Sophie Youk (sy5qm)

Professor Chao Du

Statistics 6130

4 May 2020

Air Pollution in Seoul

1. **Background**

I obtained data of *Air Pollution in Seoul* on the website Kaggle (Kim, 2020). Seoul Metropolitan Government (SMG) has collected and provided many public data including air pollution information. There are several stations measuring air pollution in South Korea including Seoul. The air pollution has been very serious issue in many countries, and of course, in South Korea too. I have raised and lived in Seoul for so long, so I have heard air pollution issues a lot of times. I wanted to see how much the pollutants are measured in the city and whether there are any relationships between them and other factors such as time or location.

1. **Summary of Data**

Measurement date: Measurement date and time

Station code: Measuring station code

Address: Address of measuring station

Latitude: Latitude of address

Longitude: Longitude of address

SO2: Sulfur dioxide

NO2: Nitrogen dioxide

O3: Ozone

CO: Carbon monoxide

PM10: Particulate matter

PM2.5: Particulate matter

The 25 mearing stations in Seoul has measured sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), carbon monoxide (CO), particulate matter (PM10), and particulate matter (PM2.5) hourly. The dataset has measurements from 12 AM on 1 January 2017 to 11 PM on 31 December 2019. Every station has its own code (101 to 125), and address, latitude, and longitude indicate where the stations are located. Originally, there were 647,511 observations. However, I used data by 8 hours which were 80,939 observations since I couldn’t perform models with the full dataset in my computer. I found additional information about PM10 and PM2.5 on Australian Government website: PM10 is particulate matter 10 micrometers or less in diameter, PM2.5 is particulate matter 2.5 micrometers or less in diameter. PM2.5 is generally described as fine particles. By way of comparison, a human hair is about 100 micrometers, so roughly 40 fine particles could be placed on its width (Particulate matter (PM10 and PM2.5), 2020).

1. **Research Problem**

I would like to focus on the data of four air pollutants and the relationship with PMs, and further, how they are related to the location (latitude, longitude, address, or stations code) and time (measurement date). Since the air pollution around the world is neither trivial nor temporary issue, researching air pollution data of Seoul can related to similar issues in other places, especially cities.

1. **Model, Method, and Analysis**

**4-1.**

I fitted multivariate regression model with measurement date, station code, latitude, and longitude as responses and SO2, NO2, O3, CO, PM10, and PM2.5 as predictors. To see whether the assumption that relationship is linear is met, I fitted a studentized residual vs. fitted value plot. Since there is a linear (U) pattern and negative association (Figure 1), I decided to use for values of SO2, NO2, O3, and CO. Table 1 shows the coefficients of new fitted model. In table 2, with response of longitude, p-value of SO2 and O3 are greater than the significance level of 0.05. I tested hypothesis of (coefficients of SO2 and O3), and I got a result of table 3. Since all p-values are very small and close to zero, the null hypothesis is rejected and, at least one coefficient of the two is not zero. When I tested and separately, I got similar result. All of the p-values are very small and close to zero, so both and are not zero. What I found interesting with coefficients is only NO2 out of other pollutants and PMs was positively related to latitude. NO2 values were recorded higher in northern counties.

**4-2.**

I fitted another multivariate regression model with PM10 and PM2.5 as responses and SO2, NO2, O3, and CO as predictors. As we can see in the figure 2, it is hard to say that there is a pattern, so I decided to use the model for further methods. As we can see from the Table 4 and table 5, all p-values were very small and close to zero. Hence, all of four pollutants were significant at the level of 0.05. Out of four air pollutants, O3 had p-values relatively greater other three. Interestingly, according to the coefficients, SO2 was negatively related to PM10 and PM2.5 values, which meant when SO2 values were high, values of PM10 and PM2.5 were recorded low.

**4-3.**

To do exploratory data analysis, I visualized the data of four air pollutants (SO2, NO2, O3, and CO). Figure 3, 4, 5, and 6 shows values of each pollutants by PM10 and PM2.5. Generally, the values of PM10 and PM2.5 are not very different. However, interestingly, 51 datasets have very high PM10 values higher than approximately 2,000. As we can see in Table 6, 50 out of 51 stations are 116, 117, or 122. I fitted a linear model with PM10 as response and SO2, NO2, O3, CO, code, and PM2.5. Except the station codes, based on p-values, four air pollutants and PM2.5 value are not significant at the level of 0.05. Since Seoul is a small city, the latitudes and longitudes are not very different, so I didn’t include them when fitting the model and focused on station codes. According to the code information, four stations (116, 117, 121, and 122) are located southern or southwestern part of Seoul. Other than that, date and times are very various, so it is hard to find what made the 51 values of PM10 extremely high.

**4-4.**

I used Mclust function from mclust package to cluster the four air pollutants (SO2, NO2, O3, and CO). As Table 8 tells, there are 9 clusters in the dataset. Then, I performed dimension reduction with and created a cluster plot (Figure 7). Also, I used classError function to see if there were any misclassified variables, but there were not misclassified variables (Table 9). To see average silhouette of dataset under 3, 5, and 7 clusters with K-mean method, I had to created new dataset by a week since R failed running with vector memory exhausted, and then standardized the dataset with scale function. According to 3 features in Figure 8, it seemed the optimal number of clusters was 3. There were also no misclassified variables (Table 10). As Figure 8 indicates, data with 3 clusters were mostly centered zero.

1. **Conclusion and Discussion**

I used several multivariate regression models with geographical data and air pollutant values. When I fitted multivariate regression model with measurement date, station code, latitude, and longitude as responses and SO2, NO2, O3, CO, PM10, and PM2.5 as predictors, with response of longitude, p-value of SO2 and O3 are greater than the significance level of 0.05. I tested hypothesis of (coefficients of SO2 and O3), and since all p-values are very small and close to zero, the null hypothesis is rejected. At least one coefficient of the two is not zero. What I found interesting with coefficients is only NO2 out of other pollutants and PMs was positively related to latitude. NO2 values were recorded higher in northern counties. Additionally, when I fitted another multivariate regression model with PM10 and PM2.5 as responses and SO2, NO2, O3, and CO as predictors, all p-values were very small and close to zero (Table 4 and Table 5). Hence, all of four pollutants were significant at the level of 0.05. Out of four air pollutants, O3 had p-values relatively greater other three. Interestingly, according to the coefficients, SO2 was negatively related to PM10 and PM2.5 values, which meant when SO2 values were high, values of PM10 and PM2.5 were recorded low.

For exploratory data analysis, I visualized the data of four air pollutants (SO2, NO2, O3, and CO). Figure 3, 4, 5, and 6 shows values of each pollutants by PM10 and PM2.5. Generally, the values of PM10 and PM2.5 are not very different, but interestingly, 51 datasets have very high PM10 values higher than approximately 2,000. As we can see in Table 6, 50 out of 51 stations are 116, 117, or 122. I fitted a linear model with PM10 as response and SO2, NO2, O3, CO, code, and PM2.5. Except the station codes, based on p-values, four air pollutants and PM2.5 value are not significant at the level of 0.05. According to the code information, four stations (116, 117, 121, and 122) are located southern or southwestern part of Seoul. Other than that, date and times are very various, so it is hard to find what made the 51 values of PM10 extremely high. As Table 8 shows, there are 9 clusters in the dataset of the four air pollutants (SO2, NO2, O3, and CO). Also, I used classError function to see if there were any misclassified variables, but there were not misclassified variables (Table 9). According to 3 average silhouette of dataset features in Figure 8, it seemed the optimal number of clusters was 3. There were also no misclassified variables (Table 10). As Figure 8 indicates, data with 3 clusters were mostly centered zero.

One county has only one measuring station, but the size, population, number of factories, and etc. are very various and random. Counties with high density of population and factories may have higher air pollutants values than other counties even though they have fewer population and factories in reality. Since there are other factors causing SO2, NO2, O3, and CO and increasing PM10 and PM2.5, more information of predictors/factors can improve analyzing the air pollution data. Also, rain sometimes decreases the air pollution measurements, so it is better to analyze data collected in the same condition of weather.

**Tables and Figures**

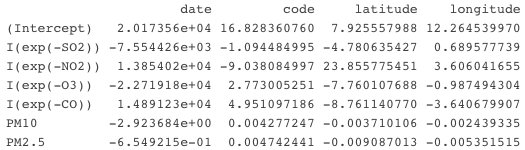


Table 1: Matrix of Estimated Coefficients

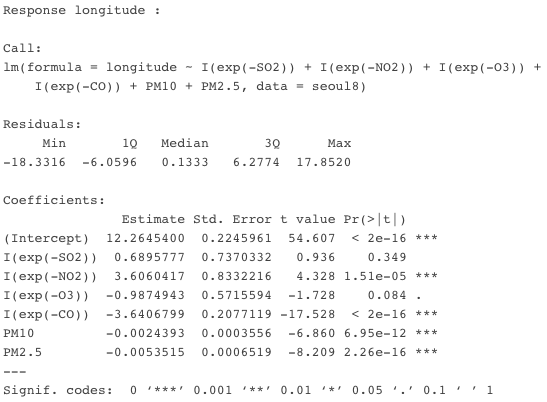


Table 2: Summary of Multivariable Regression Model

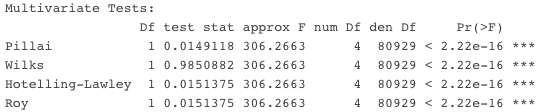


Table 3: Testing Hypothesis (Coefficients of SO2 and O3)

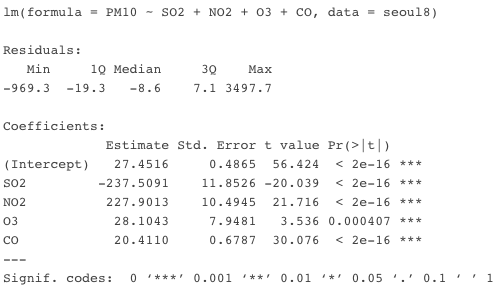


Table 4: Summary of Multivariable Regression Model (PM10)

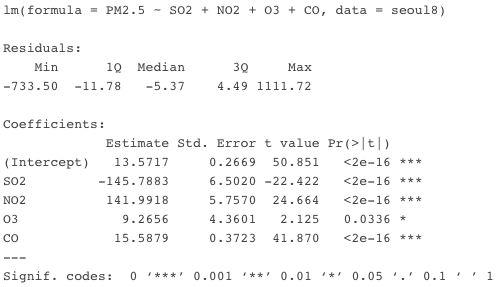


Table 5: Summary of Multivariable Regression Model (PM2.5)

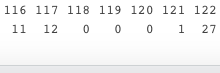


Table 6: 51 Stations with Extreme PM10 Value

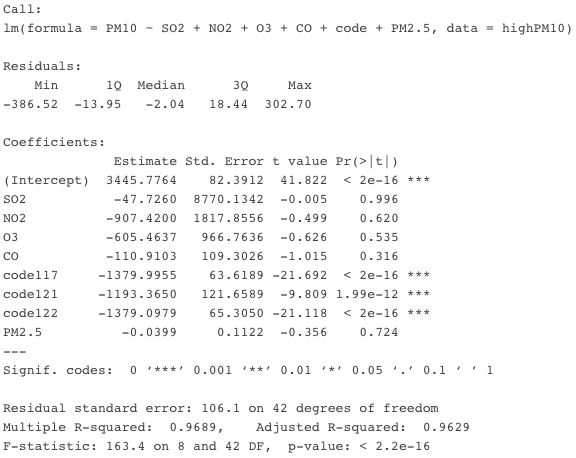


Table 7: Summary of Linear Model

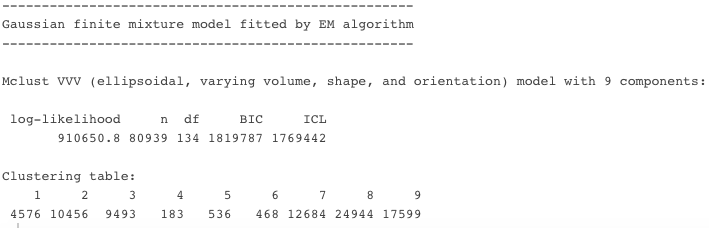


Table 8: Summary of Cluster



Table 9: R Codes 1 to Find Misclassified Variables



Table 10: R Codes 2 to Find Misclassified Variables

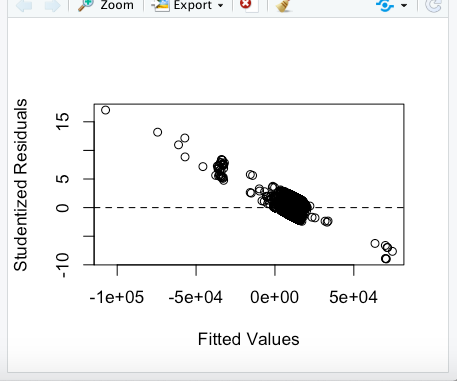


Figure 1: Studentized Residuals vs. Fitted Values 1

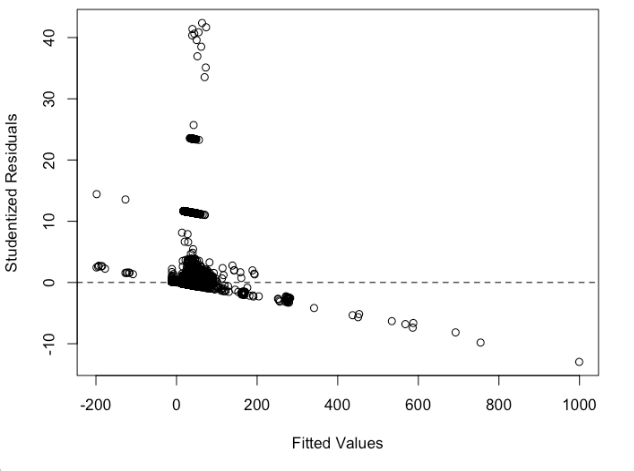


Figure 2: Studentized Residuals vs. Fitted Values 2

Information for Figure 3-6: Red – PM10 & Black – PM2.5

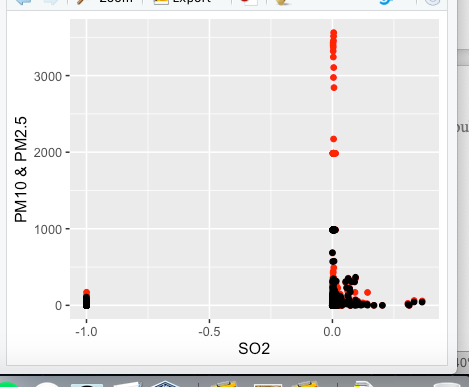
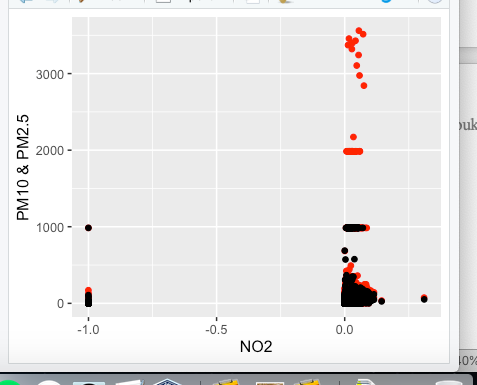
 

Figure 3: SO2 vs. PM10 & PM2.5 Figure 4: NO2 vs. PM10 & PM2.5

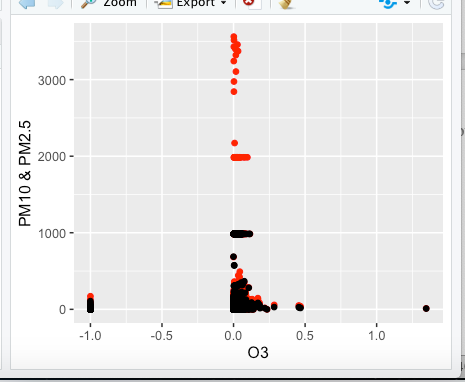
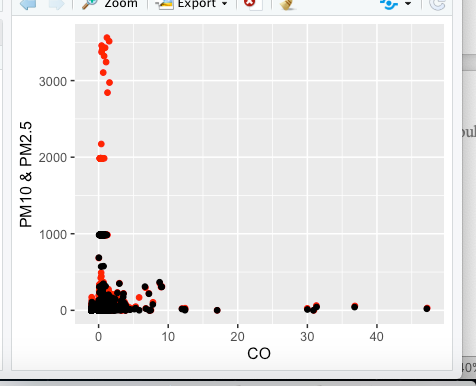
 

Figure 5: O3 vs. PM10 & PM2.5 Figure 6: CO vs. PM10 & PM2.5

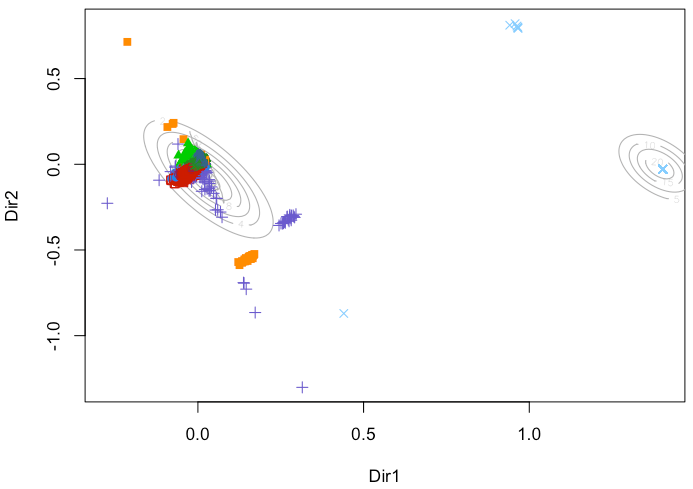


Figure 7: Plot of 9 Clusters

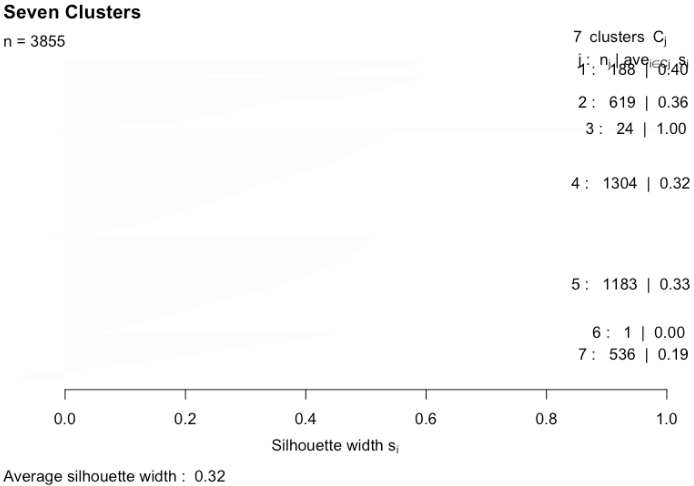
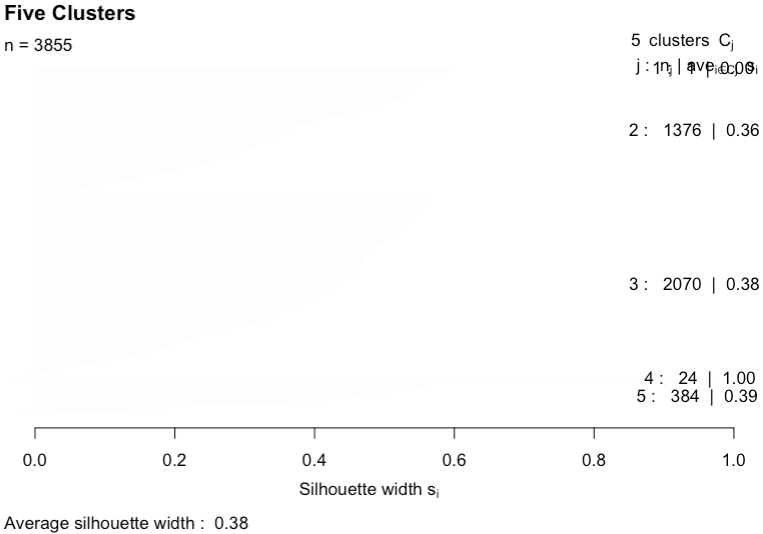
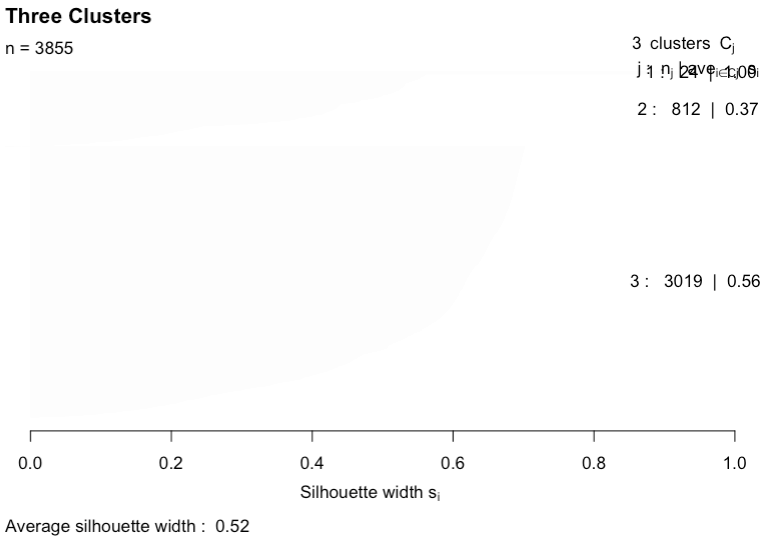


Figure 8: Average Silhouettes under 3, 5, and 7 Clusters

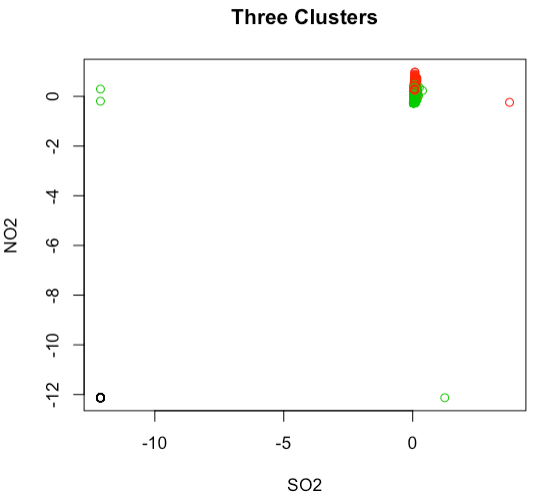


Figure 9: Plot of 3 Clusters

**Appendix**

########## Read and prepare data

seoul <- read.csv("data.csv")

seoul$code <- factor(seoul$code)

seoul$latitude <- factor(seoul$latitude)

seoul$longitude <- factor(seoul$longitude)

seq <- seq(1, 647511, 8)

seoul8 <- seoul[seq,]

########## Multivariate Regression 1

## Multivariate Regression Model

lm1 <- lm(cbind(date,code,latitude,longitude) ~ SO2+NO2+O3+CO+PM10+PM2.5, data=seoul8)

## Residuals vs. Fitted Values

library(MASS)

fitted <- fitted(lm1)

studres <- studres(lm1)

plot(fitted, studres, xlab='Fitted Values', ylab='Studentized Residuals')

abline(h=0, lty=2)

## Multivariate Regression Model

lm2 <- lm(cbind(date,code,latitude,longitude) ~

I(exp(-SO2))+I(exp(-NO2))+I(exp(-O3))+I(exp(-CO))+PM10+PM2.5, data=seoul8)

########## Hypothesis Testing

library(car)

C1 <- matrix(c(0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0), nrow=2, ncol=7)

linearHypothesis(model=lm2, hypothesis.matrix=C1)

C2 <- matrix(c(0, 1, 0, 0, 0, 0, 0), nrow=1, ncol=7)

linearHypothesis(model=lm2, hypothesis.matrix=C2)

C3 <- matrix(c(0, 0, 0, 1, 0, 0, 0), nrow=1, ncol=7)

linearHypothesis(model=lm2, hypothesis.matrix=C3)

########## Multivariate Regression

## Multivariate Regression Model

lm3 <- lm(cbind(PM10,PM2.5) ~ SO2+NO2+O3+CO, data=seoul8)

## Residuals vs. Fitted Values

fitted3 <- fitted(lm3)

studres3 <- studres(lm3)

plot(fitted3, studres3, xlab='Fitted Values', ylab='Studentized Residuals')

abline(h=0, lty=2)

lm3$coefficients

summary(lm3)

########## Exploratory Data Analysis

library(ggplot2)

## Plot SO2

ggplot(seoul8) +

geom\_jitter(aes(SO2,PM10), colour="red") +

geom\_jitter(aes(SO2,PM2.5), colour="black") +

labs(x = "SO2", y = "PM10 & PM2.5")

## Plot NO2

ggplot(seoul8) +

geom\_jitter(aes(NO2,PM10), colour="red") +

geom\_jitter(aes(NO2,PM2.5), colour="black") +

labs(x = "NO2", y = "PM10 & PM2.5")

## Plot O3

ggplot(seoul8) +

geom\_jitter(aes(O3,PM10), colour="red") +

geom\_jitter(aes(O3,PM2.5), colour="black") +

labs(x = "O3", y = "PM10 & PM2.5")

## Plot CO

ggplot(seoul8) +

geom\_jitter(aes(CO,PM10), colour="red") +

geom\_jitter(aes(CO,PM2.5), colour="black") +

labs(x = "CO", y = "PM10 & PM2.5")

## Interestingly high PM10 values

highPM10 <- seoul8[(seoul8$PM10>1500),]

table(highPM10$code)

lm4 <- lm(PM10 ~ SO2+NO2+O3+CO+code+PM2.5, data=highPM10)

summary(lm4)

**References**

Kim, Bappe. *Air Pollution in Seoul*. Kaggle, Mar. 2020, <https://www.kaggle.com/bappekim/air-pollution-in-seoul>. Accessed 3 Apr. 2020.

*Particulate matter (PM10 and PM2.5)*. Department of Agriculture, Water and the Environment. <http://www.npi.gov.au/resource/particulate-matter-pm10-and-pm25>. Accessed 3 Apr. 2020.