fieldnumber2

fieldsortinithashY

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

2023

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 proposal 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 proposal, 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 and assistance towards completing this project. We would also like to pass our vote of thanks to our coordinator *Dr. Maina Mundia*, *Dr. Cyprian Omari* and *Dr. Anthony Ngunyi* for their endless efforts in giving distinct encouragement and advice concerning our project.

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 fortheir financial support and guidance they rendered to us during our project writing.

Contents

List of Figures

List of Tables

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

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

Abstract

Weather unpredictability has risen in recent decades due to climate change, which is caused by a number of variables including growing greenhouse gas emissions and natural variability. This pressing global concerns and its far-reaching consequences makes the ability to predict temperature levels accurately paramount. Temperature is a fundamental environmental variable that has an influence on weather, human health, ecosystems, agriculture, energy usage, infrastructure, water bodies, food production, recreation, industrial processes, transportation and global climate change. current temperature prediction models often fall short in capturing the relationship between multiple predictor variables and temperature changes. This research aims to predict temperature levels through the application of the Multiple Linear Regression model. The model parameters will be estimated using the Ordinary Least Squares estimation method. Scatter plots will be used to check for linearity. Histogram, Q-Q plots and Anderson Darling test will be used to check for the normality assumption. Durbin Watson test will be used to check if the residuals are correlated. Breusch-Pagan test will be used to check for homoscedasticity. Variance inflation factor will be computed to check for multicollinearity. Significance tests such as the F-test and t-test will be conducted to evaluate the significance of estimated parameters. RMSE will be used to assess the accuracy of the predicted values. This research aims to enhance temperature predictions reliability, benefiting a wide range of industries and communities that depend on precise temperature information. The dataset chosen for this study is ”Air Pollution - Multivariate” dataset from Kaggle: https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate, comprises of 43,800 observations and seven variables: air pollution (meters per second), dew (millimetres), pressure (pascals), windspeed (meters per second), snow size (inches) and rainfall (millimetres), dependent variable is temperature (degrees celcius).

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 (CO2) into the air. Deforestation and land-use changes further exacerbate the issue, as trees play a crucial role in absorbing CO2. Additionally;industrial processes, agriculture and certain waste management practices release methane (CH4) and nitrous oxide (N2O) 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. While natural factors such as volcanic eruptions and variations in solar radiation also influence the climate, the current trend of global warming is largely attributed to human-induced factors. Understanding and addressing the sources of climate change are essential for developing effective mitigation strategies to minimize the adverse impacts on ecosystems, weather patterns and the overall stability of the planet's climate system.

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. The capacity to accurately predict temperature changes is critical to reducing the negative consequences on many industries as global temperatures fluctuate and extremes become more common. Precise temperature predictions, for example, 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. For example, 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.

However, it's important to note that multiple 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 will be to predict temperature levels using multiple linear regression model.

1.5.2 Specific objectives

The specific objectives of the study will be;

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 including businesses, government agencies, environmental conservationists and the general population.

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.

Government agencies entrusted with public well-being benefit significantly from precise temperature predictions. Timely and accurate information allows for improved public health planning particularly in anticipating and managing heatwaves or extreme cold events. Transportation infrastructure planning is also optimized as agencies can proactively address challenges posed by temperature-related impacts, such as snowfall or heat-induced stress on roads and railways.

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.

Essentially, the precise temperature predictions produced by this research serve as a pivot, enabling a diverse range of stakeholders to make knowledgeable choices, improve readiness, and aid in the more general objective of constructing resilient and sustainable communities in the face of a constantly shifting climate.

Chapter 2

Literature Review

2.1 Introduction

This section discusses important concepts and insights utilized in the research, highlighting specific theoretical contributions from prior literature. Conducting a literature review serves the purpose of deepening comprehension of previous studies relevant to the research goals. It also aids in refining the foundational concepts upon which the research is constructed. The primary objective is 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

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 (r=0.90) demonstrated the model's capability to capture patterns and relationships between the chosen weather parameters and monthly rainfall.

Anusha et al. (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 (T) and humidity (RH) as predictors. The dataset encompassed 30 years of RR, T, and RH 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 (BMKG), 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 (R2) 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.

Di Nunno 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 (R2) 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 articles collectively reveal a common gap in climate prediction research - the limited attention given to temperature prediction. While these studies offer valuable insights into rainfall prediction, temperature, a vital climate variable with extensive implications for agriculture, transportation and disaster management remains relatively understudied. Multiple linear regression, a widely used method for climate prediction has not been extensively applied to temperature prediction leaving an unexplored research opportunity. To address this gap, our research will apply multiple linear regression model to predict temperature levels. The research will include a model assessment using metrics such as the root mean squared error and confusion matrix to ensure accuracy. The research direction will have the potential to significantly enhance temperature prediction, benefiting various sectors reliant on accurate temperature predictions. Expanding the application of multiple linear regression to

encompass temperature prediction represents a promising avenue for future climate prediction studies.

Chapter 3

Methodology

3.1 Introduction

This chapter will provides an overview of the methodology underpinning the model used to predict temperature levels. It outlines the process of data analysis, encompassing analytical models and significance tests, and elucidates how conclusions were drawn from this analysis. 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 delves into the methods employed to assess the accuracy of these predictions. Lastly, Section 3.7 provides insight into the data source and its description.

3.2 Multiple Linear Regression Model

Multiple regression analysis is a statistical analysis technique model that determines the relationship between a response variable and some combination of two or more explanatory variables. The model performed in observance of five assumptions. They include, normality assumption in that the model assumes that the residuals are normally distributed. Non-normally distributed residuals(highly skewed or kurtotic variables or variables with substantial outliers) may distort the relationships and significance tests. The model also assumes that the residuals are independent that is; there is no autocorrelation between the residuals. This means that the error term of one observation is not influenced by the error term of another observation. The model assumes that there must exist a linear relationship between the response variable and each of the the explanatory variable. This assumption can best be tested with a scatter plot whereby it will aid in checking for outliers as multiple linear regression analysis is sensitive to outliers effects. Homoscedasticity is the fourth assumption of multiple linear regression where residuals are assumed to have a constant variance that is the variance of the errors are same across all levels. Multicollinearity assumption states that there is no or little multicollinearity. The model assumes that the explanatory variables are not highly correlated to each other. If there exist a correlation between the residuals, then the regression coefficients are said to be unstable. The

model describes how a response variable depends linearly on a number of predictor variables. Let Y_i be the response variable and $X_1, X_2, \ldots, X_k$ be the explanatory variables.

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,

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

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

is the vector of residuals.

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

3.3 Estimation of Parameters

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 Ordinary Least Squares Estimation Method

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 $\hat{\boldsymbol{\beta}}$ that minimizes. Let

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

Differentiating equation (3.3.1) with respect to $\boldsymbol{\beta}$ 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.

3.3.2 Properties of Least Squares Estimators

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,

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

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

3.4 Model Adequacy

After the estimation of the parameters, a diagnostic analysis will be performed to check the adequacy of the fitted regression model.

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.

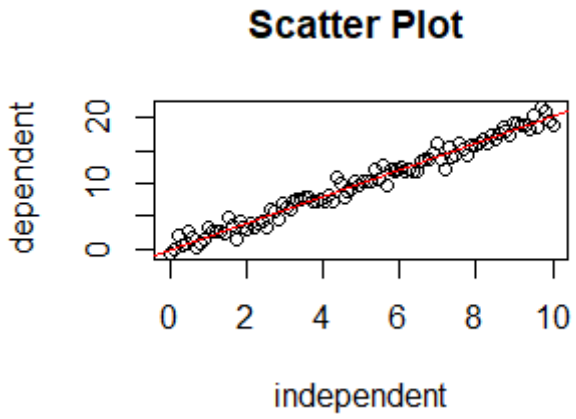


FIGURE 3.1: An illustration of a scatter plot for a linear relationship

3.4.2 Residual plots

Assumption of normality will be examined using two aspects that is, the graphical approach using the histogram and the Q-Q plot. The other approach will be using the 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

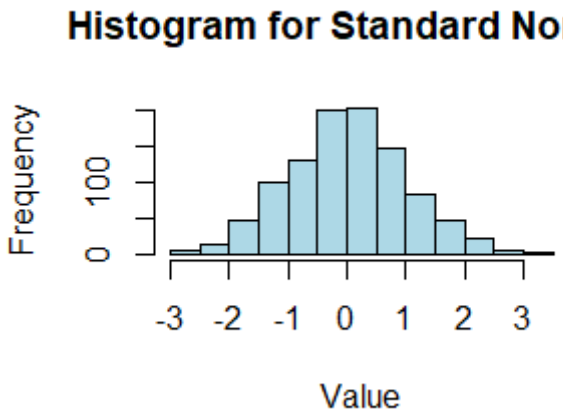


FIGURE 3.2: An illustration of a histogram for standard normal distribution

Quantile-Quantile (Q-Q) plot

The Q-Q plot is a graphical tool that is used to asses if the set of data came from some statistical distribution such as the normal distribution. A Q-Q plot is a scatter plot created by plotting two sets of quantiles against one another. If both sets of quantiles come from the same statistical distribution, the plot should form a line that is roughly straight.For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Q-Q plot to check that assumption. It allows one if the assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation. The Q-Q plot takes different shapes depending with the dataset. If all the points plotted perfectly lies on a straight line, then the Q-Q plot can be reffered as ideal in that the data in normarly distributed. A data can also be positively that is, the mean is grater than the median as the data is more towards the lower side or vice versa for a negatively skewed data that is the data has a tail on the left side. Skewness is the measure of asymmetry of a distribution. If the bottom end of a Q-Q plot deviates from the straight line but the upper end is not, then the distribution is said to have a longer tail to its left then it is said to be left skewed or negatively skewed. If the upper end of the Q-Q plot is observed to have deviated from the straight line and the lower follows a straight line then the curve has a longer tail to its right, it is said to be positively skewed or right- skewed.

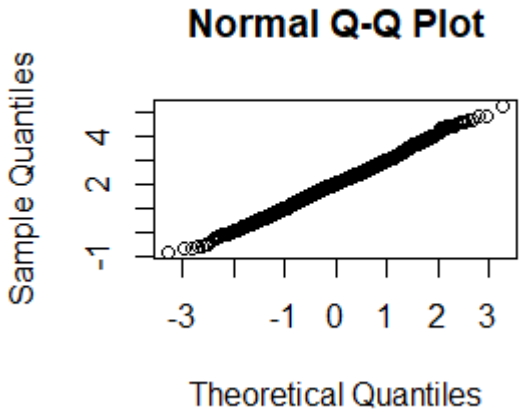


FIGURE 3.3: An illustration for normally skewed errors

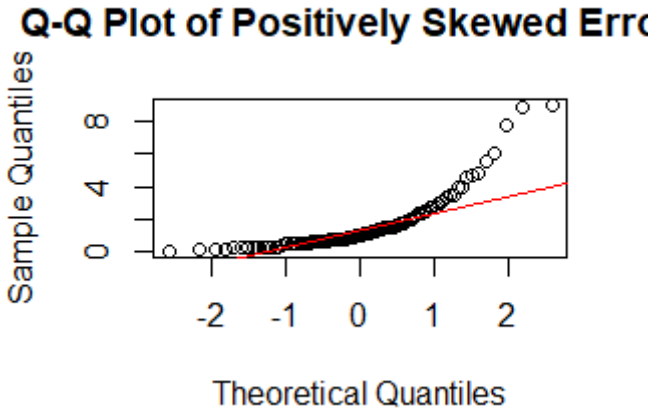


FIGURE 3.4: An illustration of a positively skewed errors

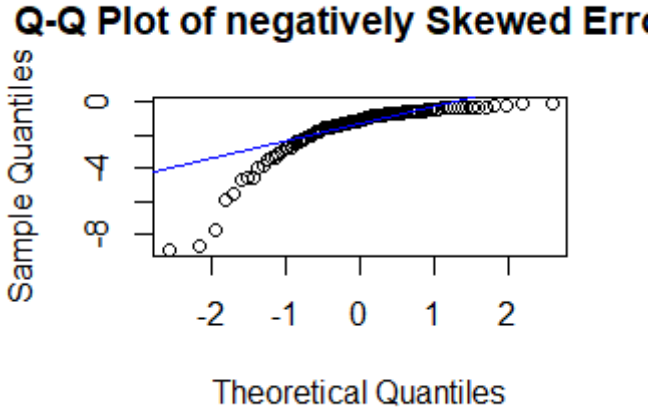


FIGURE 3.5: An illustration of a negatively skewed errors

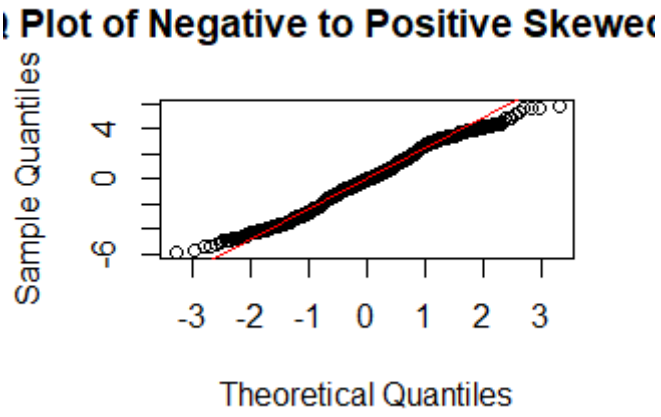


FIGURE 3.6: An illustration of negative and positive skewed errors

3.4.3 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.

3.4.4 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.

3.4.5 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.

3.4.6 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.

3.4.7 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

$$\begin{aligned} H_0 &: \beta_i = 0 \text{ for all } i \\ H_1 &: \beta_i \neq 0 \text{ for atleast one } i\text{=1,2...k} \end{aligned} \tag{3.4.10}$$

The sum of squares due to errors will be;

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{3.4.11}$$

The sum of squares due to regression is;

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{3.4.12}$$

The total sum of squares;

$$SST = SSR + SSE = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{3.4.13}$$

The null hypothesis is true,then the statistic:

$$F_0 = \frac{SSR}{k} \div \frac{SSE}{n-k} = \frac{MSR}{MSE} \sim F_{k,n-k} \tag{3.4.14}$$

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: ANOVA 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	$n-1$		

3.4.8 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.

$$R_a^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} \tag{3.4.15}$$

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.

3.4.9 t-test

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

$$\begin{aligned} &H_0 : \beta_i = 0 \\ &H_1 : \beta_i \neq 0 \text{ for } i\text{=1,2...k} \end{aligned} \tag{3.4.16}$$

the test statistic is;

$$t = \frac{\hat{\beta}_i}{Se(\hat{\beta}_i)} \tag{3.4.17}$$

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 $|t| > t_{\frac{\alpha}{2},(n-k-1)}$.

3.5 Prediction of the new observations

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

$$\mathbf{Y}\text{predict} = \hat{\boldsymbol{\beta}}^T X_0 \tag{3.5.1}$$

This is an unbiased estimator of the future response.

$$\begin{aligned} E[\mathbf{Y}\text{predict}|\mathbf{X} = X_0] &= \text{var}(X_0\boldsymbol{\beta}) + \text{var}(\boldsymbol{\epsilon}) \\ &= [X_0(\mathbf{X}\hat{T}\mathbf{X})^{-1}\mathbf{X}_0^T]\boldsymbol{\sigma}^{\mathbf{2}} + \boldsymbol{\sigma}^{\mathbf{2}} \\ &= \boldsymbol{\sigma}^{\mathbf{2}}[1 + X_0(\mathbf{X}^T\mathbf{X})^{-1}X_0^T] \end{aligned} \tag{3.5.2}$$

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

$$\frac{\mathbf{Y}\text{predict} - E[\mathbf{Y}\text{predict}]}{\sqrt{\sigma^2(1+X_0(\mathbf{X}^T\mathbf{X})^{-1}X_0^T)}} \sim t(n-p) \tag{3.5.3}$$

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

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

3.6 Accuracy of the predictions

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

3.6.1 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;

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{3.6.1}$$

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

3.7 Data Source and Description

The dataset chosen for the study, the ”Air Pollution - LSTM Multivariate” dataset from Kaggle: https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate, comprises of 43,800 observations and seven variables: air pollution(meters per second), dew(millimetres), pressure(pascals), windspeed(meters per second), snow size (inches) and rainfall(millimetres), dependent variable is temperature(degrees celcius).

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 dataset chosen for the study, the ”Air Pollution - LSTM Multivariate” dataset from Kaggle: https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate, comprises of 43,800 observations and seven variables: air pollution(meters per second), dew(degrees celcius), pressure(pascals), windspeed(meters per second), snow size (inches) and rainfall(millimetres), dependent variable is temperature(degrees celcius).

4.3 Parameter estimates of the fitted multiple linear regression model

The parameters of the fitted multiple linear regression model were estimated using the ordinary least squares method. Among the independent variables, pollution exhibited a strong negative relationship with the dependent variable, temperature, with a coefficient estimate of -0.02488. This indicated that as pollution levels increase temperature levels tend to decrease. In contrast, dew had a positive relationship with temperature, represented by a coefficient estimate of 0.4599, suggesting that higher dew levels correspond to higher temperatures. Pressure demonstrated an inverse relationship, as a decrease in pressure, represented by the coefficient estimate of -0.5005, was associated with lower temperatures. Wind speed exhibited a positive correlation with temperature, indicated by a coefficient estimate of 0.00997, meaning that higher wind speeds are linked to higher temperatures. Snow size and rainfall both had negative associations with temperature, with coefficient estimates of -0.6759

and -0.5329, respectively, implying that larger snowfall and increased rainfall contribute to lower temperatures.

The multiple linear regression model is of the form;

$$\begin{aligned} temperature =& 522.6 - 0.02488 pollution + 0.4599 dew \\ & - 0.5005 press + 0.00997 wnd\_spd \\ & - 0.6759 snow - 0.5329 rain \end{aligned} \tag{4.3.1}$$

In the multiple linear regression equation (4.2.1), all of the variables *Pollution*, *Dew*, *Press*, *Rain*, *snow* and *WindSpeed* displayed statistical significance ($p < 0.05$), which signified their substantial influence on the dependent variable Temperature.

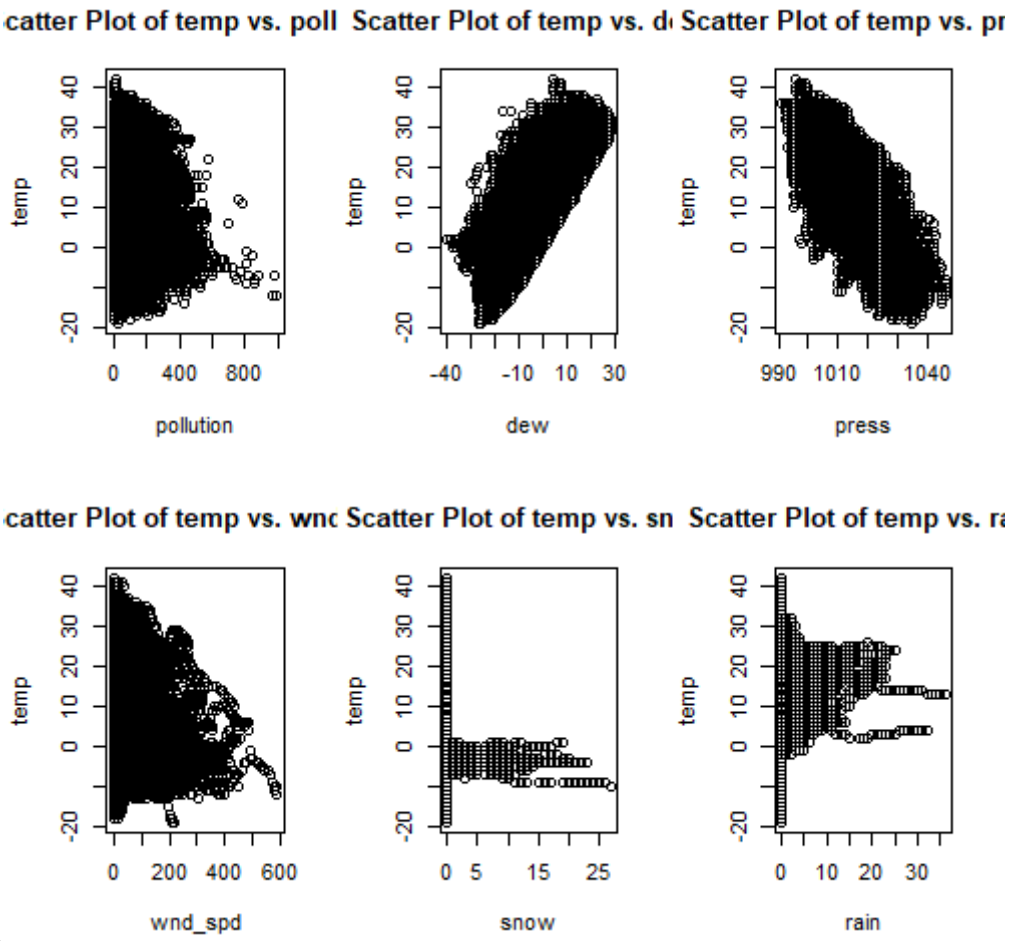
The table below is of the parameter estimates for the multiple linear regression model.

Variable	Estimate	Std. Error	Pr($>|t|$)
(Intercept)	5.23e+02	4.07e+00	$< 2e-16$
pollution	-2.49e-02	2.88e-4	$< 2e-16$
dew	4.60e-01	2.96e-3	$< 2e-16$
press	-5.01e-01	4.00e-3	$< 2e-16$
wnd_spd	9.97e-03	5.45e-4	$< 2e-16$
snow	-6.76e-01	3.36e-2	$< 2e-16$
rain	-5.33e-01	1.82e-2	$< 2e-16$

TABLE 4.1: Parameter estimates for the fitted model (reduced size)

4.4 Model adequacy

Linearity



??

FIGURE 4.1: Scatter plots

Figure 4.1 is scatter plot for the relationship between the temperature and the explanatory variables .The scatter plots visually illustrated the relationship between temperature and the explanatory variables, providing a clear representation of their associations. The data points on the plot were dispersed in a manner that suggested an approximate linear relationship, indicating that changes in the explanatory variables are accompanied by corresponding changes in temperature.

Normality

The test for normality was examined using a histogram, a Q-Q plot of residuals and using Anderson Darling test.

Histogram

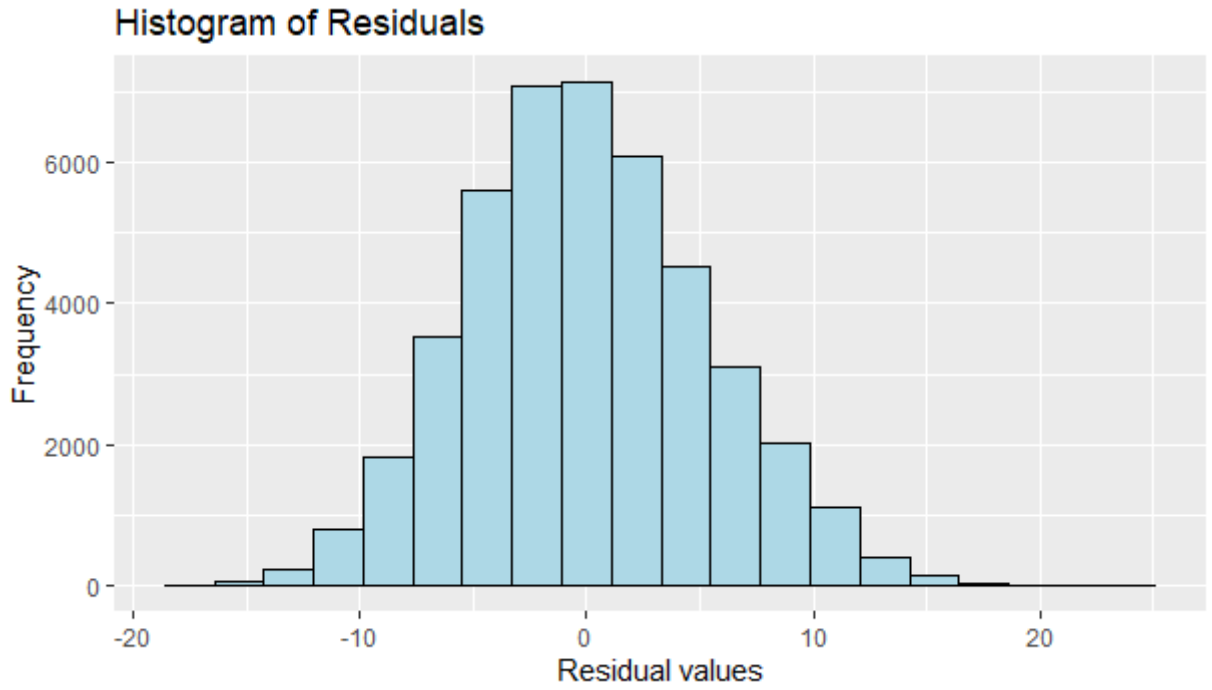


FIGURE 4.2: Histogram of models residuals

Figure 4.2 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 skewedeness 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.

Q-Q plot

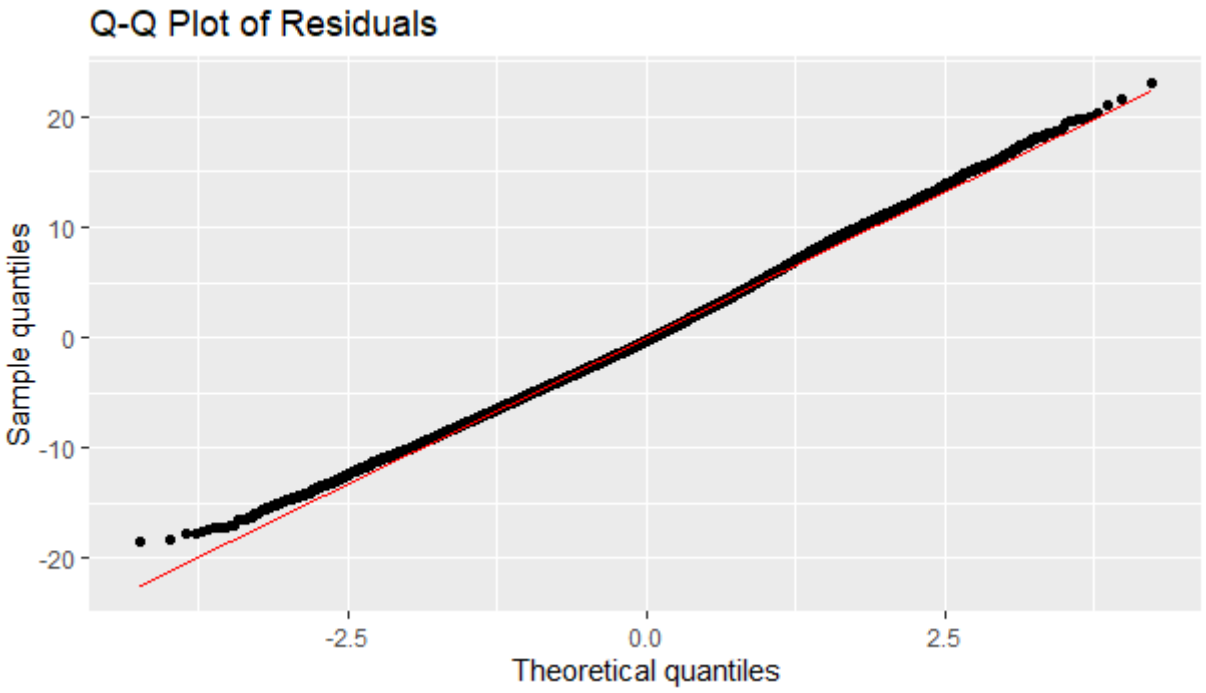


FIGURE 4.3: Q-Q plot of the error terms

Figure 4.3 is a Q-Q plot that was obtained and it enabled us to conclude that the errors exhibit a normal distribution.

Anderson darling test

The p-value of the residuals was 2.2e-16 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

The Durbin-Watson test statistic was 0.11778. This value is significantly lower than 2, suggesting the presence of positive autocorrelation in the residuals. A p-value of 2.2e-16 was obtained which lead to the rejection of the null hypothesis that there is no autocorrelation.

Breusch-Pagan test

The Breusch-Pagan test for homoscedasticity had a p-value of $2.2e-16$ 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.

Variance inflation factor

Variable	VIF
Pollution	1.087140
Dew	2.804533
Press	2.603
Wind Speed	1.145802
Snow	1.006647
Rain	1.022665

TABLE 4.2: VIF values for the independent variables in the 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 'pollution' variable exhibited little to no multicollinearity with a VIF close to 1. However, the 'dew' and 'pressure' variables had moderate levels of multicollinearity, as indicated by VIF values of 2.804533 and 2.603112, respectively. 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, indicating that the impact of correlated predictors on the variance of estimated coefficients was relatively modest.

F test

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.

Source	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**F-Ratio**
Regression	5268482	6	878080.3	30924
Error	1243493	43793	28.39479	
Total	6511975	43799		

TABLE 4.3: ANOVA Table

The critical value of the F-distribution was 3.84185. Since the F-ratio 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 and adjusted R squared

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 0.809019 which was a relatively smaller percentage than the R^2 .Those values were relatively bigger thus concluded that multiple linear regression model was the best for predicting temperature levels.

t test

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 and Accuracy of Prediction

4.5.1 RMSE

The dataset was divided into two parts, with 80% of the data used for training and the remaining 20% used for testing. The model yielded a relatively low root mean square error *RMSE* of 5.4837. This low *RMSE* value indicates the reliability of the model's temperature level predictions. As a result, the model proves 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.

5.1.1 Summary

The aim of the research was to predict temperature levels using multiple linear regression model. The factors like pollution, dew, air pressure wind spees, snow and rain were considered to be linked with the temperature levels.

The method of ordinary least squares was used to estimate the parameters of the multiple linear regression model. variables with the biggest impact on temperature levels were dew, snow, rain, and pressure while the variables with the smallest impact on temperature levels were pollution and wind speed.

The adequacy of the fitted model was assessed using the t-test and the F-test both at 5% level of significance.

Prediction of temperature levels was performed and the accuracy of the predicted results was assessed by the use of root mean squared error method $RMSE$. The results obtained implied that the multiple linear regression model is suitable for the data thus appropriate in predicting temperature.

5.2 Conclusion

An accurate prediction of temperature can help organization and governments to improve on planning, reduce risk, improve efficiency and protect human health and well-being. In the study temperature was being predicted using multiple linear regression model. From the results obtained it was concluded that multiple linear is a good model to use in temperature prediction.

5.3 Recommendations for Further Research

For further research the study recommended use of machine learning algorithms such as random forests, gradient boosting machines (GBMs), and deep neural networks (DNNs) which gives room for modeling both complex non-linearity scenarios and interactions that multiple linear regression models cannot offer.

References

Anusha, N, M Sai Chaithanya, Guru Jithendranath Reddy, and N Shanakar (2019). "Weather Prediction Using Multi Linear Regression Algorithm". In: *International Journal of Climatology* 4.1, pp. 12-18.

Di Nunno, Fabio, Francesco Granata, Quoc Bao Pham, and Giovanni de Marinis (2022). "Precipitation forecasting in Northern Bangladesh using multiple linear regression model". In: *Sustainability* 14.5, pp. 26-35.

Luthfiarta, Ardytha, Aris Febriyanto, Heru Lestiawan, and Wibowo Wicaksono (2020). "Analysis of Weather Forecast with Parameters Temperature, Humidity, Air Pressure, and Wind Speed Using Multiple Linear Regression". In: *JOINS (Journal Inf. Syst., vol. 5, no. 1, pp. 10–17, 2020, doi: 10.33633/joins*.

Mulyani, Evi Dewi Sri, Indah Septianingrum, Nisa Nurjanah, Reka Rahmawati, Syifa Nurhasani, and Kiky Milky RK (2019). "Rainfall Prediction in Majalengka Regency Using Regression Algorithm". In: *E-JOURNAL JUSITI: Journal of Information Systems and Information Technology* 8.1 (2019), 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 Predictors". In: *Journal of Meteorology, Climatology, and Geophysics* 7, pp. 6–14.

Sreehari, E and Satyajee Srivastava (2019). "Prediction of climate variable using multiple linear regression". In: *2018 4th International Conference on Computing Communication and Automation (ICCCA)*. IEEE, pp. 1–4.

Sulistiyono, Mulia, Acihmah Sidauruk, Budy Satria, Raditya Wardhana (2023). "Rainfall prediction using multiple linear regression". In: *JITK (Journal of Science and Computer Technology)* 9.1, pp. 17–22.

Yusuf, Muhammad (2022). "Analysis of rainfall prediction in the Sorong city using multiple linear regression". In: *Weather Forecasting*, pp. 34-37.

Appendix

A.1 Proposed Budget

TABLE A.1: Proposed Project Budget

Expense	**Amount (sh)**
Printing Cost	5000
Binding	100
internet	500
Total	**5600**

A.2 Project Work Plan