Alan Wong, 705773980
Doctor Zanontian
Statistics 101A, Winter 2023
March 24th, 2023

Predicting the Hourly Average Sensor Response for Ozone - Final Project

Introduction:

Within our modern world, we have seen many advancements in technology. Such advancements, which include the increased burning of fossil fuels for energy as well as the ever growing amount of people owning cars, have led to major environmental issues, which include a declining air quality. Thus, I hoped to find the opportunity to examine the air quality of the world and understand the contributing chemical compounds in the air that are prevalent in polluted areas.

We will be using the Air Quality Data Set found in the UCI Machine Learning Repository and created by Saverio De Vito of the ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development in 2016. This dataset contains 9358 observations and 15 variables, which include the hourly average sensor response of chemical compounds such as tin oxide and titania. This data was collected within an Italian city at road level from five air quality metal oxide chemical sensors from March 2004 to February 2005. For this project specifically, our possible explanatory variables will be the hourly average sensor responses for non-methane hydrocarbons (PT08.S2.NHMC), total nitrogen oxides (PT08.S3.NOx), and nitrogen dioxide (PT08.S4.NO2). Our response variable will be the hourly average sensor response for ozone (PT08.S5.O3). All of these variables are of the units $\mu g/m^3$. Through these variables, I therefore want to see if the hourly average sensor response of ozone can be predicted through the given explanatory variables stated above.

For this dataset, I chose to omit any values of the hourly average sensor responses for non-methane hydrocarbons, total nitrogen oxides, nitrogen dioxide, and ozone that were -200. This was due to the dataset explicitly stating that missing values of the dataset were tagged with a value of -200. Thus, the dataset that we will be performing our analysis on contains 8991 observations instead of 9358. Furthermore, I am also omitting the columns stating the date, time, the true hourly average concentration of the chemical compounds, the temperature, relative humidity, and absolute humidity since we are not interested in predicting the hourly average sensor response for ozone using those variables.

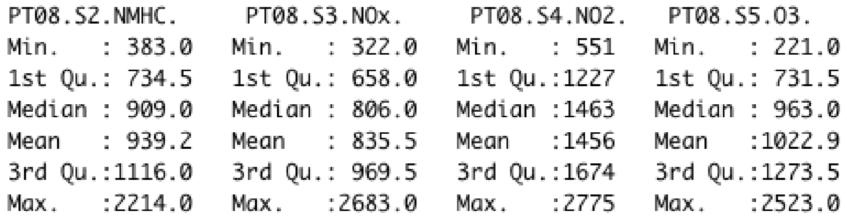
To model this relationship between my given response and explanatory variables, I will be performing multiple linear regression analysis. This is due to the fact that multiple linear regression analysis is performed when trying to find the relationship between a response variable and multiple explanatory variables at hand. By performing multiple linear regression analysis on our data, we are also able to see how statistically significant each of the explanatory variables are in regards to the response variable, thus allowing us to create the optimal regression model that would be appropriate for describing the relationship that the hourly average sensor response of ozone has among the given explanatory variables stated above.

The structure of this paper will be divided into four parts: introduction, data description, results and interpretation, and discussion. Within this current introduction section, my research question and the background and source of the dataset was stated and the justification of the method for modeling my relationship and the overview of the paper's structure was given. Within the section regarding the description of my data, my summary statistics of each of the dataset will be stated and the distribution of each of the variables and the relationships amongst the other variables will be shown through appropriate graphs. Within the section regarding my results and interpretation, I will be reporting my R results in and interpreting them in the context of the given study. I will also be suggesting two other models I have tried alongside my optimal model and will be justifying why the model I chose is the optimal one. Diagnostics will also be run to assess my model. Lastly, within the discussion section, I will be summarizing my project and discussing if my final model makes sense within a real world situation as well as the limitations faced within the analysis.

Data Description:

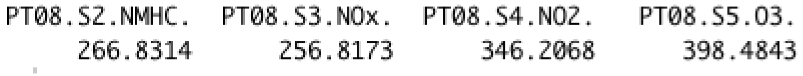
After loading my revised dataset into RStudio, I was able to obtain the summary statistics and the standard deviations of all of the variables that I was analyzing within my dataset.

Summary Statistics:



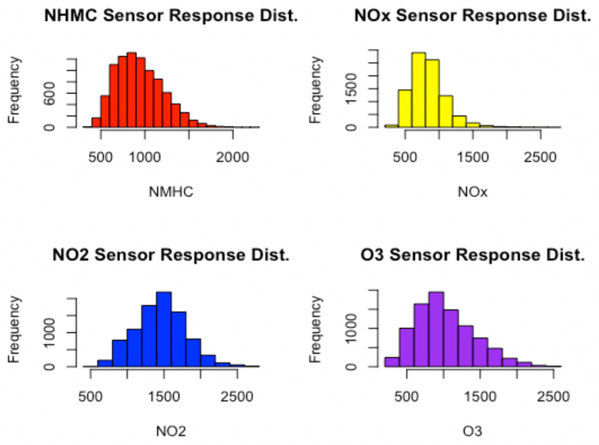
```
 PT08.S2.NMHC.     PT08.S3.NOx.     PT08.S4.NO2.   PT08.S5.O3.
 Min.   : 383.0   Min.   : 322.0   Min.   : 551   Min.   : 221.0
 1st Qu.: 734.5   1st Qu.: 658.0   1st Qu.:1227   1st Qu.: 731.5
 Median : 909.0   Median : 806.0   Median :1463   Median : 963.0
 Mean   : 939.2   Mean   : 835.5   Mean   :1456   Mean   :1022.9
 3rd Qu.:1116.0   3rd Qu.: 969.5   3rd Qu.:1674   3rd Qu.:1273.5
 Max.   :2214.0   Max.   :2683.0   Max.   :2775   Max.   :2523.0
```

Standard Deviation:



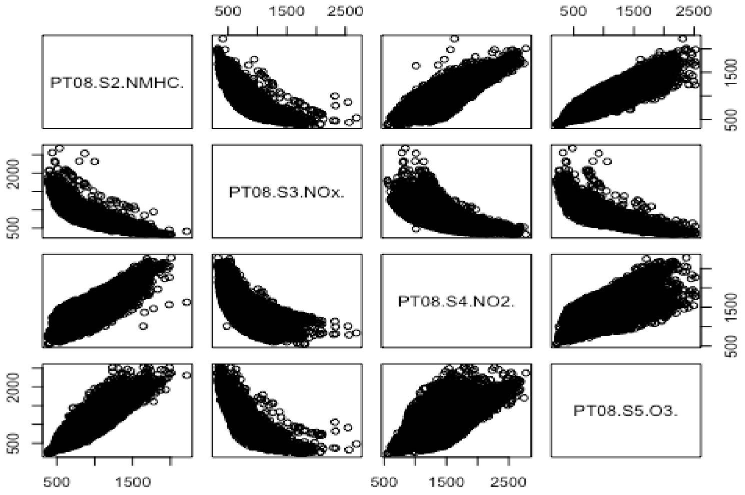
```
PT08.S2.NMHC.  PT08.S3.NOx.  PT08.S4.NO2.   PT08.S5.O3.
     266.8314      256.8173      346.2068      398.4843
```

In order to see the distribution of each of our variables, I also constructed histograms for each of the variables, in which they all seemed to be right-skewed.

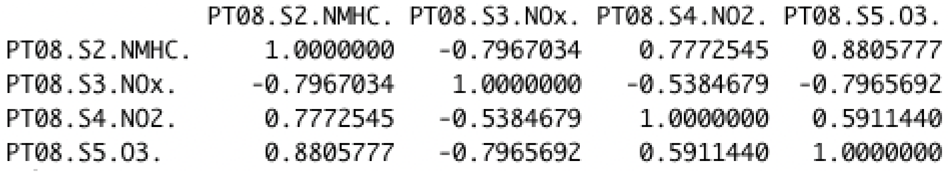


Lastly, I constructed a scatterplot matrix and a correlation matrix to both visualize and understand the relationships that each of the variables had within one another.

Scatterplot Matrix:



Correlation Matrix:

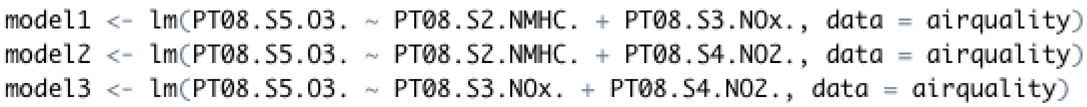


```
               PT08.S2.NMHC. PT08.S3.NOx. PT08.S4.NO2. PT08.S5.O3.
PT08.S2.NMHC.      1.0000000   -0.7967034    0.7772545   0.8805777
PT08.S3.NOx.      -0.7967034    1.0000000   -0.5384679  -0.7965692
PT08.S4.NO2.       0.7772545   -0.5384679    1.0000000   0.5911440
PT08.S5.O3.        0.8805777   -0.7965692    0.5911440   1.0000000
```

Within both these matrices, we see that there are mainly strong linear relationships seen within each of the variables, though there are weaker linear relationships present too based on how much the scatterplots resemble the graphs $y = x$ or $y =- x$. What this means is that depending on this strength, some variables are more strongly correlated and therefore associated with each other compared to other variables. Furthermore, we see that there are also both positive and negative linear relationships present. Essentially, as the value of the explanatory variable at hand increases, the value of the response variable at hand either increases or decreases depending on if the correlation coefficient is positive or negative or if the the graphs resemble $y = x$ or $y =- x$ more.

Results and Interpretation:

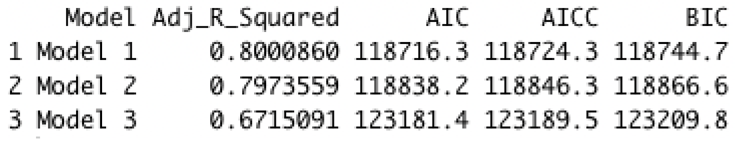
With this better grasp of the data given the summary statistics and the distributions and relationships of each of the variables, let us start creating possible models to determine the optimal linear regression model that can be used to predict the hourly average sensor response for ozone. To do so, we will have to create a total of three models that use all possible combinations that utilize two of the three of our explanatory variables. In our case, our first model will use the hourly average sensor responses for non-methane hydrocarbons and the total nitrogen oxides, the second model will use the hourly average sensor responses for non-methane hydrocarbons and nitrogen dioxide, and the third model will use the hourly average sensor responses for the total nitrogen oxides and nitrogen dioxide.



```
model1 <- lm(PT08.S5.O3. ~ PT08.S2.NMHC. + PT08.S3.NOx., data = airquality)
model2 <- lm(PT08.S5.O3. ~ PT08.S2.NMHC. + PT08.S4.NO2., data = airquality)
model3 <- lm(PT08.S5.O3. ~ PT08.S3.NOx. + PT08.S4.NO2., data = airquality)
```

To measure the goodness of fit for each of these models, we will first find each of the model's values in regards to the adjusted R^2, the Akaike Information Criterion (AIC), the AIC corrected, and the Bayes Information Criteria (BIC).

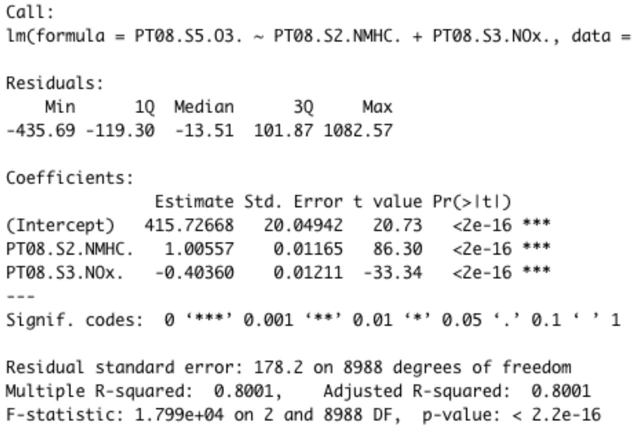
Through calculating each of the model's value in regards to the adjusted R^2, the Akaike Information Criterion (AIC), the AIC corrected, and the Bayes Information Criteria (BIC), I then created a table that stated all these values.



```
     Model Adj_R_Squared      AIC     AICC      BIC
1 Model 1     0.8000860 118716.3 118724.3 118744.7
2 Model 2     0.7973559 118838.2 118846.3 118866.6
3 Model 3     0.6715091 123181.4 123189.5 123209.8
```

In order to find the most optimal model to use for our analysis, we need to find the model that has the highest adjusted R^2 value and the lowest AIC, AICc, and BIC values. When looking at our table, we see that Model 1, which is the model that utilized the hourly average sensor responses for non-methane hydrocarbons and the total nitrogen oxides as our explanatory variables, has the highest adjusted R^2 value of 0.8000860 and the lowest AIC, AICc, and BIC values of 118716.3, 118724.3, and 118744.7, respectively. Therefore, the best model that we can use to predict the hourly average sensor response for ozone is the model that utilized the hourly average sensor responses for non-methane hydrocarbons and the total nitrogen oxides as our explanatory variables.

I then started assessing our model by first producing a summary table of the model.



```
Call:
lm(formula = PT08.S5.O3. ~ PT08.S2.NMHC. + PT08.S3.NOx., data = airquality)

Residuals:
    Min      1Q  Median      3Q     Max 
-435.69 -119.30  -13.51  101.87 1082.57 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   415.72668   20.04942   20.73   <2e-16 ***
PT08.S2.NMHC.   1.00557    0.01165   86.30   <2e-16 ***
PT08.S3.NOx.   -0.40360    0.01211  -33.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 178.2 on 8988 degrees of freedom
Multiple R-squared:  0.8001,	Adjusted R-squared:  0.8001 
F-statistic: 1.799e+04 on 2 and 8988 DF,  p-value: < 2.2e-16
```

Through looking at our summary table, we notice that our F-statistic is 1.799×10^{4} on 2 and 8988 DF, and our corresponding p-value is a value less than 2.2×10^{-16}. Since the F-statistic compares the joint effect of all the variables together and the p-value associated with the F-statistic is less than our significance level of 0.05, this means that we can reject our null hypothesis and instead say that there is at least 1 predictor variable that has a statistically significant relationship with our response variable. Furthermore, when looking at all the individual p-values for our variables, we see that they all have p-values less than our significance level of 0.05, which means that all of our given explanatory variables at hand are statistically significant in predicting values for our response variable.

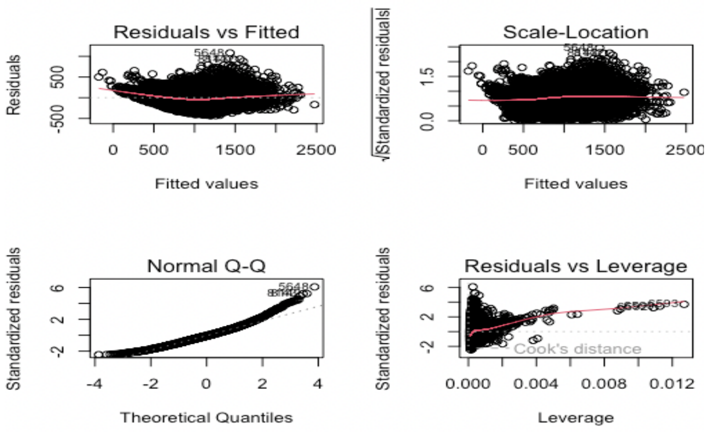
By looking at the estimated values for our intercept and explanatory variables within the summary table, we also see that the equation for our optimal model is:

Predicted PT08.S5.O3 = 415.72668 + 1.00557*PT08.S2.NHMC - 0.40360*PT08.S3.NOx

One should remember that PT08.S2.NHMC represents the hourly average sensor responses for non-methane hydrocarbons, PT08.S3.NOx represents the hourly average sensor responses for the total nitrogen oxides, and PT08.S5.O3 represents the hourly average sensor response for ozone. All of these variables are of the units $\mu g/m^{3}$.

To interpret this model in the context of our study, we can state that the hourly average sensor responses for ozone increases by 1.0057 $\mu g/m^{3}$ for every increase of 1 $\mu g/m^{3}$ in the hourly average sensor responses for non-methane hydrocarbons if the hourly average sensor responses for total nitrogen oxides is fixed. Furthermore, the hourly average sensor responses for ozone decreases by 0.40360 $\mu g/m^{3}$ for every increase of 1 $\mu g/m^{3}$ in the hourly average sensor responses for total nitrogen oxides if the hourly average sensor responses for non-methane hydrocarbons is fixed. 415.72668 $\mu g/m^{3}$ is also the predicted value of the the hourly average sensor responses for ozone if the hourly average sensor responses for non-methane hydrocarbons and total nitrogen oxides are 0 $\mu g/m^{3}$.

I then created four diagnostic plots of our model and interpreted them based on their results.



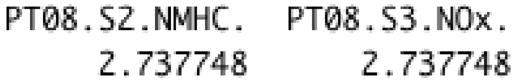
Within the Residuals vs Fitted plot, we see that the spread of the residuals seem to be consistent and a random scatter seems to be present in the graph, which means that the assumption of constant variance seems to be met. Furthermore, the linearity seems to hold relatively well since the red line within the Residuals vs Fitted plot is relatively close to the dashed line representing the residual value of 0. Therefore, there is a relatively linear relationship between the predictor variables and the outcome variable.

Within the Normal Q-Q plot, we see that points are generally aligned to the dashed straight line present in the graph. However, there seems to be a deviation from the normality present in the right tail of the distribution present, as notably seen with the point 5648. All of this means that the residuals are not approximately normally distributed and that the normality of the errors are not considered to be consistent with the normal distribution.

Within the Scale-Location plot, we see that there is barely any curvature present in the red line within the plot. This plot is used to check the assumption of equal variance and shows if residuals are spread equally along the ranges of predictors. Since a horizontal line with equally yet randomly spread points is seen within the graph, we can therefore assume that the assumption of homoscedasticity is met and that the residuals are spread equally along the ranges of predictors.

Within the Residuals vs Leverage plot, we see that most of the observations for the data are contained within the Cook's distance borders of 0.5 given how they are not present within the plot. There are also some standardized residuals that are also seen to not be at or within the values of -2 and 2 and actually greater than 2, which means that there are significant outliers present. There are a few observations with large leverage values on the far right of the plot though, such as points 6592, 6590, and 6593, which may be influential leverage points to look into.

Lastly, we can check to see if our data has any issues with multicollinearity, which can be done by calculating the variance inflation factor (VIF) values for each of our explanatory variables.



```
PT08.S2.NMHC.  PT08.S3.NOx.
     2.737748      2.737748
```

When strong correlations exist among predictor variables, the variance inflation factor (VIF) of a predictor variable is therefore high. We ideally want VIF values that are between 1 to 5 since they signal that a given predictor variable is moderately or not correlated with other predictor variables. Since all of the predictor variables have VIF values that are lower than 5, this means that all the variables are not highly correlated with one another, thereby implying that there are no issues with multicollinearity present in the data based on our chosen model.

Discussion:

Within this project, I ultimately first stated my research question and the background and source of my dataset and justified my method for modeling my relationship and gave an overview of the structure of my paper. I then provided the summary statistics of my variables, created histograms to showcase the distribution of my variables, and constructed scatterplot and correlation matrices to visualize the relationships that each of the variables had with one another. Next, candidate models were created for testing our relationship and their goodness of fit were judged depending on the criteria of the adjusted R^2, the Akaike Information Criterion (AIC), the AIC corrected, and the Bayes Information Criteria (BIC). Through these model, the model with that utilized the hourly average sensor responses for non-methane hydrocarbons and the total nitrogen oxides as the explanatory variables was found to be the optimal model due to it having the highest adjusted R^2 value and the lowest AIC, AICc, and BIC values. A summary table and diagnostic plots were then created to assess the model, and the model was interpreted using the summary table in the context of our given data. Lastly, testing for issues with multicollinearity occurred for this model, where it was ultimately found that the model had no such issues at hand.

In my opinion, my final model makes sense in the real world. Through interpreting my model, it states that both non-methane hydrocarbons and nitrogen oxides contribute to the amount of ozone present at hand. Within the article "Air-Sea Transfer: Dimethyl Sulfide, COS, CS2, NH4, Non-Methane Hydrocarbons, Organo-Halogens" by J.W. Dacey and H.J. Zemmelink within the Encyclopedia of Ocean Sciences (Second Edition), 2001, they state that "non-methane hydrocarbons (NHMCs are important reactive gases in the atmosphere since they … play key roles in the production and destruction of ozone in the troposphere." Furthermore, the United States Environmental Protection has also stated that "NO_2 along with NO_x reacts with other chemicals in the air to form both particulate matter and ozone" within their article "Basic Information about NO2." Thus, my model supports the overall conclusion that non-methane hydrocarbons and nitrogen oxides contribute to the amount of ozone present at hand.

In regards to limitations for this analysis that prevent our analysis from being more accurate, one can notice that this data collected only took place during a limited time period from March 2004 and February 2005. With this dataset being the longest freely available recordings of on field deployed air quality chemical sensor devices, it seems to be quite outdated considering how this data was collected approximately twenty years ago as of the time of this analysis being conducted. Furthermore, our model also notably only takes into account the two explanatory variables of the hourly average sensor responses for non-methane hydrocarbons and the total nitrogen oxides. Within the real world, there are most likely multiple other explanatory variables other than non-methane hydrocarbons and total nitrogen oxides that could have contributed to the hourly average sensor responses for ozone. Given such limitations, total accuracy may not be reached within our optimal model.

In order to further improve this model in the future given the limitations, the hourly average sensor response of chemical compounds present in the air should be collected all over the world rather than just in one specific Italian city. Furthermore, more variables could also be implemented and assessed within our future model and other types of regression techniques besides multiple linear regression could be utilized to see if more accurate results for predicting the hourly average sensor responses for ozone can be found. All in all though, the optimal model created today was still ultimately able to help us find the optimal way to predict the hourly average sensor responses for ozone when given the hourly average sensor responses for non-methane hydrocarbons and total nitrogen oxides specifically.