STAT 306 Final Project

Analysis of Air Quality and Pollutant Concentrations in Tianjin

Zhe (Peter) Chen
Chun Fang (Frances) Cheng
Jinlin Liu
Shiyue (Joy) Ma
Tianyi (Selena) Shao
April 4, 2017

Abstract

Toxic air has been a controversial topic in China for many years, with heavy smog in its major cities raising global concerns. The goal of this study is to investigate the main pollutants that affect air quality in China by forming a prediction equation for Air Quality Index (AQI) in Tianjin between 2015 and 2016. This project analyzed the relationship between concentrations of six main pollutants CO, NO_2, O_3, SO_2, $PM_{2.5}$ and PM_{10} and the overall AQI level. We have formulated a new prediction equation for the AQI values at a Tianjin monitor station using concentrations of NO_2, $PM_{2.5}$ and PM_{10} as the optimal number of predictor information, hence we proposed that these molecules are the major causes for the "Wumai smog" phenomenon in Tianjin.

Description of data

AQI is a numerical index commonly used to report the severity of air pollution in urban areas and to predict air quality trend (Jassim & Coskuner, 2017), with larger values representing less desirable air quality. According to the United States Environmental Protection Agency, the major pollutants that affect air quality are ground-level ozone, fine and coarse particulate matters, carbon monoxide, sulfur dioxide, and nitrogen dioxide.

In this study, AQI observations are collected from 1497 ground monitoring stations throughout China. These stations measured AQI as well as levels of six main pollutants (CO, NO_2, O_3, SO_2, $PM_{2.5}$ and PM_{10}) on an hourly basis. Our model is fitted based on the data collected at station 1017A (which is located at Tianjin) at midnight, for an average of the past 24 hours, every single day between 2015 and 2016. However, there are 33 days in which measurements for all factor levels are missing for uncontrollable reasons.

Table 1. Table of variables related to air quality in Tianjin

Variables	Explanation (unit)
AQI	Air Quality Index
CO	Concentration of carbon monoxide (parts per hundred million, pphm)
NO_2	Concentration of nitrogen dioxide (parts per hundred million, pphm)
O_3	Concentration of ozone (parts per hundred million, pphm)
SO_2	Concentration of sulphur dioxide (parts per hundred million, pphm)
$PM_{2.5}$	Fine particulate matters with a diameter of less than 2.5μm (μg/m3 micrograms per cubic metre, μg/m3)
PM_{10}	Coarse particulate matters with a diameter of between 2.5μm and 10μm (μg/m3 micrograms per cubic metre, μg/m3)

Data analysis and results

a. Statistics summary

The following is a table of the summary of the data set:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
AQI	0.0	55.0	77.5	102.9	127.2	500.0
CO_24h	0.000	0.780	1.100	1.364	1.730	7.600
NO_2_24h	0.00	28.00	40.00	44.44	57.25	155.0
O_3_24h	0.00	47.00	78.50	97.77	139.00	328.00
PM_{10}_24h	0.00	63.0	92.5	114.2	143.0	571.0
$PM_{2.5}$_24h	0.00	34.00	56.00	69.49	86.00	330.00
SO_2_24h	0.00	10.00	17.00	26.83	32.00	226.00

Where AQI is the response variable, and the others are explanatory variables.
The following is the correlation matrix. Notice that $PM_{2.5}$ and PM_{10} is strongly correlated (with 0.89 correlation). The explanatory variable that has the least correlation with AQI is O_3 with a correlation of -0.168 only.

	AQI	CO	NO_2	O_3	PM_{10}	$PM_{2.5}$	SO_2
AQI	1	0.5613	0.5974	-0.1677	0.7187	0.7255	0.4726
CO	0.5613	1	0.7563	-0.2756	0.6591	0.7230	0.5519
NO_2	0.5974	0.7563	1	-0.2936	0.6707	0.7494	0.6307
O_3	-0.1677	-0.2756	-0.2936	1	-0.1620	-0.1317	-0.3776
PM_{10}	0.7187	0.6591	0.6707	-0.1620	1	0.8899	0.5643
$PM_{2.5}$	0.7255	0.7230	0.7494	-0.1317	0.8899	1	0.5188
SO_2	0.4726	0.5519	0.6307	-0.3776	0.5643	0.5188	1

We first perform a linear regression on the row data set with all the explanatory variables included.

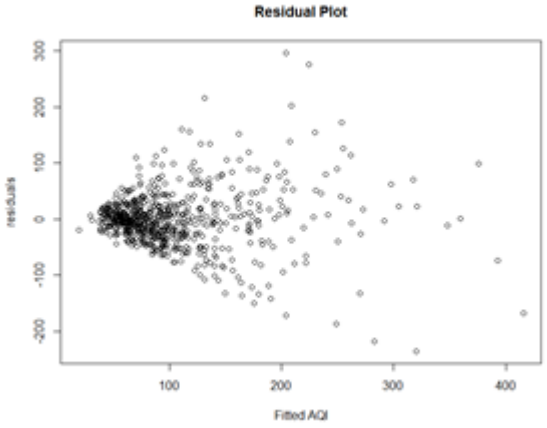


Figure1: Residual plot (fitted AQI vs residuals)

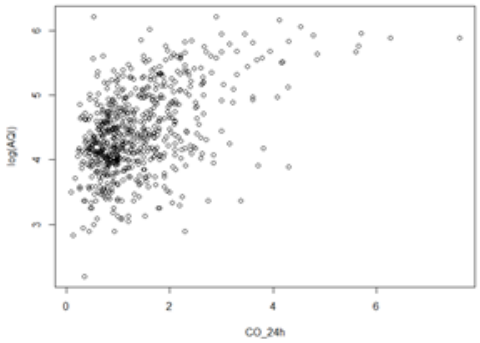
Notice that in Figure 1, the points in the residuals are scattered more outward from 0 as the fitted AQI increases, indicating increasing variance of the residuals of this linear model. This is heteroscedastic.

However, heteroscedasticity is not a desired trait. Thus we need to perform some transformation on the data in order to make it homoscedastic (i.e. constant variance of the residuals).
We will do a log transformation first. Yet, notice that we have AQI = 0 in our data set. A little close inspection reveals that the point has values equal to 0 for all the explanatory and the response variables. This is very unrealistic in the true world, and hence likely to be a sample error. We will delete this point from our data set. After the deletion, all other points have AQI > 0 and thus we can perform log transformation on AQI.

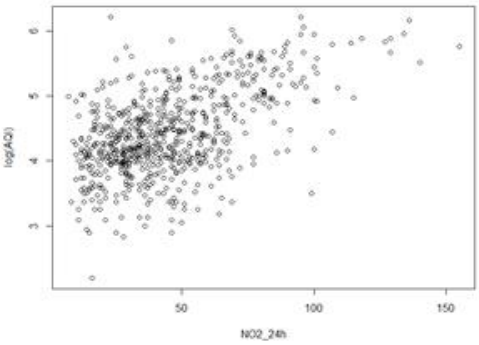
Now we need to check the plot of log(AQI) versus other explanatory variables to verify we have roughly a linear trend.

Figure 2: plot of log(AQI) VS explanatory variables

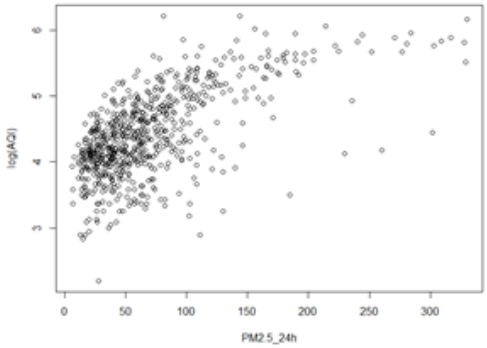
(a) log(AQI) VS CO_24h



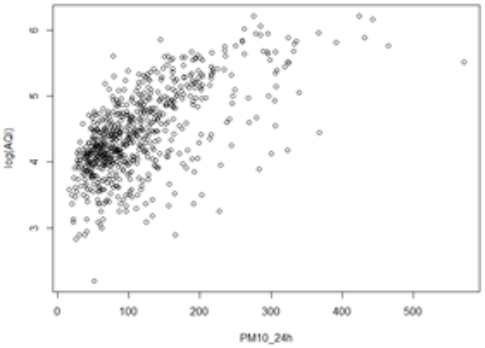
(b) log(AQI) VS NO_2_24h



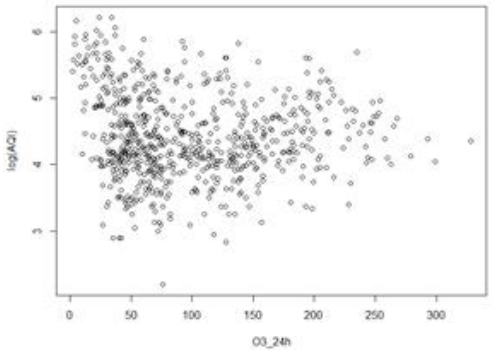
(c) log(AQI) VS $PM_{2.5}$_24h



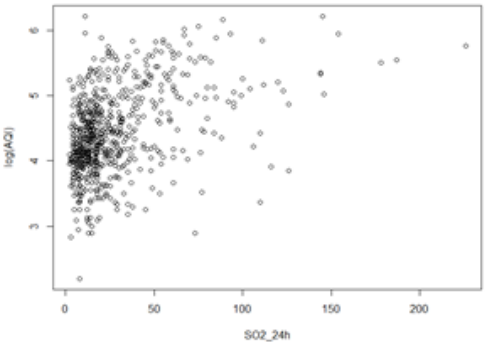
(d) log(AQI) VS PM_{10}_24h



(e) log(AQI) VS O_3_24h



(f) log(AQI) VS SO_2_24h

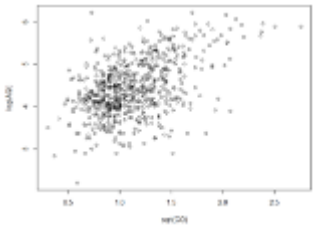


From Figure 2, we can still see some non-linear pattern.
CO: there seems to be a quadratic relation, thus we will try to square root it.
PM_{10}: exponential trend, need log.
$PM_{2.5}$: exponential trend, need log.
O_3: it seems like the points scattered less for large O3. Square root will be nice but if we plot it, we can see that the power of 0.3 will give a better spread out plot.
SO_2: exponential trend, need log.
NO_2: pretty linear trend, do not need any transformation.

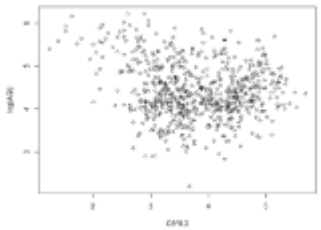
We will perform the above transformations on the explanatory variables and again plot response versus the 6 explanatory variables separately.

Figure 3: log(AQI) vs transformed explanatory variables

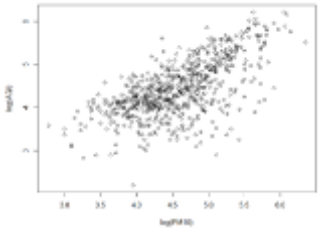
(a) log(AQI) VS sqrt(CO_24h)



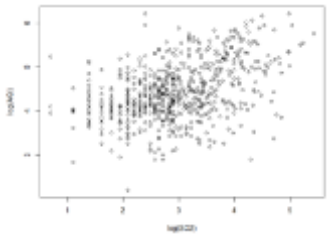
(b) log(AQI) VS (O_3_24h)^0.3



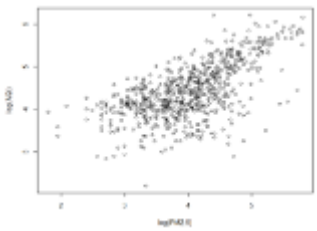
(c) log(AQI) VS log(PM_{10}_24h)



(d) log(AQI) VS log(SO_2_24h)



(e) log(AQI) VS log($PM_{2.5}$_24h)



These plots show roughly a linear relationship (no other apparent relationships are demonstrated in the plots). In all of the residual plots shown above, the points are randomly scattered, with approximately the same deviance from 0. This indicates homoscedasticity.
We will then try linear regression on log(AQI) with other transformed explanatory variables.

The following is a summary of the model:

Residuals:

Min	1Q	Median	3Q	Max
-1.83900	-0.24055	0.05821	0.30357	1.20001

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.809471	0.186410	9.707	< 2e-16 ***
sqrt(X306$CO_24h)	-0.006562	0.084316	-0.078	0.9380
X306$$NO_2$$_24h	0.006176	0.001379	4.479	8.76e-06 ***
new$$O_3$$	-0.017481	0.026468	-0.660	0.5092
log(X306$$SO_2$$_24h)	-0.027202	0.031282	-0.870	0.3848
log(X306$$PM_{10}$$_24h)	0.453176	0.065526	6.916	1.06e-11 ***
log(X306$$PM_{2.5}$$_24h)	0.106053	0.060750	1.746	0.0813 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.471 on 688 degrees of freedom
Multiple R-squared: 0.4632, Adjusted R-squared: 0.4585
F-statistic: 98.94 on 6 and 688 DF, p-value: < 2.2e-16

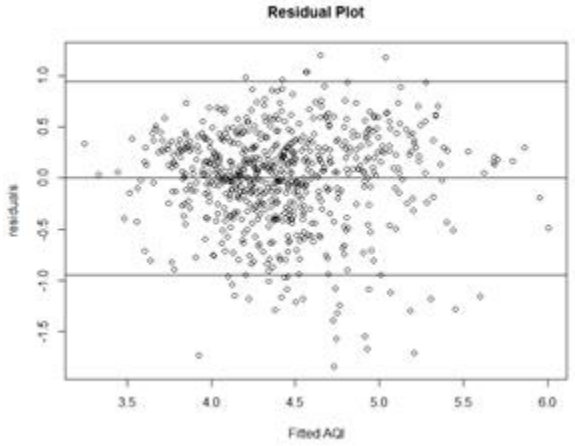
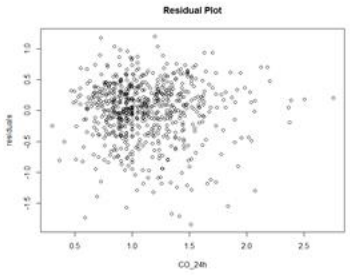


Figure 4: Residual plot of the transformed variables

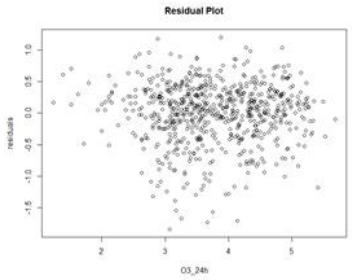
This model is scattered across the fitted AQI without any noticeable trend. And since most data points are within the sigma lines, we can conclude that the residuals are fairly random.

Figure 5: Residual plots of the transformed explanatory variables

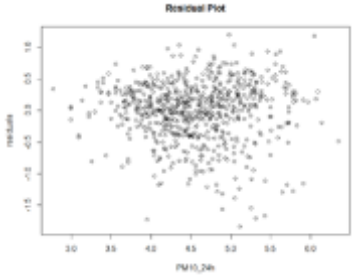
(a) residuals VS sqrt(CO_24h)



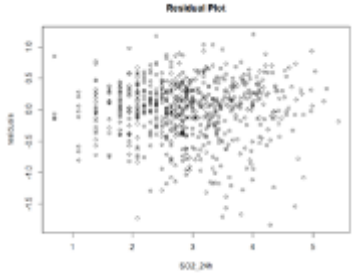
(b) residuals VS (O_3_24h)^0.3



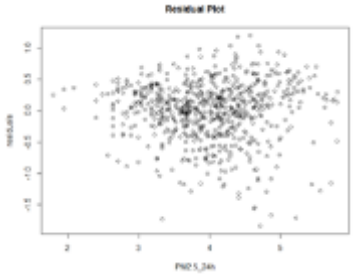
(c) residuals VS log(PM_{10}_24h)



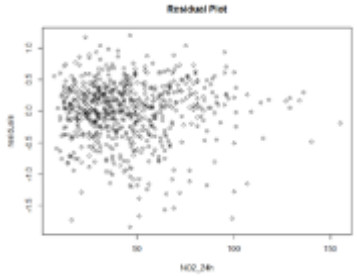
(d) residuals VS log(SO_2_24h)



(e) residuals VS log($PM_{2.5}$_24h)



(f) residuals VS NO_2_24hr



From Figure5, we can see that the points are random with no obvious trend or increasing variance (i.e. heteroscedastic). So our model is pretty good with no further transformation necessary.

From the Durbin-Watson test, our model gives a DW test statistics of 2.186 and a p-value of 0.9895. This agrees with our assumption that the residuals are independent.
Then, we select the relevant explanatory variables via variable selection from exhaustive method.

Below is a summary of the variable selection:

Six Variables (and intercept)

	Forced in	Forced out
sqrt(X306$CO_24h)	FALSE	FALSE
X306$$NO_2$_24h	FALSE	FALSE
newO_3	FALSE	FALSE
log(X306$$SO_2$_24h)	FALSE	FALSE
log(X306$ PM_{10}_24h)	FALSE	FALSE
log(X306$$PM_{2.5}$_24h)	FALSE	FALSE

1 subsets of each size up to 6

Selection Algorithm: exhaustive

	sqrt(X306$CO_24h)	X306$$NO_2$_24h	newO_3	log(X306$$SO_2$_24h)
1 (1)	" "	" "	" "	" "
2 (1)	" "	"*"	" "	" "
3 (1)	" "	"*"	" "	" "
4 (1)	" "	"*"	" "	"*"
5 (1)	" "	"*"	"*"	"*"
6 (1)	"*"	"*"	"*"	"*"

	log(X306$$PM_{10}$_24h)	log(X306$$PM_{2.5}$_24h)
1 (1)	"*"	" "
2 (1)	"*"	" "
3 (1)	"*"	"*"
4 (1)	"*"	"*"
5 (1)	"*"	"*"
6 (1)	"*"	"*"

Following is the corresponding Cps:

48.219878	3.261338	1.969188	3.436589	5.006057	7.000000

And the corresponding adjR^2:

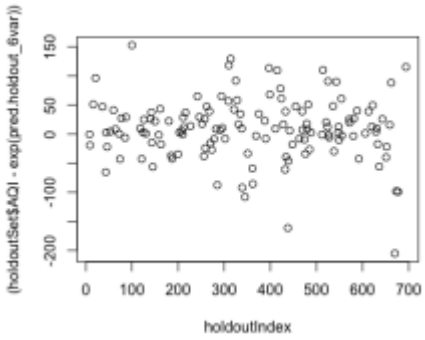
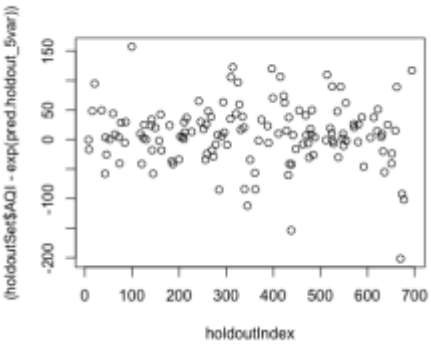
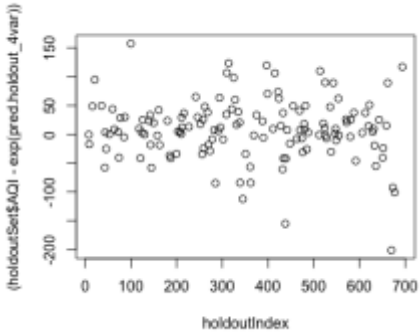
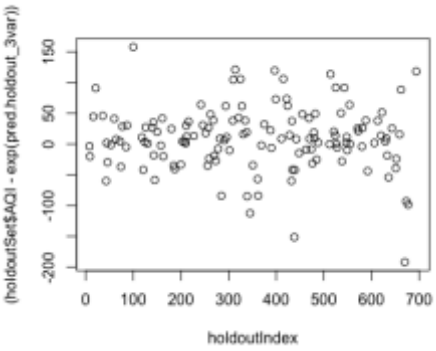
0.4223845	0.4582955	0.4600914	0.4597269	0.4592812	0.4585000

Based on the Cp and adjR^2 values, we will compare the cases with 3,4,5,6 variables since their Cp and adjR^2 values are really similar.

b. Training/holdout

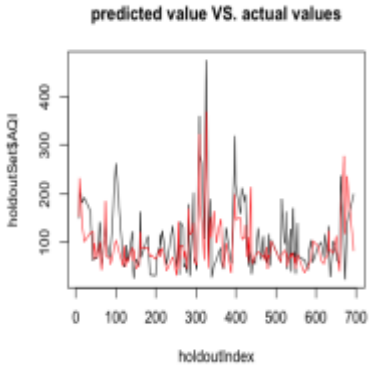
The data set is divided randomly into two parts; four fifths of the data have been selected to be the training set and the remaining one fifth of the data becomes the holdout set. As both C_p and adjusted R^2 of subsets with three, four, five, and six variables generate close results through the variable selection process, the fitting linear model will be applied to each case in this section. As a result, the regression of each case will be found and then the result of each prediction will be compared with actual values in order to choose the best fitting model.

Figure 6: Cross-validation of four regressions: (1) CO, NO_2, O_3, SO_2, $PM_{2.5}$ and PM_{10} (2) NO_2, O_3, SO_2, $PM_{2.5}$ and PM_{10} (3) NO_2, SO_2, $PM_{2.5}$ and PM_{10} (4) NO_2, $PM_{2.5}$ and PM_{10}. Assessment of prediction error was applied in four cases with training and holdout sets.

(1) Prediction on six variables	***(2) Prediction on five variables***
	
(3) Prediction on four variables	***(4) Prediction on three variables***
	

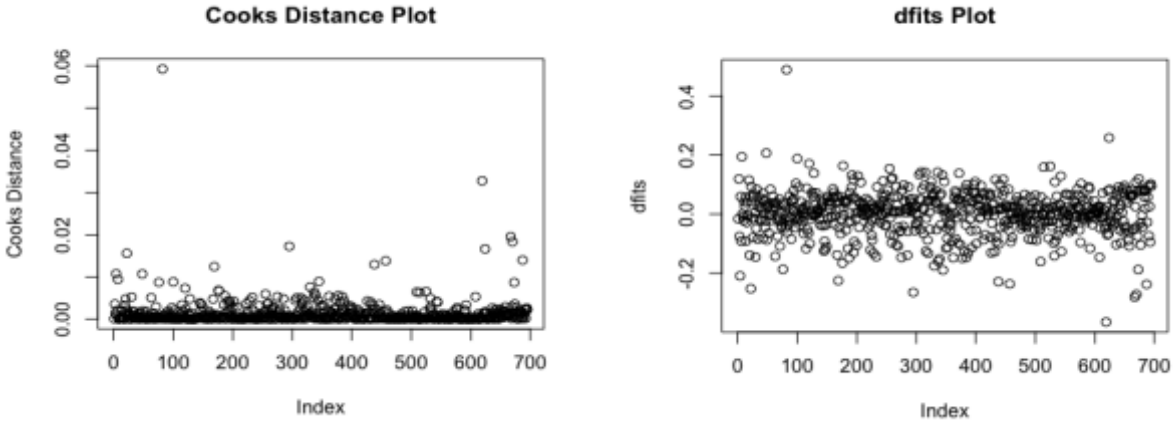
In terms of prediction on six variables (*CO, NO_2, O_3, SO_2, $PM_{2.5}$ and PM_{10}*), the prediction error on the holdout set is 53.18356. Comparatively, the corresponding prediction error of holdout set for three variables (*NO_2, $PM_{2.5}$ and PM_{10}*) decreased to 52.23173. Regarding the predictions on four variables (*NO_2, SO_2, $PM_{2.5}$ and PM_{10}*) and five variables (*NO_2, O_3, SO_2, $PM_{2.5}$ and PM_{10}*), the prediction errors are 52.48542 and 52.42269 respectively, which are pretty close to each other. The ideal regression is the three variable model which has the smallest prediction error. In Figure 7 which is based on the prediction on three variables, note that the predicted value (shown in red) has a similar trend and range of the actual value (shown in black), and this pattern in turns confirms that the regression on three variables is a good model.

Figure 7: Predicted value versus actual values



From Figure 8, both cook's distance plot and dfits plot show that the majority of the data points concentrate on the zero line, which result in low value both in cook's distance and dfits. This result also implies that there are no significant outliers or typos.

Figure 8: Cook's distance plot and dfits Plot



Brief discussion:
The values of betas are in relative small scale (10^{-3}~10^{-1}) from the summary of our model. In order to make them more interpretable, we can consider to scale up some variables in order to make the coefficients in the range of (1, 10).

Based on the definition of AQI from Wikipedia, 'an individual score (IAQI) is assigned to the level of each pollutant and the final AQI is the highest of those 6 scores'. Fitting a linear regression may not be quite useful. Since y=max{score(xi)}=max{Gi(xi)}, where Gi stands for the function converting the concentration of each explanatory variable to the 'score' determine AQI levels, we can do least square method and find the best coefficients in function Gi.

In terms of the quadratic term we used in our model above, Mousavi and Mohammadzadeh indicated that square root of the O_3 levels follow a normal distribution(2013), so using sqrt(O_3_24h) would be more reasonable as this can avoid any outlier in our explanatory variable. However, (O_3_24h)^0.3 fits our data best. Therefore, we hypothesize that sqrt(O_3_24h) may need to interact with some other terms to produce the best result. Further analysis would be involved with adding interactive terms with sqrt(O_3_24h).

Additionally, we only used the data from 2015 to 2016, which may not be sufficient to determine which explanatory variables impose the most significant effect on AQI. Especially, the air pollution in Tianjin is deteriorating in recent years. Also, the data were collected at station 1017A, which may not be representative to show the overall AQI in Tianjin. Therefore, we can improve our case study to be more persuasive by selecting data from recent 10 years and collected from 20 to 50 stations all around Tianjin.

In conclusion, we have found a good-fitting model; the residual plots look random which implies all assumptions seem reasonable, and our model does not show any heteroscedasticity and curvilinear form.

Citations

Berman, L. (2017). National AQI Stations. Harvard Dataverse. V3. doi:10.7910/DVN/NS0RPJ

Jassim, S. M., & Coskuner, G. (2017). Assessment of spatial variations of particulate matter (PM10 and PM 2.5) in Bahrain identified by air quality index (AQI). *Arabian Journal of Geosciences*. 10: 19. doi:10.1007/s12517-016-2808-9

Mousavi, S.S., & Mohammadzadeh, M. (2013). Determination of Spatial-Temporal Correlation Structure of Troposphere Ozone Data in Tehran City. *Journal of Sciences, Islamic Republic of Iran*, 24(2): 171-178. https://jsciences.ut.ac.ir/article_32084_30b3335bfabacb94ab13a127ce88498d.pdf

Zheng, M., Salmon, G. L., Schauer, J. J., Zeng, L., Kiang, C.S., Zhang, Y. H., & Cass, R. G. (2005). Seasonal trends in PM2.5 source contributions in Beijing, China. *Atmospheric Environment*, 39, 3967 - 3976. http://dx.doi.org.ezproxy.library.ubc.ca/10.1016/j.atmosenv.2005.03.036

Contributions

This project is done by Peter Chen, Frances Cheng, Yvonne Liu, Joy Ma, and Selena Shao, who all share a common interest in statistics and are concerned about the deteriorating air quality in China. The project was proposed by the team leader Joy who found the data set, and all members participated in trimming it down to the specific sections used in regression analysis. Analysis and coding was performed by Joy and Yvonne, while Frances provided the model we used in our prediction. The report was written by Peter, Frances, and Selena. All members actively participated in meetings and the critical feedback process, hence the author names are placed in alphabetical order.

Appendix

```
setwd("/Users/Joy/Desktop/STAT306_lab9/proj")

library(readxl)
library(leaps)
library(lmtest)

X306 <- read_excel("~/Desktop/STAT306_lab9/proj/306.xlsx")

## Requirement 1 (Summary of each variable and their corr)
summary(X306$AQI)
summary(X306$CO_24h)
summary(X306$NO2_24h)
summary(X306$O3_24h)
```

```
summary(X306$PM10_24h)
summary(X306$PM2.5_24h)
summary(X306$SO2_24h)
#correlation matrix
d<-data.frame(X306$AQI, X306$CO_24h, X306$NO2_24h, X306$O3_24h, X306$PM10_24h,
X306$PM2.5_24h, X306$SO2_24h)
cor(d)
#plot response variable VS. explanatory variables
plot(X306$AQI~X306$CO_24h)
plot(X306$AQI~X306$NO2_24h)
plot(X306$AQI~X306$O3_24h)
plot(X306$AQI~X306$PM10_24h)
plot(X306$AQI~X306$PM2.5_24h)
plot(X306$AQI~X306$SO2_24h)
## END Requirement 1 (Summary of each variable and their corr)

## Requirement 2 (Create a model)
X306 <- X306[-478,] #delete the row with AQI = 0
newO3 <- X306$O3_24h^0.3
fit6 <-
lm(log(X306$AQI)~sqrt(X306$CO_24h)+X306$NO2_24h+newO3+log(X306$SO2_24h)+log(X306$P
M10_24h)+log(X306$PM2.5_24h), data = X306)
summary(fit6)

#plot transformed response variable VS. transformed explanatory variables
plot(log(X306$AQI)~sqrt(X306$CO_24h))
plot(log(X306$AQI)~newO3)
plot(log(X306$AQI)~log(X306$PM2.5_24h))
plot(log(X306$AQI)~log(X306$SO2_24h))
plot(log(X306$AQI)~log(X306$PM10_24h))
#conclude that after transformation, the linear assumption is valid here
## END Requirement 2 (Create a model)

## Requirement 3 (Print out the residual)
#plot residuals vs fitted response variable
sigma <- summary(fit6)$sigma
plot(fit6$fitted.values, fit6$residuals, main = "Residual Plot", ylab = "residuals", xlab = "Fitted AQI")
abline(h=2*sigma); abline(h=-2*sigma); abline(h=0)

plot(sqrt(X306$CO_24h), fit6$residuals, main = "Residual Plot", ylab = "residuals", xlab = "CO_24h")
plot(X306$NO2_24h, fit6$residuals, main = "Residual Plot", ylab = "residuals", xlab = "NO2_24h")
plot(newO3, fit6$residuals, main = "Residual Plot", ylab = "residuals", xlab = "O3_24h")
plot(log(X306$PM10_24h), fit6$residuals, main = "Residual Plot", ylab = "residuals", xlab =
"PM10_24h")
```

```
plot(log(X306$PM2.5_24h), fit6$residuals, main = "Residual Plot", ylab = "residuals", xlab =
"PM2.5_24h")
plot(log(X306$SO2_24h), fit6$residuals, main = "Residual Plot", ylab = "residuals", xlab = "SO2_24h")
# Perform DW test
dwtest(log(X306$AQI)~sqrt(X306$CO_24h)+X306$NO2_24h+newO3+log(X306$SO2_24h)+log(X306
$PM10_24h)+log(X306$PM2.5_24h), data = X306)
#Conclude that the assumption of residuals are independent is valid here based on DW
## END Requirement 3 (Print out the residual)


# Requirement 4 (Use variable selection)
out.exh6<-
regsubsets(log(X306$AQI)~sqrt(X306$CO_24h)+X306$NO2_24h+newO3+log(X306$SO2_24h)+log(X
306$PM10_24h)+log(X306$PM2.5_24h), data = X306, nbest=1, nvmax=6)
summ.exh6<-summary(out.exh6)
cat("Cp and adjr\n")
print(summ.exh6$cp)
print(summ.exh6$adjr)
# Based on Cp and adjR^2, we decide to compare 3, 4, 5, 6 variables since their Cp and adjR^2 are really
close
# END Requirement 4 (Use variable selection)

# Requirement 5 ( Set training and hold-out set.)
## Make Train and hold out set
n <- nrow(X306)
set.seed(1)
# here select 1/5 of the data randomly to be the hold out set.
holdoutIndex <- sort(sample(1:n, round(n/5), replace = FALSE))

# Initially subset1 is the training set and subset2 is the hold-out set
holdoutSet <- X306[holdoutIndex,]
trainingSet <- X306[-holdoutIndex,]

#based on the exhaustive selection in requirement 4, fit the model with training set and predict the hold-
out set and then compare the predictions with actual values
model.trainSet_6var <-
lm(log(AQI)~sqrt(CO_24h)+NO2_24h+I(O3_24h^0.3)+log(SO2_24h)+log(PM10_24h)+log(PM2.5_24h
), data = trainingSet)
model.trainSet_3var <- lm(log(AQI)~NO2_24h+log(PM10_24h)+log(PM2.5_24h), data = trainingSet)
model.trainSet_4var <- lm(log(AQI)~NO2_24h+log(SO2_24h)+log(PM10_24h)+log(PM2.5_24h), data
= trainingSet)
model.trainSet_5var <-
lm(log(AQI)~NO2_24h+I(O3_24h^0.3)+log(SO2_24h)+log(PM10_24h)+log(PM2.5_24h), data =
trainingSet)
```

```
# Make predictions at the each hold-out data set.
pred.holdout_6var <- predict(model.trainSet_6var, holdoutSet)
holdout.err_6var <- sqrt(sum((holdoutSet$AQI - exp(pred.holdout_6var))^2)/length(pred.holdout_6var))
plot(holdoutIndex, (holdoutSet$AQI - exp(pred.holdout_6var)))
print(holdout.err_6var)

pred.holdout_3var <- predict(model.trainSet_3var, holdoutSet)
holdout.err_3var <- sqrt(sum((holdoutSet$AQI - exp(pred.holdout_3var))^2)/length(pred.holdout_3var))
plot(holdoutIndex, (holdoutSet$AQI - exp(pred.holdout_3var)))
print(holdout.err_3var)

pred.holdout_4var <- predict(model.trainSet_4var, holdoutSet)
holdout.err_4var <- sqrt(sum((holdoutSet$AQI - exp(pred.holdout_4var))^2)/length(pred.holdout_4var))
plot(holdoutIndex, (holdoutSet$AQI - exp(pred.holdout_4var)))
print(holdout.err_4var)

pred.holdout_5var <- predict(model.trainSet_5var, holdoutSet)
holdout.err_5var <- sqrt(sum((holdoutSet$AQI - exp(pred.holdout_5var))^2)/length(pred.holdout_5var))
plot(holdoutIndex, (holdoutSet$AQI - exp(pred.holdout_5var)))
print(holdout.err_5var)
#from the residual plot of the holdoutSet actual data vs predicted values based on training set
#is approx. random
#Compare the prediction errors and choose the best model: 3 variables with
NO2_24h+log(PM10_24h)+log(PM2.5_24h)

#predicted valus and actual values based on the best model chosed from requirement 4
plot(holdoutIndex, holdoutSet$AQI, type = "l", main="predicted value VS. actual values")
lines(holdoutIndex, exp(pred.holdout_3var), type = "l", col = 'red')
# END Requirement 5 ( Set training and hold-out set.)

## Requirement 6 (ls.diag(fit) Cook & dfit)
#cook's distance / dfits
fit6_new<-lm(log(AQI)~NO2_24h+log(PM10_24h)+log(PM2.5_24h), data=X306)
diagnostics <- ls.diag(fit6_new)
cookDis <- diagnostics$cooks
dfits <- diagnostics$dfits

#plot cook's distance and dfits
index <- seq(1:nrow(X306))
plot(index, cookDis, main = "Cooks Distance Plot", xlab = "Index", ylab = "Cooks Distance")
plot(index, dfits, main = "dfits Plot", xlab = "Index", ylab = "dfits")
## END Requirement 6 (ls.diag(fit) Cook & dfit)
```