Predicting Temperature Levels using Multiple Linear Regression Model

Wahome Mugane
and
John Miano

Project submitted in partial fulfilment of the requirements for the award of the Degree of

Bachelor of Science
in
Actuarial Science

Dedan Kimathi University of Technology

2024

Declaration by the Students

"We, *Wahome Mugane* and *John Miano*, declare that this project entitled, *'Predicting Temperature Levels using Multiple Linear Regression Model'* submitted in partial fulfilment of the degree of *Bachelor of Science in Actuarial Science*, is a record of original work carried out by us under the guidance of *Madam Maina*, and has not formed a basis for the award of any other degree or diploma, in this or any other Institution or University. In line with the ethical practice in reporting scientific information, due acknowledgements have been made wherever the findings of others have been cited."

WAHOME MUGANE
(S030-01-1789/2020)

Signature

Date

JOHN MIANO
(S030-01-1607/2019)

Signature

Date

Declaration by the Supervisor

This is to certify that the project entitled *'Predicting Temperature Levels using Multiple Linear Regression Model'* submitted by *Wahome Mugane* and *John Miano* to the Dedan Kimathi University of Technology, in partial fulfilment for the award of the degree of *Bachelor of Science in Actuarial Science*, is a bona-fide record of research work carried out by them under my supervision. The contents of this project, in full or in parts, have not been submitted to any other Institution or University for the award of any degree.

MADAM BEATRICE MAINA
(Supervisor)

Signature

Date

DR. SIMON MUNDIA
(Project Coordinator)

Signature

Date

Acknowledgement

We would like to express our sincere gratitude to our supervisor, *Madam Beatrice Maina* for her excellent guidance, unwavering support and invaluable insights throughout the course of this research project. Her mentorship and expertise have been instrumental in shaping the direction and quality of this project. We are also thankful for the support of our lecturers whose feedback and discussions enriched the research process.

Dedication

We dedicate this project to Almighty God for giving us the strength and inspiration towards the completion of this project. We also dedicate this work to our parents, siblings, families and friends for their financial support and guidance they rendered to us during our project writing.

Contents

Declaration by the Students i

Declaration by the Supervisor ii

Acknowledgement iii

Dedication iv

Table of Contents v

List of Figures vii

List of Tables viii

Abbreviations ix

Symbols x

Abstract xi

1 Introduction 1
- 1.1 Introduction . . . 1
- 1.2 Background of the Study . . . 2
- 1.3 Statement of the Problem . . . 3
- 1.4 Justification of the Study . . . 4
- 1.5 Objectives of the Study . . . 4
  - 1.5.1 General objective . . . 4
  - 1.5.2 Specific objectives . . . 4
- 1.6 Significance of the Study . . . 5

2 Literature Review 6
- 2.1 Introduction . . . 6
- 2.2 Empirical Review . . . 6

3 Methodology 11
- 3.1 Introduction . . . 11
- 3.2 Multiple Linear Regression Model . . . 11
- 3.3 Estimation of Parameters for the multiple linear regression model . . . 12
  - 3.3.1 The Ordinary Least Squares Estimation Method . . . 12
- 3.4 Model diagnosis . . . 14
  - 3.4.1 Linearity . . . 14
  - 3.4.2 Normality . . . 14
  - 3.4.3 Independence of errors . . . 16
  - 3.4.4 Homoskedasticity . . . 16

3.4.5 Multicollinearity . . . 17
3.5 Significance test of the Regression model . . . 18
3.6 Significance test on individual regression coefficients . . . 19
3.7 Coefficient of Determination (R^2) . . . 19
3.8 Prediction of temperature levels . . . 20
3.9 Accuracy of the predicted temperature levels . . . 20
3.9.1 Root Mean Square Error . . . 20

4 Results and Discussions **22**
4.1 Introduction . . . 22
4.2 Data Source and Description . . . 22
4.3 Fitting of a Multiple Linear Regression Model . . . 24
4.3.1 Parameter estimates of the fitted multiple linear regression model . . . 24
4.4 Checking the Adequacy of the Fitted Model . . . 24
4.5 Prediction of temperature levels . . . 29
4.6 Accuracy of the predicted temperature levels . . . 30
4.6.1 Root Mean Square Error . . . 30

5 Conclusion **31**
5.1 Introduction . . . 31
5.2 Summary . . . 31
5.3 Conclusion . . . 32
5.4 Recommendations for Further Research . . . 33

References **33**

A Appendix **36**
A.1 R Program Codes . . . 36

List of Figures

3.1 An illustration of a histogram for standard normal distribution 15

4.1 Histogram showing the distribution of the temperature level 22
4.2 Correlation Plot showing the relationship between variables 23
4.3 Scatter plots of dependent variable against independent variables 25
4.4 Histogram of residuals 26
4.5 Q-Q plot of the error terms 27

List of Tables

3.1 An illustration of analysis of variance table 19

4.1 Parameter estimates for the fitted model 24
4.2 Variance inflation factor values for the independent variables in the regression model 28
4.3 Analysis of variance table 28
4.4 Correlation Matrix between predicted and actual temperature levels 29

Abbreviations

AD	**A**derson- **D**arling test
ANOVA	**A**nalysis **O**f **V**ariance
AR	**A**uto **R**egression
Cov	**C**ovariance
DF	**D**egrees of **F**reedom
K-S	**K**olmogorov**- S**mirnov test
MAE	**M**ean **A**bsolute **E**rror
MLE	**M**aximum **L**ikelihood **E**stimation
MLR	**M**ultiple **L**inear **R**egression
MPE	**M**ean **P**ercentage **E**rror
OLS	**O**rdinary **L**east **S**quare
Q-Q	**Q**uantile **Q**uantile
RMSE	**R**oot **M**ean **S**quared **E**rror
SD	**S**tandard **D**eviation
SSE	**S**um of **S**quares due to **E**rror
SSR	**S**um of **S**quares due to **R**egression
SST	**T**otal **S**um of **S**quares
VIF	**V**ariance **I**nflation **F**actor

Symbols

Y_i	Response variable
X_i	Predictor variables
ε_i	Error term
β_i	Regression parameters for the regression model
H_0	Null hypothesis
H_1	Alternative hypothesis
k	number of regression parameters
n	Sample size
R^2	Coefficient of determination
σ	Standard deviation
$\boldsymbol{\beta}$	Parameter vector
$\boldsymbol{\epsilon}$	Error vector
$\hat{z}_i$	Predicted values for dependent variables for the i th observation
R_a^2	Adjusted coefficient of determination
α	Level of significance
σ^2	Variance
T	Temperature

Abstract

Stakeholders such as farmers predominantly depend on temperature levels for agricultural planning and decision-making. The need for more accurate prediction techniques has arisen due to the unpredictability of temperature changes brought about by the growing effects of climate change and carbon emissions. Many contemporary temperature prediction models struggle to adequately account for the complex interplay between various predictor variables and fluctuations in temperature. This research addresses this issue by employing the Multiple Linear Regression model to predict temperature levels. The model offered a framework for incorporating multiple predictors into the temperature prediction process, potentially enhancing reliability. To estimate the model parameters, the Ordinary Least Squares estimation method was utilized. Ordinary Least Squares was preferred due to its robustness in handling noise and its capability to provide unbiased estimates under certain statistical assumptions. Assessing model adequacy in multiple linear regression involved verifying several assumptions. This included examining the linearity of relationships between predictors and the response variable, ensuring that residuals are normally distributed with constant variance, checking for independence of observations, and scrutinizing for absence of multicollinearity among predictor variables. Scatter plots for the independent variables against the dependent variable were plotted to check for linearity. The normality assumption was assessed using Histograms, Q-Q plots, and the Anderson-Darling test. Additionally, the Durbin-Watson test was used to test for the autocorrelation of the residuals. The Breusch-Pagan test assessed homoscedasticity within the model, while the Variance Inflation Factor examined the degree of multicollinearity among predictor variables. Hypothesis tests, including the F-test and t-test, were used to assess the significance of estimated parameters. The study obtained an R-Squared value of 0.809045 and root mean square error of 5.339.These results provide valuable insights for stakeholders, such as environmental agencies and policymakers, enabling them to make informed decisions, adjust resource management strategies, and mitigate potential environmental risks.

Chapter 1

Introduction

1.1 Introduction

Climate change is primarily driven by human activities that release greenhouse gases into the atmosphere, amplifying the natural greenhouse effect. The combustion of fossil fuels, such as coal, oil, and natural gas for energy production is a major contributor, releasing large amounts of carbon dioxide into the air. Deforestation and land-use changes further exacerbate the issue, as trees play a crucial role in absorbing . Additionally;industrial processes, agriculture and certain waste management practices release methane and nitrous oxide potent greenhouse gases that trap heat in the atmosphere. The cumulative impact of these anthropogenic activities intensifies the greenhouse effect leading to a warming of the Earth's surface.

Climate change has had a significant influence on temperature trends in the dynamic field of climate research, underscoring the need for precise temperature predictions. The interactions between the natural processes of the Earth and human-caused climate change have added a level of complexity, making temperature predictions more difficult. Precise temperature predictions are not only research endeavours; they are essential instruments for tackling the complex consequences of global warming. Precise temperature predictions help farmers optimize crop management practices so they can adjust to shifting growth conditions and minimize production losses. Reliable temperature predictions help with energy consumption planning by facilitating the effective use of resources and minimizing the negative environmental effects of energy production. Resilience of infrastructure, health care, and environmental preservation all depend on the capacity to predict temperature fluctuations with great precision. Moreover, reliable temperature predictions are essential for developing evidence-based policy as the world struggles with the far-reaching effects of climate change. Reliable temperature predictions are critical to almost every aspect of the linked society, from directing long-term climate change plans to influencing catastrophe preparedness activities.

Essentially, the correlation between temperature predictions and climate change highlights the urgent requirement for modelling methods and predictions instruments. Improving the ability to accurately predict temperature changes helps people not only deal with the difficulties posed by a changing climate, but also gives them the opportunity to take proactive measures to address and adjust to the changing environmental scene. In light of this, the search for precise temperature predictions becomes not just a scientific project but also a vital necessity for constructing sustainability and resilience in the face of a constantly shifting environment.

1.2 Background of the Study

Regression is a statistical method used for modeling the relationship between a dependent variable (the target) and one or more independent variables (predictors or features). The primary goal of regression analysis is to understand and quantify the influence of the independent variables on the dependent variable, enabling the prediction or estimation of the dependent variable's values based on the known values of the independent variables.

Earlier on, linear regression was the most employed predictive modeling tool in practical applications. This was because parameters which had a linear relationship were easier to fit than those which were non-linearly related. Also, the statistical properties of the resulting estimators were easier to be determined. The case where only one independent variable is used to establish the linear relationship is called simple linear regression analysis also refereed to as Univariate regression. It analyzes the relationship between the dependent variable and one independent variable then formulate a linear regression analysis between these two variables. A regression analysis with one dependent variable and more than one independent variable (explanatory variable) is called multiple linear regression analysis. In this analysis, an attempt is made to account for the variation of the independent variables into the dependent variable. The purpose is to examine if the independent variables are successful in predicting the outcome variable and which of those independent variables are significant predictors for the outcome. More specifically regression analysis helps one to understand how the typical value of the dependent variable change when any one of the dependent variable or variables is varied. This analysis is an extension of simple linear regression analysis.

Most applications of the multiple linear regression fall under two broad categories. If the goal is prediction or error reduction, then multiple linear regression can be used to fit a predictive model to an observed dataset of values of the dependent and the independent variables. After the model is fitted, then if additional values of the independent variable are collected without an accompany response

variable, the fitted model can be used to make a prediction for the response variable. Secondly if the objective is to explain the variation in the independent variable that can attribute to variation in the dependent variable and in particular to determine whether some independent variable may have non linear relationship with dependent variable or the response at all or even to identify which subsets of the independent variables may contain redundant information about the dependent variable.

Analyzing data using multiple regression in the context of a weather forecasting model offers several advantages. First and foremost, it allows meteorologists to determine the relative influence of one or more predictor variables on the predicted temperature levels. Another significant advantage is the ability to identify outliers or anomalies in weather data. When studying temperature predictions, researchers might observe strong or weak correlation between variables like humidity, cloud cover, and wind speed and the predicted temperature. At times they may come across exceptional cases where all the listed predictor values are correlated with temperature, except for specific instances where temperatures deviate from the predicted patterns, suggesting local microclimates or unique weather events.

It is important to note that multiple linear regression has its disadvantages, especially when underlying assumptions fail. Meteorological data can be complex, and when assumptions related to linearity, independence, and normality are violated, the accuracy of the model may be compromised. Therefore, careful consideration and validation of the model are crucial to ensure the reliability of temperature predictions in a dynamic and multifaceted natural environment.

1.3 Statement of the Problem

One of the central challenge of the current times lies in the accurate prediction of temperature amidst the backdrop of climate change. Climate change which is fueled by various factors including rising greenhouse gas emissions and natural variability, has ushered in an era of increased weather unpredictability. Traditional methods of temperature forecasting that once served well but now struggle to maintain with the evolving dynamics of the changing climate. This challenge is compounded by the complex relationships, new variables, and non-linearities that characterize today's climate patterns. Inaccurate temperature predictions have far-reaching consequences in various sectors. For instance Agriculture is affected by disrupted planting and harvesting schedules, energy management experiences inefficiencies, public health is compromised during extreme temperatures, transportation faces delays and accidents, and infrastructure incurs unexpected maintenance costs due to the unpredictable impact of temperature fluctuations.

1.4 Justification of the Study

The ability of different sectors to perform effectively and achieve their goals is directly affected by the accuracy and timeliness of temperature projections. The complex system that is the climate is made up of a complicated network of interrelated elements that greatly affect variations in temperature. Therefore, in order to provide accurate predictions, a detailed comprehension and careful evaluation of these complex components are necessary. To do this, the research will utilize a prediction model that is intended to capture the intricate connections and interplay between these many elements. The model will carefully take into consideration the various weights and contributions of each explanatory component in an effort to balance their combined effect. In doing so, the research aims to generate temperature predictions that are more thorough and dependable, offering insightful information to a variety of industries, from public health and infrastructure development to agriculture and energy. The study's major goal is to provide decision-makers in a variety of fields with a reliable prediction tool that may help them understand the complexities of the climate system and make wise decisions even in the face of changing temperature trends.

1.5 Objectives of the Study

1.5.1 General objective

The general objective of the study was to predict temperature levels using multiple linear regression model.

1.5.2 Specific objectives

The specific objectives of the study were;

1. To fit a multiple linear regression model.
2. To assess the adequacy of the fitted model.
3. To predict temperature level using the fitted model.
4. To assess the accuracy of the predicted values.

1.6 Significance of the Study

The significance of this study extends to a broad spectrum of stakeholders. For businesses, accurate temperature predictions serve as a strategic tool for optimizing resource allocation and reducing operational costs. Industries heavily dependent on weather conditions, such as agriculture, energy and tourism can better plan and adapt their activities based on reliable temperature forecasts. This, in turn enhances overall efficiency and competitiveness in the market. Environmental conservation efforts are bolstered by accurate temperature predictions, providing critical insights into how ecosystems might be affected by changing temperature patterns. This information enables conservationists to implement targeted strategies for protecting vulnerable species and habitats, contributing to the overall preservation of biodiversity. For the general population, accurate temperature predictions translate into improved safety and well-being. Early warnings of extreme weather events, such as heatwaves or cold snaps, allow individuals and communities to take proactive measures, reducing the risks of health-related issues, property damage and other adverse impacts.

Chapter 2

Literature Review

2.1 Introduction

This section discusses important concepts and insights utilized in the project, highlighting specific theoretical contributions from prior literature. Conducting this literature review served the purpose of deepening comprehension of previous studies relevant to the research goals. It also aided in refining the foundational concepts upon which the research was constructed. The primary objective was to elucidate how the multiple linear regression model can be employed in a practical, effective, and efficient manner to address the research objectives.

2.2 Empirical Review

Rainfall predictions play a crucial role in various sectors, including agriculture, water resource management, and disaster preparedness. Accurate predictions helps mitigate risks associated with floods, droughts, and crop failure, contributing to informed decision-making and resource allocation. Mulyani et al. (2019) focused on monthly rainfall predictions based on several weather parameters, including temperature, humidity, sunshine duration, and wind speed, conducted at the Jatiwangi Majalengka Meteorological Station. The study aimed to predict monthly rainfall for the year 2019 using daily weather data from 2018 in Majalengka Regency, employing the multiple linear regression equation method. In terms of methodology, the authors utilized a multiple linear regression model to establish the relationship between the selected weather parameters and monthly rainfall. The findings of the study revealed a notable overestimation in the monthly rainfall predictions for 2019, indicating that the predicted values were higher than the actual observed values. However, it's noteworthy that the predictions performed exceptionally well for April. The evaluation metrics, such as the strong correlation coefficient demonstrated the model's capability to capture patterns and relationships between the chosen weather parameters and monthly rainfall.

Rainfall prediction techniques have garnered significant attention due to their critical role in mitigating the impact of extreme weather events. Anusha and Chaithanya (2019), investigated the challenges

and techniques involved in rainfall prediction, they highlighted the use of Multi-Linear Regression as a more accurate method compared to existing statistical approaches such as Support Vector Machine. The study focused on Uttar Pradesh, India, with a tropical monsoon climate known for extreme weather conditions. The methodology involved the collection of meteorological data from the Indian Meteorological Department over four years, encompassing parameters such as temperature, wind speed, wind direction, humidity, atmospheric pressure, and rainfall. The data was divided into training and testing sets, and Multi-Linear Regression was applied to establish relationships between these variables for rainfall prediction. The study achieved a 88 percent accuracy rate, outperforming other methods like Support Vector Machine and Bayesian Enhanced Modified Approach.

Sreehari and Srivastava (2019), researched the challenges of understanding and predicting climate phenomena, particularly in light of natural disasters and the dynamic nature of climate variables such as temperature and rainfall. They investigated a specific case study of the catastrophic flood that struck Kerala, India, in August 2019. In a country heavily reliant on agriculture, with 60 percent of its population depending on farming, the accurate prediction of rainfall is of paramount importance. The article's main methodological approach lied in the application of multiple linear regression for rainfall estimation and prediction. The method was applied to a dataset spanning six years, collected from Nellore district in Andhra Pradesh, India. The key finding of the study was that multiple linear regression offered more precise rainfall predictions compared to simpler linear regression methods.

Saragih et al. (2020), conducted a simulation of monthly rainfall prediction in Deli Serdang, North Sumatra, using regression equations with air temperature and humidity as predictors. The dataset encompassed 30 years of data from 1989 to 2018. Two regression methods, simple linear regression and multiple linear regression, were employed for predicting total monthly rainfall. The evaluation of these predictions involved the calculation of Pearson correlation values and the assessment of the deviation between predicted and actual total rainfall. The findings revealed that the simulation for total monthly rainfall predictions in 2019 for the Deli Serdang area exhibited varying degrees of accuracy based on the predictor variables. When using air humidity as the predictor, the correlation value (r) was 0.72, and the average root mean square error (RMSE) was 77.42 mm/month. The air temperature predictor resulted in a higher correlation value of 0.73 and a lower RMSE of 77.13 mm/month. The combination of air temperature and air humidity predictors yielded a correlation value of 0.70 and an RMSE of 80.53 mm/month. The study's methodology and findings indicated that both air temperature

and air humidity have potential as predictors for monthly rainfall prediction, with the air temperature predictor demonstrating slightly better performance.

Luthfiarta et al. (2020), researched on the importance of accurate weather prediction, especially in the face of changing weather patterns. They highlighted the critical role of meteorological agencies in providing early warnings for sudden and extreme weather shifts. To achieve these accurate predictions, the study adopted a supervised learning approach, specifically using multiple linear regression as the chosen algorithm. This approach aimed to predict rainfall, which served as the dependent variable, by considering four independent variables: temperature, humidity, pressure, and wind speed. The data source for the research was the Indonesian Meteorological Agency, ensuring the use of reliable and authentic meteorological information. The dataset spans three years, from 2015 to 2017, encompassing a substantial amount of data for analysis. The variables selected, such as temperature, humidity, pressure, and wind speed, are crucial in the context of predicting rainfall, making this research comprehensive in its approach. The findings of the study reveal that the coefficient of determination stands at 25.5 percent. This statistic indicates the degree to which the chosen independent variables collectively explain variations in rainfall as the dependent variable. In essence, the results suggest that the model utilizing multiple linear regression and the specified meteorological parameters can be a valuable tool for predicting rainfall accurately. The research provides insights into the reliability of this approach, offering potential benefits for decision-makers and stakeholders who heavily rely on weather forecasts for various applications.

Granata et al. (2022), conducted a precipitation forecasting in the northern region of Bangladesh, particularly in the Rangpur and Sylhet divisions, which experience a tropical monsoon climate. The study employed a multiple linear regression to develop precipitation prediction models. The performance of the prediction model was rigorously evaluated using various metrics and graphical representations such as Q-Q plots. Additionally, an analysis was conducted to assess prediction accuracy. Overall, the model Multiple lienar regression achieved high R-squared values of up to 0.87 and 0.92 for the Rangpur and Sylhet stations, respectively.

Climate change in Indonesia, a tropical region, lead to weather uncertainty, making accurate weather predictions challenging. Factors such as temperature, air pressure, wind speed, humidity, and rainfall significantly impact weather conditions. Rainfall, in particular, exhibited high diversity due to climate anomalies influenced by geographical, orographic, topographical, island orientation, and structural

factors. This lead to uneven rainfall distribution across regions. Yusuf (2022), addressed these challenges and provided daily, monthly, and yearly rainfall predictions, a statistical approach using the Multiple Linear Regression method was employed. In the study, rainfall serves as the dependent variable, while temperature and humidity act as independent variables. The research, based on data from 2017 to 2021 totaling 60 data points and analyzed using the WEKA Application, reveals a correlation coefficient of 0.8175.

Sulistiyono et al. (2023), addressesd the crucial role of rainfall in the agricultural sector of Lubuklinggau City, located in South Sumatra. They highlighted that farmers in that region traditionally relied on observational methods to determine planting times, mainly due to the lack of government-provided rainfall information. To bridge this information gap and provide farmers with more accurate data, the study focused on developing a prediction system for rainfall. The chosen method for the research was multiple linear regression analysis, a statistical approach that assesses the relationships between various climatic variables, such as temperature, humidity, wind speed, solar radiation, and rainfall. The data source and duration used in this study are not explicitly mentioned, but they are fundamental to the analysis and the subsequent prediction model. The research aims to systematically estimate future rainfall based on past and present information. The model encountered challenges in meeting the linearity assumption, as evident from the graphical representation of the residuals. However, in the subsequent model, notable improvements were observed. It yielded an increased R-squared value and a reduced regression standard error on the residuals, ultimately aligning with the regression model assumptions. As the analysis progressed to the diagnostic, it became apparent that the residuals closely adhered to the regression model assumptions, demonstrating reasonable compliance. It was observed that the rainfall data exhibited positive skewness. Furthermore, statistical tests, akin to t-tests, consistently yielded results below the significance level of 0.05, signifying the accuracy of the predictions.

The reviewed literature encompasses studies that explore various aspects of temperature elements, with some articles also delving into rainfall predictions. Utilizing Multiple Linear Regression models, these studies aim to address key challenges and gaps in predicting temperature-related patterns and fluctuations, as well as rainfall patterns. These investigations cover a range of topics, including temperature prediction, predicting temperature variations in specific regions, modeling temperature risk factors, and factors influencing temperature changes in different geographic locations, alongside rainfall prediction studies. Multiple Linear Regression emerges as a powerful tool in these inquiries,

demonstrating its effectiveness in capturing the complexities observed in temperature and rainfall data, including normally distributed variables and interactions among multiple factors. One prominent gap identified across these studies is the need for more accurate and effective methods for predicting temperature and rainfall fluctuations, considering the diverse factors influencing these changes. Multiple Linear Regression effectively addressed this gap by providing a framework for modeling temperature and rainfall-related data, enabling accurate predictions and a better understanding of the factors driving temperature and rainfall variations.

Chapter 3

Methodology

3.1 Introduction

This chapter presents the methodology of the study. Section 3.2 introduces the multiple linear regression model. Section 3.3 elaborates on how model parameters will be estimated, while Section 3.4 addresses model diagnosis. In Section 3.5, the focus is on predicting new observations, and Section 3.6 involves checking the accuracy of the predictions

3.2 Multiple Linear Regression Model

The general form of the multiple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \tag{3.2.1}$$

where $i = 1, 2, \ldots, n$. $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$ is the deterministic part, and ε_i is the stochastic part. Y_i represents the ith response variable, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are the parameters for the regression model, $X_1, X_2, \ldots, X_k$ are the independent variables, and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are the residuals. A residual is the difference between the observed value of the response variable Y and the predicted value $\hat{Y}$. This multiple linear regression can be expressed in matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

where $\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$ is a vector of response variables,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

β=$(\beta_0, \beta_1, \ldots, \beta_k)'$ is a vector of model parameters, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$

is the vector of residuals.

$$\varepsilon = \mathbf{Y} - \mathbf{X}\beta \tag{3.2.2}$$

3.3 Estimation of Parameters for the multiple linear regression model

The parameters of the model will be estimated using the least square estimation method. $\beta_0, \beta_1, \ldots, \beta_k$ are the parameters of the regression model. This method is also known as the ordinary least squares estimation method and is used in the estimation of these parameters. The least square method provides an overall rationale for the placement of the line of best fit among the data points being studied.

3.3.1 The Ordinary Least Squares Estimation Method

The ordinary Least Squares method estimates the parameters of the regression model by minimizing the sum of the squared residuals. The task is to find the vector of least squares estimators $\hat{\beta}$ that minimizes the sum of squared residuals M. Let

$$\begin{aligned} M &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \varepsilon^\top \varepsilon \\ &= (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}^\top \mathbf{Y} - \beta^\top \mathbf{X}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top \mathbf{X}\beta \\ &= \mathbf{Y}^\top \mathbf{Y} - \beta^\top \mathbf{X}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top \mathbf{X}\beta \end{aligned} \tag{3.3.1}$$

Differentiating equation (3.3.1) with respect to β and equating to zero to obtain the regression estimates:

$$\frac{\partial M}{\partial \boldsymbol{\beta}} = 0 \quad \Rightarrow \quad -2\mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = 0 \tag{3.3.2}$$

$$\Rightarrow \quad \mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y} \tag{3.3.3}$$

Multiplying both sides of the above equation by $(\mathbf{X}^\top \mathbf{X})^{-1}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y} \tag{3.3.4}$$

where $\mathbf{X}^T\mathbf{X}$ is an $n \times n$ symmetric matrix and $\mathbf{Y}$ is an $n \times 1$ matrix of the observed $\mathbf{Y}$ values.

Unbiasedness

An unbiased estimator is one for which the expected value of the estimated parameter equals the parameter itself.

$$\begin{aligned}
\mathbf{E}(\hat{\boldsymbol{\beta}}) &= \mathbf{E}\left[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}\right] \\
&= \mathbf{E}\left[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\right] \\
&= \mathbf{E}\left[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\varepsilon}\right] \\
&= \mathbf{E}\left[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\right] + \mathbf{E}\left[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\varepsilon}\right] \\
&= \boldsymbol{\beta}
\end{aligned} \tag{3.3.5}$$

Since $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ and $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\ \mathbf{X}$ is an identity matrix, thus $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

Variance

Variance explains how data points are spread from the mean.The variance of $\hat{\boldsymbol{\beta}}$ is expressed by the covariance matrix that is

$$\begin{aligned}
Cov(\hat{\boldsymbol{\beta}}) &= E[\hat{\boldsymbol{\beta}} - E[\hat{\boldsymbol{\beta}}][\hat{\boldsymbol{\beta}} - E[\hat{\boldsymbol{\beta}}^T \\
&= \mathrm{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}]
\end{aligned} \tag{3.3.6}$$

where $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$is a matrix of constants, and the variance of $\mathbf{Y}$ is $\sigma^2 I$. Thus,

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}) &= [(\text{Var}(\mathbf{X}^T\mathbf{X}))^{-1}\mathbf{X}^T\mathbf{Y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\,\text{Var}(\mathbf{Y})(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned} \tag{3.3.7}$$

Therefore, since $C_{ij} = (\mathbf{X}^T\ \mathbf{X})^{-1}$, then $\text{Var}(\hat{\beta}_j) = \sigma^2 C_{jj}$.

3.4 Model diagnosis

After the estimation of the parameters, a diagnostic analysis will be performed to check the adequacy of the fitted regression model. Graphical tools such as scatter plots, Q-Q plot and histogram will be used. Also statistical tests such Aderson Darling test, Durbin Watson test and Brewsch-Pagan test will be conducted.

3.4.1 Linearity

The linearity assumption in multiple linear regression asserts that the relationship between the dependent variable and independent variables is linear. It implies that a change in an independent variable leads to a constant change in the expected value of the dependent variable, verified through scatter plots . Violations may require transformations or interaction terms.

3.4.2 Normality

Assumption of normality will be examined using two aspects that is, the graphical approach using the histogram and the Q-Q plot and statistical method that is the Anderson-Darling test.

The histogram

The histogram and the normal probability plot are used to check whether or not it is reasonable to asses that the random errors inherent in the multiple linear regression process have been drawn from the normal distribution

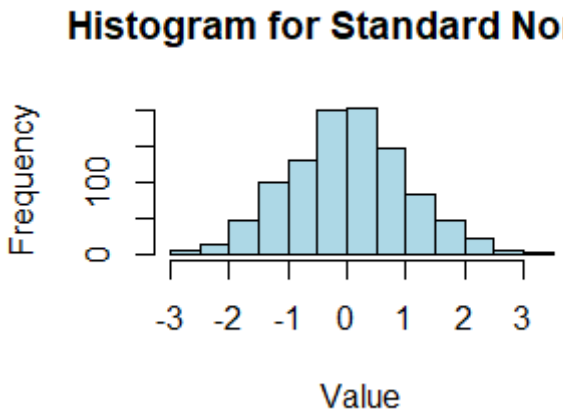

FIGURE 3.1: An illustration of a histogram for standard normal distribution

Quantile-Quantile (Q-Q) plot

The Q-Q plot is a tool used to assess if data fits a particular distribution, like the normal distribution. It compares the quantiles of the data to the quantiles of a theoretical distribution. If the data fits the distribution, the plot will show a straight line. This plot helps us check assumptions of statistical analyses. A perfect fit would result in all points lying on a straight line, indicating normal distribution. Data can be positively skewed (mean $>$ median) or negatively skewed (mean $<$ median), affecting the shape of the plot. Positive skewness shows a longer tail on the right, while negative skewness shows a longer tail on the left.

Anderson Darling test

It is used to test if a sample of a data come from population with a specific statistical distribution. Anderson Darling test makes the use of the specific statistical distribution in calculating critical values. The hypothesis for this test are;

$$\begin{aligned} H_0 &: \text{Data do not follow the normal distribution} \\ H_1 &: \text{Data follows the normal distribution} \end{aligned} \tag{3.4.1}$$

Then the test statistic is;

$$A^2 = -n - S \tag{3.4.2}$$

Where;

$$S = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2i-1}{n} - F(Y_i) - \ln(1 - F(Y_{n+1-i}))\right)^2 \tag{3.4.3}$$

and n is the sample size. F is the cumulative distribution function(CDF) of the specified distribution. Y_i is the ordered data. The null hypothesis is rejected for large values of A^2 that is if it exceeds a given critical value, smaller values of Anderson Darling indicate that the normal distribution fits the data better.

3.4.3 Independence of errors

In multiple linear regression analysis, autocorrelations are problematic. One of the key major assumptions behind multiple linear regression is that each obervation is independent. Autocorrelation invalidates that assumption, causing any predictions and insights extracted from the model to be biased.

Durbin Watson test

The Durbin-Watson (DW) test is a statistical test used to detect autocorrelation in the residuals of a linear regression model. The test statistic is a value between 0 and 4.

$$\begin{aligned} H_0 &: \text{There is correlation} \\ H_1 &: \text{There is no correlation} \end{aligned} \tag{3.4.4}$$

The test statistic is calculated using the residuals of the regression model, and it compares the differences between consecutive residuals to their average.

The test statistic is;

$$DW = \frac{\sum_{i=2}^{n}(\varepsilon_n - \varepsilon_{n-1})^2}{\sum_{i=1}^{n}\varepsilon_n^2} \tag{3.4.5}$$

To determine if a Durbin-Watson test statistic is significantly significant at a certain alpha level, this research will refer to the table of critical values. If the absolute value of the Durbin-Watson test statistic is greater than the tarbulated value in the table, then the null hypothesis of the test will be rejected and conclude that autocorrelation is not present. Also, a value close to 2 indicates no correlation, a value less than 2 indicates positive correlation and a value greater than 2 indicates negative serial correlation. Violation of the independence assumption of the error terms will lead to inefficiency of the least squares estimates.

3.4.4 Homoskedasticity

Homoskedasticity refers to a condition in which the variance of the residual, or error term, in a regression model is constant. That is, the error term does not vary much as the value of the predictor variable changes.

Breusch-Pagan test

Breusch-Pagan test tests that null hypothesis that the error variances are a multiplicative function of one or more variables. versus the alternative hypothesis that the error variances are all equal

$$BP = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \hat{z}_i \tag{3.4.6}$$

Where BP is the Breusch-Pagan test statistic n is the sample size $\hat{\varepsilon}_i$ is the estimated residual for the i th observation z^i is the predicted values for the dependent variables for the i th observation Test the hypothesis;

$$\begin{aligned} &H_0 : \text{residuals do not have a constant variance} \\ &H_1 : \text{residuals have a constant variance} \end{aligned} \tag{3.4.7}$$

The Breusch-Pagan test statistic is asymptotically distributed as chi-squared with k degrees of freedom, where k is the number of independent variables in the regression model. If the Breusch-Pagan test statistic is greater than the critical value of the chi-squared distribution with k degrees of freedom, then we reject the null hypothesis and conclude that heteroscedasticity is present.

3.4.5 Multicollinearity

Multicollinearity occurs when two or more independent variables have a high correlation with one another in a regression model, which makes it difficult to determine the individual effect of each independent variable on the dependent variable.

Variance Inflation Factor

The variance inflation factor is computed for every explanatory variable that goes into the linear model. The variance inflation factor function is of the form

$$V.I.F = \frac{1}{1-R^2} \tag{3.4.8}$$

R^2 represents the adjusted coefficient of determination for regressing the ith independent variable on the remaining ones.

$$\begin{aligned} H_0 &: \text{There is no multicollinearity} \\ H_1 &: \text{There is multicollinearity} \end{aligned} \tag{3.4.9}$$

The null hypothesis is rejected if the VIF is not equal to 1 and If the VIF results in 1, it indicates that there is no multicollinearity. This means that the predictor variable is not correlated with the other predictor variables in the model. $1 <$ VIF < 5 suggest moderate multicollinearity, while VIF exceeding 5 or 10 indicates presence of high multicollinearity between the independent variable and the others.

3.5 Significance test of the Regression model

F test

After estimating the parameters of the fitted model we will be intersted to find the evidence of a linear relation-ship between the dependent and a subset of the independent variables whereby this relationship will be used in forecasting. The hypothesis questions are

$$\begin{aligned} H_0 &: \beta_i = 0 \text{ for all } i \\ H_1 &: \beta_i \neq 0 \text{ for atleast one } i\text{=1,2...k+1} \end{aligned} \tag{3.5.1}$$

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{3.5.2}$$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{3.5.3}$$

$$SST = SSR + SSE = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{3.5.4}$$

The null hypothesis is true,then the statistic:

$$F_0 = \frac{SSR}{k} \div \frac{SSE}{n-k} = \frac{MSR}{MSE} \sim F_{k,n-k} \tag{3.5.5}$$

The F_0 statistic is compared to an F-distribution with k and $n-k-1$ degrees of freedom. If the F_0 statistic is greater than the critical value of the F-distribution, then the null hypothesis is rejected and

we conclude that the regression model is significant. These results can be displayed in an ANOVA table as shown below.

TABLE 3.1: An illustration of analysis of variance table

Source	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**F-Ratio**
Regression	SSR	$k-1$	MSR	$\frac{MSR}{MSE}$
Error	SSE	$n-k$	MSE	
Total	SST	$n-1$		

3.6 Significance test on individual regression coefficients

t-test

Conduct an individual hypothesis test for the coefficients of each independent variable in the model The hypothesis are:

$$\begin{aligned} &H_0 : \beta_i = 0 \\ &H_1 : \beta_i \neq 0 \text{ for } i\text{=1,2...k+1} \end{aligned} \tag{3.6.1}$$

the test statistic is;

$$t = \frac{\hat{\beta}_i}{Se(\hat{\beta}_i)} \tag{3.6.2}$$

The t-statistic is compared to a t-distribution with $n-k-1$ degrees of freedom, where n is the sample size and k is the number of independent variables in the regression model. The decision rule is to reject H_0 at α level of significance if $|t| > t_{\frac{\alpha}{2},(n-k-1)}$.

3.7 Coefficient of Determination (R^2)

The coefficient of determination (R^2) is the amount of variability explained by the fitted model. The coefficient of determination increases when predictor variables are added to the model. Therefore, R^2 is not a fair criterion for comparing models with a different number of predictors. To address this issue, an adjusted coefficient of determination is defined.

$$R_a^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} \tag{3.7.1}$$

where: R_a^2 is the adjusted coefficient of determination, n is the number of observations, k is the number of independent variables or predictors in the model.

The adjusted coefficient of determination (R_a^2) provides a more appropriate measure for model fitness and model comparison when there are varying numbers of independent variables. It penalizes models with excessive predictors that do not significantly contribute to explaining the variance in the dependent variable, offering a more reliable assessment of the model's performance

3.8 Prediction of temperature levels

Supposing the predictor values for the regression model at $X_0 = 1, X_{01}, X_{02}, \ldots, X_{0k}$, then the predicted value of $\mathbf{Y}$ at X_0 is

$$\mathbf{Y}\text{predict} = \hat{\boldsymbol{\beta}}^T X_0 \tag{3.8.1}$$

$$\begin{aligned} E[\mathbf{Y}\text{predict}|\mathbf{X} = X_0] &= \text{var}(X_0\boldsymbol{\beta}) + \text{var}(\boldsymbol{\epsilon}) \\ &= [X_0(\mathbf{X}\hat{T}\mathbf{X})^{-1}\mathbf{X}_0^T]\boldsymbol{\sigma}^{\mathbf{2}} + \boldsymbol{\sigma}^{\mathbf{2}} \\ &= \boldsymbol{\sigma}^{\mathbf{2}}[1 + X_0(\mathbf{X}^T\mathbf{X})^{-1}X_0^T] \end{aligned} \tag{3.8.2}$$

this statistic is used in hypothesis testing and also to construct a $100(1-\alpha)\%$ confidence interval for the predicted value.

$$\frac{\mathbf{Y}\text{predict} - E[\mathbf{Y}\text{predict}]}{\sqrt{\sigma^2(1 + X_0(\mathbf{X}^T\mathbf{X})^{-1}X_0^T)}} \sim t(n-p) \tag{3.8.3}$$

A $100(1-\alpha)\%$ prediction interval for the future observation is

$$Y_0 \pm t_{\alpha/2,(n-p-1)}\sqrt{\sigma^2(1 + X_0(\mathbf{X}^T\mathbf{X})^{-1}X_0^T)}$$

3.9 Accuracy of the predicted temperature levels

To assess the effectiveness of a model in prediction, it is important to assess the predictive accuracy based on the difference between the observed and predicted values. There are various methods that are used to assess how accurately the fitted model focuses, such as the root mean squared error (RMSE), the mean absolute error (MAE), Mean Absolute Percentage Error (MAPE) and Mean Percentage Error (MPE). The accuracy of the predicted values is determined by the difference between the observed and predicted values. This study used the root mean squared error (RMSE) to check how accurate the fitted multiple linear regression model predicted temperature level.

3.9.1 Root Mean Square Error

The root mean square error (RMSE), is defined as the standard deviation of the residuals. RMSE expresses average model predictive error in units of the response variable. It indicates how far away

each prediction is from the actual value, that is, it quantifies how different a set of values are and it expresses the average model prediction error in units of the response variable. The RMSE is always non-negative, it ranges from zero to infinity, it is indifferent to the direction of errors and it does not increase with the variance of the errors but increases with the variance of the frequency distribution of error magnitudes. RMSE is computed as;

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{3.9.1}$$

Lower values of RMSE implies a better fit thus an accurate prediction.

Chapter 4

Results and Discussions

4.1 Introduction

This chapter presents the results and discussion of research findings. Section 4.2 presents data source and description. Section 4.3 presents the results of estimation of parameters for the regression model, Section 4.4 provides inferential statistics to assess the adequacy of the model. Section 4.5 covers the prediction of new observations and the assessment of the accuracy of the predicted temperature levels.

4.2 Data Source and Description

The data used in this study was obtained from: https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate. It conists of 43,800 observations and 7 variables. The analysis involves examining temperature level as the response variable, and the independent variables considered include air pollution(micrograms per cubic metre (µg/m3)), dew(millimetres), pressure(pascals), windspeed(meters per second), snow size (inches) and rainfall(millimetres)

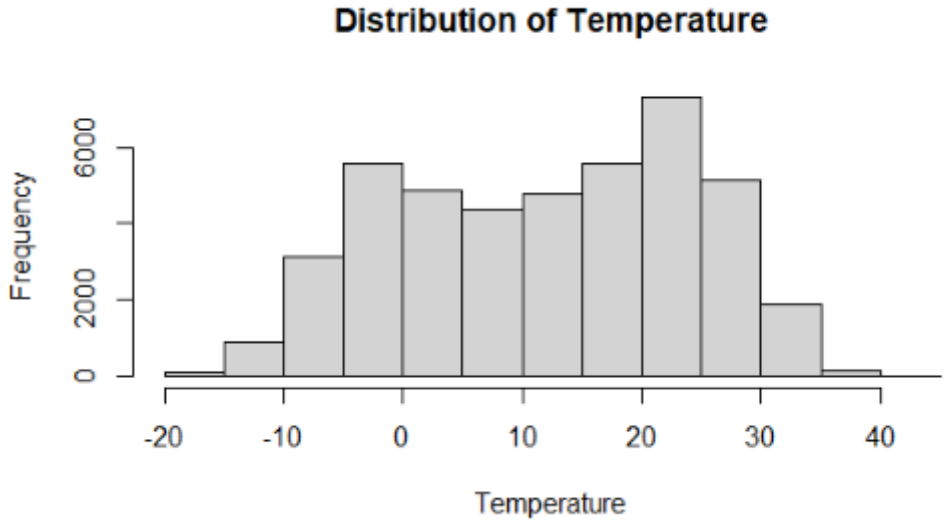

FIGURE 4.1: Histogram showing the distribution of the temperature level

Figure 4.1 provides a visual summary of the distribution of the response variable, temperature. The histogram illustrates the shape of the distribution. The histogram of the temperature levels data reveals a strong negative skew, characterized by a long left tail that extends towards lower values. This indicates greater frequency and clustering of higher temperature readings, but also the existence of some low temperature outliers on the extreme left end. The distribution highlights the non-normal, left-skewed nature of temperature, with a large mass concentrated above the mean. This suggests that the temperature data is not symmetrically distributed around the mean and that there are more extreme low temperature values than high temperature values.

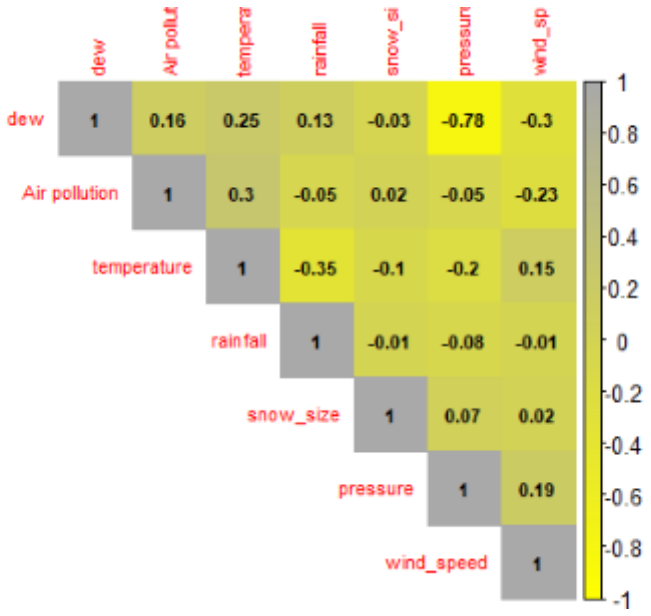

FIGURE 4.2: Correlation Plot showing the relationship between variables

Figure 4.2 is a correlation plot visualizing the relationships between the variables in the health insurance claims dataset. It displays positive or negative correlations using color shading, with strong correlations in darker colors. The diagonal of the plot shows the distributions of the individual variables. There are positive correlations between pressure and wind speed, suggesting that higher wind speeds are associated with higher air pressure. On the other hand, there are negative correlations between rainfall and dew , indicating that higher rainfall is associated with lower dew amount . Similarly, there are negative correlations between air pollution and dew , suggesting that higher air pollution levels are associated with lower dew. Overall, the correlation plot effectively summarizes the interactions and relationships between key variables related to temperature level within a single compact visual.

4.3 Fitting of a Multiple Linear Regression Model

This section covers the model fitting process which includes parameter estimation of the regression model. A Multiple linear regression model was fitted to the data to predict teperature levels. The data was split into a training set and a testing set, with 35040 observations allocated for training and 8760 observations reserved for testing.

4.3.1 Parameter estimates of the fitted multiple linear regression model

Table 4.1 is of parameter estimates for the multiple linear regression model.

TABLE 4.1: Parameter estimates for the fitted model

Variable	Estimate	Std. Error	p value
(Intercept)	522.6000	4.0700	0
Air_pollution	-0.0249	0.0003	0
dew	0.4600	0.0030	0
pressure	-0.5010	0.0040	0
wind_speed	0.0099	0.0005	0
snow size	-0.6760	0.0336	0
rainfall	-0.5330	0.0182	0

The parameters of the fitted multiple linear regression model were estimated using the ordinary least squares method. Among the independent variables, Air pollution exhibited a strong negative relationship with the dependent variable temperature, with an estimated coefficient of -0.0249. This indicated that as pollution levels increase, temperature levels tend to decrease. In contrast, dew had a positive relationship with temperature, with an estimated coefficient of 0.4600, suggesting that higher dew levels correspond to higher temperatures. Pressure demonstrated an inverse relationship with temperature, with an estimated coefficient of -0.5010. Wind speed exhibited a positive relationship with temperature, with an estimated coefficient of 0.0099, meaning that higher wind speeds are linked to higher temperatures. Snow size and rainfall both had negative associations with temperature, with estimated coefficients of -0.6760 and -0.5330 respectively. This implies that larger snowfall and increased rainfall contribute to lower temperatures.

4.4 Checking the Adequacy of the Fitted Model

This section evaluates the adequacy of fitted multiple linear regression model. Assessing model adequacy is crucial to ensure accurate representation of the data and reliable predictions. Evaluating the appropriateness of the model involved confirming a number of assumptions. This entailed assessing if

the connections between the predictors and the response variable are linear, making sure the residuals have a constant variance and are normally distributed, verifying the independence of the observations, and closely reviewing whether there is any multicollinearity among the predictor variables.

Figure 4.3 is of scatter plot for the relationship between the temperature and the explanatory variables. A scatter plot's general pattern must be examined, and factors such as direction, strength, outliers, ranges, and spread must be taken into account (Ali and Younas, 2021)

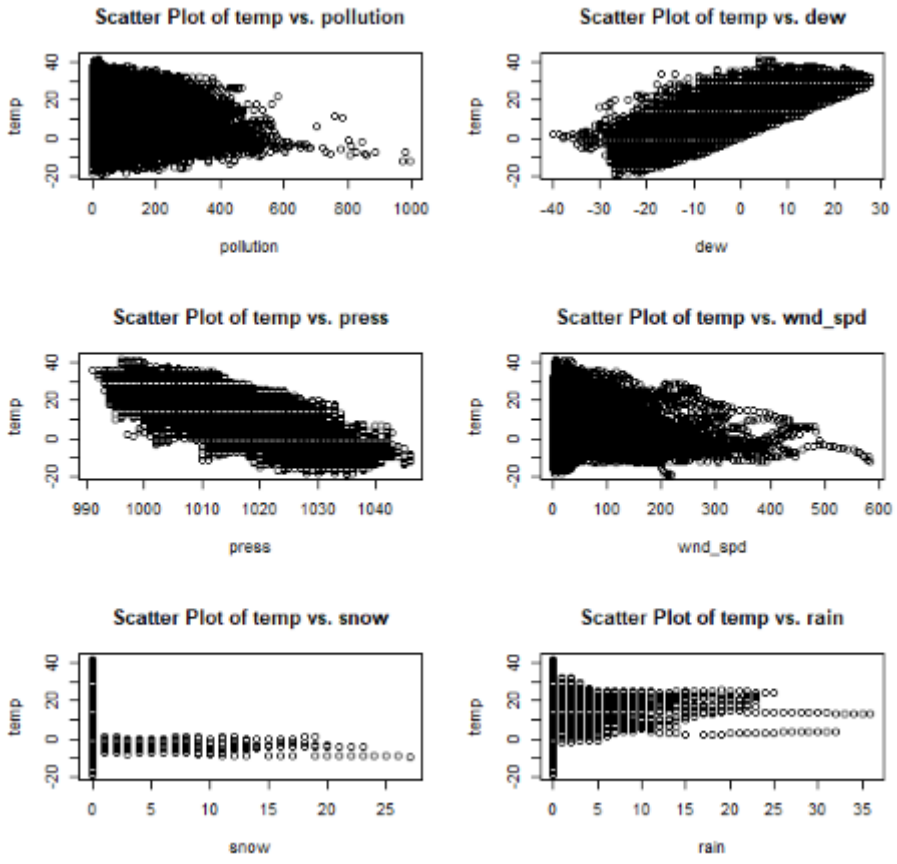

FIGURE 4.3: Scatter plots of dependent variable against independent variables

The scatter plots in Figure 4.3 visually illustrated the relationship between temperature and the explanatory variables, providing a clear representation of their associations. It was observed from the scatterplot of pollution, pressure and windspeed tend to rise along with temperature. There are a few noteworthy outliers where certain data points have significantly greater temperature levels than others, especially at higher pollutions and pressure. But the two variables showed linearity. Although there are certain data points where greater dew point values occur at higher temperatures, a general pattern of decreasing dew point values as temperature increases was seen. This can be because the dew point

is being affected by other variables. From Figure 4.3, it was seen that there is no evident association between temperature and snow depth and rainfall in this data set. Snow depth and rainfall fluctuates considerably at any given temperature, and there is no obvious trend or pattern.

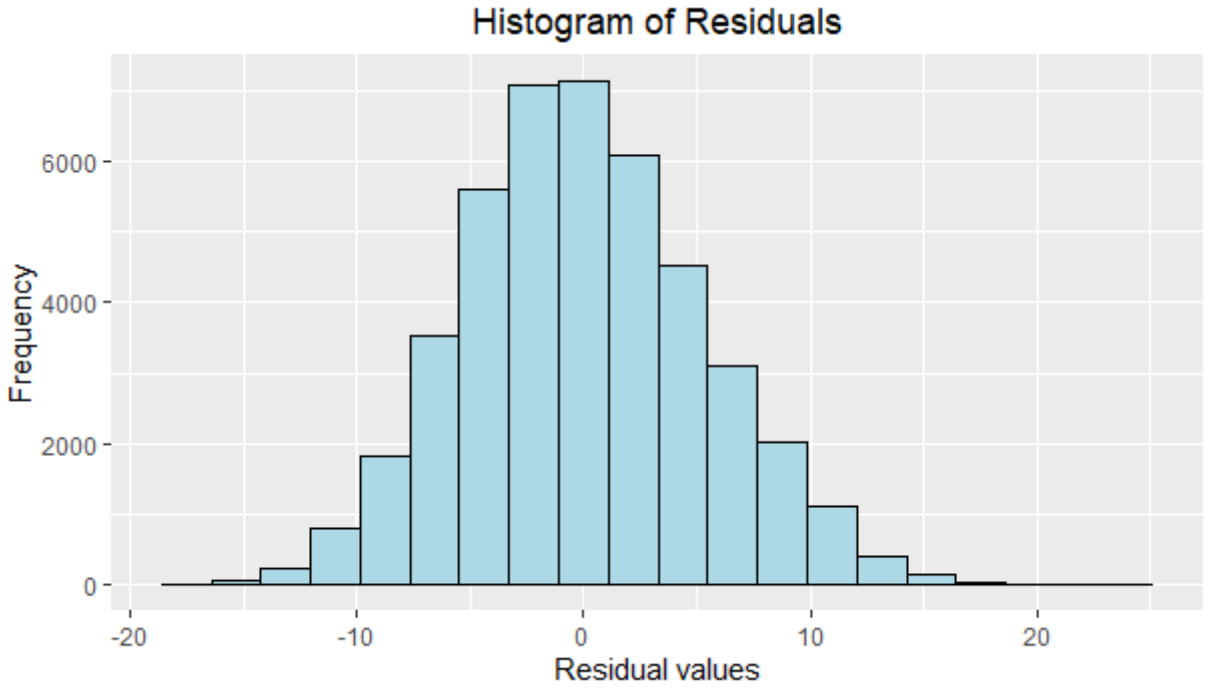

FIGURE 4.4: Histogram of residuals

Figure 4.4 is a histogram of the error terms. The figure shows that the error terms are normally distributed. For a histogram with normally distributed errors terms, the shape of the histogram is always symmetric such that the mode, mean and the median are identical. When the skewness is 0 then the error terms are said to be symmetrical, when negative, it indicate that they are negatively skewed and if positive, it indicates that the errors terms are positively skewed. A skewness of 0.000391 indicated that the distribution of residuals was approximately normal. This meant that the error terms were normally distributed.

To determine if the residuals follow a normal distribution, a Q-Q plot was employed. The residuals can be considered normally distributed if the dots on the QQ plot closely follow a straight line; deviations from the line signify deviations from this distribution (Khatun, 2021)

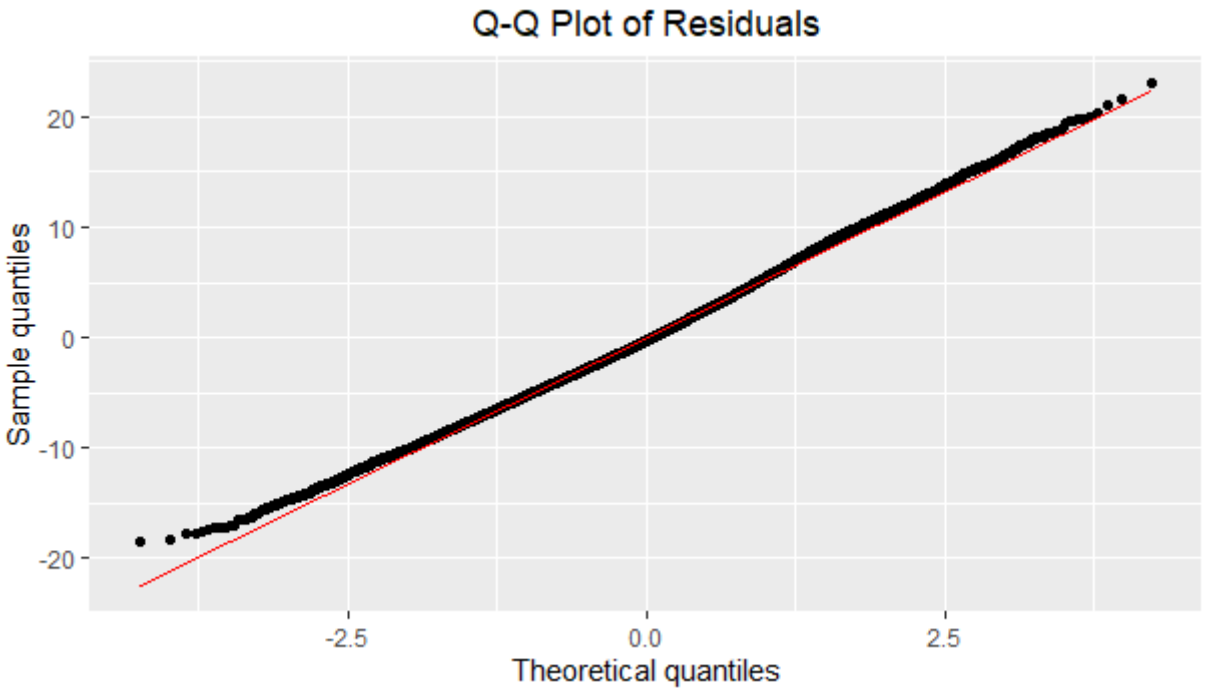

FIGURE 4.5: Q-Q plot of the error terms

Figure 4.5 is a Q-Q plot that was obtained and it enabled us to conclude that the errors exhibit a normal distribution.

The Anderson-Darling test was used to assess whether a given residuals follow the normal distribution. It evaluates the discrepancy between the observed cumulative distribution function of the sample and the expected cumulative distribution function of the hypothesized distribution, providing a quantitative measure of goodness-of-fit (Demir, 2022). The p-value of the residuals was 0 which was less than 0.05, thus the null hypothesis in equation (3.4.1) that the residuals are not normally distributed was rejected.

Durbin Watson test was used to check for autocorrelation. The Durbin-Watson statistic is a measure of autocorrelation in the residuals of a regression analysis. The Durbin-Watson test statistic was 0.11778. This value was significantly lower than 2 suggesting the presence of no autocorrelation in the residuals. A p-value of 0 was also obtained which lead to the rejection of the null hypothesis. Thus, affirming independece of residuals.

The Breusch-Pagan test for homoscedasticity had a p-value of 0 which was less than the level of significance of 0.05.This led to rejection of the null hypothesis that residuals do not have a constant

variance and the alternative hypothesis that residuals have a constant variance was accepted in equation (3.4.7). It was concluded that there was homoscedasticity in the multiple linear regression model.

The Variance Inflation Factor (VIF) values for the explanatory variables in the regression model provided insights into the extent of multicollinearity among predictors. The table below shows the different VIF values for the variables.

TABLE 4.2: Variance inflation factor values for the independent variables in the regression model

Variable	VIF
Pollution	1.0871
Dew	2.8045
Pressure	2.6030
Wind Speed	1.1458
Snow size	1.0066
Rainfall	1.0226

The 'pollution' variable exhibited little to no multicollinearity with a VIF close to 1. However, the 'dew' variable had moderate level of multicollinearity, as indicated by VIF values of 2.804533. Pressure variable had moderate level of multicollinearity, as indicated by VIF value of 2.603112. Conversely, 'wind speed,' 'snow,' and 'rain' had VIF values close to 1, suggesting minimal multicollinearity. While some level of multicollinearity is present, the VIF values did not exceed critical thresholds of 5 that's indicate severe multicollinearity (Vörösmarty and Dobos, 2020). Thus, the above values indicateded that the impact of correlated predictors on the variance of estimated coefficients was relatively modest.

TABLE 4.3: Analysis of variance table

Source	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**F-Ratio**	**p-value**
Regression	5268482	6	878080	30924	0.000
Error	1243493	43793	28.39479		
Total	6511975	43799			

Table 4.3 is the ANOVA table for the model. The P-value for the model is 0.0000 which is less than 0.05 thus implying that the model is appropriate for the data. The critical value of the F-distribution

was 3.84185. Since the computed value was 30924, which was greater than the critical value, the null hypothesis was rejected and it was concluded that the regression model was significant.

R-squared measures the proportion of variance that a regression model explains. In this study a value of 0.8090451 was obtained as the R^2 meaning that 80.90451% of the variance in the temperature levels prediction was explained by the multiple linear regression model. The R^2 increases with the increase in variables without preventing possibilities of over fitting thus it is the best to employ. Adjusted R^2 controls for each additional predictor added (to prevent from over fitting), so it may not increase as you add more variables. The value of the adjusted R^2 obtained was 0.809. These values were relatively bigger thus concluded that multiple linear regression model was the best for predicting temperature levels.

A t test was conducted and the null hypothesis was rejected for all of the independent variables, since the t-statistics for all of the independent variables in the regression model were greater than the critical value of the t-distribution at the 5% significance level. This meant that all of the independent variables were statistically significant and had a non-zero effect on the dependent variable temperature.

4.5 Prediction of temperature levels

This section presents the prediction results of the temperature levels using multiple linear regresion model The prediction equation (3.5.1) was utilized to perform the predictions for the test dataset. The resulting predictions were then combined with the actual temperature levels from the test dataset to create a data frame.

TABLE 4.4: Correlation Matrix between predicted and actual temperature levels

	Predicted	Actual
Predicted	1.00	0.89
Actual	0.89	1.00

Table 4.4 presents the correlation matrix between the predicted and the actual temperature levels. The positive correlation between the predicted values from the fitted multiple linear regression model and the actual observed temperature levels is evident. This high correlation indicates that the model provides accurate predictions that closely align with the true temperature levels. Overall, the predictive performance reflected in the correlation matrix affirms the ability of the multiple linear regression model to reliably predict temperature levels

4.6 Accuracy of the predicted temperature levels

The root mean squared error (RMSE) was used to check how accurate the fitted multiple linear regression model predicted temperature level.

4.6.1 Root Mean Square Error

The model yielded a relatively low root mean square error *RMSE* of 5.3399. This low *RMSE* value indicates the reliability of the model's temperature level predictions. As a result, the model proved to be suitable for predicting temperature levels.

Chapter 5

Conclusion

5.1 Introduction

This chapter summarized and concluded the findings of this research project. Section 5.1 provides a summary of the project, while Section 5.2 presents the conclusions drawn from the study's findings. Finally, Section 5.3 offers recommendations for future research.

5.2 Summary

The first objective of this study was to fit a Multiple Linear Regression model to analyze the data. The model parameters were estimated using the ordinary least square method, effectively revealing that each parameter exerted a significant influence on the dependent variable, temperature. This comprehensive analysis underscored the intricate relationship between predictor variables and temperature fluctuations. Consequently, it provided valuable insights into the complex dynamics of temperature prediction. The findings from this step laid a solid foundation for further investigation into temperature prediction methodologies.

Following the fitting of the Multiple Linear Regression model, the study advanced to evaluate the model's adequacy. This involved an initial examination of the distribution of error terms using the Histogram, which visually confirmed a normal distribution of errors. Statistical validation of this distribution was further conducted through the Q-Q plot and Anderson Darling test. Subsequently, the study evaluated autocorrelation of residuals using the Durbin Watson test, which revealed no significant correlation, thereby enhancing the model's reliability. Homoscedasticity, critical for regression analysis, was affirmed through the Breusch Pagan test, indicating constant variance in residuals. Additionally, the study assessed multicollinearity among predictors using the Variance Inflation Factor, suggesting a modest impact despite some predictor correlation in Dew and Pressure. Lastly, the significance of estimated parameters was evaluated through tests such as the F-test and t-test, which demonstrated the overall significance of the regression model and the statistical significance of independent variables.

With a validated Multiple Linear Regression model established, this study employed the model to predict temperature levels. By utilizing the model on the test dataset, the study generated predictions of temperature based on various environmental factors such as dew, pressure, windspeed, snow size, and rainfall. The aim of predicting temperature levels using the fitted Multiple Linear Regression model was to enable stakeholders with precise predictions of temperature fluctuations. This predictive capacity empowers stakeholders to make well-informed decisions regarding resource management, risk assessment, and environmental planning.

The final objective was to evaluate the accuracy of the predicted temperature levels generated by the Multiple Linear Regression model. Root Mean Square Error was utilized to assess the model's predictive performance by comparing the predicted temperature levels with the actual observed values. Through the evaluation of the accuracy of the predicted temperature levels, the study validated the effectiveness and reliability of the Multiple Linear Regression model in estimating temperature fluctuations, thereby enabling stakeholders with a dependable tool for making informed decisions, optimizing resource allocation, and improving overall performance in various sectors reliant on temperature forecasting.

5.3 Conclusion

It is evident from the data and analysis in this study that using Multiple Linear Regression models to predict temperature levels has shown to be a helpful and effective strategy. In addition to improving the precision of temperature predictions, the use of Multiple Linear Regression model has given stakeholders more capacity to make data-driven choices about resource allocation, risk assessment, and climate policy across a range of industries. A greater knowledge of the intricate dynamics involved in temperature prediction has also been made possible by the incorporation of MLR models, enabling stakeholders to more effectively adjust their plans and guarantee sustainable environmental management practices. In addition to streamlining decision-making procedures overall, this data-driven strategy helps organizations become more resilient and long-term sustainable in the face of climate change. The use of Multiple Linear Regression models for temperature prediction is a major advancement in the effort for more transparent and dependable environmental procedures. It also promotes an atmosphere of sustainability, responsibility, and resilience in the face of climate change. In light of this information, stakeholders should think about using Multiple Linear Regression model as an essential tool to improve stakeholder satisfaction, environmental stewardship, and operational efficiency in the dynamic field of climate-related decision-making.

5.4 Recommendations for Further Research

While Multiple Linear Regression has been widely used in temperature prediction, it faces limitations in handling non-linear variables effectively. To address this challenge and improve the accuracy of temperature predictions, this study recommends the adoption of Generalized Linear Models. Generalized Linear Models offer significant advancements in predictive modeling by allowing for non-linear relationships between temperature predictors and the response variable. These models extend the capabilities of Multiple Linear Regression by incorporating smooth functions such as splines, enabling the capture of complex non-linear patterns in temperature data. Additionally, it is crucial to consider not only the direct impact of environmental indicators on temperature but also broader contextual factors such as geographical features and land use patterns. Therefore, the inclusion of additional variables such as altitude and land cover types is essential for comprehensive temperature prediction models using Generalized Linear Models.

References

Ali, Parveen and Ahtisham Younas (2021). "Understanding and interpreting regression analysis". In: *Evidence-Based Nursing* 11.1, pp. 110-113.

Anusha, N and M Sai Chaithanya (2019). "Weather prediction using multi linear regression algorithm". In: *IOP Conference Series: Materials Science and Engineering*. Vol. 590. 1. IOP Publishing, pp. 12-34.

Demir, Süleyman (2022). "Comparison of normality tests in terms of sample sizes under different skewness and Kurtosis coefficients". In: *International Journal of Assessment Tools in Education* 9.2, pp. 397–409.

Granata, Francesco, Fabio Nunno, Quoc Bao Pham, and Giovanni de Marinis (2022). "Precipitation forecasting in Northern Bangladesh using a hybrid machine learning model". In: *Sustainability* 14.5, pp. 26-33.

Khatun, Nasrin (2021). "Applications of normality test in statistical analysis". In: *Open Journal of Statistics* 11.1, pp. 12-17.

Luthfiarta, Ardytha, Aris Febriyanto, Heru Lestiawan, and Wibowo Wicaksono (2020). "Analysis of Weather Forecasting with Parameters of Temperature, Humidity, Air Pressure, and Wind Speed Using Multiple Linear Regression". In: *Journal Information System*, pp. 10-17.

Mulyani, Evi Dewi Sri, Indah Septianingrum, Nisa Nurjanah, Reka Rahmawati, Syifa Nurhasani, and Kiky Milky RK (2019). "Rainfall Prediction in Majalengka District Using Regression Algorithm". In: *Journal of Information Systems and Technology Information* 8.1, pp. 67–77.

Saragih, Immanuel Jhonson Arizona, Inlim Rumahorbo, Ricko Yudistira, and Dedi Sucahyono (2020). "Monthly Rainfall Prediction in Deli Serdang Using Regression Equation With Temperature and Humidity Data as Predictors". In: *Journal of Meteorology, Climatology, and Geophysics* 7.2, pp. 6–14.

Sreehari, E and Satyajee Srivastava (2019). "Prediction of climate variable using multiple linear regression". In: *2019 4th International Conference on Computing Communication and Automation (ICCCA)*. 4. Pp. 1–4.

Sulistiyono, Mulia, Acihmah Sidauruk, Budy Satria, Raditya Wardhana, et al. (2023). "Rainfall Prediction Using Multiple linear Regression ALGORITHM". In: *Journal of Computer Science and Technology* 9.1, pp. 17–22.

Vörösmarty, Gergely and István Dobos (2020). "Green purchasing frameworks considering firm size: a multicollinearity analysis using variance inflation factor". In: *Supply Chain Forum: An International Journal* 21.4, pp. 290–301.

Yusuf, Muhammad (2022). "Analysis of Rainfall Prediction in the Sorong City Area Using Time Series Forecasting Method". PhD thesis. Amikom University Yogyakarta, pp. 34-37.

Appendix

A.1 R Program Codes

```
# Installing the necessary packages
install.packages("caret")
install.packages("data.table")
install.packages("data.table")
install.packages("readr")
install.packages("gridExtra")
install.packages("ggplot2")
install.packages("stats")
install.packages("stats")
install.packages("lmtest")
install.packages("car")
#loading the required libraries
library(readr)
library(gridExtra)
library(ggplot2)
library(stats)
library(lmtest)
library(car)
#loading the data
data <- read_csv("LSTM-Multivariate_pollution.csv")
# Fitting  the model
model <- lm(temp ~ pollution + dew + press +
wnd_spd + snow + rain, data = data)
# Print the summary of the model
summary(model)
#Generating scatter plots
Data is in a data frame called 'data'
independent_vars <- c("pollution", "dew", "press", "wnd_spd", "snow", "rain")
par(mfrow = c(3, 2))  # Arrange plots in a 2x3 grid
```

```
for (var in independent_vars) {
plot(data[[var]],data$temp, main =
paste("Scatter Plot of temp vs.", var),
xlab = var, ylab = "temp")}
residuals = model$residuals
print(residuals)
#generating the histogram for residuals
ggplot(data.frame(residuals = model$residuals), aes(x = residuals)) +
geom_histogram(bins = 20, fill = "lightblue", color = "black") +
labs(title = "Histogram of Residuals", x = "Residual values", y = "Frequency")
+theme(plot.title = element_text(hjust = 0.5))
#Generating the Q-Q plot of residuals
ggplot(data.frame(residuals = model$residuals), aes(sample = residuals)) +
stat_qq() +
stat_qq_line(color = "red") +
labs(title = "Q-Q Plot of Residuals", y = "Sample quantiles", x = "Theoretical quantil
# Center title
theme(plot.title = element_text(hjust = 0.5)) +
# Center axis labels
theme(axis.title.x = element_text(hjust = 0.5),
axis.title.y = element_text(hjust = 0.5))
# Durbin Watson test
dwtest(model)
#brewtch pagan test
bp_test <- bptest(model)
# Print the result
print(bp_test)
#chi square
df <- 6
# Significance level (e.g., 0.05 for a 95% confidence level)
alpha <- 0.05
# Obtain critical value
```

```
critical_value <- qchisq(1 - alpha, df)
# Print the critical value
cat("Critical value at", 1 - alpha, "quantile for df =", df, ":", critical_value, "\n"
# Calculate VIF for each predictor
vif_values <- vif(model)
# Print the VIF values
print(vif_values)
#F TEST
# Perform analysis of variance (ANOVA) to obtain the F-ratio
anova_result <- anova(model)
# Print the ANOVA table
print(anova_result)
# Extracting the F-ratio and its associated p-value
f_ratio <- anova_result$F[1]
p_value <- anova_result$`Pr(>F)`[1]
# Print the F-ratio and p-value
cat("F-ratio:", f_ratio, "\n")
cat("P-value:", p_value, "\n")
# Extract the R-squared value
r_squared <- summary_model$r.squared
# Print the R-squared value
cat("R-squared:", r_squared, "\n")
# Set seed for reproducibility
set.seed(123)
# Create splits with stratified sampling (specify target variable)
train_index <- createDataPartition(data$temp, p = 0.8, list = FALSE)
# Create training and testing sets
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
# Calculating RMSE
#  Fit the model on the training set
model <- lm(temp ~ pollution + dew + press + wnd_spd + snow + rain, data = train_data)
```

```
# Make predictions on the testing set
predictions <- predict(model, newdata = test_data)
#  Calculate RMSE
rmse <- rmse(predictions, test_data$temp)
# Print the RMSE
mae <- mean(abs(test_data$temp - predictions))
print(paste("MAE:", mae))
```