Modeling Pollutant Measures Associations with the Taiwan Air Quality Index (AQI) using an SLR model

Name: Karthik Kuppala
GitHub: https://github.com/KarthikKuppala/Taiwan-Air-Quality-Index-AQI-Analysis/tree/main

Abstract:
This report investigates the relationship between the Air Quality Index (AQI) and pollutant levels in Taiwan. pm2.5 was identified as the strongest predictor using Pearson correlation. A simple linear regression (SLR) model was constructed to analyze the association, and the model's performance was evaluated. Since the model does not follow the linearity and normality of error assumptions, in future work, we would alter the data to follow all the assumptions and investigate outlier variables.

1. Introduction:

Air Quality Index (AQI) measures how polluted the air is, with higher values indicating worse air quality. Common pollutants influencing AQI include pm2.5, PM10, and CO, all of which can affect respiratory health. The goal of this project is to build a simple linear regression model to predict AQI and identify the most important pollutant among these predictors. Using a dataset with 280 observations from Taiwan.

AQI is influenced by main factors like pm2.5, PM10, and CO. Among these, PM2.5 seems to be of the highest concern due to its ability to penetrate deep into the lungs and enter the bloodstream, causing severe health impacts. PM10 consists of larger particles but still poses health risks by irritating the respiratory tract. Carbon monoxide (CO), a colorless and odorless gas, is another important pollutant, primarily resulting from incomplete combustion, such as vehicle exhaust.

2. Methods and Results:

After conducting a correlation analysis between the training data air quality index variable and the three pollutant measures, PM2.5 was selected as the predictor variable due to its strong Pearson correlation of 0.85 with AQI among the available predictors.

The scatter plot demonstrates a positive linear relationship between PM2.5 and AQI, indicating that higher concentrations of PM2.5 are associated with poorer air quality. This is particularly concerning as PM2.5 particles can penetrate deep into the lungs and enter the bloodstream, leading to serious health risks.

For a SLR to be appropriate, the model must follow four key assumptions: linearity, independence of errors, normality of errors, and the homoscedastic assumption. Based on the standardized residual plot, we can determine the validity of linearity and homoscedasticity assumption by looking at the overall trend and spread of the residuals. Because there is no clear upwards or downwards trend in the standardized residual plot (scattered around 0), the data satisfies the linearity assumption.

However, the data does not satisfy the homoscedasticity assumption since the standardized residual plot has a cone shaped ("<"). There also seems to be an outlier in the data as at least one point has a standardized residual value greater than 3 and needs to be further investigated. The validity of the normality of errors assumption can be seen by looking at a QQPlot and determining if the data follows a 45 degree line. Since the data does not follow the 45-degree line, we can say the normality of errors assumption is not valid. The validity of the independence of errors assumption is tested by seeing if the correlation between the residuals and fits are approximately 0. Since this value is close to zero with a value of -0.0006, we can assume independence.

After conducting a linear regression on AQI and PM2.5, the final linear regression model proposed is:

$$\hat{AQI} = 30.1263 + 1.8520 * x_{pm2.5}$$

We could demonstrate that there is statistical evidence that there is an association between pm2.5 levels and AQI by running a hypothesis test. Our null hypothesis would be that $\beta 1 = 0$ and our alternative hypothesis would be $\beta 1 \neq 0$. From there, we choose a 5% significance level and assume our data follows the normality assumption. Next, we need to determine our test statistics which are based on our SLR R output for pm2.5 is 24.5 (1.8520-0/0.0756). If we find that our p-value is less than the 5% significance level, we will reject the null hypothesis. Since we get a p-value of 2e-16 < 0.05, we reject the null hypothesis, suggesting it is statistically likely that there is an association between pm2.5 levels and AQI.

Based on the SLR regression output, we also get a R^2 value of 0.72. This means that 72% of the total variance in response variable (AQI) is defined by predictor variable pm2.5. Since R^2 is 0.72 and is close to 1, it means that the model fits the data nicely but not perfectly.

While theR^2 value is high, 28% of the variability in AQI is not explained by PM2.5, indicating that other factors (such as PM10, CO, or environmental conditions) might also

influence AQI. We can also see the effective predictiveness of pm2.5 based on the graph above, where it shows the predicted value of the model vs the actual AQI index level.

How can we illustrate that the chosen model outperforms the native model with no predictors (i.e. intercept only model)?

A MAE of 12.56 indicates that the on average model performance is off by 12.56 units.

AQI Range = Max - $Min = 153 - 51= 102

and the average error is 12% of the entire range of AQI; therefore, the model seems to be performing reasonably well.

3. **Model Assumptions:**

The simple linear regression (SLR) model relies on four fundamental assumptions: linearity, independence of errors, normality of errors, and homoscedasticity.

The linearity assumption was evaluated using a scatter plot of PM2.5 versus AQI and a standardized residuals versus fitted values plot. The residuals seem to be randomly distributed about zero with no discernible systematic pattern, and the scatter plot indicates a typically positive linear connection between PM2.5 concentration and AQI. This suggests that there is a reasonable satisfaction of the linearity assumption.

The independence of errors assumption was assessed by assessing the correlation between residuals and fitted values. There is no indication of dependence among the residuals, and the independence assumption can be accepted because the correlation value was extremely near to zero (−0.0006).

The normality of errors assumption was evaluated using a Q–Q plot of the standardized residuals. It appears that the errors do not follow a normal distribution because the residuals significantly depart from the 45-degree reference line, especially in the tails. As a result, the normalcy assumption is broken.

The standardized residuals against fitted values plot was used to assess the homoscedasticity assumption. The residuals reveal a cone-shaped structure, indicating increased variation as the fitted values grow. This shows the presence of heteroscedasticity, meaning that the variance of the errors is not consistent across all levels of PM2.5.In addition, the standardized residual plot suggests the presence of at least one probable outlier, with a standardized residual surpassing an

absolute value of 3. This finding needs more research because it can have a disproportionate impact on the fitted model.

4. Discussion:

The fitted SLR model shows a significant positive correlation between PM2.5 concentration and AQI, implying that higher PM2.5 levels correspond with deteriorating air quality. The projected regression coefficient suggests that, on average, the AQI rises by roughly 1.85 units for each one-unit rise in PM2.5 concentration. The hypothesis test for the slope parameter offered strong statistical support against the null hypothesis, bolstering the conclusion that PM2.5 is an important predictor of AQI.

The model attained an R^2 value of 0.72, suggesting that around 72% of the variation in AQI is attributable to PM2.5 alone. Although this indicates a good model fit, it implies that a significant amount of AQI variability is still unexplained, probably due to additional pollutants, weather conditions, or non-linear effects not represented in the basic linear regression model

5. Limitations

Despite its advantages, the model has drawbacks. Breaches of the normality and homoscedasticity assumptions suggest that the SLR model might not be entirely suitable for inference or prediction without adjustments. The existence of possible outliers indicates that the model's estimates could be influenced by extreme values.

Future research might tackle these shortcomings by transforming the response or predictor variables, examining and possibly eliminating influential outliers, or broadening the analysis to a multiple linear regression model that incorporates other pollutants like PM10 and CO. Including meteorological variables could enhance model performance and yield a more thorough understanding of the elements affecting AQI

6. Conclusion

This study establishes a significant positive correlation between PM2.5 and AQI in Taiwan. While the SLR model provides a strong baseline, the breach of key statistical assumptions and the presence of unexplained variance highlight the complexity of atmospheric modeling. [cite_start]Future research should explore data transformations to address heteroscedasticity, investigate the impact of influential outliers, and expand the framework into a Multiple Linear Regression (MLR) model incorporating PM10, CO, and meteorological variables.

7. References

https://github.com/KarthikKuppala/Taiwan-Air-Quality-Index-AQI-Analysis/tree/main