UNIVERSITI TEKNOLOGI MARA

COMPARISON OF REGRESSION MODELS FOR PM_{10} CONCENTRATION IN KLANG WITH PRESENCE OF OUTLIERS

FATEN HANNANI BINTI FADZEL
NURUL NADIAH BINTI OMAR
NABILAH BINTI ABDUL RAHMAN

BSc

December 2019

UNIVERSITI TEKNOLOGI MARA

COMPARISON OF REGRESSION MODELS FOR PM_{10} CONCENTRATION IN KLANG WITH PRESENCE OF OUTLIERS

FATEN HANNANI BINTI FADZEL
NURUL NADIAH BINTI OMAR
NABILAH BINTI ABDUL RAHMAN

Report submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science (Hons.) Statistics

Faculty of Computer and Mathematical Sciences

December 2019

ABSTRACT

PM_{10} concentration is a type of air pollutant which is a mixture of solid and liquid particles dangle in the air with diameter at most to 10 micrometer (mm). Measurements of air pollution often contain outliers that can significantly affect further analysis. However, outliers may contain valuable information. This study focuses on comparing the performance of regression models in modelling PM_{10} concentration in Klang area with the presence of outliers. The dataset were obtained from Department of Environment (DOE) and the variables involve are meteorological factors (temperature, humidity, wind direction) and gaseous concentration (NO_2, SO_2, CO, O_3). The method used in this study are robust regression methods (Huber and Bisquare) and Ordinary Least Square (OLS) method. The comparison between method using Coefficient of Determination (R^2) and Mean Square Error (MSE). Model comparison shows that MSE for Bisquare is 270.379 which is the lowest compare to Huber and OLS, 592.199 and 8643.889 respectively. Meanwhile the R^2 value for Bisquare is the highest compared to Huber and OLS which is 0.297, 0.198 and 0.096. Thus, the result shows robust regression method (Bisquare) is the best model with temperature, wind direction, NO_2, CO and O_3 as significant factors that contribute towards PM_{10} concentration in Klang.

ACKNOWLEDGEMENT

Foremost, our greatest and ultimate gratitude to Allah S.W.T the Most Merciful for giving us the opportunity to complete this long and challenging project promptly. Our gratitude and thanks go to our supervisor Miss Zuraidah Binti Derasit for consistent guidance, helping and support throughout developing this project. Without her help and pressure on us, we may not be able to complete the course work within the required time.

Furthermore, our appreciation goes to a few lecturers who has been giving us guidance in completing the analysis. In addition, we would to express our thanks to staff member of Department of Environment (DOE) for providing us a complete set of data to run this project successfully.

Next, we would like to thank to our fellow group members for the stimulating discussions and for all the fun we have had in order to complete this final year project on time. Thanks for all the moral support and countless help while we are finishing this project.

Finally, yet importantly, our deepest gratitude goes to our parents, family members and friends for vision and determination to educate us. Without prayer and support from them, we may not be strong enough to complete this task.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbols

°C	Degree Celsius
%	Percentage
°	Degrees
ppm	Parts per million
mg/m^3	Microgram per cubic meter
x_j	Predictor variables
β_j	Partial regression coefficient
r_i^2	Squared residuals
ρ	Symmetric function
$\hat{\theta}_j$	Regression coefficients
ψ	Derivative of ρ
x_i	Row vector of explanatory variables of the ith case
σ	Standard Deviation
$\hat{\sigma}$	Standardize the residuals
n	Number of data points
Y_i	Observed values
$\hat{Y}_i$	Predicted values

LIST OF ABBREVIATIONS

Abbreviations

PM	Particulate Matter
DOE	Department of Environment
RMAAQG	Malaysia Ambient Air Quality Guideline
OLS	Ordinary Least Square
WHO	World Health Organization
API	Air Pollutant Index
PCA	Principal Component Analysis
CO	Carbon Monoxide
SO_2	Sulphur Dioxide
NO_2	Nitrogen Dioxide
O_3	Ozone
LOESS	Locally Weighted Scatterplot Smoothing
MAAQS	Malaysian Ambient Air Quality Standard
NAAQS	National Ambient Air Quality Standard
USEPA	United States Environmental Protection Agency
OSPM	Operational Street Pollution Model
GDP	Gross Domestic Product
IRLS	Iterative Reweighted Least Squares

EKC	Environmental Kuznets Curve
VIF	Variance Influence Factor
LTS	Least Trimmed Squares
LMS	Least Median Squares
MSE	Mean Squared Error
R^2	Coefficient of Determination
MLR	Multiple Linear Regression

CHAPTER ONE
INTRODUCTION

1.1 Background of Study

The reduction in air quality where in the countries constantly progressing lead to climate changes due to fuel combustion, industrial growth and rapid urbanization (Tiwari et al. 2012). In Malaysia, manufacturing industries have served as the main contribution of economic growth for years now. However, industrial activities like combustion of burning fuel and road traffics are the main cause of the rise in number of air pollution.

Particular Matter (PM) is defined as particle pollution commonly usually cannot be seen with naked eyes. To imprecise, PM is an element mix of air particles and liquid droplets in the air. PM comes in many different sources in a different atmosphere with complex mixtures of both organic and inorganic particles. Cars, trucks, fires and burning activities are some types of gas emission that the particles are released without deviation into the atmosphere. The gas emission leads to the chemical reaction that formed in the air involves a mixture of pioneer chemicals such as sulphate, nitrate and carbon. These particles are varied in sizes and shape that could be seen through microscopic vision.

The air concentrations are divided into two groups of sizes, $PM_{2.5}$ and PM_{10} concentrations which differs in term of aerodynamic diameters. $PM_{2.5}$ concentration contains fine particles microns while PM_{10} includes coarse particles with aerodynamic diameter up to 2.5 microns for $PM_{2.5}$ concentration and 10 microns for PM_{10} concentration respectively. These two concentrations are differed as in the measurement of air because of their association in the chemical composition of particles. The larger particles are known as the coarse particles produced by the break-up process of larger solid particles from unpaved roads, mining operation or traffic dust.

Smaller particles categorized as the fine particles formed from the emission of gases such as sulphate and carbon. According to Department of Environment (DOE), vehicular emissions, power stations and industrial sectors are the three main contributors of PM_{10} in Malaysia. In Malaysia, seventy-six percent emission of PM_{10} concentration emitted from motor vehicles, fifteen percent emitted from power plant

and four percent from the industrial sector.

Due to some specific scientific reason, whether long or short exposure to these particulate matters can cause serious health effects to the extent of severe. The inhalable particulate matters are very small in size and can go deep into lungs that cause adverse cardiovascular effects including heart attack.

To extent, the most significant pollutant in Southeast Asia including Peninsular Malaysia is PM_{10} concentration (Juneng et al., 2011). PM_{10} concentration recorded higher than the safe value detected during the dry season due to vast quantity of smoke released from biomass burning. Malaysia Ambient Air Quality Guideline (RMAAQG) recorded the highest value of PM_{10} concentration frequently during dry season (Azmi et al., 2010). Thus, the rapid transformation of the small industrial area into a wide urban has contributed to most of the environmental issues especially air pollution.

Regression is common tools applied in science field studies. Previous study conducted by Ahmad Zia et al. in 2012 focuses on modelling for predicting PM_{10} concentration in industrial area using Robust Regression in Pulau Pinang. Method used in the study are OLS and Robust Regression. OLS method commonly used because of the simple computation and analysing. However, OLS will become unreliable when the data sets gross in errors due to outliers. While Robust Regression is the opposite of OLS. Robust regression can be generalized with presence of outliers which is the data sets are assume to linear and normally distributed with errors.

In specific, this paper used Robust Regression and OLS method to make comparison for modelling PM_{10} concentration in Klang with the presence of outliers.

1.2 Problem Statement

Asia region encounters a major problem due to air particulate matter pollution. According to the Department of Environment Malaysia, particulate matter specifically PM_{10} is generated from local anthropogenic activities including the use of motor vehicles, heat and power generation plants, industrial emission and open burning activities. Measurements of air pollution concentrations often contain outliers that can significantly affect further analysis (Čampulová, et al., 2018). However, outliers may contain valuable information and sometimes can be the most important observation. Thus, the aim of this study is to suggest an alternative method that insensitive to outliers in the air quality data.

Regression is one of the statistical techniques that most widely used. Out of many potential regression techniques, the ordinary least squares (OLS) method has usually been adopted due to tradition and computational convenience, whereas robust regression method can minimize the influence of outliers directly (Rousseeuw & Leroy, 1987). However, OLS method becomes unreliable if the data sets are influenced by the occurrence of outliers. Robust regression presents answer alike to least square regression, thus can act as a complement to the OLS method when the data for OLS are normally distributed errors and linear (Mahajan et al., 1984).

There are very limited studies in Malaysia that compared using OLS and robust regression method using air quality data. Most of the modelling for air quality use OLS and robust regression only applied in foreign countries. Hence, this study specifically compares the performance of regression models in modelling PM_{10} concentration in Klang with the presence of outliers.

1.3 Research Questions

The research questions for our study on the PM_{10} concentration in Klang are:

a) Is there any correlation between PM_{10} concentration with meteorological parameters (Temperature, Humidity and Wind Direction) and gaseous concentration (Carbon Monoxide, Nitrogen Dioxide, Sulphur Dioxide and Ozone)?
b) What is the difference in the performance of regression models, OLS and Robust regression for PM_{10} concentration in Klang?
c) What is the most significant factor that contribute to PM_{10} concentration?

1.4 Research Objectives

The research objectives for this study on PM_{10} concentration in Klang are:

a) To investigate the correlation between PM_{10} concentration with meteorological parameters (Temperature, Humidity and Wind Direction) and gaseous concentration (Carbon Monoxide, Nitrogen Dioxide, Sulphur Dioxide and Ozone).
b) To compare the performance of several regression models which is OLS and Robust regression for PM_{10} concentration in Klang.
c) To identify the most significant factor that contribute to PM_{10} concentration using the best regression method obtained.

1.5 Research Hypotheses

The research hypotheses for this study on PM_{10} concentration in Klang are:

a) H_0: There is no significant relationship between PM_{10} concentration and meteorological parameters (Temperature, Humidity and Wind Direction).

H_1: There is a significant relationship between PM_{10} concentration and meteorological parameters (Temperature, Humidity and Wind Direction).

b) H_0: There is no significant relationship between PM_{10} concentration and gaseous concentration (Carbon Monoxide, Nitrogen Dioxide, Sulphur Dioxide and Ozone).

H_1: There is a significant relationship between PM_{10} concentration and gaseous concentration (Carbon Monoxide, Nitrogen Dioxide, Sulphur Dioxide and Ozone).

c) H_0: Robust regression model is not different from OLS.

H_1: Robust regression model is different from OLS.

d) H_0: There is no significant relationship between all predictor variables.

H_1: There is a significant relationship between all predictor variables.

1.6 Significance of the Study

The findings of this study will be beneficial to the society since the topic will be discussed may be applied to all residents in Klang. Besides, the goals of such studies are to protect residents' health and to decrease the inevitable economic losses associated with the health problems caused by PM_{10} concentration. Other than that, the study also shows that residents need to know about PM_{10} concentration which occurs nowadays that getting more problematic. It can help them increase awareness about the environment that may be causing harm for them. As this study only focusing in the Klang area, hence the result of this study will benefit residents in those areas. They can use the result to increase the consciousness and goals based on the factors that may contribute to PM_{10} concentration.

Besides that, it is also very beneficial to academicians. Academician could be either a group of an artistic, literary or scientific academy. They are important in updating the study to be relevant and significant during their time for the future. For example, most things in this world, industries and other market are improving themselves to evolve better. As industries grow and change, technologies, processes and other more could once be viable for the outcome as it becomes outdated. Therefore, academician is the one will be updating the study as it benefits both ways such that the study will be updated and the researcher will be referring to the previous study. Next, this study is also important for the future in the usage of the model for predicting the future. For example, if any researcher is studying any subject related to our topic, they could use this model as their reference. Therefore, this study could help anyone to continue the study since this topic is vital for the environment and people's health.

Finally, the responsibility to enhance the Quality Act of 1974 and its subsidiary legislation has been charged to the Department of Environment (DOE) of Malaysia which was established since 1975. By increasing the awareness of the study about PM_{10}, hopefully the DOE will monitor more about it as it could be hazardous and maybe could affect other areas.

1.7 Scope and Limitation of the Study

This study focuses on finding the best regression in modelling PM_{10} concentration in the presence of outliers. Only with the presence of outliers, this study can be generalized to make conclusion. Furthermore, this study will not cover other problems that are not considerably in the same factors involved and area. The area covered in this study is only in Klang which can be generalizable only in the targeted area. Further ado, the result of this study cannot be concluded for whole Malaysia as general. The factors involved is gaseous concentration (Carbon Monoxide, Nitrogen Dioxide, Sulphur Dioxide and Ozone) and meteorological factors (Temperature, Humidity and Wind Direction) are used to measure along with PM_{10} concentration.

A set of data provided by the Department of Environment (DOE) to run the analysis. The set of data were given according to our demand on which variables to be provided. The data given only from January 2017 until December 2018. The latest data cannot be obtained as they have not completed their annual data collection. Data collected from 2017 until 2018 are used due to the similar meteorological environment across these years. Therefore, there are only a total 730 observations at each variable received.

CHAPTER TWO
LITERATURE REVIEW

2.1 Study on Air Quality

The impact of climate change on air quality is difficult to assess because it is not directed by the dominant factor. The impacts of climate change in air quality, the emission of pollutants have effects on aerosols is the dominant factors. Atmospheric aerosol gives impact of air quality on human and ecosystem as the air has polluted (WHO, 2013a). Climate change commonly depends on the response of changes in air pollutant among chemical complex interaction, land surfaces and other meteorological factors i.e temperature and relative humidity. According to study done by S. Fuzzi et al. (2015) objectively summarizing the most recent results within the field of aerosol studies. For instance, in meteorology, temperature will affect the chemical rate in determining pollutant concentration, changes in precipitation can amend aerosol concentration and so do changes in the atmospheric circulation will affect the pollutant distribution (Pausata et al., 2015). Even Jacob and Winner (2015) reviewed the analysis with multiple approaches regarding the observed correlations between both $PM_{2.5}$ and PM_{10} concentrations with meteorological factors.

There is major research in relations to air pollutants. Air pollutants depend on ambient concentration of emission rate of meteorology and geography as well (Seaman, 2003). Meteorology plays an important role in air quality because the concentration of air pollutants be the helm of meteorological domain (Seaman, 2000). Ambient concentrations are either measured or modelled. Measuring concentration gives information about air quality level at a point for current scenario while modelling can provide the information about air quality level for a region for both current and future scenarios. The requirement of meteorological data for air quality modelling can be attained by either onsite monitoring or meteorological modelling. The number of onsite meteorological measurements is fatally limited in most regions of the world. Therefore, meteorological model can help to generate onsite meteorological data to be use in air quality models. Meteorological and air quality models have been applied in many studies with different objectives and addressed various scientific research questions across the world.

Kumar et al. (2017) has done a survey on article in India but did not produce neither new research nor modelling studies. In the article, they include only the introduction of meteorological, air quality models and model applications. The main idea of the article is to support the application of WRF data in dispersion modelling. Besides, he also highlighted there is many researches done in early nineties regarding to meteorological modelling on the air pollutants related to air quality. Mainly because commonly meteorological models calculate a three-dimensional meteorological using mathematical equations to simulate atmospheric processes of the variation in temperature and winds over time. At the end of the result, the development group on air quality has developed methods of meteorological forecast in predicting the atmospheric dispersion.

2.2 Air Quality in Malaysia

In Malaysia, air pollution continuously had been an overwhelming issue. Sepang and Kuala Selangor were recorded as Malaysia's worst episode of agricultural and open burning involving 500 hectares of agricultural waste in 2002, resulting in an unhealthy level of Klang Valley Air Pollutant Index (API) (Mahmud, 2005).

Jamalani et al. (2016) conducted a study to establish the spatial-temporal relationship between PM_{10} and its characteristic of each location. The air quality data of PM_{10} were obtained from the Malaysian Department of Environment (DOE), with a total of 480 observations data of PM_{10} at the selected stations within the year 2000 until 2009. The stations were located at Klang, Petaling Jaya, Kajang and Shah Alam. All the stations within the Klang Valley region were influenced by heavy traffic because of vehicles emission. The result by using principal component analysis (PCA) indicated all four locations impacted by PM_{10} which were clarified as one of the pollutants that worsened the air quality.

According to Abdullah et al. (2011), a research was carried out to determine the statistical characteristics of the PM_{10} concentration in Shah Alam, Selangor for the industrialized area. Shah Alam is chosen as the area of study since it is located in the high traffic and population industrialized area. The air quality data of PM_{10} was used for this research with one-year period hourly average data for 2006 and 2007. The results showed the mean values for the area in 2006 and 2007 are higher than their

respective median which shows that the pollutants distributions are positively skewed. In addition, PM_{10} concentration had minor exceedances from the end of September to the mid-October 2006. According to the Department of Environment, it is due to the haze episodes in Malaysia in 2006. However, no exceedance was observed for 2007.

Kamaruzzaman et al. (2017) carried out a study to examine the air quality trend, the relationship between air pollutants and API and investigate the significant pollutant in Putrajaya. Putrajaya is one of the main economic zones in Malaysia. Secondary air quality data was obtained from the Air Quality Division of the Department of Environment (DOE) from 2011 to 2013. The result for principal component analysis and factor analysis shows only five parameters with strong positive loadings have strong variation which are wind speed, wind direction, sulphur dioxide, nitrogen dioxide and carbon monoxide. Besides, statistical process control shows sulphur dioxide had been identified as the most influenced the quality of the air compared to other pollutants.

2.3 Factors in Modelling PM_{10} Concentration

2.3.1 Temperature

The increased mortality directly associated with the modification of temperature patterns induced by global climate change has been reported by the previous studies. (Ren et al., 2011, Vardoulakis et al., 2014, Carmona et al., 2016, Lee et al., 2018). Additionally, global climate change affects public health indirectly through its influence on air quality due to its direct impacts on temperature patterns and extreme weather events. Current studies have advised that two-way interactions between weather variables (e.g., temperature) and air pollution should be carefully thought-about to characterize synergistic effects on health (Stafoggia et al., 2008, Chen et al., 2018b). Typically, risk estimates from epidemiologic studies model temperature as a confounder whereas few studies have examined the role of temperature as an impact modifier to short-term exposure to pollutants like ozone, whose dynamics powerfully rely on temperature too (Jhun et al., 2014).

2.3.2 Humidity

The previous research paper has done tests on a few meteorological conditions including relative humidity towards PM_{10} Concentrations. Meteorological conditions consisting of wind speed, relative humidity, temperature, and solar intensity were measured in correlation analysis to understand the inhalable particulate exposure risk in humans. This study investigated the potential exposure risk levels of PM_{10} and $PM_{2.5}$ concentrations at two different road configuration sites in Bangkok, Thailand, which is between covered and open roadside areas. Thus, the results show the exposure to particulate matters at the covered areas has higher potential risk for human compared to open areas (Sahanavin et al, 2016). Moreover, the significant negative correlation was found in the relationship between Particulate Matters (PM) and relative humidity at open areas. This result is comparable with previous study done by Akyuz and Cabuk (2009) has similar result in negative correlation. Contrarily, there is correlation showed an insignificant relationship between Particulate Matters (PM) and relative humidity at covered areas this is because the atmospheric particulate can be conveyed by the relative humidity and diminish the motion of particle (Hien et al, 2002).

2.3.3 Wind Direction

Effects of airspeed and wind direction on a human's thermal conditions and of air distribution round the body (Oh & Kato, 2018). The wind direction additionally affects air pollution. If the wind is blowing towards an urban area from an industrial area then pollution levels are probably to be higher within the city or town than if the air is blowing from another direction of for example, open farmland. Sunray may have an effect on pollution levels. Street directions are often optimized to bring the city's dominant wind direction to the preferred wind speed without disrupting large-scale urban ventilation corridors, which might advance the dispersion of air pollutants within the whole area.

2.3.4 Carbon Monoxide Concentration

Short time variation of Carbon Monoxide (CO) concentration depends on traffic dust, wind speed and wind direction (Moseholm et al., 1996). In his paper shows the result of neural network model in predicting the maximum CO concentration at a monitoring station located at Santiago. The station located in observably zone of worst air conditions. There are other methods used in the study as well which is to measure the relationship between Carbon Monoxide and PM_{10} Concentration including linear regression and neural network. At the end of the study, neural network suitable as tool for predicting CO concentration as significantly depends on traffic pollutant.

2.3.5 Sulphur Dioxide Concentration

Kamaruzzaman et al. (2017) conducted a study to examine the significant pollutant in Putrajaya. Secondary air quality data was obtained from the Air Quality Division of the Department of Environment (DOE) from 2011 to 2013. The result for principal component analysis and factor analysis shows five parameters affected the air quality which are wind speed, wind direction, sulphur dioxide, nitrogen dioxide and carbon monoxide. Besides, the result shows sulphur dioxide had been identified as the most influenced the quality of the air compared to other pollutants.

2.3.6 Nitrogen Dioxide Concentration

Kuo et al. (2008) conducted a study to determine factors affecting PM_{10} concentration in central Taiwan. In this analysis, the types of synoptic weather that are extremely likely to produce PM_{10} is referred as HPE and low probability is referred as LPE. Multiple linear regressions analysis showed that NO_2 the most affected to the PM_{10} concentration with 35.61% for HPE weather. For LPE weather, the season factor had the highest contribution to PM_{10} concentration with 48.11%. The result of the correlation coefficient showed the increase in PM_{10} concentration due to the rise of SO_2 and NO_2, which was about 51% of the total increase in PM_{10} from LPE to HPE.

2.3.7 Ozone Concentration

For ozone, the American Cancer Society Cancer Prevention Study (CPS) Jerrett et al. (2009) reported a 4.0% increase (95% CI: 1.0%, 6.7%) within the risk of death from respiratory causes per 10 ppb increase in daily maximum O_3 concentrations (in the April – September period) among individuals age 30 and older in the United State, though no significant association with all-cause mortality was found (RR of 1.001 (0.996, 1.007) in this study. Many cities in Latin America have poor air quality, experimenting with unhealthy concentrations of particles and ozone (O_3). These cities multitude nearly 43 million individuals and pollution levels were higher than the World Health Organization (WHO) guidelines. For PM_{10}, the percentage increase in risk of death due to the respiratory diseases in infants during a fixed-effect model was 0.47% (0.09–0.85). For respiratory deaths in kids 1–5 years old, the increase in risk was 0.58% (0.08–1.08) whereas the higher effect was determined for lower respiratory infections (LRI) in kids 1–14 years old [1.38% (0.91–1.85)]. For O_3, the only summarized estimate statistically significant was for LRI in infants. Analysis by season showed that effects of O_3 within the warm season for respiratory diseases in infants, whereas negative effects were determined for respiratory and LRI deaths in kids (Romieu et al., 2012).

2.4 Method Used to Study Air Quality

According to Mohtar et al. (2018), time series analysis can be used to study the spatial and temporal variation of air pollution at four different locations selected for the studies. The aim of the study is to observe the patterns of air pollutants with mean values of each month calculated by averaging hourly concentration measurements method. The data were first clean using Locally Weighted Scatterplot Smoothing (LOESS). The method used to examine the non-linearity trends of the data (Jang et al., 2017). The calculated mean values are normalised in order to investigate the long-term trend pattern of pollutants. The daily maximum concentration of pollutants with averaging hour of Malaysian Ambient Air Quality Standard (MAAQS) was calculated and plotted in a graph. Thus, the running average between the factors measured for O_3 were 8 hours and PM_{10} were 24 hours while CO, NO_2 and SO_2 were 1 hour. Therefore, other factors (wind speed and wind direction) analysed using statistical software (Carslaw, 2015).

The importance in developing and deploy methods in obtaining on-road micro-scaled measurements of vehicle emissions to estimate the level of pollutants. Based on a case study done in Malaysia by Fadzil (2013), few of the high traffic roads in Selangor were selected for air quality measurement and were analysed using the Operational Street Pollution Model (OSPM). The study shows the result that there were no serious of air pollution recorded in the period of January 2012. Field measurements data along with the traffic information are collected in each week in January 2012. Comparison with simulated data showed good agreement thus indicating the suitability of the model to be used in Malaysia condition. The air quality trend patterns for the pollutants in January 2012 generally shows downward trends or stable trends well below the level of the Malaysian Ambient Air Quality Guideline (RMAAQG) standard value. However, PM_{10} and O_3 concentration are the dangerous pollutants in Selangor. The comprehensive review has revealed that moving vehicles creates a significant impact on air quality on several locations located in Selangor. Meanwhile, the latest data concerning carbon monoxide, nitrogen monoxide, hydrocarbon and particulate emissions are obtained from the Department of Environments (DOE) Malaysia. Following with the study used simulation analysis to conduct analysis using traffic software based on Operational Street Pollution Model (OSPM) lead to a good result indicating the suitability of analysis.

Another study was conducted in Klang Valley, Peninsular Malaysia in four selected air monitoring stations monitored by Department of Environment (DOE) in the period of 10 years, 2000-2009 respectively. The aim of this study is to identify the spatial-temporal relationship of PM_{10}. Spearman Correlation Test was used in this study via XLStat software as a statistical approach to achieve the objective. Spearman Correlation Test is used to determine the relationship between two variables. This test is one of the non-parametric approaches which suitable for not normally distributed data and preferred monotonic graph instead of linear graph. According to Chua (2013), Spearman correlation only shows the strength of correlation between two variables without showing cause and consequence between those two variables. The correlation values are in the range between -1 and +1. Thus, this study showed strong significant relationship between all stations and fair relationship between Petaling Jaya-Kajang and Kajang-Shah Alam.

To be conclude, any statistical analysis can be used to prove the relationship of the various locations on the concentration of air quality since the nature of study are not

normally distributed.

2.5 Ordinary Least Square (OLS) in Study

Regression is one of the most commonly used statistical method. Based from Alma (2011), Ordinary Least Square (OLS) is a type of regression can be known as "global" because of the spatial static of coefficient estimates. It means that it can be applied on different areas of interest equally by using a single model. Among other many possibilities when comes to regression technique, ordinary least squares (OLS) method are generally been adopted due to its tradition and easy to conduct or compute. If errors are normally, independently, and identically distributed, then OLS is more efficient than any other unbiased estimator. If errors are not normal or not independent, other unbiased estimators may perform better than OLS. OLS aims to identify outliers, at the expense of the rest of the sample matching them. The OLS is used to find the best parameter estimation with the least-squares criterion which minimizes the total number of squared distances of all points from the actual observation to the regression surface. (Fox et al, 1997). Nevertheless, OLS estimation of regression weights in the multiple regression is influenced by the occurrence of outliers, non-normality, multicollinearity, and missing data (Ho et al, 2000).

Referred to the Environmental Kuznets Curve (EKC), most studies proved that emissions from air pollution exhibit correlation to both economic development and urbanization (Dietz et al., 2012). In spite of this general trend, most parts of China did not reach the points of inflection of an EKC (Song, 2013). The levels of sulphur dioxide (SO_2) and particulate pollution plummet, whilst nitrogen dioxide (NO_2) levels have increased (Brajer et al., 2011). This shows that the existence of variability or volatility in space and uncertainty is demonstrated by both air quality and the effects of urbanization. In addition, from existing experience, there is a clear trans-regional character to air pollution. All these features break with the basic requirement for classical regression analysis, which holds that the samples analysed must be independent. In these circumstances, if we make an OLS estimate, the results are likely to be biased. (Anselin & Rey, 2015). Therefore, OLS is compared to the other methods in terms of efficiency, breakdown, leverage points and coefficient of determination. The coefficient of determination (R-square) is one of the effective performance statistics. R-square is the statistic that will provide information on the goodness of fit of the model.

To compare OLS with robust regression estimates, the coefficient is used in linear regression models.

Another study was conducted in which the researchers focused on studying the effects of financial resource redistribution obtained by applying the ' green ' fiscal policy to dependent variables. (Bostan, 2016). Researchers incorporate variables into complex models examined through multiple regression, both standard and robust, as well as the panel of fixed effects in 20 European countries that represent the various effects on environmental policy and the expenses it incurred. The main purpose of the analysis to be carried out is to assess the impact of the environment policy for environment expenditure tenable within the European framework on reduction of air pollution. These effects are identified using the statistical analysis include means of regression equations (OLS), robust regression (M method), fixed and random effects, using panel data from 18 EU countries, as well as Switzerland and Turkey due to their position in relation to the community block.

Further to the application of multiple regression statistical methods (OLS and robust M), the results show that overall environmental investment in mining and quarrying and gross domestic product (GDP) expenditure percentage, played a major role in carbon monoxide reduction. These are the total investments made in the mining sector, when the expenditure increased by 1% of the GDP value, there was a decrease of 11 628.3 thousand tons Cot at the level of the European countries analysed, based on the result of the OLS analysis and robust M estimation.

Thus, some statistical methods used in research on environmental inequality may lead to biased findings (Miao Q et al., 2015). Although a variety of methods are used to evaluate inequality, most researchers use a regression-based approach to quantify the magnitude and direction of the inequality. Based on the studies reviewed here, air pollution is the outcome or dependent variable. Ordinary least squares (OLS) regression assumes that outcomes are independent. Due to air pollution frequently shows a pattern of spatial autocorrelation, it is important to determine spatial autocorrelation and if autocorrelation is present, use a spatial analytical technique. This will ensure that the independence of observations assumption is not violated.

2.6 Robust Regression in Air Quality

Robust method is less sensitive than ordinary least squares (OLS) to large changes in small parts of the data. A study by Ul-Saufie et al. (2012) was done in Pulau Pinang to identify the best robust regression models for long term prediction of PM_{10} concentration based on gaseous concentration (SO_2, NO_2, CO) and meteorological parameters. In this study, the researcher used robust regression methods (weighted least square) for predicting PM_{10} concentration in an industrial area in Pulau Pinang because it is among the densest populated states in Malaysia. Regression works by distribution a weight to every data point. The result showed robust regression is most suitable method in predicting PM_{10} concentration for data with outliers. On contrary, this research found that robust regression procedure provides a suitable mechanism for outliers' identification as well as evaluating their relative influence on the final regression coefficient estimations. The reason why robust regression was used is that it has a lot of benefits than OLS. For example, robust regression model will minimize the influence of outlier in the data sets. Thus, analysis of residuals will eventually detect outliers in order to have the best performance.

According to Bostan et al. (2016), the objective of the study was to assess the effect of the policy for environment expenditure tenable within the European framework on reducing nuisance air pollution. These effects are identified using the statistical analysis include means of regression equations (OLS), robust regression (M method), fixed and random effects, using panel data from 20 European countries. In this study, the researcher used robust regression because the extreme values are more sensitive than the OLS. The results indicated that the application of multiple regression statistical methods which are OLS and robust M shows that sum up environmental investments for mining and quarrying. A major role for the reduction of carbon monoxide is by the percentage of gross domestic product (GDP) expenses.

A prime method for analyzing data that are contaminated with outliers is robust regression. It can be used to provide resistant results in the presence of outliers and to detect outliers. In order to explain the behavior of outliers in linear regression and to compare some of robust regression methods, Alma (2011) has conducted a research study via simulation study. The used of simulation study is to determine which method is the best in all of the linear regression scenarios. Comparing the robust regression methods between LTS, M estimate, Yohai MM estimate, and S estimate against OLS

regression estimation using determination of coefficient is the objective of the study. As a result, robust methods are highly regard in terms of efficiency, breakdown, leverage points and coefficient of determination.

Pope et al. (2013) has conducted a research study about the fine particle of air pollution, expectancies of life in the United States and the role of influential observations. In order to ensure that the results are resistant to the presence of outliers and influential observations, standard robust regression procedures are used by the researchers. The examples of standard robust regression procedures used in this study are M-estimation, S-estimation and MM-estimation. The used of robust regression procedures give some of the largest and most statistically significant estimates of the LEPM2.5 associations. The result specified that when human is exposed to fine particulate matter air pollution, it gives harmful effects on human health.

2.7 Method of Estimation

According to Bostan et al. (2016), the main purpose of the study was to establish the effect on air pollution against the policy for environment expenditure tenable. The goal of statistical analysis is to identify these effects by means of regression equations (OLS), robust regression (M-estimation method), fixed and random effects. This study used panel information from 18 EU countries, as well as Turkey and Switzerland due to their position in relation to the group block. The results from OLS and robust M-estimation shows that total environmental investments by Mining and quarrying sector as Percentage of Gross Domestic Product (GDP) expenses played a major role in reducing carbon monoxide. This means that when the total investment increased by 1% of the GDP value, Carbon Monoxide was reduced by 11 628.3 thousand tons at European countries level.

According to Alma (2011), the study was conducted to determine the pattern of outliers in regression models and to compare robust regression methods which are S-estimation, least trimmed squares (LTS), M-estimation and Yohai MM-estimation against OLS regression estimation method. These methods are also measure the percentage difference of outliers in response variable that have 10%, 25% and 40% outliers for each model. The program code for the outlier generation and data simulation was adapted in SAS. The results showed that the OLS and LTS method was the least efficient method. Huber M-estimation is a good estimator when the data has outliers

and no leverage points compared to other methods. For the outliers with low percentage in response variables, S estimator method performs better. However, when the percentage of outliers are increasing, the performance for all methods are decreasing.

Yulita et al. (2018) conducted a study to model the response data which contain the outliers using the methods of M-estimation with Bisquare, Hampel, Huber, and Welsch weight function. Simulation data and HDI (Human Development Index) data in West Java Province were used in this study. The result showed for HDI data, Welsch function is best to use with R^2 89.28% followed by Bisquare, Huber and Hampel. While the simulation data, when contaminations percentage is 5% and 15% of the data, the four functions are considered good enough to model the data. However, when contaminations percentage reach 20%, Hampel, Huber, and Welsch functions considered no longer effective, thus the recommended function to use is Bisquare.

A study done by Rasheed et al. (2014) to estimate state-wide crime model parameters in the United States in 1993. This study used robust M-estimation of Huber and Tukey Bisquare function using iterative reweighted least-squares (IRWLS) and least trimmed squares. Robust regression can be used to predict regression parameters in the presence of outliers and heteroscedasticity. The result shows M-estimation and the least trimmed method estimators are more efficient than OLS. Tukey bisquare has the least standard errors with the largest t-values compared to the t-value obtain from IRWLS estimator using Huber function, RWLTS and OLS

In summary, this study will compare the performance of OLS and robust regression method of PM_{10} concentration with the presence of outliers. The data consists of concentration with meteorological parameters in Klang (Temperature, Humidity and Wind Direction) and gaseous concentration (Carbon Monoxide, Nitrogen Dioxide, Sulphur Dioxide and Ozone).

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Chapter Overview

This chapter provides comprehensive discussions on the methods adopted by this research. This method will mention every component involved in conducting this research from the data collection, data preparation and data cleaning. Finally, this chapter provides a detailed explanation of the data fitting model using different estimation approaches and the method used for selecting the best model.

3.2 Source and Location of Data

The source of data used in this study is a secondary data on air quality. These data were obtained from Air Quality Division, Department of Environment Malaysia (DOE) in Selangor. The data can be divided into two categories of variables; meteorological variables (Temperature, Humidity and Wind Direction) and gaseous concentration variables (Carbon Monoxide, Nitrogen Dioxide, Sulphur Dioxide and Ozone). These two variables are the main factors that affect the Particulate Matter (PM_{10}) concentration. Hence, this study will be focusing on which variables contribute to PM_{10} the most.

Klang Valley is known as the center of Malaysia's industry and commerce area adjoining cities and towns. Klang is located within Klang Valley, Peninsular Malaysia which the air quality monitoring stations is located. Klang air quality monitoring station is located at Sekolah Menengah Kebangsaan (P) Raja Zarina, which is in the most industrialized sites, near to a busy port and power plant and surrounded by a congested main road.

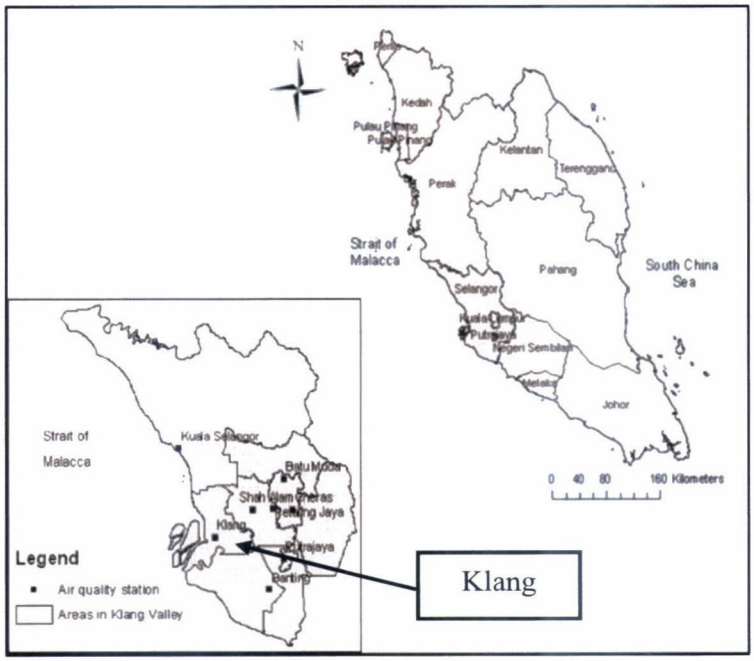


Figure 3.1 Map of the Air Station Klang Valley in Selangor

The Figure 3.1 above shows the map of peninsular Malaysia. The grey shaded part shows the location of air quality monitoring stations located in Klang Valley. This study will only be focused on the Klang area.

3.3 Data Description

The information of the data provided by the Department of Environment, Selangor. Table 3.1 shows the variables, description of each of the variables and types of the measurement scale of the data in this study.

Table 3.1 The Description of the Variables

Variables	Description	Measure
Temperature	Temperature at each location (° C)	Numeric
Humidity	Humidity of each location (%)	Numeric
Wind Direction	The direction of the wind (°)	Numeric
Carbon Monoxide	Concentration of Carbon Monoxide (CO) (ppm)	Numeric
Nitrogen Dioxide	Concentration of Nitrogen Dioxide (NO_2) (ppm)	Numeric
Sulphur Dioxide	Concentration of Sulphur Dioxide (SO_2) (ppm)	Numeric
Ozone	Concentration of Ozone (O_3) (ppm)	Numeric
PM_{10} Concentration	Concentration of PM_{10} (mg/m^3)	Numeric

The data contain a series of daily observation at the meteorological station at Klang. There are 730 observations used in this study covered from January 2017 until December 2018.

3.4 Flow Chart Process of Research

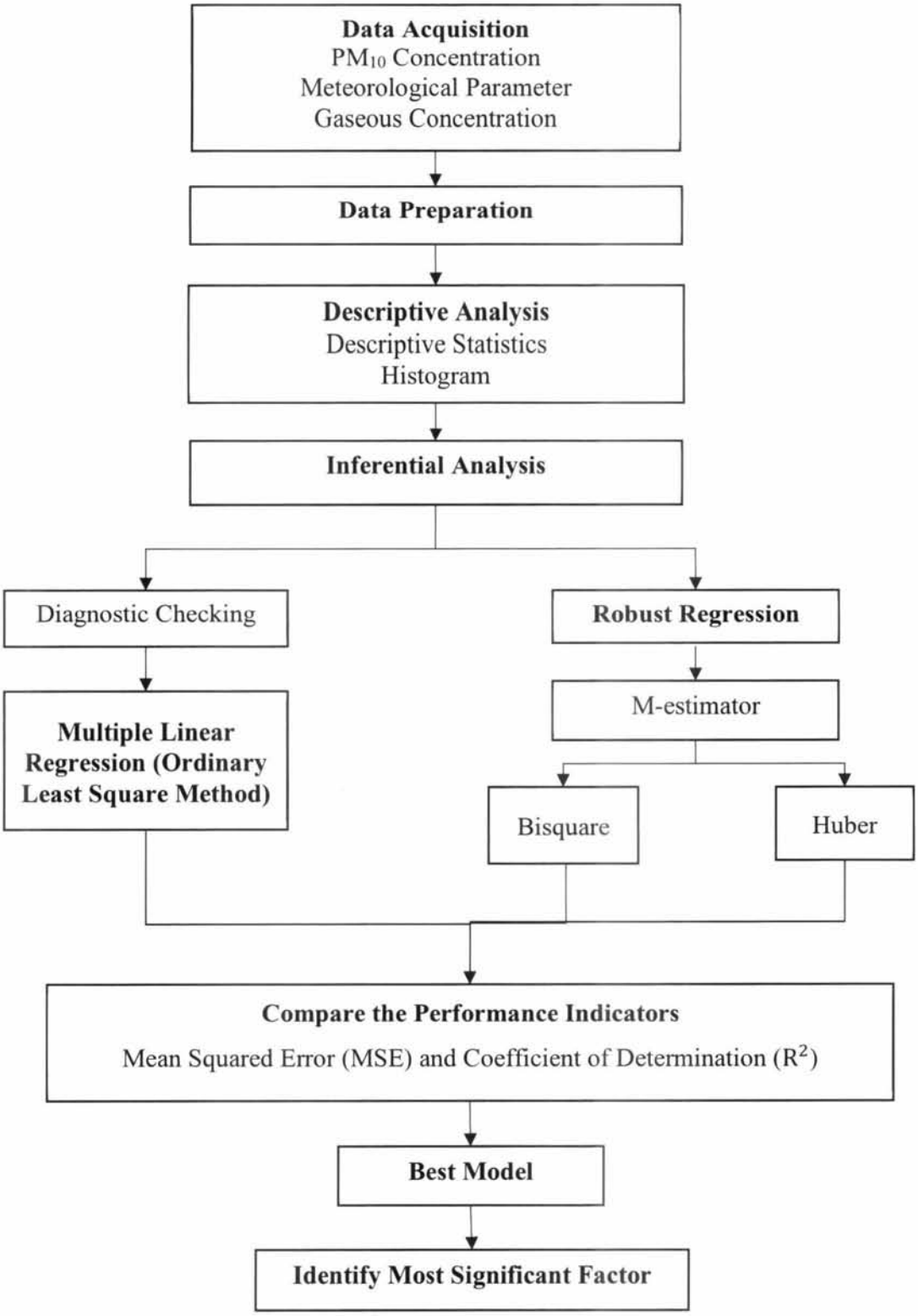


Figure 3.2 Flow Chart Process of Research

3.5 Conceptual Framework

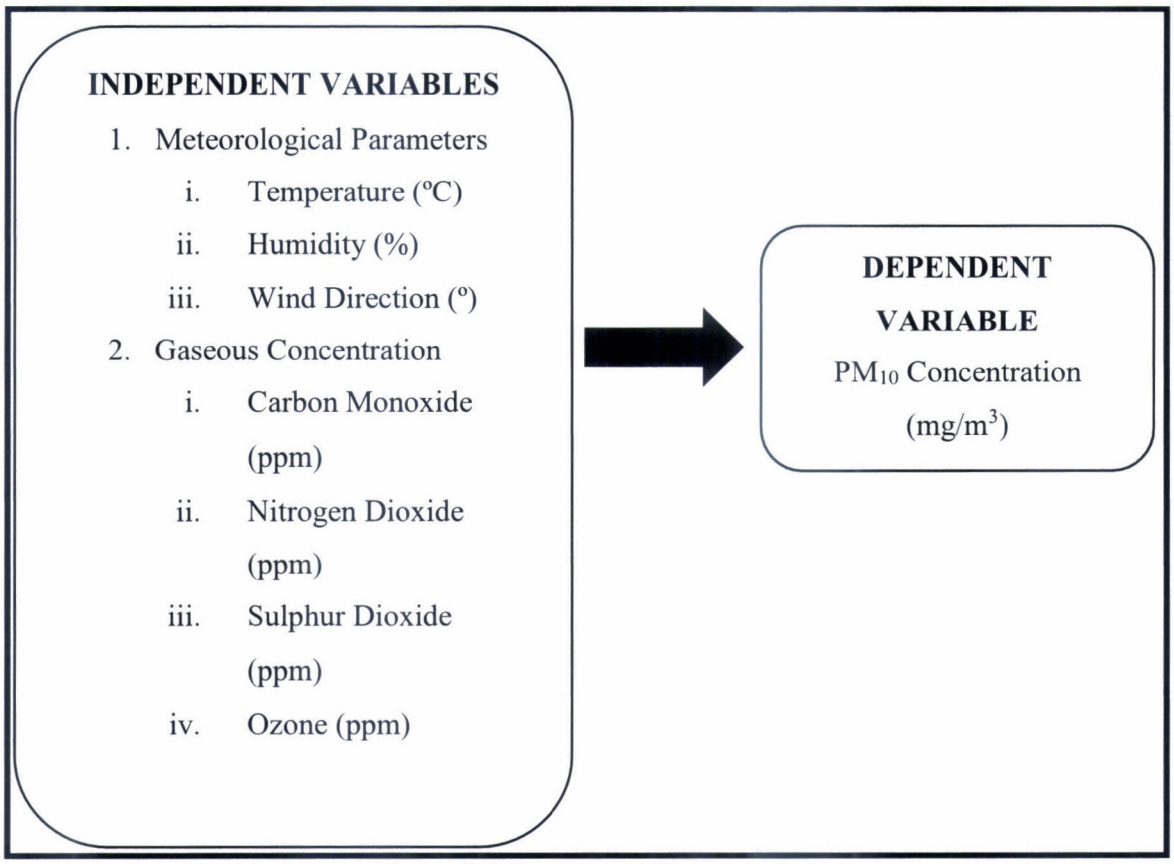


Figure 3.3 Conceptual Framework

The figure above shows the conceptual framework for this study. Meteorological factors contribute to the amount of particulate matter in the region. PM_{10} particulate matter has been classified as the most significant pollutant in Southeast Asia including Peninsular Malaysia (Juneng et al., 2011). Malaysia Ambient Air Quality Guideline (RMAAQG) recorded PM_{10} concentration was frequently exceeded the safe value of recommended, especially during the dry season (Azmi et al., 2010). High PM_{10} concentration was detected during the dry season or also known as summer monsoon due to the vast quantities of smoke released by biomass burning from regional sources. Thus, the rapid transformation of the small industrial area into a wide urban has contributed most of the environmental issues especially air pollution.

3.6 Data Preparation

In this section, the data of temperature, humidity, wind direction, carbon monoxide, nitrogen dioxide, sulphur dioxide and ozone were performed to see whether or not there was a missing value of all variables selected. If the data contain many missing values and outliers, the data is not suitable and appropriate to use for multiple linear regression hence different sources required. After that, data cleaning process was involved which fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data.

3.6.1 Missing Data Imputation

To impute, the missing value on incomplete data, the SPSS was used in this study. The variables that have missing value problem has been identified and ready to impute the missing value by replacing with other values. SPSS result will show the number of missing values problems in variables of this study. The SPSS was used a default technique of imputation that replace the missing value with the mean value for continuous data and replace the missing value with mode value for categorical data. The data consist of 4.9% of missing values.

3.6.2 Multicollinearity

Multicollinearity is can be defined as the situation where the predictor variables are highly correlated with one another (Paul, 2010). When the independent variables are found to be to correlate with the other independent variable, therefore multicollinearity is said to be exist in the dataset. To proceed with the analysis, we have to confirm the independent variable is not related to each other. To ensure that, the data will be checked by looking at the Variance Influence Factor (VIF) values or Tolerance value. If VIF value is more than 10 and Tolerance value less than 0.1, it indicates that there exists multicollinearity problem. If multicollinearity exists, the most correlated variable will be drop in order to produce an accurate result.

3.6.3 Outlier Detection

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. For the purpose of this study, the data will test for the outlier. If the data contain the outlier, the observation will not be removed instead this study will continue with the analysis by including the outliers. Boxplot is one of the methods used to check for the existence of outliers. There are no outliers exist in the data if no point lies outside the box. Outlying points were determined using the values of studentized deleted residual to check outlier with respect to response variable, Leverage to check outlier with respect to all predictor variables and Cook's D to determine the influential points. Cook's D plot examine the outlier that influence all fitted values and its cut off point is $D_i > |\frac{4}{n}|$.

3.6.4 Normality of Distribution

Next, for checking the constant of error variance and the normality of the distribution, the scatter plot of residuals versus predicted variable has been plotted. If the scatter plot shows no obvious pattern, hence it indicates that the residual has constant variance and the distribution is normal. Besides that, normality also can be examined by using the histogram of residual and Normal Q-Q plot. If the histogram is symmetric, it indicates that the distribution of residual is normal. For checking normality using normal Q-Q plot, if the majority of the point lies approximately along the straight line, then the error term is reasonably normally distributed. In order to confirm for the normality assumption, therefore Kolmogorov-Smirnov Test or Shapiro-Wilk must be conducted.

3.6.5 Independence of the Data

The residual is independent and has no potential problem with dependency if the distribution of the error term has no pattern over time. Testing independence of error terms can also be checked by using Durbin-Watson Statistic. If Durbin-Watson value is near to 2, therefore the residuals are not correlated. This indicates that the error term is independent.

3.7 Method of Analysis

As mentioned in an earlier chapter, in order to satisfy the objectives of the study that is to find the best regression model, two types of regression will be used in this study which are Multiple Linear Regression and Robust Regression. Thus, the mathematical formulation of the method will be further discussed. Besides, performance indicator will be clearly explained in this section.

3.7.1 Multiple Linear Regression Model

Statistical technique uses to analyse the relationship between two or more quantitative variables is regression analysis. So that a dependent variable (response or outcome variable) can be predicted from the independent variable (explanatory variable). Regression analysis is helpful in forming the equations. Regression models provide the scientist with a powerful tool, allowing predictions about past or present event.

According to Min & Min (2019), the most widely used tool in econometrics are the multiple linear regression model and its estimation using ordinary least squares (OLS). It allows to estimate the relationship between a set of explanatory variables and a dependent variable.

Multiple linear regression models are a regression model that involves more than one regressors variable. Multiple linear regression models can be written as follows:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon \quad (3.1)$$

where :

y = The value of the response variable

X = The value of the independent variable on (X_1, X_2, X_3, X_4, X_5, X_6, X_7)

X_1 = Wind Direction

X_2 = Temperature

X_3 = Humidity

X_4 = SO_2 Concentration

X_5 = NO_2 Concentration

X_6 = O_3 Concentration

X_7 = CO Concentration

β_j = Measures the expected change in the response y per unit change in x_j when all regressors variables are held constant. β_j is also called partial regression coefficient

j = 0, 1, 2, …, k is called regression coefficients

ε = Random error with mean

The assumptions of the model are:

1. The dependent variable should be measurable.
2. The relationship between the independent variables and dependent variable is linear relationship.
3. The errors between observed and predicted values (i.e. the residuals of the regression) should be normally distributed.
4. There is no multicollinearity in the data.
5. The data must homoscedasticity.
6. There is no outlier exist in the data.

3.7.2 Robust Regression

Robust regression analysis offers an alternative to a least squares regression when assumptions are not fulfilled. This method is very useful in dealing with the outliers, as all the outliers are detected in a single blow by simply running a robust estimator. As there are diagnostics, there are almost as many robust estimations. M-estimation, Least Trimmed Squares (LTS), Least Median Squares (LMS) and S-estimation are among the robust estimations used in estimating the parameters of the regression line. In this study, the M-estimation is used and will be briefly described in the next sections.

3.7.2.1 M-Estimator

M-estimation is the most common general method of robust regression that is nearly as efficient as OLS (Huber, 1992). M-estimator models contain all models that are derived to be maximum likelihood models. The aim is to minimize the function ρ

of the errors with M-estimate instead of minimize the sum of squared errors. This method is based on the idea of replacing the squared residuals r_i^2 by another function of the residuals.

$$\underset{\hat{\theta}}{\text{minimize}} \sum_{i=1}^{n} \rho(r_i) \tag{3.2}$$

where ρ is a symmetric function [i.e., $\rho(-t) = \rho(t)$ for all t] with a unique minimum at zero. Differentiating this expression with respect to the regression coefficients $\hat{\theta}_j$.

$$\sum_{i=1}^{n} \psi(r_i) x_i = 0 \tag{3.3}$$

where ψ is the derivative of ρ, and x_i is the row vector of explanatory variables of the ith case:

$$\begin{aligned} x_i &= (x_{i1}, \ldots, x_{ip}) \\ 0 &= (0,\ldots,0) \end{aligned} \tag{3.4}$$

Therefore, 3.2 is really a system of ρ equations, the solution of which is not always easy to find. In practice, one uses iteration schemes based on reweighted LS (Holland and Welsch 1977) or the so-called H-algorithm (Huber and Dutter 1974, Dutter 1977, Marazzi 1980). However, the solution of equation 3.2 is not transformed properly with respect to a magnification of the y-axis. Therefore, one has to standardize the residuals by means of some estimate of σ.

$$\sum_{i=1}^{n} \psi(r_i/\hat{\sigma}) x_i = 0 \tag{3.5}$$

where $\hat{\sigma}$ must be estimated simultaneously.

The choice of the ψ function is related to the choice of how much weight outliers are to be defined. A monotone ψ function does not consider weight on outliers as much as least squares (e.g. 10σ outlier would receive the same weight as a 3σ outlier). A descending ψ function raises the weight specify to an outlier until a specified distance and then reduces the weight to 0 as the outlying distance becomes significant.

i) Huber M-Estimator

This function was proposed by Peter Huber in 1964 is given by

$$H(\varepsilon) = \begin{cases} 1 & \text{for } |\varepsilon| \leq k \\ k/|\varepsilon| & \text{for } |\varepsilon| > k \end{cases} \tag{3.6}$$

The value k=1.345 indicates it is 95% as efficient as least squares if the true distribution is normal.

ii) Bisquare M-Estimator

This function was proposed by Tukey and known as Tukey's bisquare or Bisquare, is given by

$$B(\varepsilon) = \begin{cases} \left[1 - \left(\frac{\varepsilon}{k}\right)^2\right]^2 & \text{for } |\varepsilon| \leq k \\ 0 & \text{for } |\varepsilon| > k \end{cases} \tag{3.7}$$

The value k=4.685 indicates it is 95% as efficient as least squares if the true distribution is normal.

The values of k for the Bisquare and Huber estimators are called a tuning constant, which smaller k values provide more resistance to outliers, but when the errors are normally distributed at the expense of lower efficiency. In the normal case, the tuning constant is usually chosen to provide reasonable high efficiency.

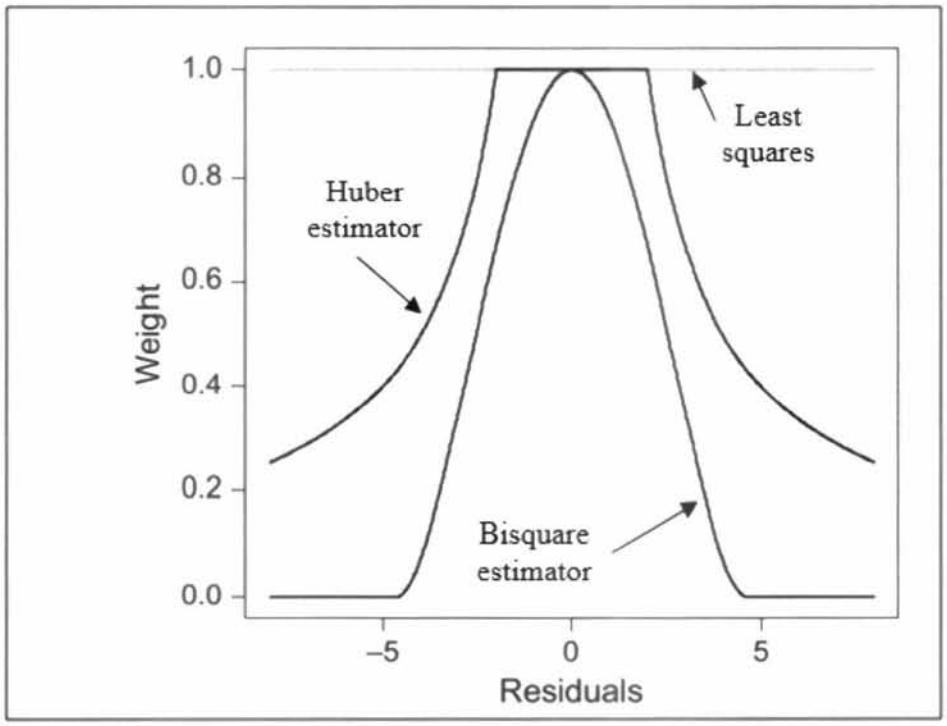


Figure 3.4 The weight function for Huber estimator and Bisquare estimator compared with least squares estimation

These robust estimators often reduce the weight of outliers compared with the traditional least-squares approach. Figure 3.4 shows the function for Bisquare estimator and Huber estimator. Huber function works like least squares until the residuals reach certain magnitude, and then their influence drops off. The residual effect for Bisquare estimator drop off immediately, and they have no influence after a point (Fox, 2002). Least-squares assign equal weights to each case, while the weights for Huber estimator decay when $| \varepsilon | > k$ and the weights for Huber estimator decay once ε departs from 0, and are 0 for $| \varepsilon | > k$.

According to Yu and Yao (2017), Bisquare function can entirely eliminate the effect of large outliers, while Huber function can only lessen the effect of large outliers.

3.7.2.2 Iteratively Reweighted Least Square

An estimator may be defined through a set of altered normal equations that cannot be solved explicitly in closed form, as well as the weighted least squares

problem:

$$\sum_{i} w_i r_i x_i = 0 \tag{3.8}$$

M-estimator is one of this type. Therefore, an iterative solution called iteratively reweighted least-squares is required to obtain this estimator in which $w_i = \psi(r_i)/(r_i)$. To solve the weighted least squares problem, begins with an initial estimator, the IRLS procedure recursively updates the estimate until convergence is reached (The & Some, 1992).

3.7.3 Performance Indicators

Performance indicators were used to evaluate the goodness of fit to determine which regression model is the best to represents the PM_{10} concentration in Klang. Two types of performance indicators, Mean Squared Error (MSE) and Coefficient of Determination (R^2) for each method were considered to identify the best regression method. In this study, performance indicators were calculated for each model using R programming.

3.7.3.1 Mean Squared Error (MSE)

Mean Squared Error measures the average of the squares of the errors, which is the average squared difference between the actual value and the estimated value. An occurrence of large error would significantly affect the value of MSE. The closer the MSE value to zero indicates a better method. The formula is as given below:

$$\text{MSE} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n} \tag{3.9}$$

where n is the number of data points, Y_i represents the observed values and $\hat{Y}_i$ represents the predicted values.

3.7.3.2 Coefficient of Determination (R^2)

The value of the coefficient of determination, R^2, measures of the total variation in the response variable that is explained by changes in the independent variable. It will give information about the goodness of fit of the model. The value of R^2 is between 0 (extremely a poor fit) and 1 (perfect fit), that is usually expressed in percentage from i.e. $0 \leq R^2 \leq 100$ percent. The formula is as given below:

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}} \tag{3.10}$$

3.8 Software

The analysis of modelling for PM_{10} concentration were analysed using several software. The analysis took places using SPSS and R Programming for different analysis purposes. In this section we discussed on the uses of these software.

3.8.1 SPSS

SPSS is short form for Statistical Package for the Social Sciences and it can be used in various kinds of research for complex statistical data analysis. SPSS is a software has easy capabilities to produce outputs. In this study, SPSS were used for data imputation, check on multicollinearity, data description and Pearson correlation.

3.8.2 R Programming

R is a programming language and environmentally used in statistical programming, data analytics and scientific research. There are many packages were used to produce outputs using command from the package. Packages used are 'olsrr' for outliers checking, 'robustreg' for robust regression and 'dvmisc' to get MSE for OLS method. However, other analysis included does not involve in packages, instead we used common command to get the outputs.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 Introduction

This chapter discusses the results in light of the research questions under study. Initially, this chapter will describe the relationship between predictor variables and response variable. The overview of data imputation will be described followed by descriptive analysis and further justification regression models used which are OLS, Huber and Bisquare by comparing the different types of models. From the best regression model, the most significant factor that effect PM_{10} concentration will be identified.

4.2 Data Preparation

In this section, the data preparation of temperature, humidity, wind direction, Carbon Monoxide, Nitrogen Dioxide, Sulphur Dioxide and Ozone were performed to see whether or not there was a missing value of all variables selected.

4.2.1 Data Imputation

The data that have missing value problems have been identified and ready to impute missing value by replacing with the mean value because the data is continuous. The data consist of 4.9% of missing values.

4.2.1.1 Replace Missing Value

Table 4.1 Replacing Missing Value

Variables	No of Replaced Missing Values (%)	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
		First	Last		
SO_2 Concentration	29 (4.5%)	1	649	649	SMEAN(SO2_Conc)
NO_2 Concentration	110 (16.9%)	1	649	649	SMEAN(NO2_Conc)
O_3 Concentration	114 (17.6%)	1	649	649	SMEAN(O3_Conc)
CO Concentration	4 (0.6%)	1	649	649	SMEAN(CO_Conc)

Table 4.1 above shows the result after imputation. From the table, all of the missing values were replaced with the mean value using mean function at each respective variable. Thus, every variable now has a total of 649 valid observations each. Originally, the data has in total of 730 observations at each variable received. However, due to unavoidable circumstance, there are particularly missing data continuously for 3 months at every variable. Thus, the missing data omitted from the study in order to avoid inaccuracy in data analyzation. Overall, the total data used in this study is 649 observations.

4.2.1.2 Detect Missing Value

Table 4.2 Number of Missing Values

Variables	Valid	Missing (%)
Wind Direction	649	0
Temperature	649	0
Humidity	649	0
SO_2 Concentration	620	29 (4.5%)
NO_2 Concentration	539	110 (16.9%)
O_3 Concentration	535	114 (17.6%)
CO Concentration	645	4 (0.6%)
PM_{10} Concentration	649	0

Each of the variables has a total of 649 observations. The table above shows each of the variables has missing value. SO_2 Concentration reportedly 29 missing values whiles NO_2 Concentration has 110 missing values. CO Concentration and O_3 Concentration has missing values of 4 and 114 respectively. Thus, since all the missing values are numerical, it can be replaced with the mean average of each variable through imputation.

4.3 Descriptive Statistics

Table 4.3 Descriptive Analysis

	N	Range	Minimum	Maximum	Mean	Standard Deviation	Variance
Wind Direction (°)	649	208	152	360	311.19	58.034	3368.000
Temperature (°C)	649	10.4100	24.600	35.0100	31.933632	1.3845353	1.917
Humidity (%)	649	39.000	60.000	99.000	89.592664	5.9483410	35.383
PM_{10} Concentration (mg/m^3)	649	1453.9810	0.000	1453.981	82.666643	97.838975	9572.465
SO_2 Concentration (ppm)	649	9.8990	0.0010	9.900	2.990978	2.6043210	6.782
NO_2 Concentration (ppm)	649	7.8799	0.000	7.8799	2.843722	1.5944915	2.542
O_3 Concentration (ppm)	649	9.2399	0.000	9.2399	3.377270	2.0045189	4.018
CO Concentration (ppm)	649	5.4100	0.000	5.4100	1.705823	0. 7696754	0. 592

Table above shows the descriptive analysis for each variable. All variables have 649 total observations each. Wind direction has 152° at the lowest and 360° at the highest with an average mean of 311.19 °. Temperature has 24.6°C at the lowest and 35.016°C at the highest with average mean approximately 32°C. Humidity has 60% at the lowest and 99% at the highest with an average mean of 89.59 %. PM_{10} Concentration has 0.0 mg/m^3 at the lowest and 991453.98 mg/m^3 at the highest with an average mean of 82.666643 mg/m^3. SO_2 Concentration has 0.01 ppm at the lowest and 9.9000 ppm at the highest with average mean of 2.990978 ppm. NO_2 Concentration has 0.0 ppm at the lowest and 7.8799 ppm at the highest with average mean of 2.84 ppm. O_3 Concentration has 0.0 ppm at the lowest and 9.24 ppm at the highest with average mean of 3.37 ppm. CO Concentration has 0.0 ppm at the lowest and 5.41 ppm at the highest with an average mean of 1.71 ppm.

Among the variables, the temperature has the lowest variance with 1.917 while wind direction has the highest variance for meteorological factors with 3368. Among gaseous concentration, CO concentration has the lowest variance with 0.592 and SO_2 Concentration has the highest variance with 6.78.

4.3.1 Preliminary Data Analysis

4.3.1.1 Scatter Plot Matrix between Independent Variables

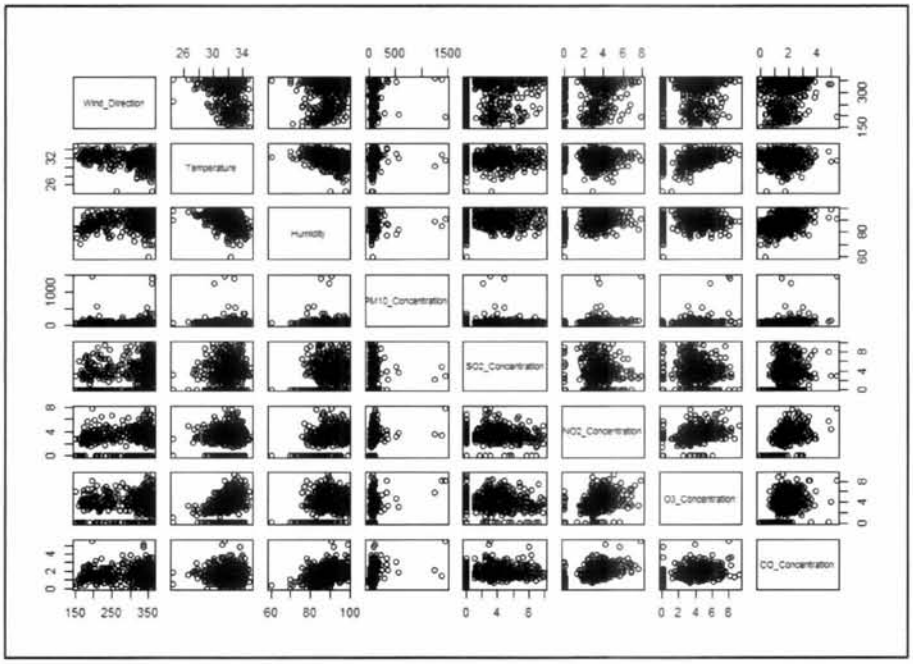
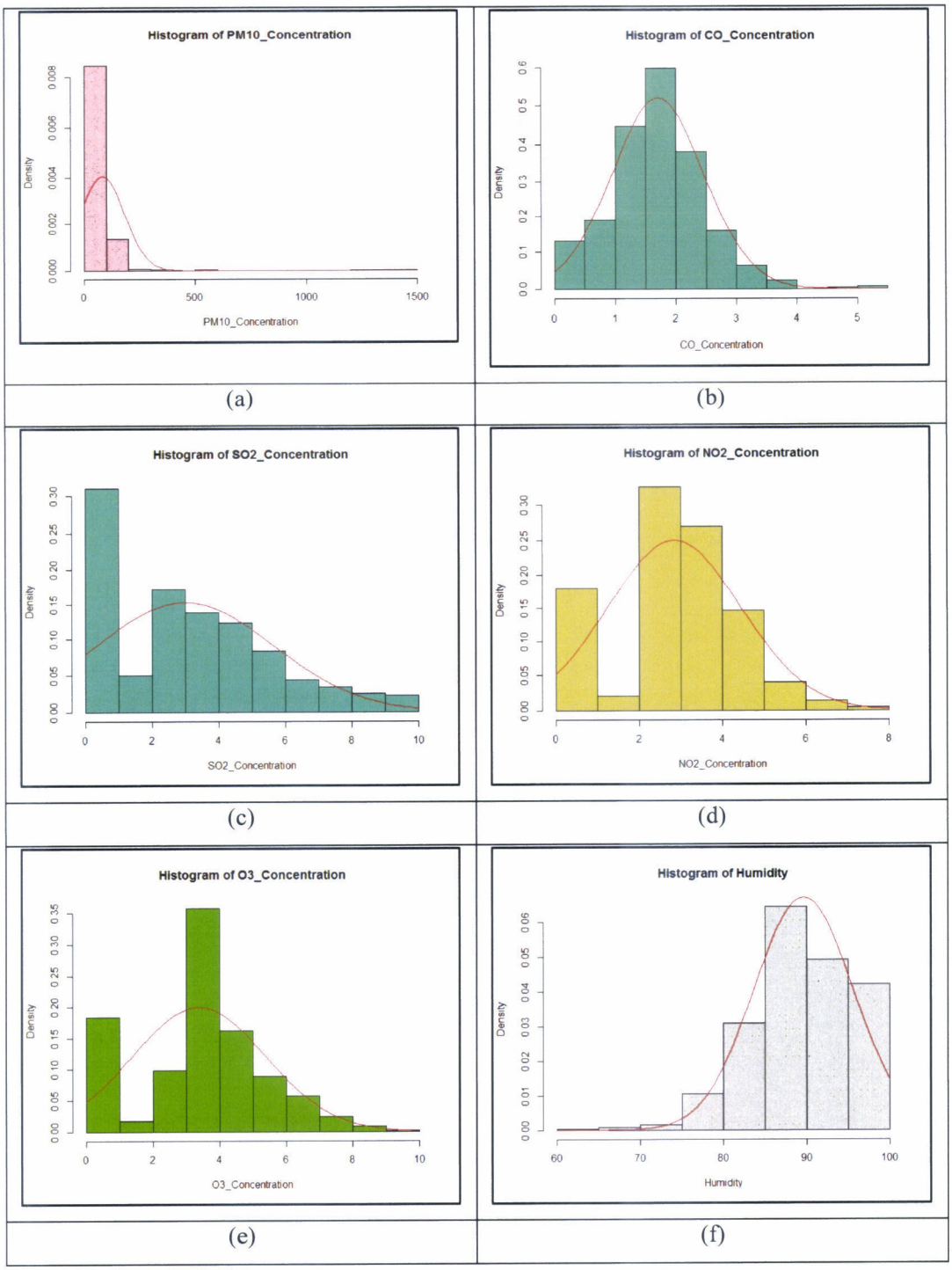
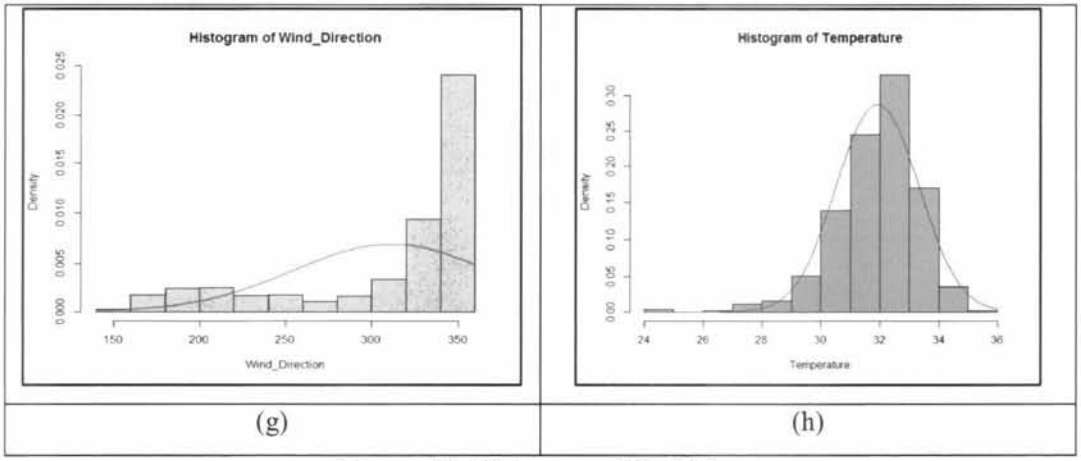


Figure 4.1 Scatter Plot Matrix

The scatter plot matrix represents the relationship between predictor variables. The predictor variables are divided into two groups, gaseous concentration and meteorological parameters. Gaseous concentration includes Carbon Monoxide (CO), Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2) and Ozone concentration (O_3) while meteorological parameters include temperature, humidity and wind direction. Hence in this scatter plot matrix, all pair of predictor variables shows the variables are not correlated to each other since the plot shows every pair of variables are scatter and does not have fixed pattern. Thus, this shows that the predictor variables are independent towards each other.

4.3.1.2 Histogram of Variables





(g) (h)

Figure 4.2 Histogram of Variables

Figure above shows the histogram of all variables. Figure (a) shows the histogram of PM_{10} concentration, it shows that the value of density is between 0 mg/m^3 and 500 mg/m^3. Therefore, the distribution of PM_{10} concentration is positively skewed or skewed to the right. Figure (b) above shows the histogram of CO Concentration. There is unusual small density value in between 4 ppm and 5 ppm. The histogram is slightly symmetric but due to small density, the histogram for humidity is skewed to the left. Figure (c) showed the histogram of SO_2 concentration. The histogram is skewed to the right, which means positively skewed. There is unusual small density value on NO_2 concentration showed in Figure (d). The value is between 1 ppm and 2 ppm, then large density value from 2 ppm and 3 ppm. Therefore, the distribution is positively skewed or skewed to the right.

Figure above (e) showed the histogram of O_3 concentration. There is unusual small density value on O_3 concentration in between 1 ppm and 2 ppm, then large density value from 3 ppm and 4 ppm, hence the distribution is positively skewed or skewed to the right. Refer to Figure (f), the histogram of air humidity. There is unusual small density value in between 60% and 75%. The histogram is slightly symmetric but due to small density value for first 60 ppm, the histogram for humidity is skewed to the left.

Figure (g) showed the histogram of wind direction. There is large density value increment in between 300° and 350°. The data is dispersed to the left. Meanwhile the histogram of temperature in Figure (h) is slightly symmetric due to the rise between 28°C and 32°C the drop in between 32°C and 36°C. Thus, due to small value starting first 24°C, the histogram is skewed to the left.

4.3.1.3 Outliers

Boxplot is used to check for the existence of outliers.

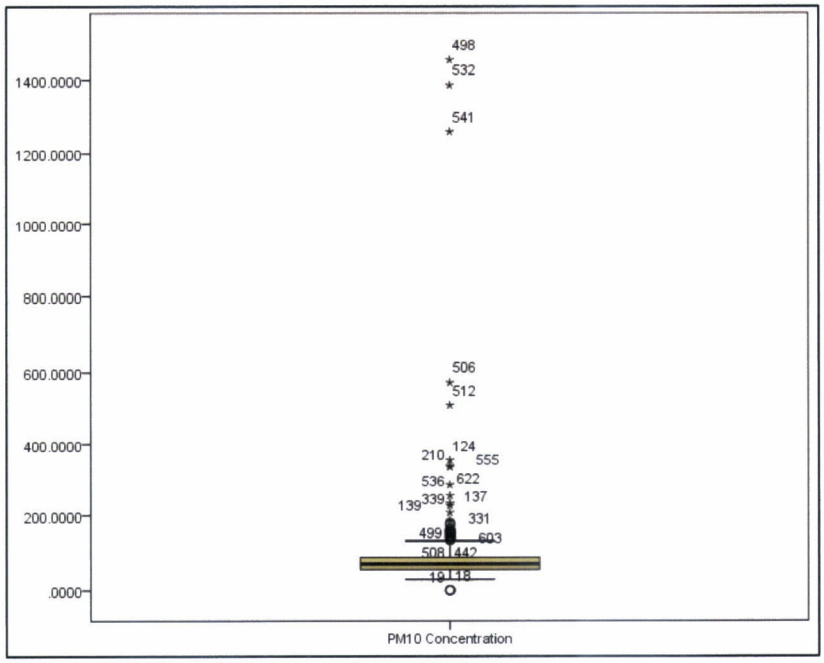


Figure 4.3 Boxplot of PM_{10} Concentration

Figure 4.3 shows that there are many outliers exist in the data since many points lie outside the box.

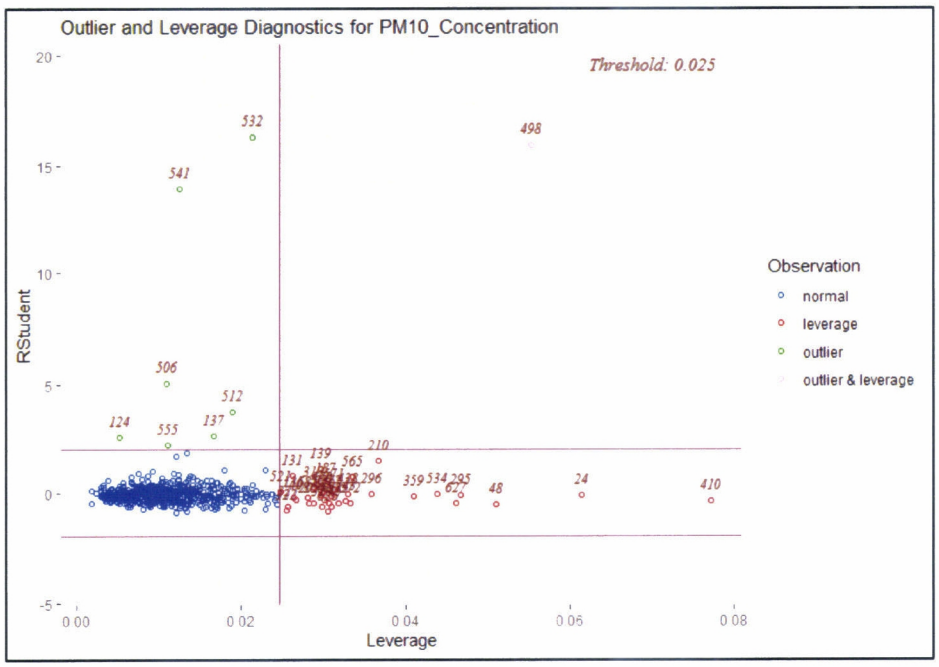


Figure 4.4 Outlier and Leverage Diagnostics for PM_{10} Concentration

Figure 4.4 shows that there are many outliers exist in the data consist of outliers in predictor variables. When the observation is greater than cut-off point value which is 0.025, the observation is considered as the outlier. Thus, the green points show there are outliers exist for observation 124th, 137th, 506th, 512th, 555th, 541st and 532nd.

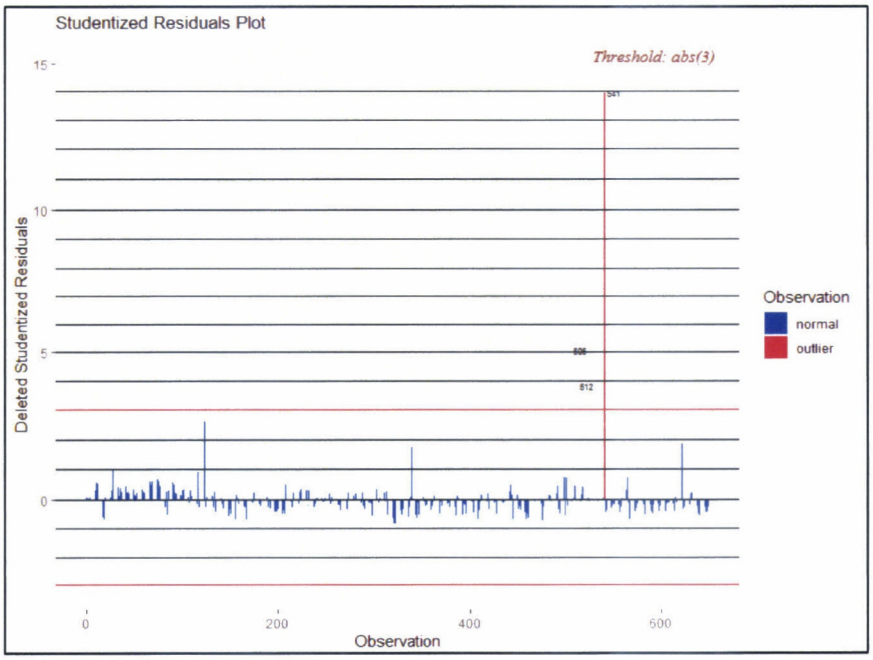


Figure 4.5 Plot of Studentized Residuals

Figure 4.5 shows there are exist few outliers in the data consist of outliers in the response variable. The observation is called an outlier if the externally studentized residual value is greater than the cut-off point which is 3. The points lie outside the red line is the outliers which is observation 506th, 512th and 541st.

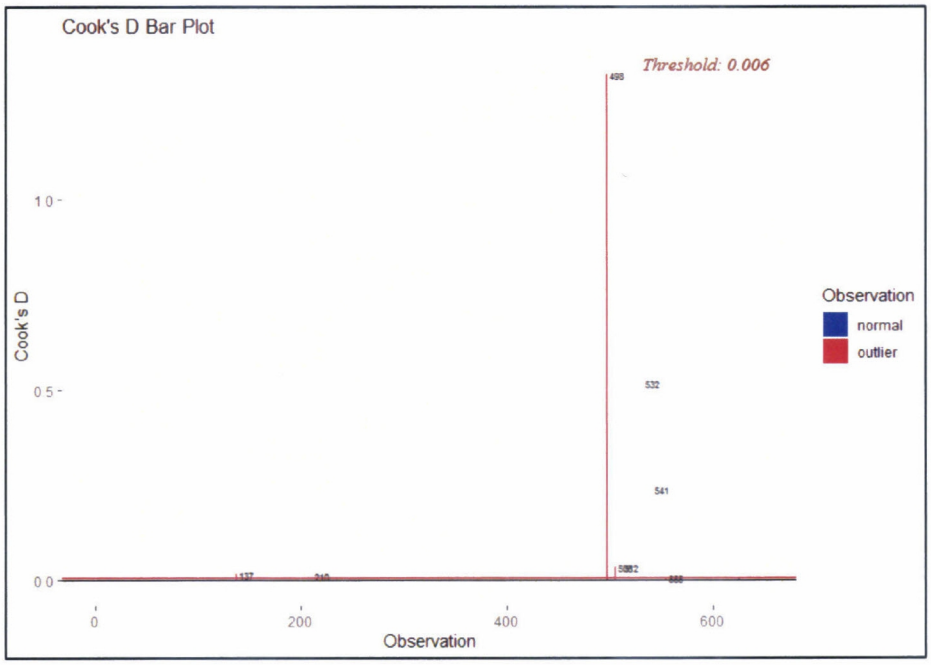


Figure 4.6 Bar Plot of Cook's D of PM_{10} Concentration

Figure 4.6 shows the Cook's D plot in detecting observation that strongly influence fitted values in the model. The points that lie outside the red line with the value greater than the cut-off point value, 0.006 is considered as the outliers. There are 8 observations considered as outliers in the model which are 137th, 210th, 498th, 506th, 512th, 532th, 541st and 555th.

According to the three measures which are Outliers versus Leverage plot, Studentized residual plot and Cook's D plot, there are three points of observation frequently appear as outliers in the response variable, predictor variable and models. The three points are 506th, 512th and 541st. Thus, these three points are considered as the influential points in the model.

4.4 Inferential Analysis

4.4.1 The relationship between predictor variables and PM_{10} Concentration

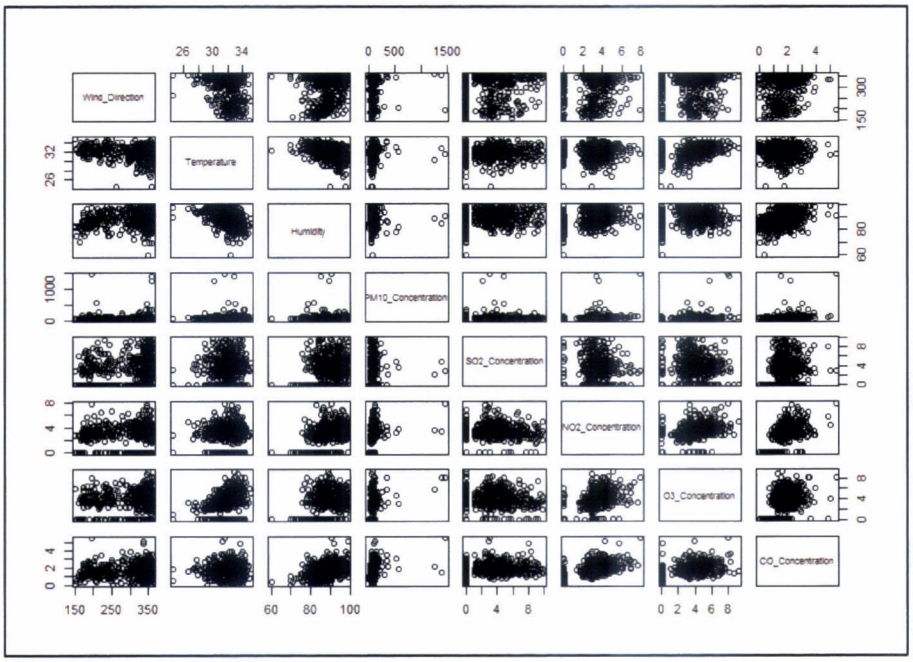


Figure 4.7 Scatter Plot Matrix

Figure 4.7 of scatter plot matrix indicates the bivariate relationship among the variables. The plot shows a negative weak relationship between PM_{10} concentration and wind direction (-0.016). Next, there is a positive weak relationship between PM_{10} concentration and temperature (0.054). The relationship between PM_{10} concentration and humidity (-0.068) shows it has negative weak relationship. Furthermore, PM_{10} concentration and SO_2 concentration (0.051) shows a positive weak relationship. There is a positive weak relationship between PM_{10} concentration and NO_2 concentration (0.18). Meanwhile, there is a positive weak relationship between PM_{10} concentration and O_3 (0.197) and PM_{10} concentration and CO (0.24).

4.4.2 Ordinary Least Square

Table 4.4 Test Hypotheses of Model Coefficient for OLS

Variables	Hypotheses	Estimate	P-Value	Decision	Conclusion
Wind Direction	$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$	0.0017	0.979	P-value = 0.979 > 0.05, fail to reject H_0.	Wind direction is not significant in explaining PM_{10} concentration.
Temperature	$H_0: \beta_2 = 0$ $H_1: \beta_2 \neq 0$	-4.9372	0.122	P-value = 0.122 > 0.05, fail to reject H_0.	Temperature is not significant in explaining PM_{10} concentration.
Humidity	$H_0: \beta_3 = 0$ $H_1: \beta_3 \neq 0$	-3.4448	0.000	P-value = 0.000 < 0.05, reject H_0.	Humidity is significant in explaining PM_{10} concentration.
SO_2 concentration	$H_0: \beta_4 = 0$ $H_1: \beta_4 \neq 0$	-1.3431	0.377	P-value = 0.377 > 0.05, fail to reject H_0.	SO_2 concentration is not significant in explaining PM_{10} concentration.
NO_2 concentration	$H_0: \beta_5 = 0$ $H_1: \beta_5 \neq 0$	1.3897	0.706	P-value = 0.706 > 0.05, fail to reject H_0.	NO_2 concentration is not significant in explaining PM_{10} concentration.
O_3 concentration	$H_0: \beta_6 = 0$ $H_1: \beta_6 \neq 0$	6.4388	0.018	P-value = 0.018 < 0.05, reject H_0.	O_3 concentration is significant in explaining PM_{10} concentration.
CO concentration	$H_0: \beta_7 = 0$ $H_1: \beta_7 \neq 0$	30.1498	0.000	P-value = 0.000 < 0.05, reject H_0.	CO concentration is significant in explaining PM_{10} concentration.

Based on Table 4.4, variables humidity, O_3 concentration and CO concentration are significant since p-value are less than alpha value which is 0.05. Therefore, humidity, O_3 concentration and CO concentration are significantly contributed to the PM_{10} concentration.

4.4.2.1 Checking on OLS Model Assumption

a) Multicollinearity

Table 4.5 Multicollinearity

Model	Unstandardized Coefficients		Standardized Coefficients	t	Significant	Collinearity Statistics	
	Estimate	Standard Error	Beta			Tolerance	VIF
Constant	475.313	144.182		3.297	0.001		
Wind Direction	0.002	0.066	0.001	0.026	0.979	0.910	1.099
Temperature	-4.937	3.185	-0.070	-1.550	0.122	0.694	1.440
Humidity	-3.445	0.772	-0.209	-4.460	0.000	0.640	1.563
SO_2 Concentration	-1.343	1.519	-0.036	-0.884	0.377	0.863	1.159
NO_2 Concentration	1.390	3.685	0.023	0.377	0.706	0.391	2.556
O_3 Concentration	6.439	2.707	0.132	2.378	0.018	0.459	2.181
CO Concentration	30.150	6.608	0.237	4.563	0.000	0.522	1.915

Based on the value of variance inflation and tolerance, the predictor variables are not highly correlated towards each other since the values of VIF are less than 10 and tolerance are more than 0.1. Therefore, no multicollinearity exists.

b) Normality of Distribution

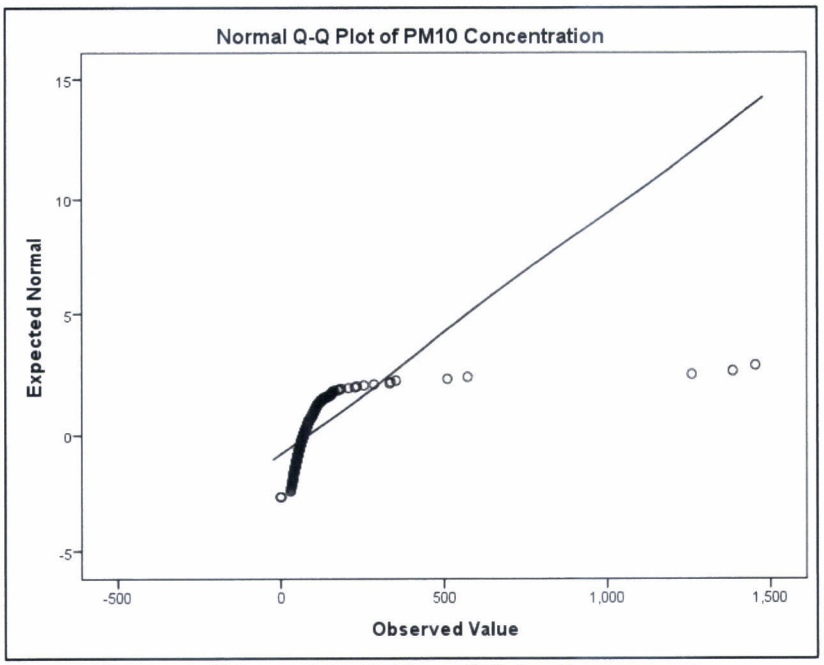


Figure 4.8 Normal Q-Q Plot of PM_{10} Concentration

Figure 4.8 shows a Q-Q plot where the points do not lie approximately along the straight line. This figure indicates that the residuals might be assumed to be not normally distributed. In order to confirm for the normality assumption, therefore Kolmogorov-Smirnov Test is conducted.

Table 4.6 Test of Normality

	Kolmogorov-Smirnov		
	Statistic	**df**	**Significant**
PM_{10} Concentration	0.298	649	0.000

The null hypothesis for this test is that the distribution of the residuals is normal, while the alternative hypothesis is that the distribution of the residuals is not normal. Since the p-value is 0.000 is less than alpha value which is 0.05, therefore there is enough evidence to reject the null hypothesis. This indicates that the distribution of the residuals is not normal.

c) Constant Variance

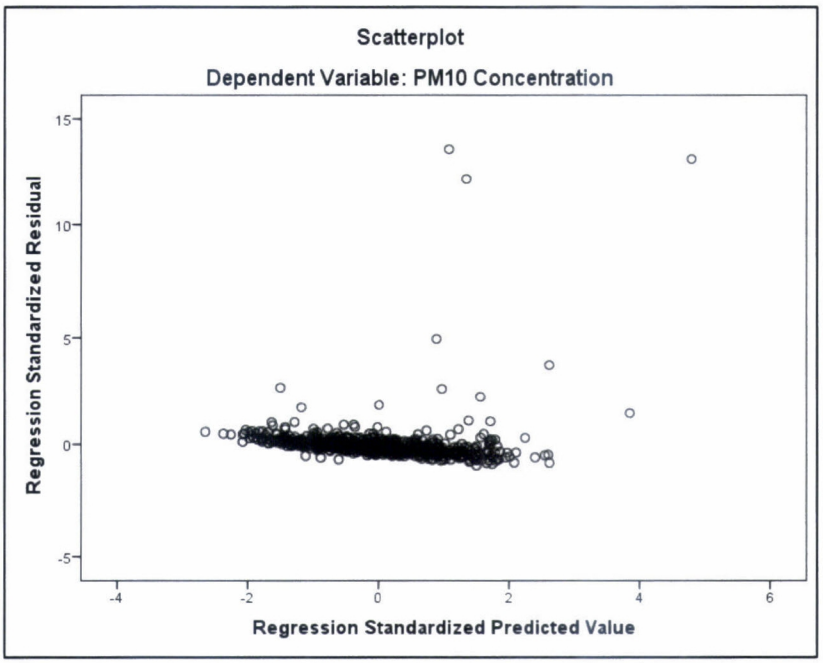


Figure 4.9 Plot of Residuals versus Predicted Value

Figure 4.9 shows that there is a pattern in the distribution of the plots. Therefore, by analyzing the pattern of the plot, it can be concluded that the residuals has no constant variance.

d) Independence

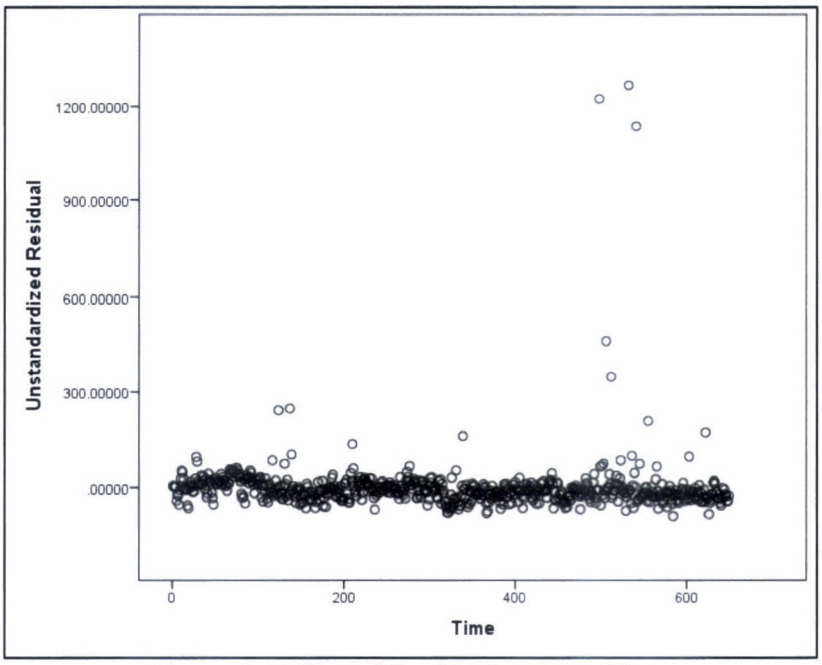


Figure 4.10 Plot of Residuals versus Time

Figure 4.10 shows that the distribution of the error term has no pattern over time. Therefore, it can be concluded that the residual is independent and has no potential problem with dependency. Testing independence of error terms can also be checked by using Durbin-Watson Statistic.

Table 4.7 Durbin-Watson Statistic for Independence of Error Terms

Durbin-Watson	1.925

The null hypothesis for this test is that the residuals are not correlated, while the alternative hypothesis is the residuals are correlated. Since Durbin-Watson is 1.925 is almost to 2, therefore the residuals are uncorrelated. This indicates that the error term is independent.

4.4.3 Robust Regression

Based on the above analysis, it showed that the outliers violate the assumption of Ordinary Least Squares (OLS). It is suggested to use a robust regression when we have an outlier in a data since it works with less restrictive assumption. Below are the analysis of the robust regression using two different method of estimations (Huber and Bisquare).

4.4.3.1 Bisquare

Table 4.8 Test Hypotheses of Model Coefficient for Bisquare

Variables	Hypotheses	Estimate	P-Value	Decision	Conclusion
Wind Direction	$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$	0.0389	0.011	P-value = 0.011 < 0.05, reject H_0.	Wind direction is significant in explaining PM_{10} concentration.
Temperature	$H_0: \beta_2 = 0$ $H_1: \beta_2 \neq 0$	1.8614	0.011	P-value = 0.011 < 0.05, reject H_0.	Temperature is significant in explaining PM_{10} concentration.
Humidity	$H_0: \beta_3 = 0$ $H_1: \beta_3 \neq 0$	-1.2322	0.000	P-value = 0.000 < 0.05, reject H_0.	Humidity is significant in explaining PM_{10} concentration.
SO_2 concentration	$H_0: \beta_4 = 0$ $H_1: \beta_4 \neq 0$	-0.1222	0.727	P-value = 0.727 > 0.05, fail to reject H_0.	SO_2 concentration is not significant in explaining PM_{10} concentration.
NO_2 concentration	$H_0: \beta_5 = 0$ $H_1: \beta_5 \neq 0$	4.1008	0.000	P-value = 0.000 < 0.05, reject H_0.	NO_2 concentration is significant in explaining PM_{10} concentration.
O_3 concentration	$H_0: \beta_6 = 0$ $H_1: \beta_6 \neq 0$	-0.8695	0.161	P-value = 0.161 > 0.05, fail to reject H_0.	O_3 concentration is not significant in explaining PM_{10} concentration.
CO concentration	$H_0: \beta_7 = 0$ $H_1: \beta_7 \neq 0$	9.2047	0.000	P-value = 0.000 < 0.05, reject H_0.	CO concentration is significant in explaining PM_{10} concentration.

Based on Table 4.8, variables wind direction, temperature, humidity, NO_2 concentration and CO concentration are significant since p-value are less than alpha value which is 0.05. Therefore, wind direction, temperature, humidity, NO_2 concentration and CO concentration contributed to the PM_{10} concentration.

4.4.3.2 Huber

Table 4.9 Test Hypotheses of Model Coefficient for Huber

Variables	Hypotheses	Estimate	P-Value	Decision	Conclusion
Wind Direction	$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$	0.0327	0.037	P-value = 0.037 < 0.05, reject H_0.	Wind direction is significant in explaining PM_{10} concentration.
Temperature	$H_0: \beta_2 = 0$ $H_1: \beta_2 \neq 0$	1.7344	0.020	P-value = 0.020 < 0.05, reject H_0.	Temperature is significant in explaining PM_{10} concentration.
Humidity	$H_0: \beta_3 = 0$ $H_1: \beta_3 \neq 0$	-1.3465	0.000	P-value = 0.000 < 0.05, reject H_0.	Humidity is significant in explaining PM_{10} concentration.
SO_2 concentration	$H_0: \beta_4 = 0$ $H_1: \beta_4 \neq 0$	0.1174	0.742	P-value = 0.742 > 0.05, fail to reject H_0.	SO_2 concentration is not significant in explaining PM_{10} concentration.
NO_2 concentration	$H_0: \beta_5 = 0$ $H_1: \beta_5 \neq 0$	3.7984	0.000	P-value = 0.000 < 0.05, reject H_0.	NO_2 concentration is significant in explaining PM_{10} concentration.
O_3 concentration	$H_0: \beta_6 = 0$ $H_1: \beta_6 \neq 0$	-0.7024	0.268	P-value = 0.268 > 0.05, fail to reject H_0.	O_3 concentration is not significant in explaining PM_{10} concentration.
CO concentration	$H_0: \beta_7 = 0$ $H_1: \beta_7 \neq 0$	11.3770	0.000	P-value = 0.000 < 0.05, reject H_0.	CO concentration is significant in explaining PM_{10} concentration.

Based on Table 4.9, variables wind direction, temperature, humidity, NO_2 concentration and CO concentration are significant since p-values are less than alpha value which is 0.05. Therefore, wind direction, temperature, humidity, NO_2 concentration and CO concentration are significantly contributed to the PM_{10} concentration.

4.4.4 Model Performance Comparison

Table 4.10 Summary of the Variables for Each Models

Methods	OLS		Huber		Bisquare	
	Standard Error	t-value	Standard Error	t-value	Standard Error	t-value
Intercept	144.1822	3.297	33.7931	2.94	33.1096	2.57
Wind Direction	0.0664	0.026	0.0156	2.09	0.0153	2.54
Temperature	3.1855	-1.550	0.7427	2.34	0.7264	2.56
Humidity	0.7723	-4.460	0.1830	-7.36	0.1802	-6.84
SO_2 Concentration	1.5191	-0.884	0.3561	-0.33	0.3502	-0.35
NO_2 Concentration	3.6847	0.377	0.8664	4.38	0.8479	4.84
O_3 Concentration	2.7073	2.378	0.6332	-1.11	0.6189	-1.40
CO Concentration	6.6077	4.563	1.6039	7.09	1.5854	5.81

Table 4.10 show the summary result of statistics which include the standard errors and t-values for each variable at each method. Based on the result obtained, it shows that all of the Standard Error (SE) value of each variables is the smallest in Bisquare compared to Huber and OLS. However, the t-value for each variable majority has the highest value in Huber.

Table 4.11 Model Performance Comparison

Models	MSE	R^2
OLS	8643.889	0.096
Bisquare	270.379	0.297
Huber	592.199	0.198

Table 4.11 shows the performance indicators between OLS method and robust regression. The result shows that OLS method has the highest MSE value with 8643.889 while between Bisquare and Huber, Bisquare has the lowest MSE value with 270.379. For R^2, Bisquare and Huber methods has better value compared to OLS which is 0.297 and 0.198 respectively. This indicates that MLR (OLS) is less suitable method compared to robust regression method in modelling PM_{10} concentration for data with outliers. Bisquare robust regression is better than MLR (OLS) and Huber robust regression for modelling PM_{10} concentration in Klang. However, the R^2 value of Bisquare does not reach the appropriate value which is close to 1. Therefore, Bisquare able to compare between these models but less generalizable to apply as suitable model for modelling PM_{10} concentration.

According to Zaman and Bulut (2019), they propose new regression-type estimators by considering Tukey-M, Hampel M, Huber MM, LTS, LMS and LAD robust methods and MCD and MVE robust covariance matrices in stratified sampling. They obtain the mean square error (MSE) for these estimators to evaluate the performance. Thus, this study used MSE to evaluate the performance of regression models that represents the PM_{10} concentration in Klang.

4.4.5 The Best Model

Table 4.12 Parameter Estimate for Bisquare

Variables	Estimate	Standard Error	t-value	P-value
Intercept	85.1875	33.1096	2.57	0.010
Wind Direction	0.0389	0.0153	2.54	0.011
Temperature	1.8614	0.7264	2.56	0.011
Humidity	-1.2322	0.1802	-6.84	0.000
SO_2 Concentration	-0.1222	0.3502	-0.35	0.727
NO_2 Concentration	4.1008	0.8479	4.84	0.000
O_3 Concentration	-0.8695	0.6189	-1.40	0.161
CO Concentration	9.2047	1.5854	5.81	0.000

The general equation of the model is:

$\hat{y}$ = 85.1875 + 0.0389 wind direction + 1.8614 temperature – 1.2322 humidity – 0.1222 SO_2 concentration + 4.1008 NO_2 concentration – 0.8695 O_3 concentration + 9.2047 CO concentration

Table 4.12 shows seven predictor variables based on several criterions. Two possible variables that not significant because of p-value > alpha value =0.05 will be chosen for comparison purpose among them. Next, between those two variables, variable with highest p-value that not significant were determined. In this case, the highest p-value that not significant is SO_2 concentration. Therefore, the most appropriate variable to be removed in the model is SO_2 concentration because it satisfies most of the criterion compare to other variables.

STEP 1: Backward Elimination after variable SO_2 Concentration is removed

Table 4.13 Parameter Estimate for Bisquare after variable SO_2 Concentration is removed

Variables	Estimate	Standard Error	t-value	P-value
Intercept	85.0769	33.1040	2.57	0.010
Wind Direction	0.0392	0.0153	2.56	0.010
Temperature	1.8689	0.7263	2.57	0.010
Humidity	-1.2355	0.1800	-6.86	0.000
NO_2 Concentration	4.0932	0.8466	4.83	0.000
O_3 Concentration	-0.8957	0.6144	-1.46	0.145
CO Concentration	9.0961	1.5707	5.79	0.000

The table above shows six predictor variables based on several criterions. One possible variable that not significant because p-value > alpha value = 0.05 was determined. In this case, the highest p-value that not significant is variables O_3 concentration. Therefore, the most appropriate variable to be removed in the model is O_3 concentration because it satisfies most of the criterion compare to other variables.

STEP 2: Backward Elimination after variable O_3 Concentration is removed

Table 4.14 Parameter Estimate for Bisquare after variable O_3 is removed

Variables	Estimate	Standard Error	t-value	P-value
Intercept	98.3708	31.8807	3.09	0.002
Wind Direction	0.0409	0.0153	2.67	0.007
Temperature	1.5094	0.6864	2.20	0.028
Humidity	-1.2694	0.1786	-7.11	0.000
NO_2 Concentration	3.4559	0.7334	4.71	0.000
CO Concentration	8.7453	1.5588	5.61	0.000

The result also shows that the best model is:

$\hat{y}$ = 98.3708 + 0.0409 wind direction + 1.5094 temperature – 1.2694 humidity + 3.4559 NO_2 concentration + 8.7453 CO concentration

Interpretation of all coefficients:

$\beta_0 = 98.3708$

When there is no change in the wind direction, temperature, humidity, NO_2 concentration and CO concentration, the PM_{10} concentration will remain at 98.3708 mg/m^3.

$\beta_1 = 0.0409$

If wind direction is increased by one unit (°) while temperature, humidity, NO_2 concentration and CO concentration are held constant, the expected PM_{10} concentration will decrease by 0.0409 mg/m^3.

$\beta_2 = 1.5094$

If temperature is increased by one unit (°C) while wind direction, humidity, NO_2 concentration and CO concentration are held constant, the expected PM_{10}

concentration will increase by 1.5094 mg/m^3.

β_3 = - 1.2694

If humidity is increased by one unit (%) while wind direction, temperature, NO_2 concentration and CO concentration are held constant, the expected PM_{10} concentration will decrease by 1.2694 mg/m^3.

β_4 = 3.4559

If NO_2 concentration is increased by one unit (ppm) while wind direction, temperature, humidity and CO concentration are held constant, the expected PM_{10} concentration will increase by 3.4559 mg/m^3.

β_5= 8.7453

If CO concentration is increased by one unit (ppm) while wind direction, temperature, NO_2 concentration and humidity are held constant, the expected PM_{10} concentration will increase by 8.7453 mg/m^3.

Model performance:

Table 4.15 The Regression Model

MSE	**R^2**
262.025	0.296

The value of R^2 show that the 29.6% of total variation in PM_{10} concentration was explained by wind direction, temperature, humidity, NO_2 concentration and CO concentration. The balance of 70.4% will be explained by other factors. Since the R^2 value is 0.296<0.8, therefore the model does not fit the data well.

According to Alma (2011), R^2 value for all robust methods (Least Trimmed Squares, Huber M-estimation, MM Estimation and S-Estimation) are decreasing when the percentage of outliers in dependent variable are increasing. Thus, the R^2 value in this study is small due to the presence of outliers.

Table 4.16 Test Hypotheses of model coefficient for Bisquare

Variables	Hypotheses	P-Value	α	Decision	Conclusion
Wind Direction	$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$	0.007	0.05	P-value = 0.007 < 0.05, reject H_0.	Wind direction is significant in explaining PM_{10} concentration.
Temperature	$H_0: \beta_2 = 0$ $H_1: \beta_2 \neq 0$	0.028	0.05	P-value = 0.028 < 0.05, reject H_0.	Temperature is significant in explaining PM_{10} concentration.
Humidity	$H_0: \beta_3 = 0$ $H_1: \beta_3 \neq 0$	0.000	0.05	P-value = 0.000 < 0.05, reject H_0.	Humidity is significant in explaining PM_{10} concentration.
NO_2 Concentration	$H_0: \beta_4 = 0$ $H_1: \beta_4 \neq 0$	0.000	0.05	P-value = 0.000 < 0.05, reject H_0.	NO_2 concentration is significant in explaining PM_{10} concentration.
CO Concentration	$H_0: \beta_5 = 0$ $H_1: \beta_5 \neq 0$	0.000	0.05	P-value = 0.000 < 0.05, reject H_0.	CO concentration is significant in explaining PM_{10} concentration.

The final model only includes five predictor variables which are wind direction, temperature, humidity, NO_2 concentration and CO concentration.

Hence the final model is:

$\hat{y}$ = 98.3708 + 0.0409 wind direction + 1.5094 temperature – 1.2694 humidity + 3.4559 NO_2 concentration + 8.7453 CO concentration

4.5 Conclusion

Based on past literature, Ul-Saufie et al. (2012) conducted a study by using robust regression and OLS to identify the best robust regression models for prediction of PM_{10} concentration in Pulau Pinang. The result showed the best method for handling data with outliers is robust regression. This past study shows that the results obtained in this analysis is quite similar with the same area of study and robust regression method as the best regression model.

On the other hand, Kamaruzzaman et al. (2017) has done a study to investigate the significant pollutant in Putrajaya. The result shows five variables significantly contributed the air quality which are wind speed, wind direction, Sulphur Dioxide, Nitrogen Dioxide and Carbon Monoxide. Among those five variables, wind direction, Nitrogen Dioxide and Carbon Monoxide are significant in our study.

A few studies have examined the role of temperature as an impact modifier to short-term exposure to pollutants like Ozone, whose dynamics powerfully rely on temperature (Jhun et al., 2014). Furthermore, a study investigated the potential exposure risk levels of PM_{10} and $PM_{2.5}$ concentrations at two different road configuration sites in Bangkok, Thailand, which is between covered and open roadside areas. Thus, the results show the exposure to particulate matters at the covered areas has higher potential risk for human compared to open areas (Sahanavin et al, 2016). In conclusion, these past studies show that the results obtained in this study quite similar with all significant variables. It shows that the results can be used for future researchers in further study. However, the obtained model cannot be used for prediction due to a low value of R^2 value which is 29.6%.

CHAPTER FIVE
CONCLUSION AND RECOMMENDATIONS

5.1 Chapter Overview

This chapter concludes the overall of the study comparing the regression models for PM_{10} concentration in Klang with the presence of outliers. Thus, a summary and brief finding of the study will be presented. Furthermore, a few suggestions also will be discussed on modelling PM_{10} concentration by using a robust regression method for future purposes. The future researchers might gain insight and make further improvements based on this study.

5.2 Conclusion

Measurements of air pollution concentrations often contain outliers that can significantly affect the analysis. However, outliers may provide valuable information and sometimes can be the most important observation. Therefore, regression models were compared for PM_{10} concentration in Klang with the presence of outliers. The best robust regression model will be suggested as an alternative method that insensitive to outliers in the air quality data.

Based on the scatter plot matrix between the response variable and predictor variables, it shows that there is positive weak relationship between PM_{10} concentration with O_3 concentration, NO_2 concentration and CO concentration. While, there is a no linear relationship between PM_{10} concentrations with wind direction, temperature, humidity, SO_2 concentration. This is due to the presence of outliers.

Robust regression technique should be taken into consideration as an alternative approach for modelling PM_{10} concentration in Klang. Assessment of model performance suggest MLR (OLS) is much less suitable technique compared to the robust regression method in modelling PM_{10} concentration for data with outliers. There are two estimation approaches in robust regression used which are Huber and Bisquare. Among these two robust estimations, Bisquare robust regression is better than Huber robust regression by comparing the MSE and R^2 value. Therefore, Bisquare robust regression is preferable model for PM_{10} concentration in Klang with the presence of

outliers.

Based on the overall result, by using Bisquare robust regression, the final model only includes five predictor variables which are wind direction, temperature, humidity, NO_2 concentration and CO concentration. This is because all these variables are significant in explaining PM_{10} concentration.

5.3 Recommendations

Similar to other research studies, this study also recommends further analysis. The same process of comparing the regression models for PM_{10} concentration can be implemented in any other areas besides Klang. Furthermore, regarding the outliers in OLS method, future researcher might be interested in handling the outliers to improve the quality of the research. Due to data limitation, it is suggested to include the wind speed into the model as one of the factors that may affect PM_{10} concentration in future study.

Besides, regarding the missing values in PM_{10} concentration data, future researcher might be interested in handling the missing values by using other missing data treatment methods to preserve the quality of the original data. Despite of using R programming, they can also use Matlab, Stata or SAS to analyze the data by using robust regression method. In addition, the future researcher may use another robust regression estimator which fits with the data sets suitability for further information.

REFERENCES

Abdullah, M. M. A., Tan, C. Y., Ramli, N. A., Yahaya, A. S., & Fitri, N. F. M. Y. (2011). Modelling of PM_{10} concentration for industrialized area in Malaysia: A case study in Shah Alam. *Physics Procedia*, 22, 318-324.

Akyüz, M., & Çabuk, H. (2009). Meteorological variations of PM2. 5/PM10 concentrations and particle-associated polycyclic aromatic hydrocarbons in the atmospheric environment of Zonguldak, Turkey. Journal of hazardous materials, 170(1), 13-21.

Alma, Ö. G. (2011). *Comparison of Robust Regression Methods in Linear Regression*. *6*(9), 409–421.

Anselin, L., & Rey, S. J. (2010). Perspectives on spatial data analysis. In *Perspectives on Spatial Data Analysis* (pp. 1-20). Springer, Berlin, Heidelberg.

Azmi, S. Z., Latif, M. T., Ismail, A. S., Juneng, L., & Jemain, A. A. (2010). Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. *Air Quality, Atmosphere & Health*, *3*(1), 53-64.

Bostan, I., Onofrei, M., Dascălu, E. D., Fîrțescu, B., & Toderaşcu, C. (2016). Impact of Sustainable Environmental Expenditures Policy on Air Pollution Reduction, During European Integration Framework. *Amfiteatru Economic Journal*, 18(42), 286-302.

Brajer, V., Mead, R. W., & Xiao, F. (2011). Searching for an Environmental Kuznets Curve in China's air pollution. *China Economic Review*, *22*(3), 383-397.

Čampulová, M., Grochová, L., & Michálek, J. (2018). Outlier detection in PM10 aerosols by generalised linear model. In *AIP Conference Proceedings* (Vol. 1978, No. 1, p. 090003). AIP Publishing.

Carmona, R., Díaz, J., Mirón, I.J., Ortiz, C., Luna, M.Y., Linares, C., 2016. Mortality attributable to extreme temperatures in Spain: a comparative analysis by city. Environ. Int. 91, 22–28. https://doi.org/10.1016/j.envint.2016.02.018.

Carslaw, D. C. (2015) The Openair manual-open-source tools for analysing air pollution data. *Manual for Version 1*.1–4, King's College London.

Chen, K., Wolf, K., Breitner, S., Gasparrini, A., Stafoggia, M., Samoli, E., Andersen, Z.J., Bero-Bedada, G., Bellander, T., Hennig, F., Jacquemin, B., Pekkanen, J., Hampel, R., Cyrys, J., Peters, A., Schneider, A., 2018b. Two-way effect modifications of air pollution and air temperature on total natural and

cardiovascular mortality in eight European urban areas. Environ. Int. 116, 186–196. https://doi.org/10.1016/j.envint. 2018.04.021.

Chua, Y. P. (2013). Spearman correlation test; Book 3, basic research statistic, analysis of likert scale data.

Dietz, T.; Rosa, E.A.; York, R. Environmentally efficient well-being: Is there a Kuznets curve? Appl. Geogr. 2012, 32, 21–28.

Europe, W. H. O. (2013). *Review of evidence on health aspects of air pollution—REVIHAAP project*. Technical Report, World Health Organization, Regional Office for Europe, Copenhagen, Denmark.

Fuzzi, S., Baltensperger, U., Carslaw, K., Decesari, S., Denier van der Gon, H., Facchini, M. C., ... & Nemitz, E. (2015). Particulate matter, air quality and climate: lessons learned and future needs. *Atmospheric chemistry and physics*, *15*(14), 8217-8299.

Hien, P. D., Bac, V. T., Tham, H. C., Nhan, D. D., & Vinh, L. D. (2002). Influence of meteorological conditions on PM2. 5 and PM2. 5− 10 concentrations during the monsoon season in Hanoi, Vietnam. *Atmospheric Environment*, *36*(21), 3473-3484.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics* (pp. 492-518). Springer, New York, NY.

J. Fox, Applied Regression Analysis, *Linear Models and Related Methods, 3th ed.*, Sage Publication, USA, 1997.

Jacob, D. J., & Winner, D. A. (2009). Effect of climate change on air quality. *Atmospheric environment*, *43*(1), 51-63.

Jamalani, M. A., Abdullah, A. M., Azid, A., Ramli, M. F., Baharudin, M. R., Bose, M. M., & Gumel, D. Y. (2016). Monthly analysis of PM_{10} in ambient air of Klang Valley, Malaysia. *Malaysian Journal of Analytical Sciences*, 20(5), 1159-1170.

Jang, E., Do, W., Park, G., Kim, M., & Yoo, E. (2017). Spatial and temporal variation of urban air pollutants and their concentrations in relation to meteorological conditions at four sites in Busan, South Korea. *Atmospheric Pollution Research*, *8*(1), 89-100.

Jerrett, M., Burnett, R.T., Pope, C.A., Ito, K., Thurston, G., Krewski, D., Shi, Y., Calle, E., Thun,M.,2009.Long-termozoneexposureandmortality.N.Engl.J.Med.360,1085. https://doi.org/10.1056/NEJMoa0803894

Jhun, I., Fann, N., Zanobetti, A., Hubbell, B., 2014. Eff ect modification of ozone-related mortality risks by temperature in 97 US cities. Environ. Int. 73, 128–134. https://doi. org/10.1016/j.envint.2014.07.009.

Juneng, L., Latif, M. T., & Tangang, F. (2011). Factors influencing the variations of PM10 aerosol dust in Klang Valley, Malaysia during the summer. *Atmospheric Environment*, *45*(26), 4370-4378.

K. Ho, J. Naugher, Outliers lie: An illustrative example of identifying outliers and applying robust models. multiple linear regression viewpoints, 26(2) (2000), 2-6.

Kamaruzzaman, A., Saudi, A. S. M., Azid, A., Balakrishnan, A., Abu, I. F., Amin, N. A., & Rizman, Z. I. (2017). Assessment on air quality pattern: A case study in Putrajaya, Malaysia. *Journal of Fundamental and Applied Sciences*, 9(4S), 789-800.

Kumar, A., Patil, R. S., Dikshit, A. K., & Kumar, R. (2017). Application of WRF model for air quality modelling and AERMOD–a survey. *Aerosol and Air Quality Research*, *17*(7), 1925-37.

Kuo, C. Y., Chen, P. T., Lin, Y. C., Lin, C. Y., Chen, H. H., & Shih, J. F. (2008). Factors affecting the concentrations of PM10 in central Taiwan. *Chemosphere*, *70*(7), 1273-1279.

Lee, W., Bell, M.L., Gasparrini, A., Armstrong, B.G., Sera, F., Hwang, S., Lavigne, E., Zanobetti, A., Coelho, M. de S.Z.S., Saldiva, P.H.N., Osorio, S., Tobias, A., Zeka, A., Goodman, P.G., Forsberg, B., Rocklöv, J., Hashizume, M., Honda, Y., Guo, Y.-L.L., Seposo, X., Van Dung, D., Dang, T.N., Tong, S., Guo, Y., Kim, H., 2018. Mortality burdenofdiurnaltemperature rangeanditstemporalchanges: amulti-countrystudy. Environ. Int. 110, 123–130. https://doi.org/10.1016/j.envint.2017.10.018.

Mahajan V., Sharma S., and Wind Y., (1984) Parameter estimation in marketing models in the presence of influential response Data: Robust Regression and applications, *Journal of Marketing Research*, vol. XXI, pp 288-277

Mahmud, M. (2005). Active fire and hotspot emissions in Peninsular Malaysia during 2002. *Geografia*, *1*(1), 1-17.

Miao, Q., Chen, D., Buzzelli, M., & Aronson, K. J. (2015). Environmental equity research: review with focus on outdoor air pollution research methods and analytic tools. *Archives of environmental & occupational health*, *70*(1), 47-55.

Min, C., & Min, C. (2019). Multiple linear regression models. *Applied Econometrics*, 43–68. https://doi.org/10.4324/9780429024429-3

Mohtar, A. A. A., Latif, M. T., Baharudin, N. H., Ahamad, F., Chung, J. X., Othman, M., & Juneng, L. (2018). Variation of major air pollutants in different seasonal conditions in an urban environment in Malaysia. *Geoscience Letters*, *5*(1), 21.

Moseholm, L., Silva, J., & Larson, T. (1996). Forecasting carbon monoxide concentrations near a sheltered intersection using video traffic surveillance and neural networks. *Transportation Research Part D: Transport and Environment*, *1*(1), 15-28.

Oh, W., & Kato, S. (2018). The effect of airspeed and wind direction on human's thermal conditions and air distribution around the body. *Building and Environment*, *141*, 103-116.

Paul, R. K. (2010). Multicollinearity: causes, effects and remedies. *Indian Agricultural Statistics Research Institute*, (4405), 14. https://doi.org/10.1111/j.1755-148X.2008.00460.x

Pausata, F. S. R., Gaetani, M., Messori, G., Kloster, S., & Dentener, F. J. (2015). The role of aerosol in altering North Atlantic atmospheric circulation in winter and its impact on air quality. *Atmospheric Chemistry and Physics*, *15*(4), 1725-1743.

Pope III, C. A., Ezzati, M., & Dockery, D. W. (2013). Fine particulate air pollution and life expectancies in the United States: the role of influential observations. *Journal of the air & waste management association*, *63*(2), 129-132.

Rasheed, B. A., Adnan, R., Saffari, S. E., & Pati, D. K. (2014). Robust weighted least squares estimation of regression parameter in the presence of outliers and heteroscedastic errors. *Jurnal Teknologi*, *71*(1).

Regression, M. (n.d.). *Robust Regression*. 1–25.

Ren, C., O'Neill, M.S., Park, S.K., vSparrow, D., Vokonas, P., Schwartz, J., 2011. Ambient temperature, air pollution, and heart rate variability in an aging population. Am. J. Epidemiol. 173, 1013–1021. https://doi.org/10.1093/aje/kwq477. Saiz-Lopez, A., Borge, R., Notario, A., Adame, J.A., Paz, D.D.L., Querol, X., Artíñano, B.,

Romieu, I., Gouveia, N., Cifuentes, L.A., de Leon, A.P., Junger, W., Vera, J., et al., 2012. Multicity Study of Air Pollution and Mortality in Latin America (the ESCALA Study). Res. Rep. Health Eff. Inst 171, 5-86.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection* (Vol. 1). New York: Wiley.

Sahanavin, N., Tantrakarnapa, K., & Prueksasit, T. (2016). Ambient PM10 and PM2. 5 concentrations at different high traffic-related street configurations in Bangkok, Thailand. *Southeast Asian Journal of Tropical Medicine and Public Health*, *47*(3), 528-535.

Seaman, N. L. (2000). Meteorological modeling for air-quality assessments. *Atmospheric environment*, *34*(12-14), 2231-2259.

Seaman, N. L. (2003). Future directions of meteorology related to air-quality research. *Environment international*, *29*(2-3), 245-252.

Shuhaili, A., Fadzil, A., Ihsan, S. I., & Faris, W. F. (2013). Air pollution study of vehiclesemission in high volume traffic: Selangor, Malaysia as a case study. *WSEAS Transactions on Systems*, *12*(2), 67-84.

Song, M.-L.; Zhang, W.; Wang, S.-H. Inflection point of environmental Kuznets curve in mainland China. *Energy Policy 2013*, 57, 14–20.

Stafoggia, M., Schwartz, J., Forastiere, F., Perucci, C.A., 2008. Does temperature modify the association between air pollution and mortality? A multicity case-crossover analysis in Italy. Am. J. Epidemiol. 167, 1476. https://doi.org/10.1093/aje/kwn074.

The, I., & Some, S. (1992). *Low Breakdown Regression Procedures*. (1979), 18–50.

Tiwari, S., Chate, D. M., Srivastaua, A. K., Bisht, D. S., & Padmanabhamurty, B. (2012). Assessments of PM1, PM2. 5 and PM10 concentrations in Delhi at different mean cycles. *Geofizika*, *29*(2), 125-141.

Ul-Saufie, A. Z., Yahaya, A. S., Ramli, N. A., & Hamid, H. A. (2012). Robust regression models for predicting PM_{10} concentration in an industrial area. *International Journal of Engineering and Technology*, 2(3), 364-370.

Vardoulakis, S., Dear, K., Hajat, S., Heaviside, C., Eggen, B., 2014. Comparative assessmentoftheeff ectsofclimatechangeonheat-andcold-related mortalityintheUnited Kingdom and Australia, 122. http://doi.org/10.1289/ehp.1307524.

World Health Organization. (2013). Review of evidence on health aspects of air pollution–REVIHAAP Project. World Health Organization, Copenhagen, Denmark.

Yu, C., & Yao, W. (2017). Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation*, *46*(8), 6261-6282.

Yulita, T., Notodiputro, K. A., & Sadik, K. (2018). M-Estimation Use Bisquare, Hampel, Huber, and Welsch Weight Functions in Robust Regression.

Zaman, T., & Bulut, H. (2019). Modified regression estimators using robust regression methods and covariance matrices in stratified random sampling. *Communications in Statistics-Theory and Methods*, 1-14.

APPENDICES

APPENDIX 1

Data observation

1	Location	Year	Wind_direction	Temperature	Humidity	So2_Conc	No2_Conc	O3_Conc	Co_Conc	PM10_Conc
2	Klang	2017	345.000	30.100	90.000	0.004	0.024	0.051	1.170	58.000
3	Klang	2017	341.000	28.900	90.000	0.004	0.012	0.033	0.720	48.000
4	Klang	2017	336.000	33.000	87.000	0.004	0.024	0.062	0.810	40.000
5	Klang	2017	338.000	31.000	90.000	0.003	0.019	0.033	0.630	36.000
6	Klang	2017	310.000	31.900	73.000	0.003	0.020	0.049	0.720	48.000
7	Klang	2017	337.000	33.200	74.000	0.005	0.019	0.043	0.720	58.000
8	Klang	2017	341.000	32.600	70.000	0.004	0.014	0.040	0.000	50.000
9	Klang	2017	353.000	33.600	70.000	0.002	0.024	0.042	1.080	48.000
10	Klang	2017	350.000	34.200	75.000	0.009	0.034	0.063	1.260	80.000
11	Klang	2017	353.000	33.600	80.000	0.009	0.046	0.103	1.080	96.000
12	Klang	2017	358.000	34.200	79.000	0.007	0.037	0.063	0.630	106.000
13	Klang	2017	358.000	32.000	85.000	0.006	0.037	0.084	0.900	100.000
14	Klang	2017	321.000	32.200	79.000	0.007	0.029	0.037	0.900	100.000
15	Klang	2017	272.000	32.800	77.000	0.004	0.029	0.042	0.900	74.000
16	Klang	2017	162.000	32.100	83.000	0.004	0.015	0.042	0.900	56.000
17	Klang	2017	322.000	30.900	88.000	0.002	0.019	0.038	0.720	35.000
18	Klang	2017	211.000	32.700	85.000	0.004	0.031	0.041	1.080	48.000
19	Klang	2017	288.000	32.400	85.000	0.006	0.039	0.049	1.080	0.000
20	Klang	2017	348.000	31.600	83.000	0.007	0.046	0.055	0.990	0.000
21	Klang	2017	345.000	31.300	86.000	0.006	0.031	0.043	0.810	56.000
22	Klang	2017	295.000	30.300	86.000	0.004	0.027	0.065	0.810	54.000
23	Klang	2017	352.000	33.500	88.000	0.005	0.027	0.055	1.080	58.000
24	Klang	2017	335.000	29.300	90.000	0.005	0.019	0.034	0.720	50.000
25	Klang	2017	264.000	24.600	90.000	0.003	0.019	0.015	0.450	45.000
26	Klang	2017	243.000	29.400	90.000	0.004	0.031	0.030	0.900	36.000
27	Klang	2017	256.000	30.500	91.000	0.005	0.024	0.041	0.900	45.000
28	Klang	2017	323.000	30.600	91.000	0.006	0.032	0.000	0.450	52.000

711	Klang	2018	319.857	30.462	93.817	0.0031	0.0243	0.028	1.652	48.041
712	Klang	2018	330.018	32.274	99	0.0023	0.0427	0.0412	2.159	52.916
713	Klang	2018	321.324	32.425	99	0.0057	0.0423	0.0405	1.848	69.339
714	Klang	2018	319.267	29.125	99	0.0054	0.0325	0.0192	1.691	71.575
715	Klang	2018	353.232	30.975	99	0.0134	0.0233	0.0309	1.847	70.953
716	Klang	2018	350.725	31.593	92.183	0.007	0.0297	0.0321	1.53	52.255
717	Klang	2018	357.046	32.147	93.875	0.0058	0.041	0.043	1.453	49.947
718	Klang	2018	194.183	33.328	92.85	0.0028	0.0378	0.0553	1.837	52.929
719	Klang	2018	327.385	32.898	92.167	0.0038	0.0406	0.0609	2.231	74.475
720	Klang	2018	350.425	32.322	88.978	0.0107	0.0359	0.0608	1.793	68.056
721	Klang	2018	351.479	32.573	85.55	0.0102	0.0356	0.0394	1.899	57.393
722	Klang	2018	359.738	32.334	86.991	0.0029	0.0378	0.0381	2.907	73.262
723	Klang	2018	350.926	32.417	89.917	0.0037	0.0348	0.049	2.173	66.335
724	Klang	2018	354.14	32.176	93.105	0.0077	0.0301	0.04	2.298	57.274
725	Klang	2018	356.096	31.441	89	0.0218	0.028	0.0363	1.34	46.962
726	Klang	2018	356.818	30.668	99	0.0052	0.0224	0.029	1.446	36.357
727	Klang	2018	357.732	29.545	98	0.0125	0.032	0.0294	2.243	61.176
728	Klang	2018	358.734	31.433	91.567	0.0049	0.0313	0.0286	2.057	59.582
729	Klang	2018	356.749	31.057	86.3	0.0078	0.0241	0.0311	1.741	48.574
730	Klang	2018	357.17	31.467	91.417	0.0051	0.0212	0.0426	2.239	56.114
731	Klang	2018	359.399	31.357	93.934	0.003	0.0188	0.0259	1.384	29.289

APPENDIX 2

Codes of R Programming

Retrieve data from excel:

```
data=read.csv("klang.csv")
names(data)
attach(data)
str(data)
```

Outliers:

```
library(olsrr)
ols_plot_resid_lev(ols)
ols_plot_cooksd_bar(ols)
ols_plot_resid_stud(ols)
```

Histogram:

```
hist(PM10_Concentration, prob=T, col="pink")
curve(dnorm(x, mean=mean(PM10_Concentration),
sd=sd(PM10_Concentration)), add=TRUE, col="red")

hist(Wind_Direction, prob=T, col="light blue")
curve(dnorm(x, mean=mean(Wind_Direction),
sd=sd(Wind_Direction)), add=TRUE, col="red")

hist(Temperature, prob=T, col="light green")
curve(dnorm(x, mean=mean(Temperature),
sd=sd(Temperature)), add=TRUE, col="red")

hist(Humidity, prob=T, col="light grey")
curve(dnorm(x, mean=mean(Humidity), sd=sd(Humidity)),
add=TRUE, col="red")

hist(SO2_Concentration, prob=T, col="turquoise")
curve(dnorm(x, mean=mean(SO2_Concentration),
```

```
sd=sd(SO2_Concentration)), add=TRUE, col="red")

hist(NO2_Concentration, prob=T, col="khaki")
curve(dnorm(x, mean=mean(NO2_Concentration),
sd=sd(NO2_Concentration)), add=TRUE, col="red")

hist(O3_Concentration, prob=T, col="lawngreen")
curve(dnorm(x, mean=mean(O3_Concentration),
sd=sd(O3_Concentration)), add=TRUE, col="red")

hist(CO_Concentration, prob=T, col="turquoise")
curve(dnorm(x, mean=mean(CO_Concentration),
sd=sd(CO_Concentration)), add=TRUE, col="red")
```

Scatter plot matrix:

```
pairs(data)
```

Ordinary Least Square Equation:

```
ols <- lm(PM10_Concentration ~ Wind_Direction +
Temperature + Humidity + SO2_Concentration +
NO2_Concentration + O3_Concentration + CO_Concentration ,
data = data)
summary(ols)
```

Robust Regression Equation:

```
library(robustreg)
robustBS <- robustRegBS(PM10_Concentration ~
Wind_Direction + Temperature + Humidity +
SO2_Concentration + NO2_Concentration + O3_Concentration
+ CO_Concentration, data = data, tune=4.685, m=TRUE,
anova.table=TRUE)
```

```
robustH <- robustRegH(PM10_Concentration ~ Wind_Direction
+ Temperature + Humidity + SO2_Concentration +
NO2_Concentration + O3_Concentration + CO_Concentration,
data = data, tune=1.345, m=TRUE, anova.table=TRUE)
```

The best model for Bisquare using backward elimination:

i. Full model

```
robustBS <- robustRegBS(PM10_Concentration ~
Wind_Direction + Temperature + Humidity +
SO2_Concentration + NO2_Concentration + O3_Concentration
+ CO_Concentration, data = data, tune=4.685, m=TRUE,
anova.table=TRUE)
```

ii. Step 1

```
robustBS <- robustRegBS(PM10_Concentration ~
Wind_Direction + Temperature + Humidity +
NO2_Concentration + O3_Concentration + CO_Concentration,
data = data, tune=4.685, m=TRUE, anova.table=TRUE)
```

iii. Step 2

```
robustBS <- robustRegBS(PM10_Concentration ~
Wind_Direction + Temperature + Humidity +
NO2_Concentration + CO_Concentration, data = data,
tune=4.685, m=TRUE, anova.table=TRUE)
```

MSE value for OLS:

```
library(dvmisc)
get_mse(ols)
```

APPENDIX 3

Correlation Table between Predictor Variables and Response Variable

Correlations

		PM10 Concentration	Wind Direction	Temperature	Humidity	SO2 Concentration	NO2 Concentration	O3 Concentration	CO Concentration
PM10 Concentration	Pearson Correlation	1	-.016	.054	-.068	.051	.180**	.197**	.240**
	Sig. (2-tailed)		.679	.172	.086	.193	.000	.000	.000
	N	649	649	649	649	649	649	649	649
Wind Direction	Pearson Correlation	-.016	1	-.205**	.241**	.026	.011	-.047	.109**
	Sig. (2-tailed)	.679		.000	.000	.513	.783	.234	.006
	N	649	649	649	649	649	649	649	649
Temperature	Pearson Correlation	.054	-.205**	1	-.417**	.010	.020	.241**	.019
	Sig. (2-tailed)	.172	.000		.000	.795	.609	.000	.631
	N	649	649	649	649	649	649	649	649
Humidity	Pearson Correlation	-.068	.241**	-.417**	1	.173**	.363**	.236**	.334**
	Sig. (2-tailed)	.086	.000	.000		.000	.000	.000	.000
	N	649	649	649	649	649	649	649	649
SO2 Concentration	Pearson Correlation	.051	.026	.010	.173**	1	.327**	.309**	.319**
	Sig. (2-tailed)	.193	.513	.795	.000		.000	.000	.000
	N	649	649	649	649	649	649	649	649
NO2 Concentration	Pearson Correlation	.180**	.011	.020	.363**	.327**	1	.680**	.662**
	Sig. (2-tailed)	.000	.783	.609	.000	.000		.000	.000
	N	649	649	649	649	649	649	649	649
	Pearson Correlation	.197**	-.047	.241**	.236**	.309**	.680**	1	.536**

O3 Concentration	Sig. (2-tailed)	.000	.234	.000	.000	.000	.000		.000
	N	649	649	649	649	649	649	649	649
CO Concentration	Pearson Correlation	.240**	.109**	.019	.334**	.319**	.662**	.536**	1
	Sig. (2-tailed)	.000	.006	.631	.000	.000	.000	.000	
	N	649	649	649	649	649	649	649	649

**. Correlation is significant at the 0.01 level (2-tailed).