Menemui Matematik
(Discovering Mathematics)

journal homepage: https://persama.org.my/dismath/home

Exploring the Relationship Between Environmental Variables and Air Quality: A Statistical Perspective

Lim Fong Peng[1*], Wan Maryam Hazirah binti Wan Mohamad Sukri[2], Yap Hong Keat[3] and Kek Sie Long[4]

[1,2]*Department of Mathematics and Statistics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.*
[3]*Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Selangor, Malaysia.*
[4]*Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 84600 Pagoh, Muar, Johor, Malaysia.*

[1*]fongpeng@upm.edu.my, [2]mrymhzrh@gmail.com, [3]yaphk@utar.edu.my, [4]slkek@uthm.edu.my
*Corresponding author

Received: 8 September 2025
Accepted : 4 December 2025

ABSTRACT

In recent years, growing concerns about air quality have emerged, often linked to a lack of consideration for sustainable development. Various environmental and anthropogenic factors contribute to air quality degradation, yet identifying the most influential ones remains a challenge. This study aims to determine the key factors affecting air quality using multiple linear regression (MLR) analysis and to develop the best-fitted predictive model through a data set obtained from Saverio De Vito, ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development. Statistical analysis such as multicollinearity checks, stepwise regression, best subset regression, and residual analysis are conducted for all factors of air quality that we consider in this study. A preliminary model is constructed and refined, ensuring all regression assumptions are met. The final model highlights significant predictor variables and offering valuable insights into the factors most strongly associated with air quality.

Keywords: air quality, environmental variables, multiple linear regression

INTRODUCTION

Air pollution is a critical environmental and public health concern, characterized by the presence of harmful or undesirable substances in the atmosphere. Key air pollutants, such as carbon monoxide (CO), carbon dioxide (CO_2), nitrogen oxides (NO_x), and particulate matter, are released from both natural and anthropogenic sources. While natural contributors include volcanic activity, wildfires, and methane emissions from livestock, human activities are the predominant cause of air pollution, particularly in urban and industrial areas. Major anthropogenic sources include vehicular emissions, industrial operations, land clearing, and construction-related activities.

Among these, construction sites and diesel-powered engines are notable contributors to airborne pollutants. Construction activities release large volumes of dust and suspended particulates, while diesel engines emit toxic gases such as hydrocarbons and nitrogen oxides. These emissions not only degrade environmental quality but also pose significant health risks. Prolonged exposure to such pollutants has been associated with respiratory illnesses,

cardiovascular disease, and an increased incidence of cancer. Vulnerable populations, particularly pregnant women and children are at greater risk due to the potential for developmental and long-term health complications.

According to the World Health Organization (WHO), air pollution is responsible for more than two million premature deaths each year worldwide. Brunekreef (2007) noted that nitrogen oxides, frequently used as proxies for combustion-related pollutants, are closely linked to traffic emissions and serve as indicators of broader exposure to harmful compounds. Additionally, air pollution contributes to global environmental issues such as acid rain and climate change. Greenhouse gases, including methane, ozone, and water vapor, trap heat in the Earth's atmosphere, thereby accelerating global warming.

The adverse effects of acute pollution events have also been documented. For instance, the collapse of the Twin Towers on September 11, 2001, released large quantities of hazardous materials, including asbestos, lead, and mercury, into the atmosphere. The resulting exposure led to increased cases of respiratory illnesses and cancer among first responders, volunteers, and nearby residents, highlighting the severity of short-term, high-intensity pollution exposure.
In response to the growing awareness of these health and environmental impacts, agencies such as the Environmental Protection Agency (EPA) have intensified efforts to monitor and manage air quality. Advances in sensor technology and increased understanding of the interactions between air quality and climate change have supported the development of evidence-based policies. However, to effectively mitigate air pollution, it is essential to identify and quantify the specific factors contributing to its variation.

Multiple linear regression (MLR) analysis is a widely used statistical method for modelling the relationship between a dependent variable and multiple independent variables. As noted by Zsuzanna and Marian (2012), MLR allows researchers to assess the simultaneous influence of several predictors and serves as a powerful tool for forecasting. Uyanik and Guler (2013) emphasized its value in establishing causal relationships and noted the importance of fulfilling key assumptions, including normality, linearity, the absence of outliers, and low multicollinearity. Recent studies demonstrate that multiple linear regression remains a reliable and interpretable statistical tool for analysing the complex interplay of environmental and anthropogenic variables affecting air quality. Its predictive capacity and ability to isolate significant contributors make it especially valuable in urban pollution management and regulatory planning (Lim et al, 2016, 2019; Oh, 2022, 2023; Ali & Rahman, 2024; Jiang, 2025).

This study applies multiple linear regression analysis to investigate the primary factors influencing air quality. By developing a statistically sound predictive model, this research aims to enhance understanding of the key contributors to air pollution and support data-driven strategies for air quality management and environmental policy development.

METHODOLOGY

The dataset used in this study was provided by Saverio De Vito of ENEA – the Italian National Agency for New Technologies, Energy, and Sustainable Economic Development. It comprises 13 variables collected over 9,358 instances, representing hourly averaged responses from an array of five metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was deployed at road level in a heavily polluted urban area within an Italian city. Data were continuously recorded from March 2004 to February 2005, constituting one of the longest publicly

available datasets of real-world air quality sensor measurements. All variables analyzed in this study are listed in the table below:

Table 1: Description of Data Sets

Variable	Description
PT08.S1(CO)	Hourly averaged sensor response (tin oxide)
CO(GT)	True hourly averaged concentration CO in mg/m^3
NMHC(GT)	True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3
C_6H_6(GT)	True hourly averaged Benzene concentration in microg/m^3
PT08.S2(NMHC)	Hourly averaged sensor response (titania)
NO_x(GT)	True hourly averaged NO_x concentration in ppb
PT08.S3(NOX)	Hourly averaged sensor response (tungsten oxide)
NO_2(GT)	True hourly averaged NO_2 concentration in microg/m^3
PT08.S4(NO_2)	Hourly averaged sensor response (tungsten oxide)
PT08.S5(O_3)	Hourly averaged sensor response (indium oxide)
T	Temperature in Â°C
RH	Relative Humidity (%)
AH	Absolute Humidity

In statistical analysis, we initially investigate the issues of multicolinearity among the variables that we consider, as listed in Table 1, by calculating the Pearson correlation and variance inflation factor (VIF) of the variables. High values of Pearson correlation coefficient and VIF show the variables are highly correlated among each other, and cause the multicolinearity problem. Excluding those variables with multicolinearity problem, we proceed to determine the best-fitted model by using stepwise regression at the significant level of 0.05. Residual analysis then is conducted for assessing the quality of the best fitted model that we propose. It may helps determine if the model's assumptions are met and if the model adequately captures the relationship between variables.

RESULTS

In identifying the existence of the issues of multicolinearity, Figure 1 shows that the correlation coefficients are high for the variables T vs C_6H_6(GT), AH vs C_6H_6(GT), AH vs T and AH vs RH, that is 0.971, 0.985, 0.981 and 0.944 respectively. Their high correlation relationship can be further justified by the significant results in correlation tests at the significant level of 0.10. When fitting all variables we consider in a regression model, Figure 2 points out that the variance inflation factor (VIF) values of C_6H_6(GT), PT08.S2(NMHC), PT08.S4(NO_2), T, RH and AH are greater than 10, that is 1302.14, 103.29, 21.54, 195.05, 42.50 and 1490.97 respectively, which tend to the occurrence of multicollinearity. These variables are then removed from the regression model. Removal of variables are carried out until there is no multicollinearity occurrence. The remaining variables then are used to proceed to the stepwise regression in order to gain the best fitted regression model at the significant level of 0.05. Thus, the results found that the air quality is significantly affected by PT08.S5(O_3), CO(GT), NMHC(GT), NO_x(GT), NO_2(GT), PT08.S4(NO_2) and RH. Futhermore, the histogram in Figure 3 - 4 show that the air quality data is normal distributed.

Correlation: PT08.S1(CO), CO(GT), NMHC(GT), C6H6(GT), PT08.S2(NMHC), ...

	PT08.S1(CO)	CO(GT)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)
CO(GT)	0.041 0.000				
NMHC(GT)	0.170 0.000	0.128 0.000			
C6H6(GT)	0.853 0.000	-0.031 0.002	0.037 0.000		
PT08.S2(NMHC)	0.933 0.000	0.030 0.004	0.110 0.000	0.767 0.000	
NOx(GT)	0.278 0.000	0.526 0.000	-0.004 0.670	-0.001 0.910	0.331 0.000
PT08.S3(NOx)	0.087 0.000	-0.090 0.000	0.049 0.000	0.512 0.000	-0.074 0.000
NO2(GT)	0.154 0.000	0.671 0.000	0.103 0.000	-0.011 0.289	0.177 0.000
PT08.S4(NO2)	0.845 0.000	-0.074 0.000	0.163 0.000	0.775 0.000	0.875 0.000
PT08.S5(O3)	0.892 0.000	0.080 0.000	0.101 0.000	0.641 0.000	0.910 0.000
T	0.755 0.000	-0.069 0.000	-0.000 0.999	0.971 0.000	0.669 0.000
RH	0.745 0.000	-0.048 0.000	0.008 0.423	0.925 0.000	0.586 0.000
AH	0.765 0.000	-0.046 0.000	0.013 0.227	0.985 0.000	0.647 0.000

	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)
PT08.S3(NOx)	-0.436 0.000				
NO2(GT)	0.817 0.000	-0.256 0.000			
PT08.S4(NO2)	0.036 0.001	0.123 0.000	-0.022 0.033		
PT08.S5(O3)	0.462 0.000	-0.209 0.000	0.253 0.000	0.724 0.000	
T	-0.138 0.000	0.588 0.000	-0.084 0.000	0.755 0.000	0.504 0.000
RH	-0.053 0.000	0.574 0.000	-0.081 0.000	0.641 0.000	0.525 0.000
AH	-0.096 0.000	0.622 0.000	-0.060 0.000	0.692 0.000	0.519 0.000

	T	RH
RH	0.886 0.000	
AH	0.981 0.000	0.944 0.000

Cell Contents: Pearson correlation
P-Value

Figure 1: Pearson Correlation Coefficient Between Variables

Regression Analysis: PT08.S1(CO) versus CO(GT), NMHC(GT), C6H6(GT), PT08.S2(NMHC, ...

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	12	970772043	80897670	16094.53	0.000
CO(GT)	1	2251	2251	0.45	0.503
NMHC(GT)	1	6365568	6365568	1266.43	0.000
C6H6(GT)	1	748491	748491	148.91	0.000
PT08.S2(NMHC)	1	40673	40673	8.09	0.004
NOx(GT)	1	129965	129965	25.86	0.000
PT08.S3(NOx)	1	2230775	2230775	443.81	0.000
NO2(GT)	1	311	311	0.06	0.804
PT08.S4(NO2)	1	486745	486745	96.84	0.000
PT08.S5(O3)	1	7292651	7292651	1450.87	0.000
T	1	84844	84844	16.88	0.000
RH	1	364716	364716	72.56	0.000
AH	1	87293	87293	17.37	0.000
Error	9344	46966742	5026		
Lack-of-Fit	9313	46966742	5043	*	*
Pure Error	31	0	0		
Total	9356	1017738786			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
70.8972	95.39%	95.38%	95.37%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	783.6	17.8	43.93	0.000	
CO(GT)	-0.0087	0.0130	-0.67	0.503	1.90
NMHC(GT)	0.21990	0.00618	35.59	0.000	1.39
C6H6(GT)	7.800	0.639	12.20	0.000	1302.14
PT08.S2(NMHC)	0.0619	0.0218	2.84	0.004	103.29
NOx(GT)	0.03908	0.00769	5.08	0.000	7.28
PT08.S3(NOx)	-0.13451	0.00638	-21.07	0.000	7.87
NO2(GT)	-0.0033	0.0134	-0.25	0.804	5.42
PT08.S4(NO2)	0.07161	0.00728	9.84	0.000	21.52
PT08.S5(O3)	0.19324	0.00507	38.09	0.000	10.00
T	-0.973	0.237	-4.11	0.000	195.05
RH	0.7948	0.0933	8.52	0.000	42.50
AH	-3.026	0.726	-4.17	0.000	1490.97

Regression Equation

PT08.S1(CO) = 783.6 - 0.0087 CO(GT) + 0.21990 NMHC(GT) + 7.800 C6H6(GT)
+ 0.0619 PT08.S2(NMHC) + 0.03908 NOx(GT) - 0.13451 PT08.S3(NOx)
- 0.0033 NO2(GT) + 0.07161 PT08.S4(NO2) + 0.19324 PT08.S5(O3) - 0.973 T
+ 0.7948 RH - 3.026 AH

Figure 2: Pearson Correlation Coefficient Between Variables

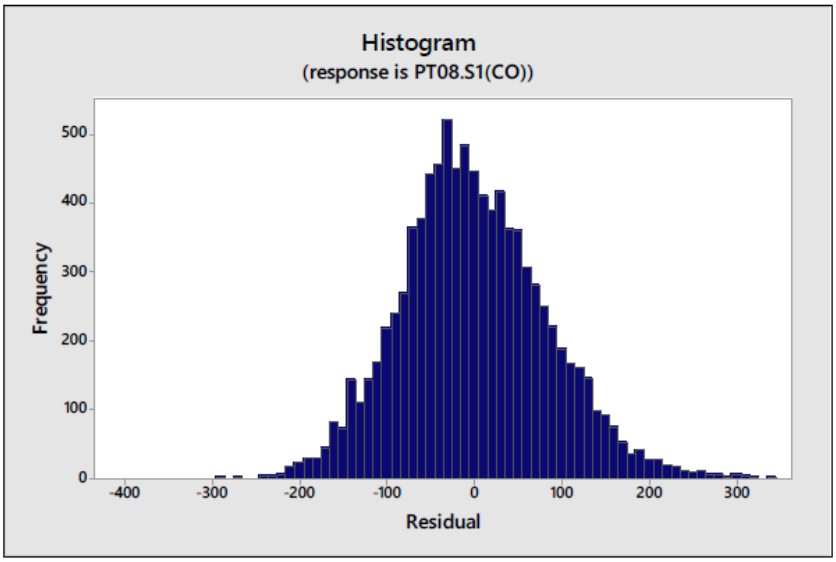

Figure 3: Histrogram of the Residual for Air Quality Data

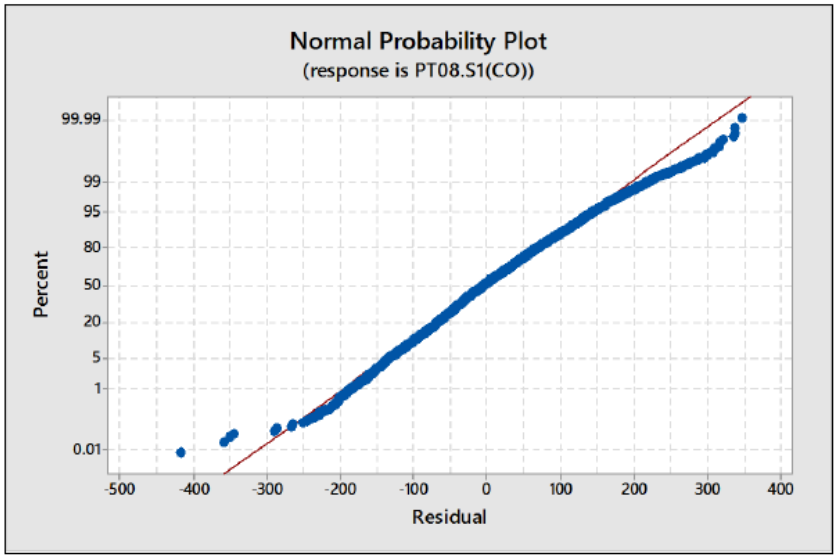

Figure 4: Normal Probability Plot for Air Quality Data

CONCLUSION

In this study, the multiple linear regression analysis is conducted for determining the key factors of affecting air quality through a data set obtained from Saverio De Vito, ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development. We carry out the initial checking of multicollinearity among the variables involved. Those remaining variables then are used to develop the best-fitted air quality predictive model by using stepwise regression. It is observed that hourly averaged sensor response (indium oxide), true hourly averaged concentration CO (in mg/m^3), true hourly averaged overall Non Metanic HydroCarbons concentration (in microg/m^3), true hourly averaged NO_x concentration (in ppb), true hourly averaged NO_2 concentration (in microg/m^3), hourly averaged sensor response (tungsten oxide), and relative humidity (%) are the significant factors that affect the air quality. Perhaps the findings of this study may contribute to predict the air quality in future.

REFERENCES

Ali, T. and Rahman, S. (2024). *Comparative modeling of urban air quality using MLR and machine learning algorithms*. *Environmental Statistics Review*, **34(2)**: 112–124.

Brunekreef, B. (2002). Association between mortality and indicators of traffic-related air pollution: A cohort study. *The Lancet*, **360(9341)**: 1203–1209.

Brunekreef, B. (2001). Estimation of long-term average exposure to outdoor air pollution for a cohort study on mortality. *Journal of Exposure Analysis and Environmental Epidemiology*, **11(6)**: 459–469.

De Vito, S., Massera, E., Piga, M., Martinotto, L., & Di Francesco, D. (2008). Air quality dataset [Data set]. UCI Machine Learning Repository. Retrieved September 8, 2025, from https://archive.ics.uci.edu/dataset/360/air+quality

Environmental Protection Agency (EPA). (n.d.). *Air research: Air sensor toolbox for citizen scientists, researchers and developers*. U.S. Environmental Protection Agency. https://www.epa.gov/air-research

Jiang, R. (2025). *Multiple linear regression in environmental forecasting: Air quality prediction in urban China*. *Theoretical and Natural Science*, **101**: 45–52.

Lim, F.P., Mohamed, I., Daud, N. and Goh, S.L. (2016). Comparison of outlier detection methods in standard 2×2 crossover design. *Sains Malaysiana* **45(3)**: 499-506.

Lim, F.P., Mohamed, I., Ibrahim, A.I.N., Goh, S.L., Mohamed, N.A. and Rahman, A. (2019). Outlier detection in 2 × 2 crossover design using Bayesian framework. *Sains Malaysiana* **48(4)**: 893-899.

Oh, Y.L., Lim, F.P., Chen, C.Y., Ling, W.S.Y. and Loh, Y.F. (2022). Exponentiated Weibull Burr type X distribution's properties and its applications. *Electronic Journal of Applied Statistical Analysis* **15(3)**: 553–573.

Oh, Y.L., Lim, F.P., Chen, C.Y., Ling, W.S.Y. and Loh, Y.F. (2023). A new exponentiated beta Burr type X distribution: model, theory, and applications. *Sains Malaysiana* **52(1)**: 281–294.

Uyanik, G. K., & Guler, N. (2013). A study on multiple linear regression analysis. *Procedia - Social and Behavioral Sciences*, **106**: 234–240.

World Health Organization (WHO). (n.d.). *Air pollution*. https://www.who.int/health-topics/air-pollution

Zsuzsanna, T., & Marian, L. (2012). Multiple regression analysis of performance indicators in the ceramics industry. *Procedia Economics and Finance*, **3**: 509–514.