EB5101 Foundation Business Analytics:

Linear Regression Assignment

Prediction of Atmospheric Particulate Matter (PM2.5) in Northern Taiwan by Multiple Linear Regression

Lecturer: Ms. Catherine Khaw

Student ID	Name	E-mail
A0178551X	Choo Ming Hui Raymond	e0267862@u.nus.edu
A0178431A	Huang Qingyi	e0267742@u.nus.edu
A0178415Y	Jiang Zhiyuan	e0267726@u.nus.edu
A0178365R	Wang Jingli	e0267676@u.nus.edu
A0178500J	Yang Chia Lieh	e0267811@u.nus.edu

Table of Contents

1. Introduction

An air quality monitoring dataset for Northern Taiwan was obtained from the Environmental Protection Administration of Taiwan. This dataset contains measurements from 25 observation stations located in Northern Taiwan, and comprises of 21 meteorological and environmental parameters measured in 2015.

2. Objective

The aim of this report is to build a predictive model using Multiple Linear Regression to predict the concentration of atmospheric particulate matter with diameter less than 2.5 micrometer (PM2.5) based on other measured environmental and meteorological data.

3. Data Preparation

Prior to using the dataset for analysis, a series of data preparation activities were carried out. This includes a sanity check of the data reported for each variable to ensure that the range of values reported for each variable were logical. The steps involved in data preparation are detailed in the following sub sections.

3.1 Invalid values

As identified in the data dictionary (included in the Appendix), invalid values due to equipment inspection, program inspection and human inspection were appended with "#,* and x" respectively. As these values could not be imputed reliably, listwise deletion was used to remove all rows containing such invalid values.

3.2 Data Imputation

As indicated in the data dictionary, three variables (PH_RAIN, RAINFALL and RAIN_COND) had "NR" (No Rainfall) indicated. While those under RAINFALL could be imputed as "0", as no rainfall means that 0mm of rainfall was measured, such observations under PH_RAIN and RAIN_COND had to be removed, as it's not possible to reflect these values numerically.

3.3 Removal of Date/Time and Location

Considering that the 21 environmental variables contains information such as humidity, ambient temperature etc, it was determined that these variables were sufficient to describe the conditions of the location and the time at which these data were collected. As such, Date/Time and Location were excluded in our analysis.

3.4 Missing Values

The remaining observations which had missing values after the above steps were removed due to the lack of domain knowledge. It was decided that imputation without the necessary domain knowledge may skew the dataset, and make the modelling less accurate.

4. Multiple Linear Regression Model Building

4.1 Visual Inspection of the Scatterplot of PM2.5

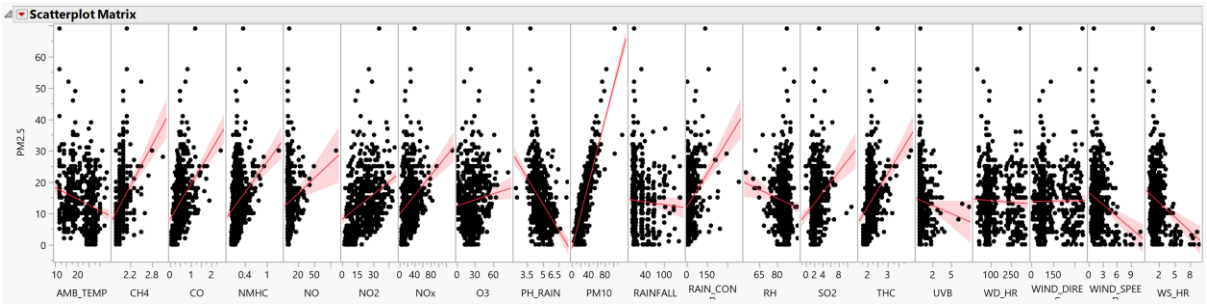

Figure 1: Scatterplot of PM2.5 versus all other variables

After data preparation was completed, a scatterplot of PM2.5 against all other variables was generated to investigate if there is any evidence of linearity between other predictors. From Figure 1, its noted that there's a very strong evidence of linearity between PM10 and PM2.5. This is logical, because the definition of PM10 (Particulate Matter of less than 10 micrometer in diameter) also overlaps the definition of PM2.5. Apart from this, PH_RAIN and NO_2 also appears to be following the trend of the fitted line generated. As such, this dataset was deemed suitable for Multiple Linear Regression.

4.2 Approach

To investigate if the 20 variables are sufficiently significant in explaining the variance of the dataset, these variables would be used in model fitting with their p values evaluated against a level of significance of 0.05, and the adjusted R^2 reviewed. Once the predictors have been finalized, their Variance Inflationary Factor (VIF) would be checked to ensure that no multi-collinearity exists between themselves.

4.3 Training and Validation Data Split

After the Data Preparation stage, the final dataset contains a total of 583 samples with 1 target variable and 20 predictors. Considering that the number of observations versus the number of predictors exceed the 10:1 ratio, the size of the dataset was deemed sufficient for Multiple Linear Regression. The dataset was then split into training data and test data based on a ratio of 70:30, so that sufficient data was available to train and validate the model.

4.4 Training Data Model

4.4.1 Model Building

From the training dataset obtained in section 4.3, a linear model was fitted with PM2.5 set as the dependent variable. Based on the results generated in R, the least significant variable was iteratively dropped, with the model evaluated after each removal of variables. This process was completed once there were no insignificant variables remaining.

Table 1: Removal of Variables and Resulting Adjusted R^2

Sequence of Removal	Variable Dropped	Adjusted R^2
initial model	-	0.8106
1	NO	0.8111
2	NMHC	0.8116
3	RAINFALL	0.812
4	RAIN_COND	0.8125
5	WIND_SPEED	0.8128
6	WIND_DIREC	0.813
7	CH4	0.8129
8	NO2	0.8122
9	THC	0.8117
10	WD_HR	0.8112
11	NOx	0.8105

The significant variables remaining are listed in Table 2. At this stage, the adjusted R^2 of the model is 0.8105. In addition, by inspecting the values of Estimate ± Standard Error, it's verified that the coefficients of these variables are non-zero. As such, the finalised variables are considered sufficient for further analysis.

Table 2: List of Variables, Estimate, Standard Error and P Value

Variable	Interpretation	Estimate	Standard Error	P value
AMB_TEMP	Ambient Air Temperature	-0.24038	0.04669	0.000000414
CO	Carbon Monoxide	4.93266	0.93706	0.000000231
O_3	Ozone	0.06573	0.02042	0.001388
PH_RAIN	pH of Rain	-1.35644	0.34374	0.0000938
PM10	Particulate Matter ≤ 10µm	0.47179	0.01566	< 2e-16
RH	Relative Humidity	0.18825	0.04953	0.000167
SO_2	Sulphur Dioxide	0.40099	0.18984	0.035289
UVB	Ultraviolet B	0.66109	0.29199	0.024104
WS_HR	Average Wind Speed per Hour	-0.59653	0.17656	0.0008

4.4.2 Residual Analysis

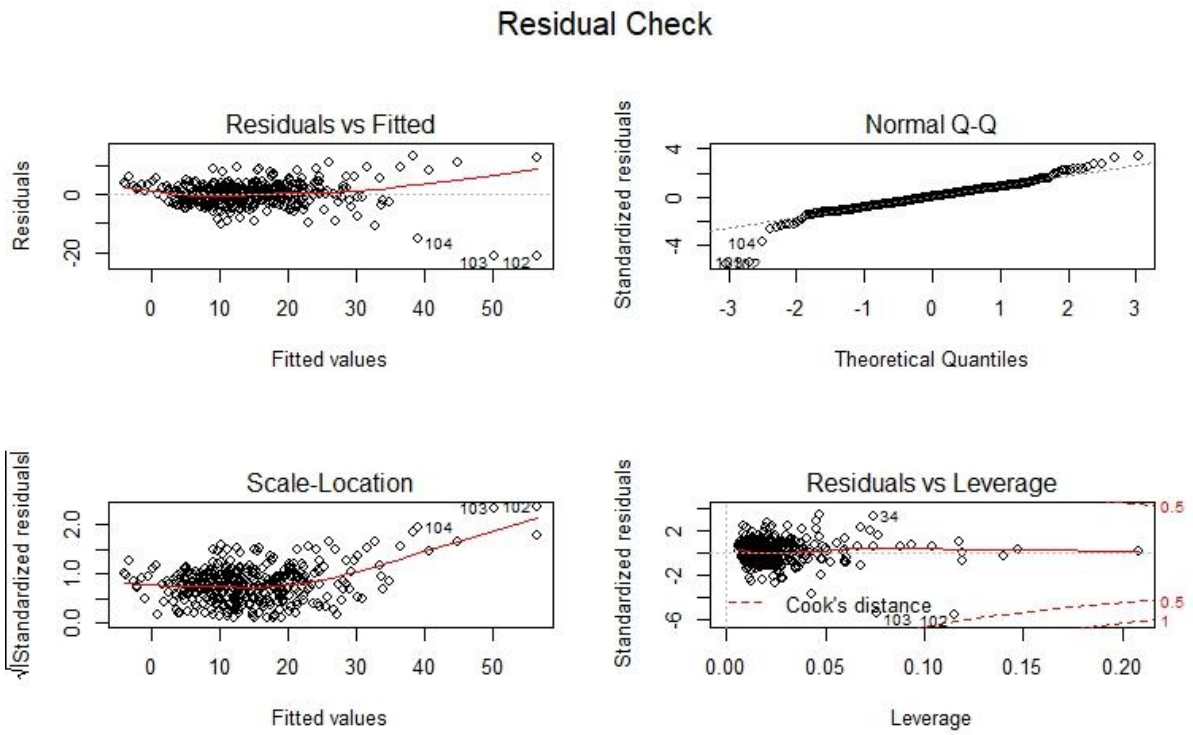

Figure 2: Diagnostic Plots for Linear Regression Analysis

4.4.2.1 Non-Linearity between Dependent and Predictor Variables

From Figure 2, it's observed from the Residuals versus Fitted Plot that there is no non-linear relationship between the predictors and the dependent variable. This means that the multiple linear regression model is capable of capturing the data in the trained dataset.

4.4.2.2 Normality of Residuals

From the Normal Q-Q Plot, a straight line is observed. This satisfies the assumption that the residuals are normally distributed.

4.4.2.3 Homoscedasticity of Residuals

From the Scale-Location Plot, there appears to be evidence of heteroscedasticity, as the red line is moving upwards as the fitted values increases beyond 30. Further verification through the Breusch-Pagan Test in R confirms that heteroscedasticity is present, with a p-value of less than 2.2e-16. This violates the assumption of linear regression that the variance of the residuals is constant.

4.4.2.4 Effect of Outliers on Model

Lastly, the Residuals versus Leverage Plot indicates that there are no outliers that have a high Cook's distance. This means that any outliers in our dataset, if any, are not significant in influencing the linear regression model built.

4.4.3 Correcting the model for Heteroscedasticity

Since heteroscedasticity is observed in our model as indicated in section 4.4.2.3, transformation of the dependent variable and application of a weighted least squares model were two approaches explored in an attempt to correct the model.

4.4.3.1 Dependent Variable Transformation

Various approaches (e.g. applying a natural logarithmic function to PM2.5) were explored to transform the dependent variable before rebuilding a new model. However, the approach that yielded the best result appeared to be a Box Cox Transformation. The intent of Box Cos Transformation is to find a value of λ such that the dependent variable is transformed to the following:

$$PM2.5_{New} = \frac{PM2.5^{\lambda} - 1}{\lambda}$$

Due to the limitation of the transformation, 13 instances of the training dataset where PM2.5 = 0 was removed. Based on the transformation, λ = 0.4 was obtained. Building a new model

based on this new transformed dependent variable yielded the results in Table 3, with an adjusted R^2 of 0.7577.

Table 3: List of Variables, Estimate, Standard Error and P Value (Transformed Dependent Variable)

Variable	Interpretation	Estimate	Standard Error	P value
AMB_TEMP	Ambient Air Temperature	-0.05957	0.009197	2.82E-10
CO	Carbon Monoxide	0.76531	0.3363	0.02341
NMHC	Non-Methane Hydrocarbon	1.38527	0.65308	0.03454
O3	Ozone	0.01689	0.00424	8.2E-05
PH_RAIN	pH of Rain	-0.3407	0.07574	9.1E-06
PM10	Particulate Matter ≤ 10µm	0.08149	0.00337	< 2E-16
SO2	Sulphur Dioxide	0.08564	0.04279	0.04606
WS_HR	Average Wind Speed per Hour	-0.1588	0.04859	0.00118

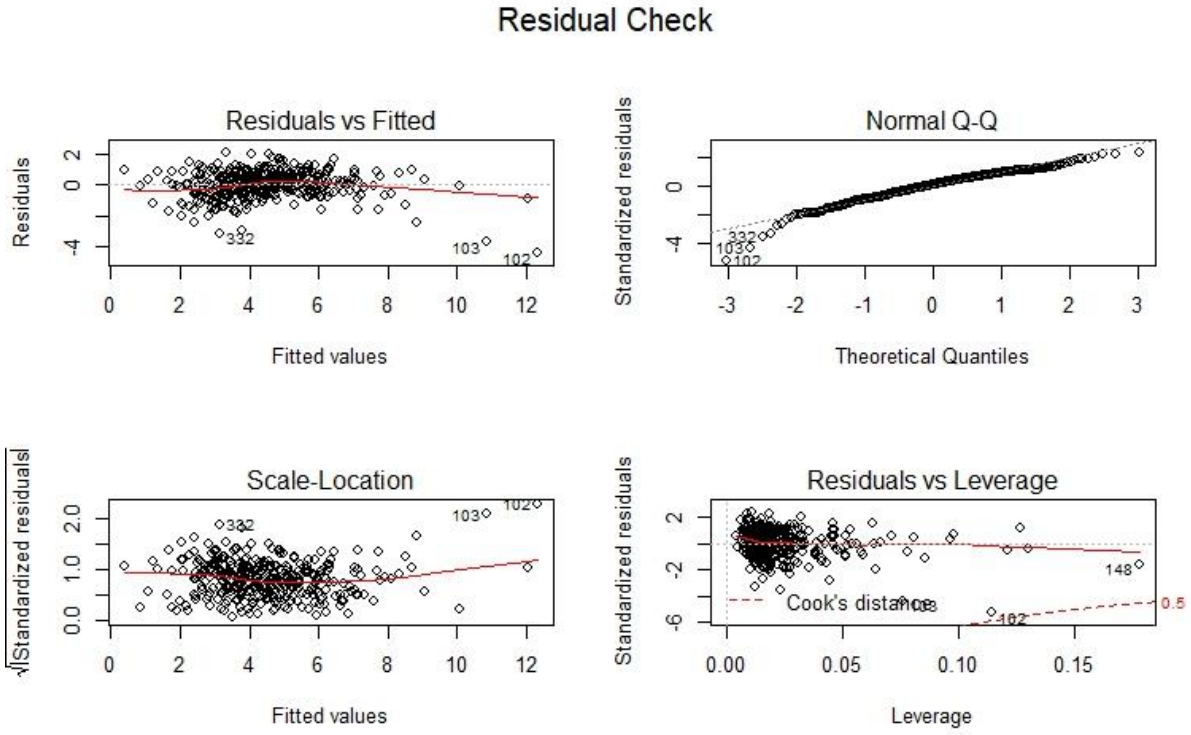

Figure 3:Diagnostic Plots for Linear Regression Analysis (Transformed Dependent Variable)

On inspecting the residual diagnostic plots of the newly built model, it's noted that there is an improvement in the Scale-Location Plot. In addition, the characteristics of the other three plots remain unchanged. However, on verifying the model with the Breusch -Pagan Test, it's noted that the residuals are still considered as heteroscedastic, with a p-value of 4.766e-08 rejecting the null hypothesis that the residuals are homoscedastic.

4.4.3.2 Weighted Least Squares

Since the initial model was built based on ordinary least square estimators, heteroscedasticity was likely to occur due to incorrect computation of standard errors. Accordingly, the initial model was transformed into one with homoscedastic errors, using generalized least squares estimator.

It is noted that this transformation requires a logarithmic function to be applied on all predictors. As such, observations wherein the variables contained zero value were removed (2 occurrences of SO2 and 1 occurrence of WS_HR), while UVB was removed from the dataset altogether as there were 233 occurrences of zero values for this variable, amounting to more than half of the training dataset. This resulted in a training dataset of 407 observations with 19 variables used to build a new model applying weighted least squares.

The fitted models are as follows:

Ordinary Least Squares:

$$PM2.5 = -8.83 - 0.24AMB_{TEMP} + 4.93CO + 0.07O3 - 1.36PH\_RAIN + 0.47PM10 + 0.19RH + 0.40SO2 + 0.66UVB - 0.59WS\_HR$$

Generalised Least Squares:

$$PM2.5 = -3.95 - 0.22AMB_{TEMP} + 5.21CO + 0.07O3 - 1.18PH\_RAIN + 0.46PM10 + 0.12RH + 0.41SO2 - 0.42WS\_HR$$

However, the adjusted R^2 of the Generalised Least Squares model dropped from 0.8105 to 0.8047, while the p-value of the Breusch-Pagan Test did not improve. Hence, this approach was not considered to be appropriate, and the dependent variable transformation approach was adopted instead.

4.5 Multi-Collinearity

Based on the model built in section 4.4.3.1, a VIF test was done to inspect the finalized dataset for multi-collinearity.

Table 4: VIF Results

AMB_TEMP	CO	NMHC	O3	PH_RAIN	PM10	SO2	WS_HR
1.200704	5.088495	5.868661	1.933892	1.282449	1.304998	1.247557	1.276881

From the results, it was noted that CO and NMHC exhibit multi-collinearity against each other. NMHC was dropped due to its lower significance compared to CO, resulting in a model without multi-collinearity. At this point, the adjusted R^2 is 0.7556, which is a negligible drop from the earlier model. In addition, the residual diagnostic plots did not exhibit any difference due to the removal of NMHC from the model.

4.6 Model Accuracy

With the model finalized, the validation dataset was used to evaluate the prediction accuracy of the model. Predicted values of the transformed dependent variables were generated with $\lambda = 0.4$ as per section 4.4.3.1, and these were compared with the actual values of the transformed dependent variables using the Mean Absolute Percentage Error (MAPE).

The formula of MAPE is given by:

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

where n is number of observations, A_t is the actual value, and F_t is the forecasted value.

Since the limitation of MAPE is that Actual values cannot be zero (due to division by 0), these two occurrences were removed from the test dataset first. This gives us a MAPE of 25.8%. While this result is not ideal, there's a marked improvement in accuracy as compared to the original model built in section 4.1, before dependent variable transformation. If that model was used, the MAPE of that model comes to 34.3%.

5. Final Model

The formula of the final model is given by:

$$\frac{PM2.5^{0.4}-1}{0.4} = 3.738034 - -0.055182AMB_{TEMP} + 1.323396CO + 0.01435503 - 0.341518PH\_RAIN + 0.081491PM10 + 0.103794SO2 - 0.175049WS\_HR$$

6. Conclusion

In this report, an air quality monitoring dataset for Northern Taiwan was used to build a Multiple Linear Regression model to predict the concentration of PM2.5 based on other measured environmental and meteorological data. While the dataset was assessed to be suitable for Linear Regression, heteroscedasticity of the residuals was observed during the initial residual analysis. The dependent variable was transformed using Box Cox Transformation to correct for heteroscedasticity, resulting in an approximately 10% improvement in model accuracy as compared to the original approach, with the final model having a MAPE of 25.8% based on the validation dataset. Moving forward, an alternative predictive model such as logistic regression through categorizing PM2.5 into dichotomous levels (e.g. Healthy, Unhealthy) could be explored to improve the accuracy of the predictive model, as homoscedasticity is not required for error terms in logistic regression.

Appendix A

The following metadata is provided by the Environmental Protection Administration of Taiwan together with the raw data.

The columns in csv file are:

- time - The first column is observation time of 2015
- station - The second column is station name, there is 25 observation stations
    - [Banqiao, Cailiao, Datong, Dayuan, Guanyin, Guting, Keelung, Linkou, Longtan, Pingzhen, Sanchong, Shilin, Songshan, Tamsui, Taoyuan, Tucheng, Wanhua, Wanli, Xindian, Xinzhuang, Xizhi, Yangming, Yonghe, Zhongli, Zhongshan]
- items - From the third column to the last one
    - item - unit - description
    - SO2 - ppb - Sulfur dioxide
    - CO - ppm - Carbon monoxide
    - O3 - ppb - ozone
    - PM10 - μg/m3 - Particulate matter
    - PM2.5 - μg/m3 - Particulate matter
    - NOx - ppb - Nitrogen oxides
    - NO - ppb - Nitric oxide
    - NO2 - ppb - Nitrogen dioxide
    - THC - ppm - Total Hydrocarbons
    - NMHC - ppm - Non-Methane Hydrocarbon
    - CH4 - ppm - Methane
    - UVB - UVI - Ultraviolet index
    - AMB_TEMP - Celsius - Ambient air temperature
    - RAINFALL - mm
    - RH - % - Relative humidity
    - WIND_SPEED - m/sec - The average of last ten minutes per hour
    - WIND_DIREC - degress - The average of last ten minutes per hour
    - WS_HR - m/sec - The average of hour
    - WD_HR - degress - The average of hour

  - PH_RAIN - PH - Acid rain
  - RAIN_COND - µS/cm - Conductivity of acid rain

Data mark

- # indicates invalid value by equipment inspection
- indicates invalid value by program inspection
- x indicates invalid value by human inspection
- NR indicates no rainfall
- blank indicates no data