancements affecting emissions. This limitation could affect the model's applicability over time or across different regulatory environments. **Predictive Scope:** The model primarily focuses on linear relationships, which might overlook more complex, non-linear interactions between variables. Future models could benefit from incorporating non-linear modeling techniques to capture a broader spectrum of data dynamics.

## 4.3 Enhanced Recommendations for Model Improvement

**Integration of Additional Variables:** Expanding the model to include more environmental factors, such as atmospheric pressure and solar radiation, could provide deeper insights into the dynamics influencing NH3 emissions.

**Advanced Statistical Techniques:** Employing more sophisticated statistical methods such as mixed-effects models or machine learning algorithms like random forests could help capture non-linear relationships and interactions between variables more effectively.

**Robustness Checks:** Implementing robustness checks through techniques like bootstrapping and sensitivity analysis to assess the stability of the model predictions under various scenarios.

**Temporal Dynamics:** Considering temporal dynamics in the model could account for seasonal variations and trends over time, providing a more accurate and realistic prediction of NH3 emissions.

## 4.4 Future Work

**Investigating alternative transformation techniques:** Future research could focus on exploring various transformation techniques, such as Box-Cox or Yeo-Johnson transformations, to further improve model performance. These techniques may help in stabilizing variance and making the data more normally distributed, enhancing the accuracy of the regression models.

**Applying more advanced regression methods:** Another area of future work involves applying more advanced regression methods, such as Lasso or Elastic Net, to handle multicollinearity. These methods not only address multicollinearity but also perform variable selection, leading to more parsimonious and interpretable models.

**Validating the model with different datasets:** To ensure the generalizability of the model, it is crucial to validate it using different datasets. This step can help in verifying the robustness of the model's performance across various contexts and identifying any potential overfitting issues.

**Integration with IoT for Real-time Monitoring:** Integrating the predictive models with IoT devices for real-time monitoring and prediction of NH3 emissions can significantly benefit environmental management practices. This would allow for immediate corrective actions in areas where pollution levels are predicted to exceed safe thresholds.

**Incorporating additional predictors:** Future studies could consider incorporating additional predictors that may influence $NH_3$ emissions for a more comprehensive model. This could include variables related to meteorological conditions, land use patterns, and other environmental factors that could impact ammonia emissions.

# References


[1] Rejeph, "Air pollution emissions dataset," https://www.kaggle.com/datasets/rejeph/air-pollution, 2023.

[2] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models*, 4th ed. Chicago, IL: Irwin, 1996.

[3] P. Bruce and A. Bruce, *Practical Statistics for Data Scientists: 50 Essential Concepts*, 1st ed. Sebastopol, CA: O'Reilly Media, 2017.

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 1st ed. New York: Springer, 2014.

[5] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed. Hoboken, NJ: Wiley, 2009.

[6] W. H. Greene, *Econometric Analysis*, 7th ed. Boston: Pearson, 2012.

[7] J. Durbin and G. S. Watson, "Testing for serial correlation in least squares regression. iii," *Biometrika*, vol. 58, no. 1, pp. 1–19, 1971.