Ian Beller (1006837620)

STA238 Final Project

March 19th, 2022

**Pollution levels in Beijing, China from 2010 to 2015**

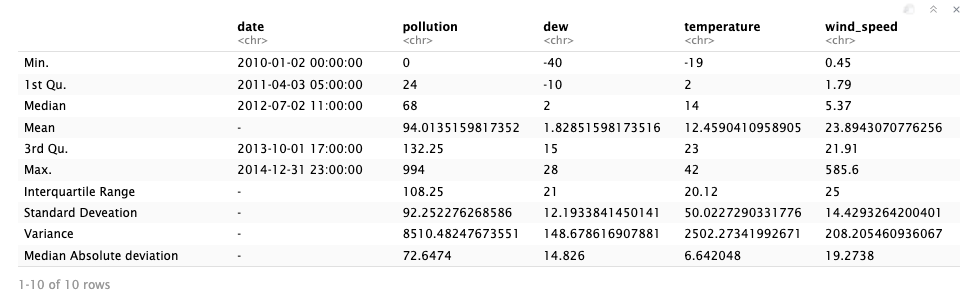
Introduction

Beijing is one of the most populated cities in the world and has been heavily criticized for its pollution levels. We want to invetsgate if the trend in the amount of pollution in the air between the 2010 and 2015 changed between seasons in Beijing, China? We will be using the dataset from the U.S. embassy which reports on pollution levels and weather conditions in Beijing, China for five years from January 2nd, 2010, to December 31st, 2014.[[1]](#footnote-0) Variables collected were pollution level (PM 2.5 levels), Dew points (DEWP), temperature (in Celsius), pressure (newtons per square meter), wind diretction, cumulated wind speed (in knots), cumulated hours of snow, and cumulated hours of rain. This dataset does cover numerous variables, but we will only consider temperature, wind speed, and rain since the National Weather Service[[2]](#footnote-1) state that these are key factors that affect PM2.5 levels. Since we cannot correctly quantify rainfall as its measurement in the dataset is ambiguous, we will use Dew points instead, as it measures the temperature needed for air to cool down to achieve a relative humidity, hence higher dew point results in more effectiveness in removing pollutants, a similar effect as rain has on pollution.

Data and Exploratory Data Analysis (EDA)

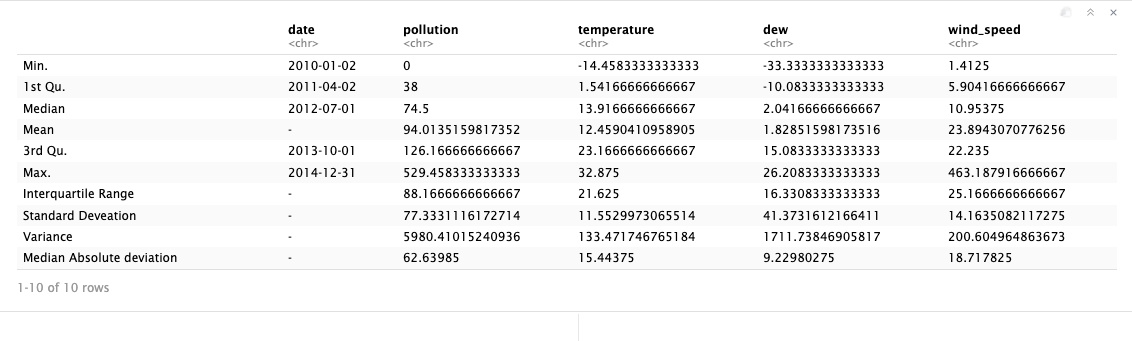
To clean our data, we dropped our unwanted variables to focus on the important variables.[[3]](#footnote-2) Then we performed an EDA[[4]](#footnote-3) and constructed a summary table shown in Figure 1 that depicts key aspects and values from our data.

*Figure 1*



We found that the data set has 43,800 data entries that have a large variation due to the change in conditions throughout the day. To have more consistent data we calculated the average for each variable, for each day in our dataset and created a new table with these calculations[[5]](#footnote-4). Figure 2 shows the same EDA analysis done in Figure 1, however, in our new EDA we used our average per day calculations [[6]](#footnote-5).

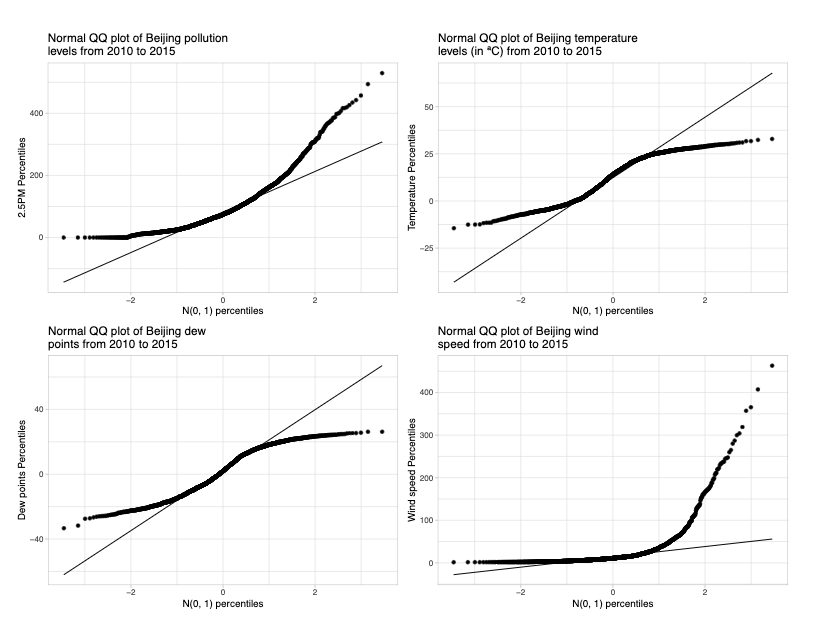
*Figure 2*



We see that the means were kept the same across all variables, but the variance and standard deviation decreased in all variables in the averaged summary table since outliers have less of an effect on our values since grouping our data per day, resulting in more consistent data. We also notice there is a minor difference between mean and median in temperature and dew, while we see a more significant difference in wind speed and pollution. We see that pollution has a far larger median absolute deviation comprare to the other variables which suggest a much larger variation of data.

We then checked for normality of each variable by plotting. QQ plots for each variable as shonw in Figure 3. We observe that the pollution levels have a heavy right tail while the left tail of the distribution is light, which means the distribution of pollution is right-skewed. Temperature and dew points have a bimodal distribution since the first quantiles have a percentile close to zero, then the points align with the normal distribution in the center but then cross the line around the x=1 which indicates that the data quantiles are not approaching zero as normal distribution does. Wind speed has a similar distribution to pollution by being a right-skewed normal distribution. However, compared to pollution distribution, we see that wind speed is, even more, right-skewed as it has a heavier right tail and a lighter left tail.

*Figure 3 [[7]](#footnote-6)*



We will now construct a simple linear regression to understand trends in pollution in Beijing.

Simple Linear regression

We will follow the standard regression model for determining pollution in different seasons.

where y = 2.5PM levels, = temperature

= wind speed = dew points

= y-intercept = temperature coefficient/slope, = wind speed coefficient/slope, and = dew points coefficient/slope

Using our standard regression model model our estimated linear regression model we see a summary[[8]](#footnote-7) for pollution levels all over Beijing. Hence our model equation is

The model has a high SSE and low adjusted because Beijing experiences a vast variety of weather conditions throughout the year that differently affects pollution. This makes a linear regression difficult to be a good fit. For that reason, we will compare Beijing during the same seasons, to keep weather and other factors relatively consistent[[9]](#footnote-8). We separated the data into four seasons[[10]](#footnote-9).

Using the model summary and calculations[[11]](#footnote-10), we estimate that the linear regression model equation for pollution during spring is

Using the model summary and calculations[[12]](#footnote-11), we estimate that the linear regression model equation for pollution during winter is

Using the model summary and calculations[[13]](#footnote-12), we estimate that the linear regression model equation for pollution during autumn is

Note that by eliminating the wind speed parameter, we were able to increase , decrease SSE, and also decrease SE of all variables, which makes the regression model a better fit.

Using the model summary and calculations[[14]](#footnote-13), we estimate that the linear regression model equation for pollution during summer is

We should note the y-intercept or is 0 since when calculating our estimate we found that , however, we cannot have a negative PM2.5 level, hence .

From our regression models we found that in Beijing when all our variables were 0, the air pollution was at 181.42 PM2.5 levels which are considered dangerous. However, we saw that temperature and wind speed had a negative correlation with pollution, with temperature having a stronger negative relationship. However, dew points have a positive correlation with pollution levels. These correlations are persistent throughout all seasons of the year.

Based on our regression models, we see that when all variables are equal to zero, winter has the highest pollution levels while autumn is closely second, then comes spring, and lastly summer that would have no pollution. Ttemperature is a determinant of pollution levels in all seasons except for summer. It is also the most determinant factor of pollution during autumn as a 1ºC increase, pollution decreases by at least twice as much as in any other season. Wind speed is the only a determinant of pollution during spring and winter, however, it is the least determining variable in those seasons, as its effects on pollution levels are minor compared to other variables. Lastly, we see that dew points are a determining factor for pollution throughout all seasons in Beijing, and it is the second most important variable in determining pollution throughout all seasons except during summer when it is the most as it is the only variable determining pollution. We then determine that dew points are the most determining variable for any random season of the year since it is the most or second most important variable for determining pollution levels for any season in Beijing. However, given a year where Beijing has gone through all four seasons, we determine temperature being the most important factor. We should note that all our regression models had a which means that any of these models can at most exokaun 42.3% of the variation in pollution is based on the variables in the model. Hence there exist other factors that we did not study that also influence pollution.

Confidence intervals

Our regression models are estimates for a dataset and that these estimates will have some sort of error. We want to investigate an interval in which each coefficient/ value could be in with a 95% confidence interval as any value in this interval would be considered an accurate estimation.

For any year in Beijing between 2010 and 2015, we estimate with 95% confident[[15]](#footnote-14) the following:

* When all our variables (temp, wnd\_spd, dew) are equal to 0, pollution in Beijing is between and PM2.5 levels.
* A 1ºC increase in temperature would decrease pollution levels on average by to PM2.5 levels
* A one-knot increase in wind speed would decrease pollution levels on average by to PM2.5 levels
* A point increase in dew would increase pollution levels on average by PM2.5 levels.

For any year between 2010 and 2015 in Beijing, China during the spring season, we estimate with 95% confidence the following [[16]](#footnote-15):

* When all our variables (temp, wnd\_spd, dew) are equal to 0, pollution in Beijing is between PM2.5 levels.
* A 1ºC increase in temperature would decrease pollution levels on average by to PM2.5 levels
* A one-knot increase in wind speed would decrease pollution levels on average by to PM2.5 levels
* A point increase in dew would increase pollution levels on average by PM2.5 levels.

For any year between 2010 and 2015 in Beijing, China during the winter season, we estimate with 95% confidence[[17]](#footnote-16) the following:

* When all our variables (temp, wnd\_spd, dew) are equal to 0, pollution in Beijing is between PM2.5 levels.
* A 1ºC increase in temperature would decrease pollution levels on average by to PM2.5 levels
* A one-knot increase in wind speed would decrease pollution levels on average by to PM2.5 levels
* A point increase in dew would increase pollution levels on average by PM2.5 levels.

For any year between 2010 and 2015 in Beijing, China during the autumn season, we estimate with 95% confidence[[18]](#footnote-17) the following:

* When all our variables (temp, dew) are equal to 0, pollution in Beijing is between PM2.5 levels.
* A 1ºC increase in temperature would decrease pollution levels on average by to PM2.5 levels
* A point increase in dew would increase pollution levels on average by PM2.5 levels.

For any year between 2010 and 2015 in Beijing, China during the summer season, we estimate with 95% confidence[[19]](#footnote-18) the following:

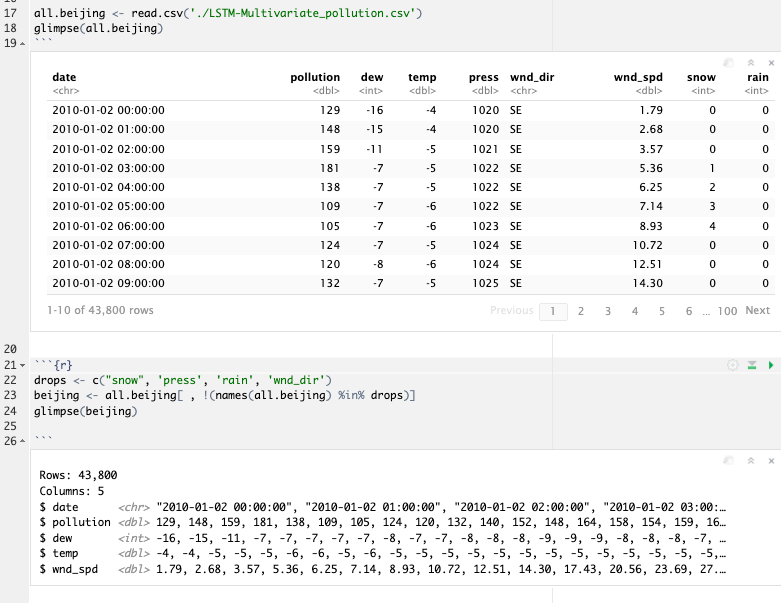
* When dew is equal to 0, pollution in Beijing is between PM2.5 levels. However as discussed before, we cannot have negative PM2.5 levels, hence pollution would be 0.
* A point increase in dew would increase pollution levels on average by PM2.5 levels.

Conclusion

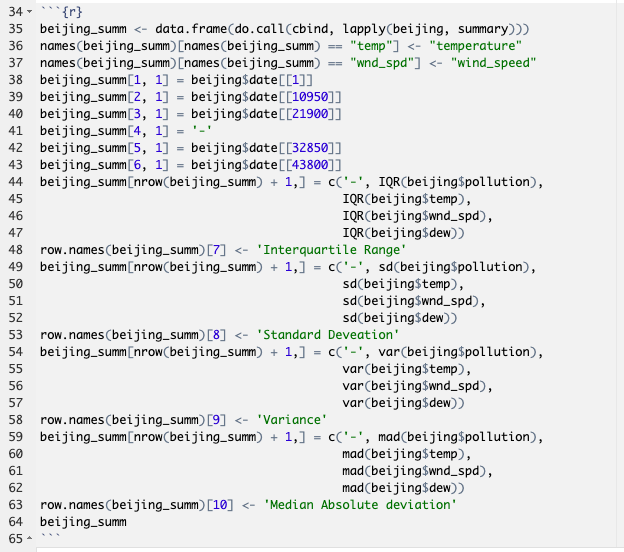
We found that pollution has a negative correlation of some degree with temperature and wind speed, however, a stronger negative correlation with temperature than with wind speed. We also found that dew points have a positive correlation with pollution. The results suggest that temperature is the most determining factor for pollution, however, dew points was also a key factor in predicting pollution levels at any given season of the year. Although wind speed was a factor to consider, it had minor effects on pollution. Throughout the whole year, all the variables studied were considered factors that affected pollution, however the importance of factors in determining pollution varied depending on the season. We found that weather conditions and the season of the year are detrimental to pollution levels. We saw that given the same dew point in all seasons (the only variable all regression models share) the order from highest to lowest pollution levels was: winter, autumn, spring, and summer which is persistent with other findings.2 We should note that all our regression models could at most explain 42.6 % of the variation in pollution which suggests there are other factors that we did not consider that have a significant effect on pollution levels and should be further studied. To further improve our estimations, we could have considered non-linear, multivariate regressions or regression trees that could have had a better fit for our data. Given these points, our findings are not conclusive as they may be considered potential answers since we did not consider all variables that determine pollution and should be further explored.

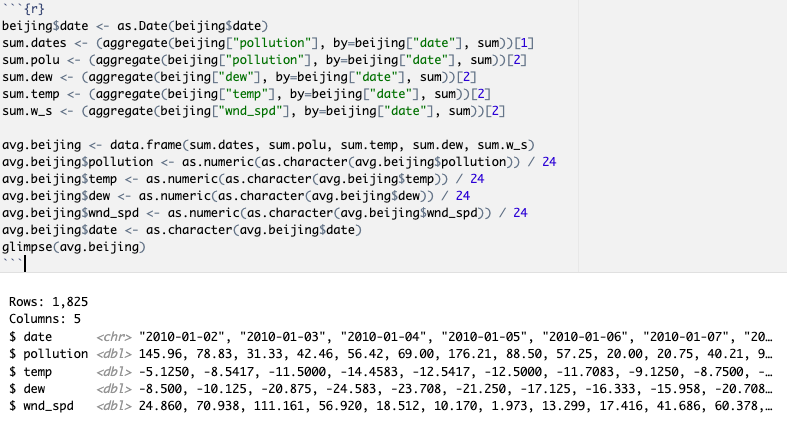
Appendix

Appendix A

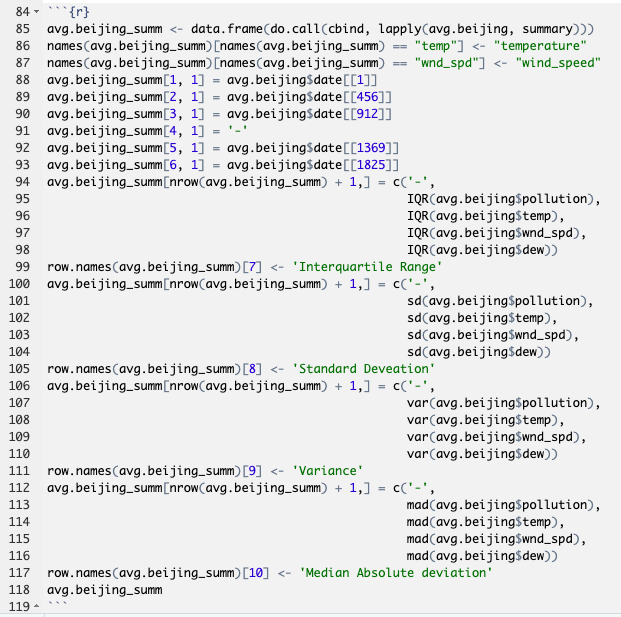


Appendix B



Appendix C

Appendix D

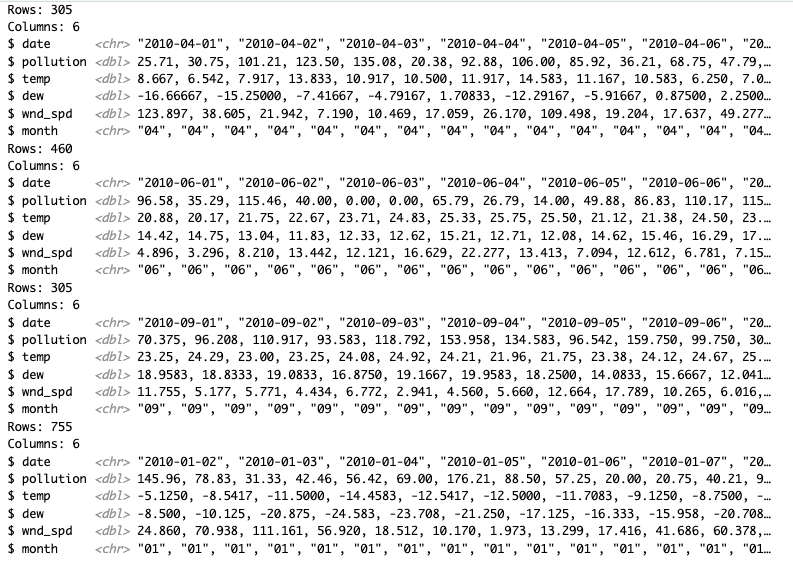


Appendix E

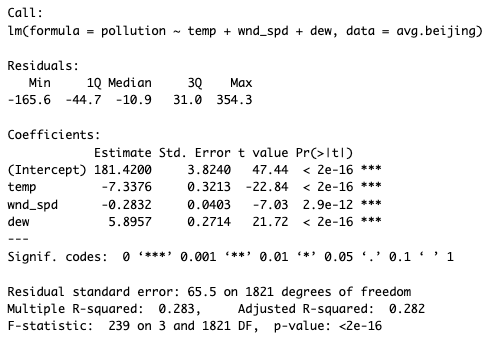
Appendix E.1



Appendix E.2



Appendix F

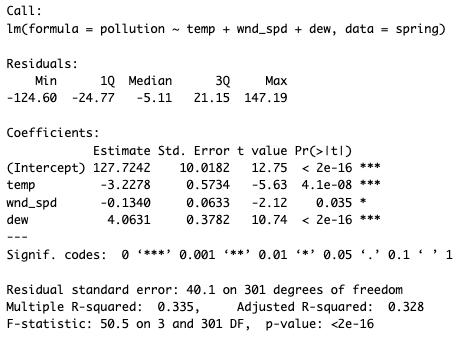


Since we see that all of our variables are statistically significant since they have a p-value below 0.05,we can use this regression model as its statistically significant. Hence the estimates for our parameters are



We see that the residual standard error is 65.5 with a degrees of freedom (df) of 1821, hence the sum of squared errors (SSE) is and that the adjusted . We also see that the standard error (SE) for each parameter is

Appendix G



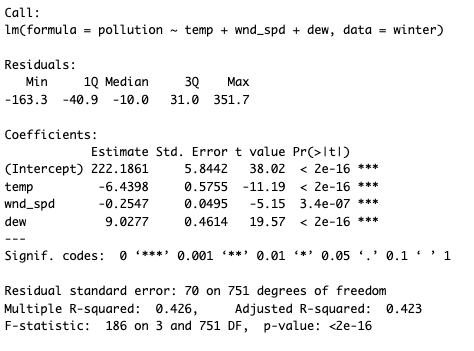
Since we see that all of our variables are statistically significant since they have a p-value below 0.05,we can use this regression model as its statistically significant. Hence the estimates for our parameters are

* -

We see that the residual standard error is 40.1 with a degrees of freedom of 301, hence the is and that the adjusted . We also see that the standard error (SE) for each parameter is

* -
* -

Appendix I



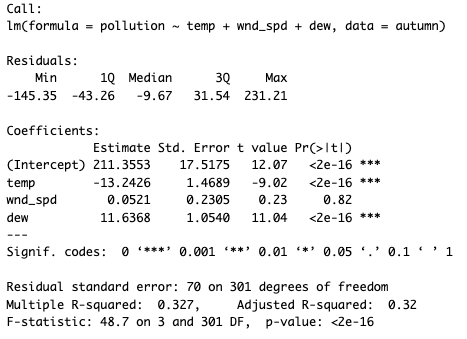
Since we see that all of our variables are statistically significant since they have a p-value below 0.05,we can use this regression model as its statistically significant. Hence the estimates for our parameters are

* -

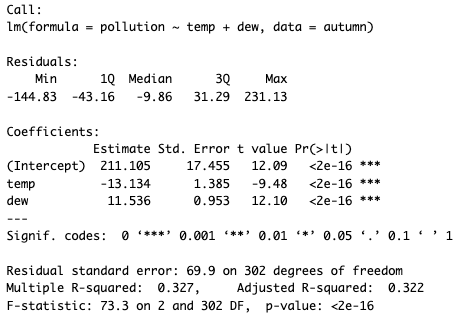
We see that the residual standard error is 70 with a degrees of freedom of 751, hence the is and that the adjusted . We also see that the standard error (SE) for each parameter is

* -
* -

Appendix J



We notice that wnd\_spd parameter is statistically insignificant since it has a p-value above 0.05. Hence to have the model with the most precision when estimating pollution levels in autumn, we will disregard the parameter wnd\_spd. Hence our statistical summary for our new model is



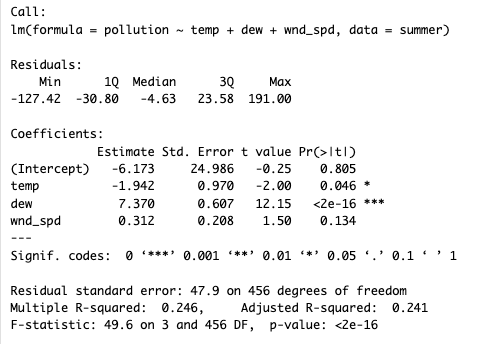
When disregarding the parameter wnd\_spd, we see that all of our variables are statistically significant, hence we can use this regression model as its statistically significant. The estimates for our parameters are



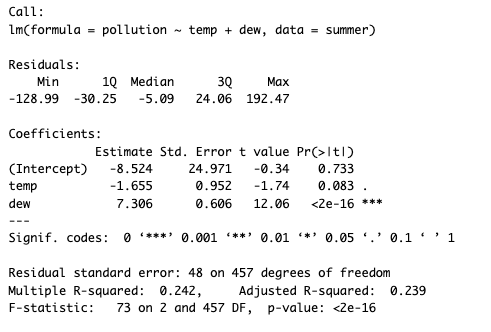
We see that the residual standard error is 69.9 with a degrees of freedom of 302, hence the is and that the adjusted . We also see that the standard error (SE) for each parameter is

* -

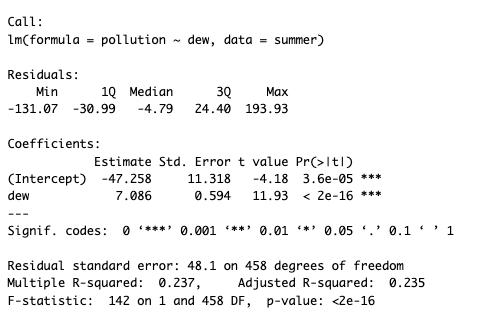
Appendix K



We notice that wnd\_spd parameter is statistically insignificant since it has a p-value above 0.05. We also see that the intercept value is also statistically insignificant and this skews with our model since we do not have enough evidence to reject the null hypothesis for our variables which means that we do not have enough evidence to assume our regression line is not horizontal/ there is no relationship. Hence to have the model with the most precision when estimating pollution levels in summer, we will disregard the parameter wnd\_spd. Hence our statistical summary for our new model is



In the new regression model, we have the same issue as our regression model but now the temperature is statistically insignificant which keeps the intercept also statistically insignificant. Hence we will re-run another regression model but now disregard temperature as well.



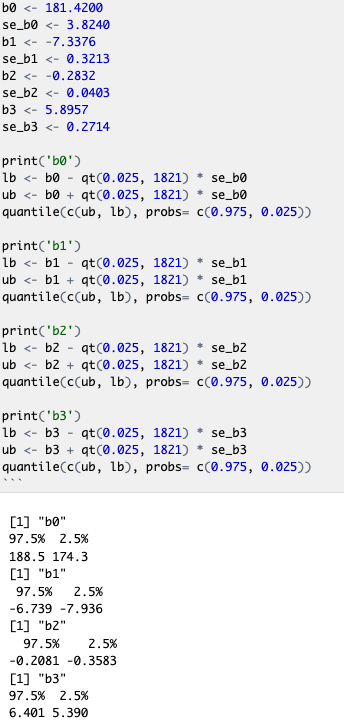
When disregarding the parameters wnd\_spd and temp, we see that all of our variables are statistically significant, hence we can use this regression model as its statistically significant. The estimates for our parameters are



We see that the residual standard error is 69.9 with a degrees of freedom of 302, hence the is and that the adjusted . We also see that the standard error (SE) for each parameter is

* -

Appendix L



We used the parameter values found in the linear regression model for all of Beijing which can be found in

Appendix F

We will use the following the equation

to find the confidence interval where is the estimated beta value is the standard error of and is the critical t-values. We know our regression model has a degree of freedom is 1821 and we want to find a confidence interval of 95% hence our critical T value is

as calculated in R.

The critical t-value are the same for all values.

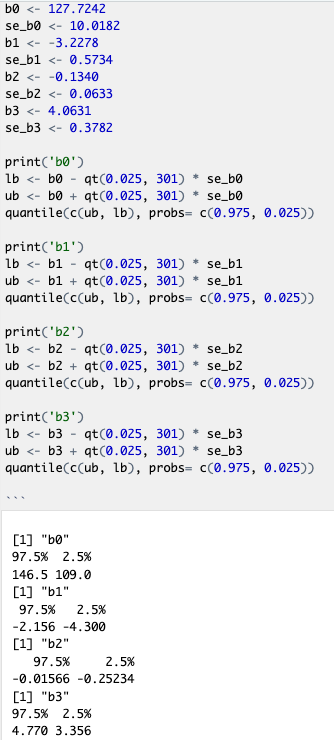
CI for is which results in an interval of .

CI for is which results in an interval of .

CI for is which results in an interval of .

CI for is which results in an interval of .

Appendix M



We will use the following the equation

to find the confidence interval where is the estimated beta value is the standard error of and is the critical t-values. We know our regression model has a degree of freedom is 301 and we want to find a confidence interval of 95% hence our critical T value is

as calculated in R.

The critical t-value are the same for all values.

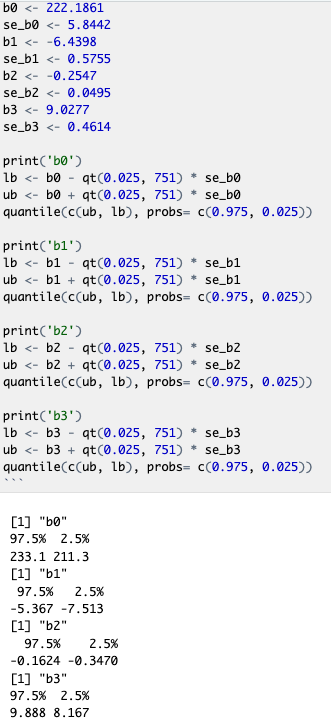
CI for is which results in an interval of .

CI for is which results in an interval of .

CI for is which results in an interval of .

CI for is which results in an interval of .

Appendix N



We will use the following the equation

to find the confidence interval where is the estimated beta value is the standard error of and is the critical t-values. We know our regression model has a degree of freedom is 751 and we want to find a confidence interval of 95% hence our critical T value is

as calculated in R.

The critical t-value are the same for all values.

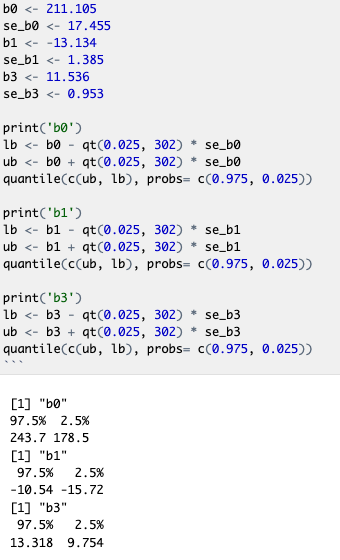
CI for is which results in an interval of .

CI for is which results in an interval of .

CI for is which results in an interval of .

CI for is which results in an interval of .

Appendix O



We will use the following the equation

to find the confidence interval where is the estimated beta value is the standard error of and is the critical t-values. We know our regression model has a degree of freedom is 302 and we want to find a confidence interval of 95% hence our critical T value is

as calculated in R.

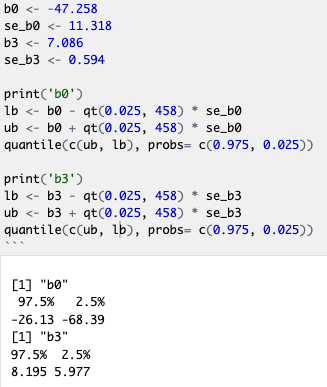
The critical t-value are the same for all values.

CI for is which results in an interval of .

CI for is which results in an interval of .

CI for is which results in an interval of .

Appendix P



We will use the following the equation

to find the confidence interval where is the estimated beta value is the standard error of and is the critical t-values. We know our regression model has a degree of freedom is 458 and we want to find a confidence interval of 95% hence our critical T value is

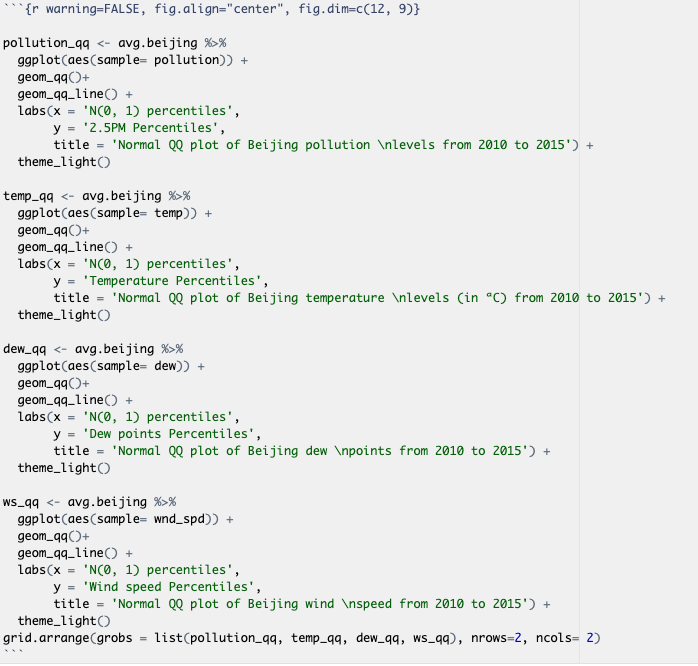
as calculated in R.

The critical t-value are the same for all values.

CI for is which results in an interval of .

CI for is which results in an interval of .

Appendix Q



Reference

*Clearing the Air on Weather and Air Quality*. Weather.gov. Retrieved 19 March 2022, from<https://www.weather.gov/wrn/summer-article-clearing-the-air>.

Roy, R., & Embassy of the United State of America in Beijing, China. (2022). Air Pollution Forecasting - LSTM Multivariate. Kaggle.com. Retrieved 18 March 2022, from<https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate?select=LSTMMultivariate_pollution.csv>.

Travel China Guide. (2022). *Beijing Weather: 7-Day Forecast, Best Time to Visit, Monthly Climate*. Travelchinaguide.com. Retrieved 18 March 2022, from<https://www.travelchinaguide.com/climate/beijing.htm>.

1. Roy, R., & Embassy of the United State of America in Beijing, China. (2022). Air Pollution Forecasting - LSTM Multivariate. Kaggle.com. Retrieved 18 March 2022, from <https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate?select=LSTMMultivariate_pollution.csv>. [↑](#footnote-ref-0)
2. *Clearing the Air on Weather and Air Quality*. Weather.gov. Retrieved 19 March 2022, from <https://www.weather.gov/wrn/summer-article-clearing-the-air> . [↑](#footnote-ref-1)
3. Appendix A [↑](#footnote-ref-2)
4. Appendix B [↑](#footnote-ref-3)
5. Appendix C [↑](#footnote-ref-4)
6. Appendix D [↑](#footnote-ref-5)
7. Appendix Q [↑](#footnote-ref-6)
8. Appendix F [↑](#footnote-ref-7)
9. According toTravel China Guide , spring is from April-May, summer extends from June-August, autumn goes from September-October and winter is from November-March. [↑](#footnote-ref-8)
10. Appendix E (Table in Appendix E.2) [↑](#footnote-ref-9)
11. Appendix G [↑](#footnote-ref-10)
12. Appendix I [↑](#footnote-ref-11)
13. Appendix J [↑](#footnote-ref-12)
14. Appendix K [↑](#footnote-ref-13)
15. Appendix L [↑](#footnote-ref-14)
16. Appendix M [↑](#footnote-ref-15)
17. Appendix N [↑](#footnote-ref-16)
18. Appendix O [↑](#footnote-ref-17)
19. Appendix P [↑](#footnote-ref-18)