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] 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] 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.

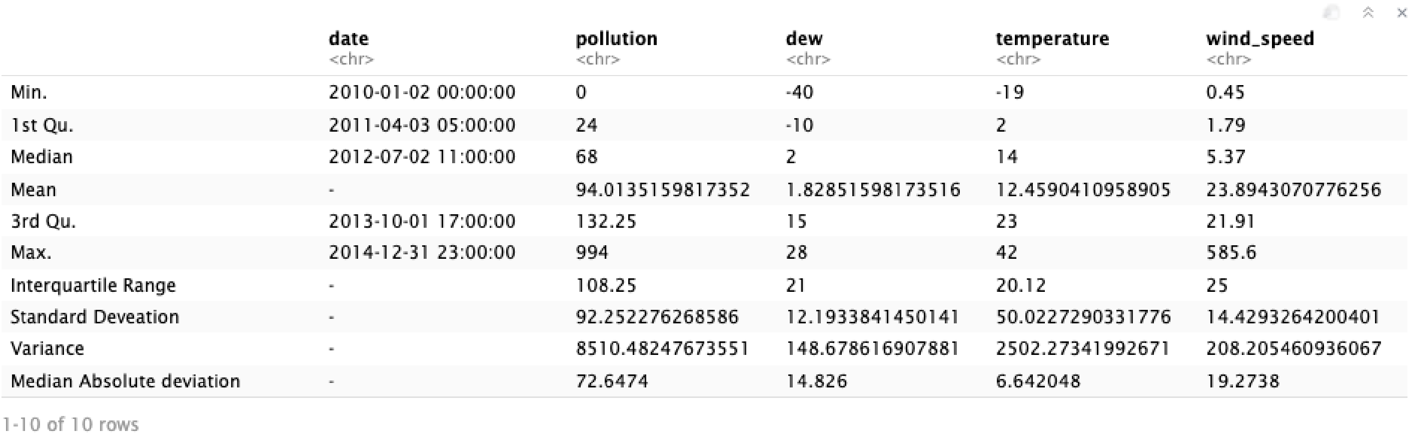
[1] Roy, R., & Embassy of the United State of America in Beijing, China. (2022). Air Pollution Forecasting - LSTM Multivariate. Kaggle.com. Retrieved 18 March 2022, from https://www.kaggle.com/datasets/rupakroy/lstm-datasets-multivariate-univariate?select=LSTMMultivariate_pollution.csv.

[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 .

Data and Exploratory Data Analysis (EDA)

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

Figure 1



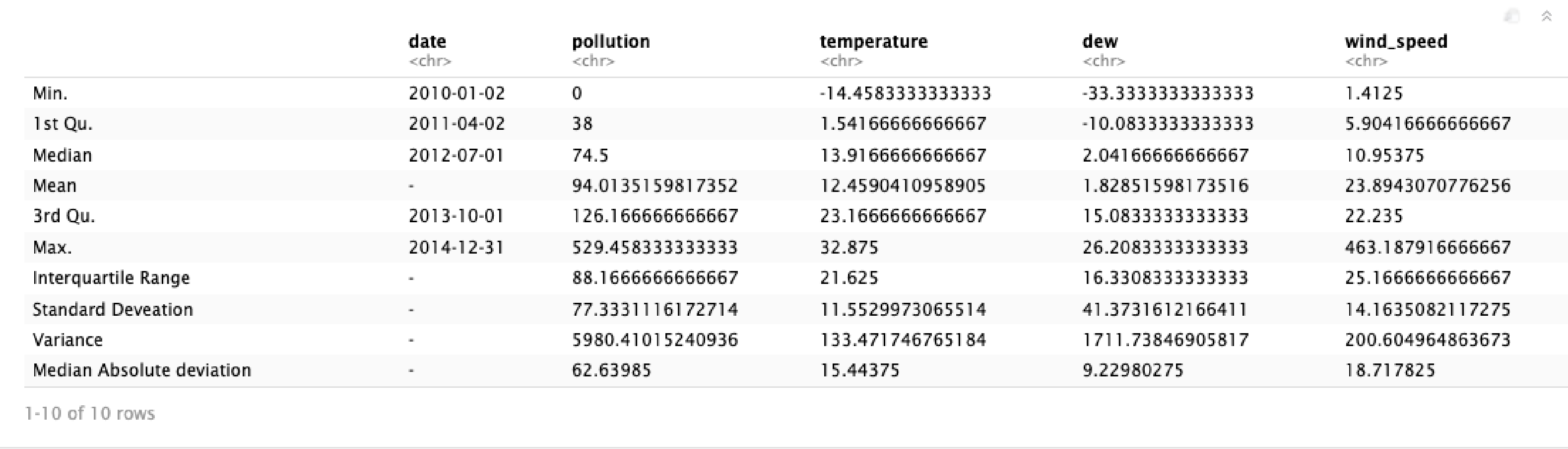
	date <chr>	pollution <chr>	dew <chr>	temperature <chr>	wind_speed <chr>
Min.	2010-01-02 00:00:00	0	-40	-19	0.45
1st Qu.	2011-04-03 05:00:00	24	-10	2	1.79
Median	2012-07-02 11:00:00	68	2	14	5.37
Mean	-	94.0135159817352	1.82851598173516	12.4590410958905	23.8943070776256
3rd Qu.	2013-10-01 17:00:00	132.25	15	23	21.91
Max.	2014-12-31 23:00:00	994	28	42	585.6
Interquartile Range	-	108.25	21	20.12	25
Standard Deveation	-	92.252276268586	12.1933841450141	50.0227290331776	14.4293264200401
Variance	-	8510.48247673551	148.678616907881	2502.27341992671	208.205460936067
Median Absolute deviation	-	72.6474	14.826	6.642048	19.2738

1-10 of 10 rows

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]. Figure 2 shows the same EDA analysis done in Figure 1, however, in our new EDA we used our average per day calculations [6].

[3] Appendix A
[4] Appendix B
[5] Appendix C
[6] Appendix D

Figure 2



	date	pollution	temperature	dew	wind_speed
	<chr>	<chr>	<chr>	<chr>	<chr>
Min.	2010-01-02	0	-14.4583333333333	-33.3333333333333	1.4125
1st Qu.	2011-04-02	38	1.54166666666667	-10.0833333333333	5.90416666666667
Median	2012-07-01	74.5	13.9166666666667	2.04166666666667	10.95375
Mean	-	94.0135159817352	12.4590410958905	1.82851598173516	23.8943070776256
3rd Qu.	2013-10-01	126.166666666667	23.1666666666667	15.0833333333333	22.235
Max.	2014-12-31	529.458333333333	32.875	26.2083333333333	463.187916666667
Interquartile Range	-	88.1666666666667	21.625	16.3308333333333	25.1666666666667
Standard Deveation	-	77.3331116172714	11.5529973065514	41.3731612166411	14.1635082117275
Variance	-	5980.41015240936	133.471746765184	1711.73846905817	200.604964863673
Median Absolute deviation	-	62.63985	15.44375	9.22980275	18.717825

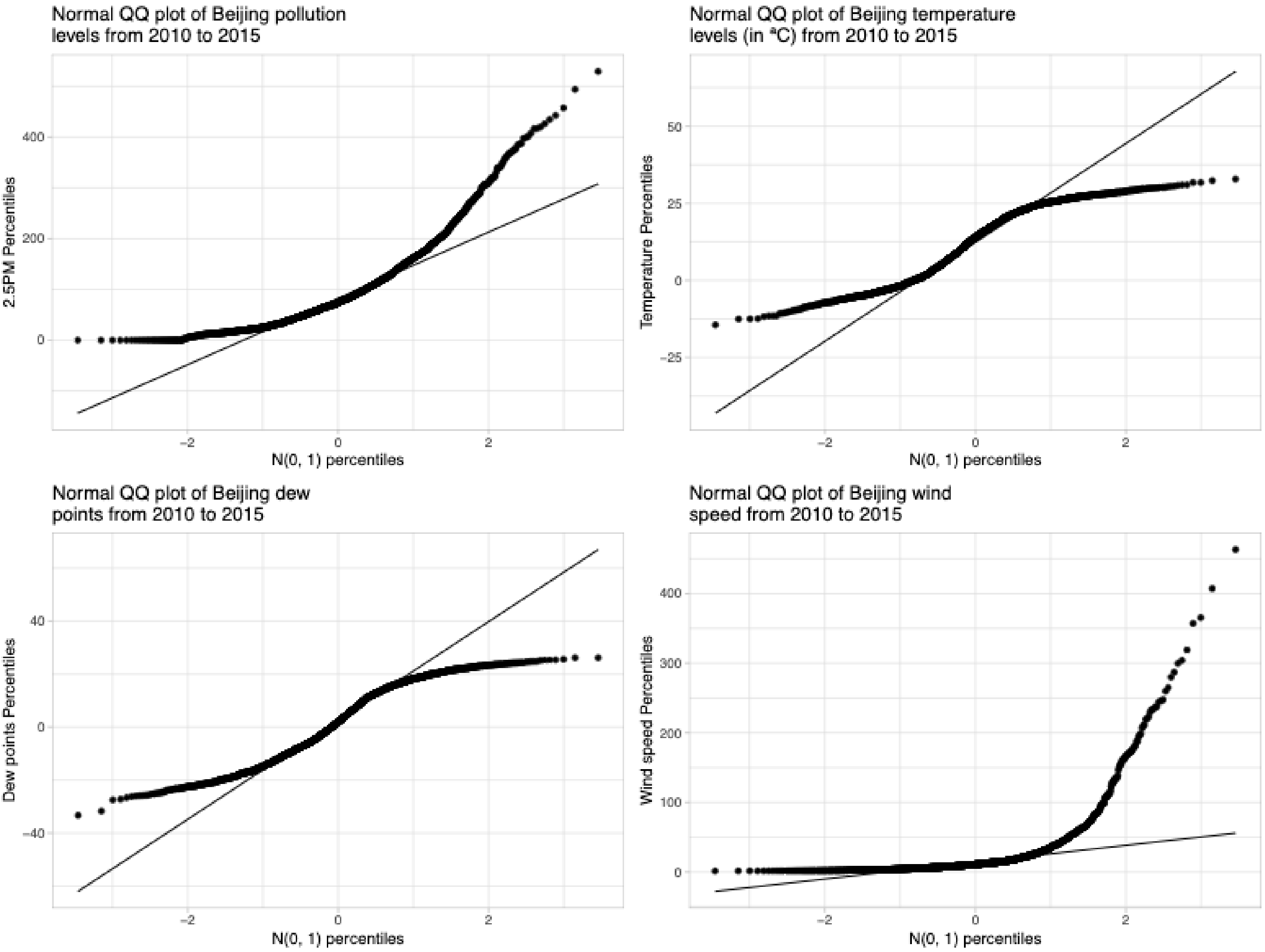
1-10 of 10 rows

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]



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

[7] Appendix Q

Simple Linear regression

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

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3$$

where y = 2.5PM levels, x_1= temperature

x_2 = wind speed x_3= dew points

β_0= y-intercept β_1= temperature coefficient/slope,

β_2= wind speed coefficient/slope, and β_3= dew points coefficient/slope

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

$$y = 181.42 - 7.3376\, x_1 - 0.2832x_2 + 5.8957x_3$$

The model has a high SSE and low adjusted R^2 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]. We separated the data into four seasons[10].

Using the model summary and calculations[11], we estimate that the linear regression model equation for pollution during spring is

$$y = 127.7242 - 3.2278\, x_1 - 0.1340x_2 + 4.0631x_3$$

[8] Appendix F

[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.

[10] Appendix E (Table in Appendix E.2)

[11] Appendix G

Using the model summary and calculations[12], we estimate that the linear regression model equation for pollution during winter is

$$y\ =\ 222.1861\ -\ 6.4398x_1 - 0.2547x_2 +\ 9.0277x_3$$

Using the model summary and calculations[13], we estimate that the linear regression model equation for pollution during autumn is

$$y\ =\ 211.105\ -\ 13.134\,x_1+\ 11.536x_3$$

Note that by eliminating the wind speed parameter, we were able to increase R^2 , decrease SSE, and also decrease SE of all variables, which makes the regression model a better fit.

Using the model summary and calculations[14], we estimate that the linear regression model equation for pollution during summer is

$$y\ = 7.086x_3$$

We should note the y-intercept or β_0 is 0 since when calculating our estimate we found that $\beta_0\ =-\ 47.258$, however, we cannot have a negative PM2.5 level, hence $\beta_0\ =\ 0$.

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.

[12] Appendix I
[13] Appendix J
[14] Appendix K

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 $0.235 \leq adjusted\, R^2 \leq 0.423$ 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/ β 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] the following:

- When all our variables (temp, wnd_spd, dew) are equal to 0, pollution in Beijing is between 174.3 and 188.5 PM2.5 levels.
- A 1ºC increase in temperature would decrease pollution levels on average by 6.739 to 7.936 PM2.5 levels
- A one-knot increase in wind speed would decrease pollution levels on average by 0.2081 to 0.3583 PM2.5 levels
- A point increase in dew would increase pollution levels on average by $5.390\ to\ 6.401$ 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]:

- When all our variables (temp, wnd_spd, dew) are equal to 0, pollution in Beijing is between $109.0\ to\ 146.5$ PM2.5 levels.
- A 1ºC increase in temperature would decrease pollution levels on average by 2.156 to 4.300 PM2.5 levels
- A one-knot increase in wind speed would decrease pollution levels on average by 0.01566 to 0.25234 PM2.5 levels
- A point increase in dew would increase pollution levels on average by $3.356\ to\ 4.770)$ PM2.5 levels.

[15] Appendix L

[16] Appendix M

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

- When all our variables (temp, wnd_spd, dew) are equal to 0, pollution in Beijing is between $211.3\ to\ 233.1$ PM2.5 levels.
- A 1ºC increase in temperature would decrease pollution levels on average by 5.367 to 7.513 PM2.5 levels
- A one-knot increase in wind speed would decrease pollution levels on average by 0.1624 to 0.3470 PM2.5 levels
- A point increase in dew would increase pollution levels on average by $8.167\ to\ 9.888$ PM2.5 levels.

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

- When all our variables (temp, dew) are equal to 0, pollution in Beijing is between $178.5\ to\ 243.7$ PM2.5 levels.
- A 1ºC increase in temperature would decrease pollution levels on average by 10.54 to 15.72 PM2.5 levels
- A point increase in dew would increase pollution levels on average by $9.754\ to\ 13.318$ PM2.5 levels.

[17] Appendix N
[18] Appendix O

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

- When dew is equal to 0, pollution in Beijing is between $-\ 68.39\ to\ -\ 26.13$ 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 $5.977\ to\ 8.195$ PM2.5 levels.

Conclusion

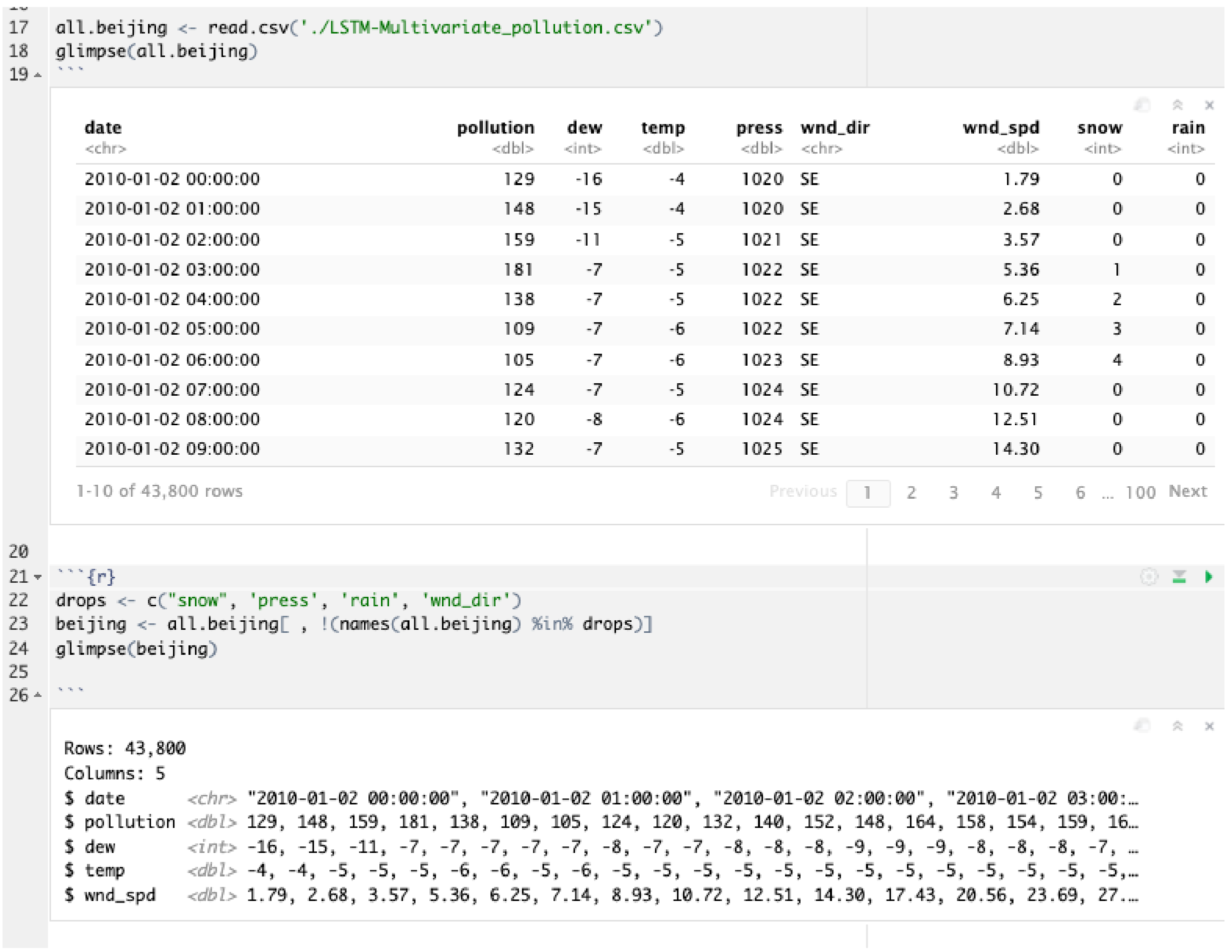
We found that pollution has a negative correlation of some degree with temperature and wind speed, however, a stronger negative correlation with temperature than with wind speed. We also found that dew points have a positive correlation with pollution. The results suggest that temperature is the most 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

[19] Appendix P

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.

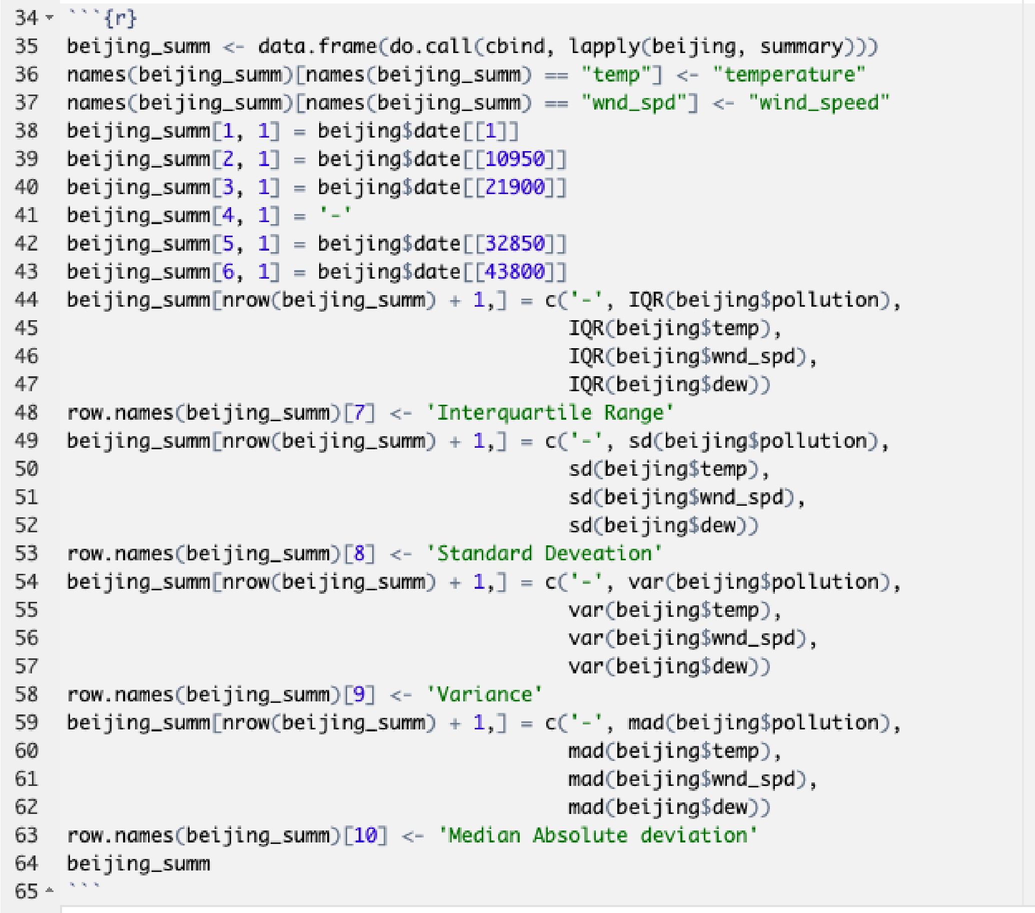
Appendix

Appendix A



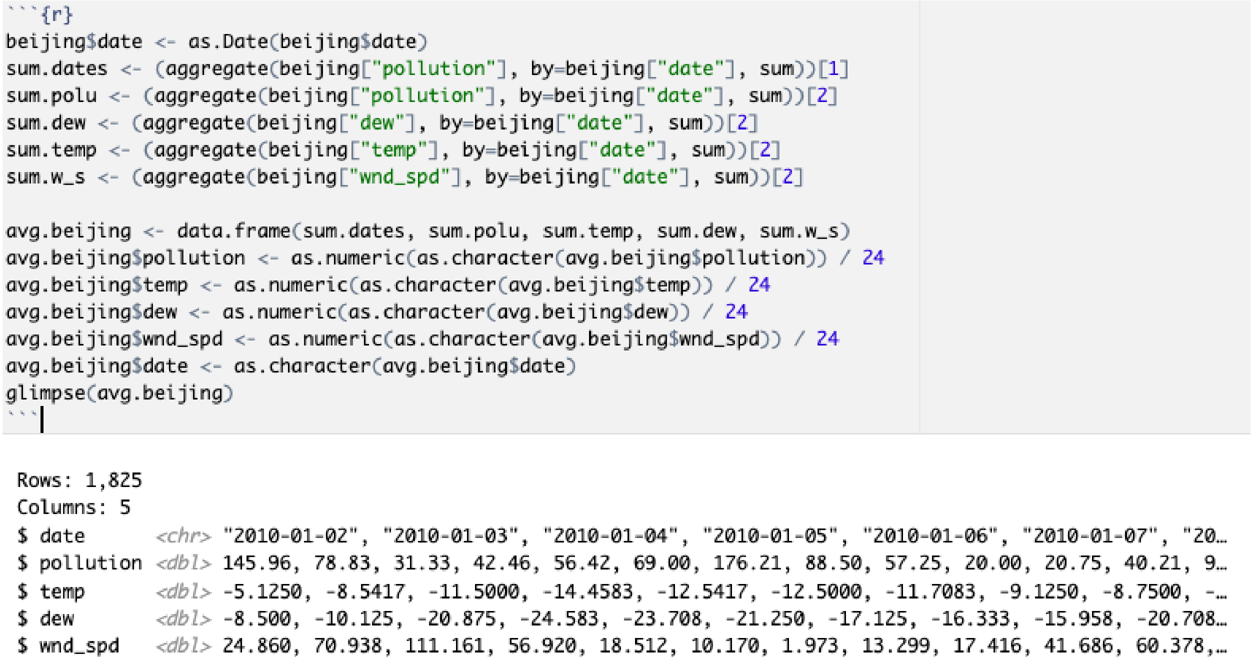
```r
all.beijing <- read.csv('./LSTM-Multivariate_pollution.csv')
glimpse(all.beijing)
```
~~~

date <chr>	pollution <dbl>	dew <int>	temp <dbl>	press <dbl>	wnd_dir <chr>	wnd_spd <dbl>	snow <int>	rain <int>
2010-01-02 00:00:00	129	-16	-4	1020	SE	1.79	0	0
2010-01-02 01:00:00	148	-15	-4	1020	SE	2.68	0	0
2010-01-02 02:00:00	159	-11	-5	1021	SE	3.57	0	0
2010-01-02 03:00:00	181	-7	-5	1022	SE	5.36	1	0
2010-01-02 04:00:00	138	-7	-5	1022	SE	6.25	2	0
2010-01-02 05:00:00	109	-7	-6	1022	SE	7.14	3	0
2010-01-02 06:00:00	105	-7	-6	1023	SE	8.93	4	0
2010-01-02 07:00:00	124	-7	-5	1024	SE	10.72	0	0
2010-01-02 08:00:00	120	-8	-6	1024	SE	12.51	0	0
2010-01-02 09:00:00	132	-7	-5	1025	SE	14.30	0	0

1-10 of 43,800 rows

Previous
1
2
3
4
5
6
...
100
Next

```r
```{r}
drops <- c("snow", 'press', 'rain', 'wnd_dir')
beijing <- all.beijing[ , !(names(all.beijing) %in% drops)]
glimpse(beijing)

~~~
```

```
Rows: 43,800
Columns: 5
$ date      <chr> "2010-01-02 00:00:00", "2010-01-02 01:00:00", "2010-01-02 02:00:00", "2010-01-02 03:00:…
$ pollution <dbl> 129, 148, 159, 181, 138, 109, 105, 124, 120, 132, 140, 152, 148, 164, 158, 154, 159, 16…
$ dew       <int> -16, -15, -11, -7, -7, -7, -7, -7, -8, -7, -7, -8, -8, -8, -9, -9, -9, -8, -8, -8, -7, …
$ temp      <dbl> -4, -4, -5, -5, -5, -6, -6, -5, -6, -5, -5, -5, -5, -5, -5, -5, -5, -5, -5, -5, -5, -5,…
$ wnd_spd   <dbl> 1.79, 2.68, 3.57, 5.36, 6.25, 7.14, 8.93, 10.72, 12.51, 14.30, 17.43, 20.56, 23.69, 27.…
```
```

Appendix B



~~~
```{r}
beijing_summ <- data.frame(do.call(cbind, lapply(beijing, summary)))
names(beijing_summ)[names(beijing_summ) == "temp"] <- "temperature"
names(beijing_summ)[names(beijing_summ) == "wnd_spd"] <- "wind_speed"
beijing_summ[1, 1] = beijing$date[[1]]
beijing_summ[2, 1] = beijing$date[[10950]]
beijing_summ[3, 1] = beijing$date[[21900]]
beijing_summ[4, 1] = '-'
beijing_summ[5, 1] = beijing$date[[32850]]
beijing_summ[6, 1] = beijing$date[[43800]]
beijing_summ[nrow(beijing_summ) + 1,] = c('-', IQR(beijing$pollution),
                                          IQR(beijing$temp),
                                          IQR(beijing$wnd_spd),
                                          IQR(beijing$dew))
row.names(beijing_summ)[7] <- 'Interquartile Range'
beijing_summ[nrow(beijing_summ) + 1,] = c('-', sd(beijing$pollution),
                                          sd(beijing$temp),
                                          sd(beijing$wnd_spd),
                                          sd(beijing$dew))
row.names(beijing_summ)[8] <- 'Standard Deveation'
beijing_summ[nrow(beijing_summ) + 1,] = c('-', var(beijing$pollution),
                                          var(beijing$temp),
                                          var(beijing$wnd_spd),
                                          var(beijing$dew))
row.names(beijing_summ)[9] <- 'Variance'
beijing_summ[nrow(beijing_summ) + 1,] = c('-', mad(beijing$pollution),
                                          mad(beijing$temp),
                                          mad(beijing$wnd_spd),
                                          mad(beijing$dew))
row.names(beijing_summ)[10] <- 'Median Absolute deviation'
beijing_summ
```
~~~

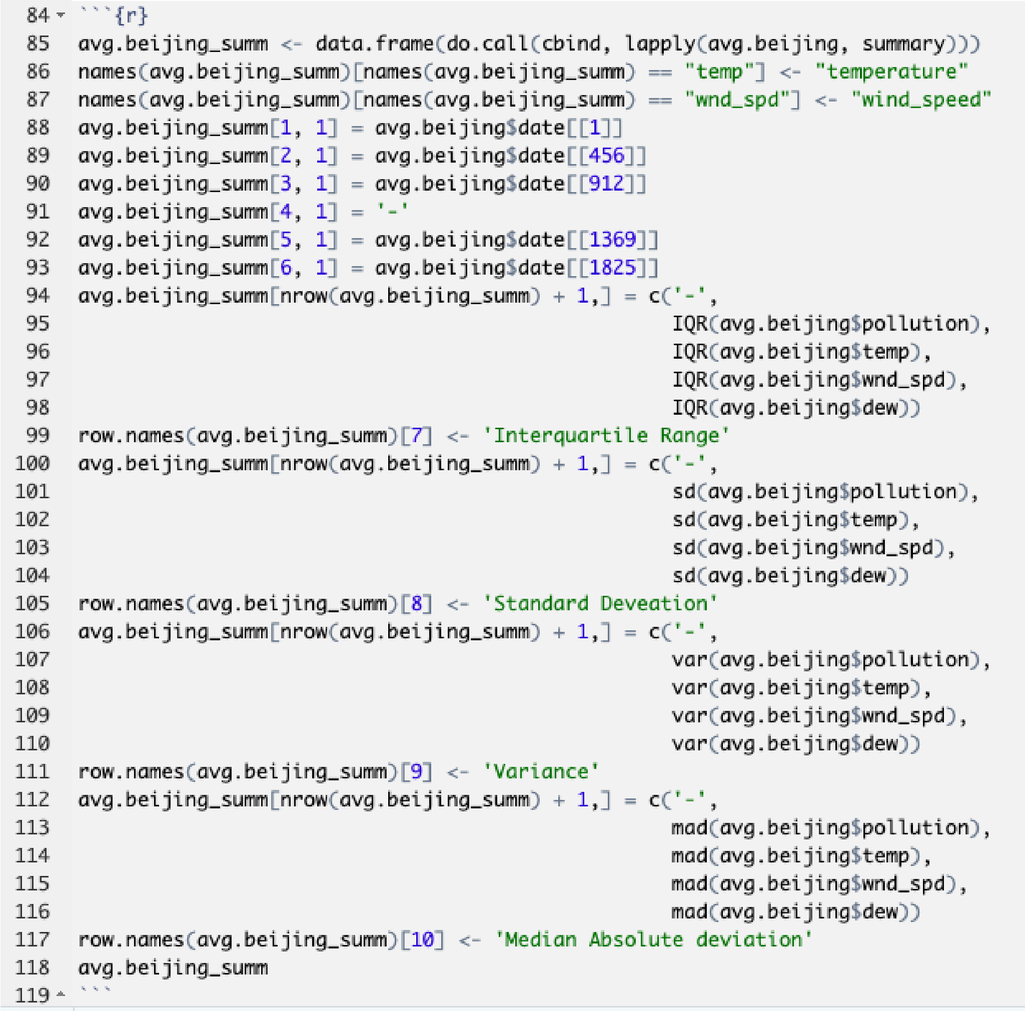
Appendix C



~~~
```{r}
beijing$date <- as.Date(beijing$date)
sum.dates <- (aggregate(beijing["pollution"], by=beijing["date"], sum))[1]
sum.polu <- (aggregate(beijing["pollution"], by=beijing["date"], sum))[2]
sum.dew <- (aggregate(beijing["dew"], by=beijing["date"], sum))[2]
sum.temp <- (aggregate(beijing["temp"], by=beijing["date"], sum))[2]
sum.w_s <- (aggregate(beijing["wnd_spd"], by=beijing["date"], sum))[2]

avg.beijing <- data.frame(sum.dates, sum.polu, sum.temp, sum.dew, sum.w_s)
avg.beijing$pollution <- as.numeric(as.character(avg.beijing$pollution)) / 24
avg.beijing$temp <- as.numeric(as.character(avg.beijing$temp)) / 24
avg.beijing$dew <- as.numeric(as.character(avg.beijing$dew)) / 24
avg.beijing$wnd_spd <- as.numeric(as.character(avg.beijing$wnd_spd)) / 24
avg.beijing$date <- as.character(avg.beijing$date)
glimpse(avg.beijing)
```
~~~

```
Rows: 1,825
Columns: 5
$ date      <chr> "2010-01-02", "2010-01-03", "2010-01-04", "2010-01-05", "2010-01-06", "2010-01-07", "20…
$ pollution <dbl> 145.96, 78.83, 31.33, 42.46, 56.42, 69.00, 176.21, 88.50, 57.25, 20.00, 20.75, 40.21, 9…
$ temp      <dbl> -5.1250, -8.5417, -11.5000, -14.4583, -12.5417, -12.5000, -11.7083, -9.1250, -8.7500, -…
$ dew       <dbl> -8.500, -10.125, -20.875, -24.583, -23.708, -21.250, -17.125, -16.333, -15.958, -20.708…
$ wnd_spd   <dbl> 24.860, 70.938, 111.161, 56.920, 18.512, 10.170, 1.973, 13.299, 17.416, 41.686, 60.378,…
```

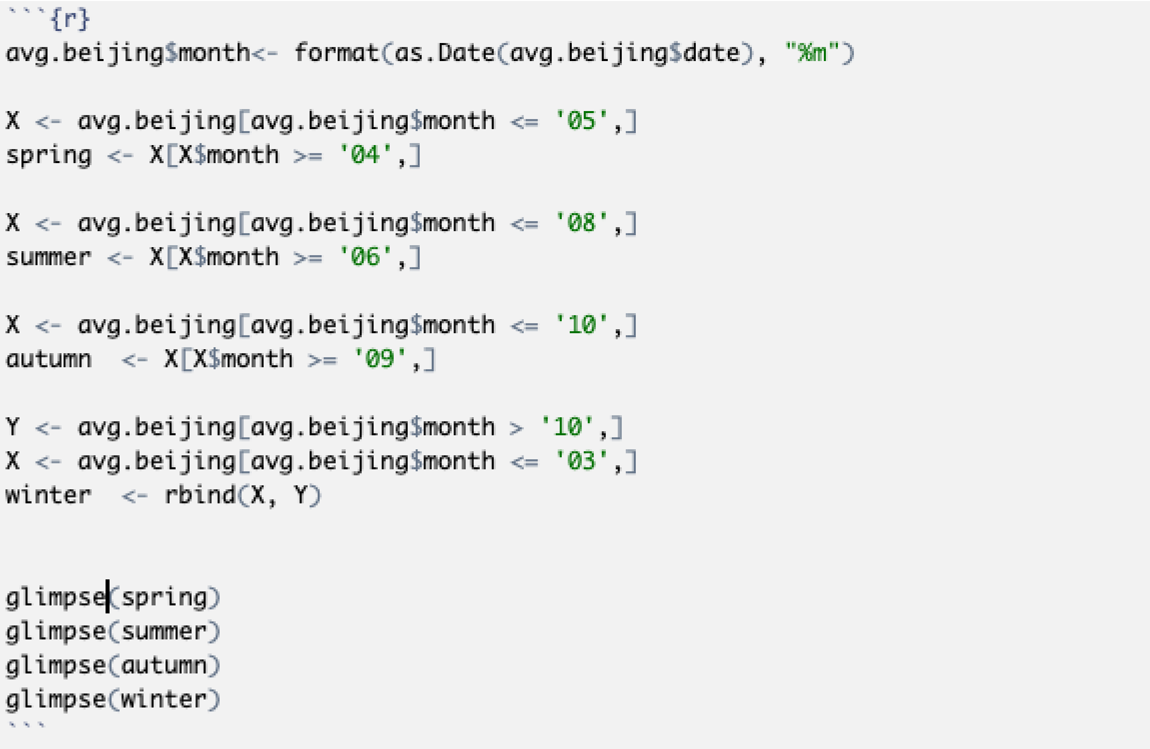
Appendix D



~~~
```{r}
avg.beijing_summ <- data.frame(do.call(cbind, lapply(avg.beijing, summary)))
names(avg.beijing_summ)[names(avg.beijing_summ) == "temp"] <- "temperature"
names(avg.beijing_summ)[names(avg.beijing_summ) == "wnd_spd"] <- "wind_speed"
avg.beijing_summ[1, 1] = avg.beijing$date[[1]]
avg.beijing_summ[2, 1] = avg.beijing$date[[456]]
avg.beijing_summ[3, 1] = avg.beijing$date[[912]]
avg.beijing_summ[4, 1] = '-'
avg.beijing_summ[5, 1] = avg.beijing$date[[1369]]
avg.beijing_summ[6, 1] = avg.beijing$date[[1825]]
avg.beijing_summ[nrow(avg.beijing_summ) + 1,] = c('-',
                                                  IQR(avg.beijing$pollution),
                                                  IQR(avg.beijing$temp),
                                                  IQR(avg.beijing$wnd_spd),
                                                  IQR(avg.beijing$dew))
row.names(avg.beijing_summ)[7] <- 'Interquartile Range'
avg.beijing_summ[nrow(avg.beijing_summ) + 1,] = c('-',
                                                  sd(avg.beijing$pollution),
                                                  sd(avg.beijing$temp),
                                                  sd(avg.beijing$wnd_spd),
                                                  sd(avg.beijing$dew))
row.names(avg.beijing_summ)[8] <- 'Standard Deveation'
avg.beijing_summ[nrow(avg.beijing_summ) + 1,] = c('-',
                                                  var(avg.beijing$pollution),
                                                  var(avg.beijing$temp),
                                                  var(avg.beijing$wnd_spd),
                                                  var(avg.beijing$dew))
row.names(avg.beijing_summ)[9] <- 'Variance'
avg.beijing_summ[nrow(avg.beijing_summ) + 1,] = c('-',
                                                  mad(avg.beijing$pollution),
                                                  mad(avg.beijing$temp),
                                                  mad(avg.beijing$wnd_spd),
                                                  mad(avg.beijing$dew))
row.names(avg.beijing_summ)[10] <- 'Median Absolute deviation'
avg.beijing_summ
```
~~~

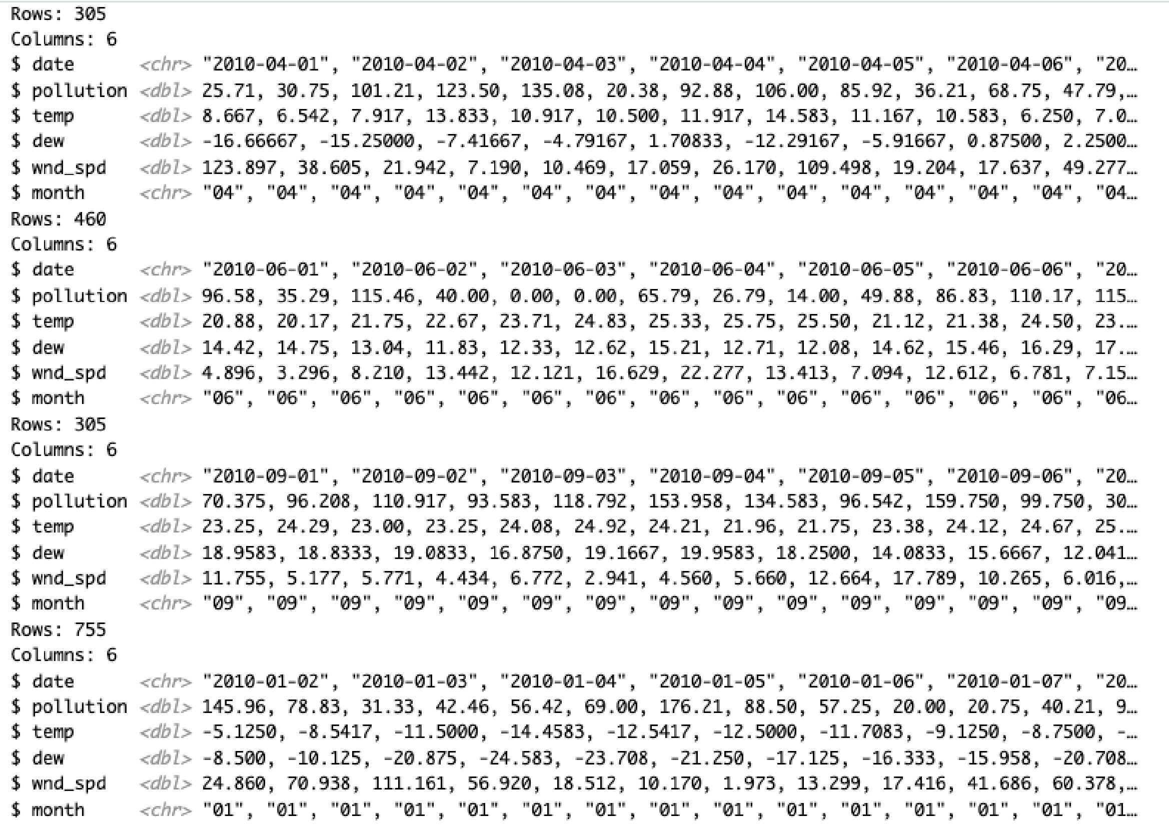
Appendix E

Appendix E.1



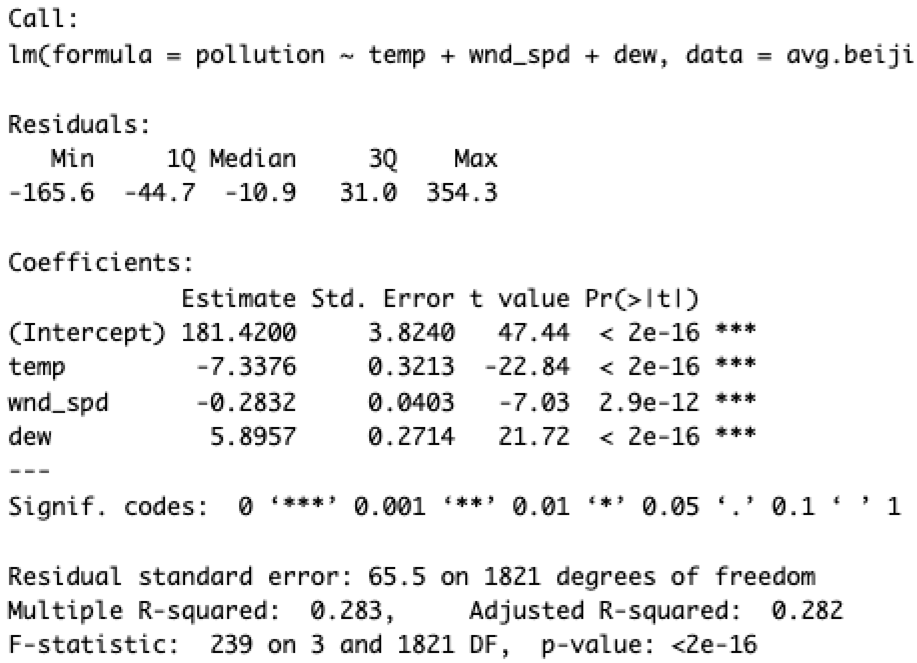
~~~
```{r}
avg.beijing$month<- format(as.Date(avg.beijing$date), "%m")

X <- avg.beijing[avg.beijing$month <= '05',]
spring <- X[X$month >= '04',]

X <- avg.beijing[avg.beijing$month <= '08',]
summer <- X[X$month >= '06',]

X <- avg.beijing[avg.beijing$month <= '10',]
autumn  <- X[X$month >= '09',]

Y <- avg.beijing[avg.beijing$month > '10',]
X <- avg.beijing[avg.beijing$month <= '03',]
winter  <- rbind(X, Y)


glimpse(spring)
glimpse(summer)
glimpse(autumn)
glimpse(winter)
```
~~~

Appendix E.2



```
Rows: 305
Columns: 6
$ date      <chr> "2010-04-01", "2010-04-02", "2010-04-03", "2010-04-04", "2010-04-05", "2010-04-06", "20…
$ pollution <dbl> 25.71, 30.75, 101.21, 123.50, 135.08, 20.38, 92.88, 106.00, 85.92, 36.21, 68.75, 47.79,…
$ temp      <dbl> 8.667, 6.542, 7.917, 13.833, 10.917, 10.500, 11.917, 14.583, 11.167, 10.583, 6.250, 7.0…
$ dew       <dbl> -16.66667, -15.25000, -7.41667, -4.79167, 1.70833, -12.29167, -5.91667, 0.87500, 2.2500…
$ wnd_spd   <dbl> 123.897, 38.605, 21.942, 7.190, 10.469, 17.059, 26.170, 109.498, 19.204, 17.637, 49.277…
$ month     <chr> "04", "04", "04", "04", "04", "04", "04", "04", "04", "04", "04", "04", "04", "04", "04…
Rows: 460
Columns: 6
$ date      <chr> "2010-06-01", "2010-06-02", "2010-06-03", "2010-06-04", "2010-06-05", "2010-06-06", "20…
$ pollution <dbl> 96.58, 35.29, 115.46, 40.00, 0.00, 0.00, 65.79, 26.79, 14.00, 49.88, 86.83, 110.17, 115…
$ temp      <dbl> 20.88, 20.17, 21.75, 22.67, 23.71, 24.83, 25.33, 25.75, 25.50, 21.12, 21.38, 24.50, 23.…
$ dew       <dbl> 14.42, 14.75, 13.04, 11.83, 12.33, 12.62, 15.21, 12.71, 12.08, 14.62, 15.46, 16.29, 17.…
$ wnd_spd   <dbl> 4.896, 3.296, 8.210, 13.442, 12.121, 16.629, 22.277, 13.413, 7.094, 12.612, 6.781, 7.15…
$ month     <chr> "06", "06", "06", "06", "06", "06", "06", "06", "06", "06", "06", "06", "06", "06", "06…
Rows: 305
Columns: 6
$ date      <chr> "2010-09-01", "2010-09-02", "2010-09-03", "2010-09-04", "2010-09-05", "2010-09-06", "20…
$ pollution <dbl> 70.375, 96.208, 110.917, 93.583, 118.792, 153.958, 134.583, 96.542, 159.750, 99.750, 30…
$ temp      <dbl> 23.25, 24.29, 23.00, 23.25, 24.08, 24.92, 24.21, 21.96, 21.75, 23.38, 24.12, 24.67, 25.…
$ dew       <dbl> 18.9583, 18.8333, 19.0833, 16.8750, 19.1667, 19.9583, 18.2500, 14.0833, 15.6667, 12.041…
$ wnd_spd   <dbl> 11.755, 5.177, 5.771, 4.434, 6.772, 2.941, 4.560, 5.660, 12.664, 17.789, 10.265, 6.016,…
$ month     <chr> "09", "09", "09", "09", "09", "09", "09", "09", "09", "09", "09", "09", "09", "09", "09…
Rows: 755
Columns: 6
$ date      <chr> "2010-01-02", "2010-01-03", "2010-01-04", "2010-01-05", "2010-01-06", "2010-01-07", "20…
$ pollution <dbl> 145.96, 78.83, 31.33, 42.46, 56.42, 69.00, 176.21, 88.50, 57.25, 20.00, 20.75, 40.21, 9…
$ temp      <dbl> -5.1250, -8.5417, -11.5000, -14.4583, -12.5417, -12.5000, -11.7083, -9.1250, -8.7500, -…
$ dew       <dbl> -8.500, -10.125, -20.875, -24.583, -23.708, -21.250, -17.125, -16.333, -15.958, -20.708…
$ wnd_spd   <dbl> 24.860, 70.938, 111.161, 56.920, 18.512, 10.170, 1.973, 13.299, 17.416, 41.686, 60.378,…
$ month     <chr> "01", "01", "01", "01", "01", "01", "01", "01", "01", "01", "01", "01", "01", "01", "01…
```

Appendix F



```
Call:
lm(formula = pollution ~ temp + wnd_spd + dew, data = avg.beijing)

Residuals:
   Min     1Q Median     3Q    Max 
-165.6  -44.7  -10.9   31.0  354.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 181.4200     3.8240   47.44  < 2e-16 ***
temp         -7.3376     0.3213  -22.84  < 2e-16 ***
wnd_spd      -0.2832     0.0403   -7.03  2.9e-12 ***
dew           5.8957     0.2714   21.72  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.5 on 1821 degrees of freedom
Multiple R-squared:  0.283,	Adjusted R-squared:  0.282 
F-statistic:  239 on 3 and 1821 DF,  p-value: <2e-16
```

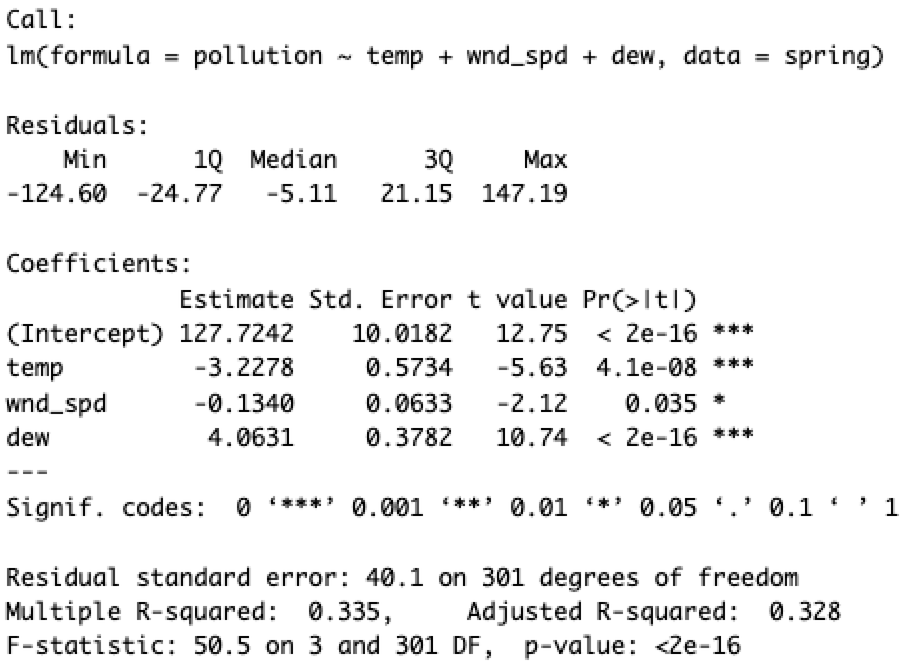
Since we see that all of our variables are statistically significant since they have a p-value below 0.05,we can use this regression model as its statistically significant. Hence the estimates for our parameters are

- $\widehat{\beta}_0 = 181.42$
- $\widehat{\beta}_1 = -7.3376$
- $\widehat{\beta}_2 = -0.2832$
- $\widehat{\beta}_3 = 5.8957$

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 $SSE = (65.5)^2 \cdot (1823 - 2) = 7812545.25$ and that the adjusted $R^2 = 0.282$. We also see that the standard error (SE) for each parameter is

- $SE(\widehat{\beta}_0) = 3.8240$
- $SE(\widehat{\beta}_2) = 0.0403$
- $SE(\widehat{\beta}_1) = 0.3213$
- $SE(\widehat{\beta}_3) = 0.2714$

Appendix G



```
Call:
lm(formula = pollution ~ temp + wnd_spd + dew, data = spring)

Residuals:
    Min      1Q  Median      3Q     Max 
-124.60  -24.77   -5.11   21.15  147.19 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 127.7242    10.0182   12.75  < 2e-16 ***
temp         -3.2278     0.5734   -5.63  4.1e-08 ***
wnd_spd      -0.1340     0.0633   -2.12    0.035 *  
dew           4.0631     0.3782   10.74  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.1 on 301 degrees of freedom
Multiple R-squared:  0.335,     Adjusted R-squared:  0.328 
F-statistic: 50.5 on 3 and 301 DF,  p-value: <2e-16
```

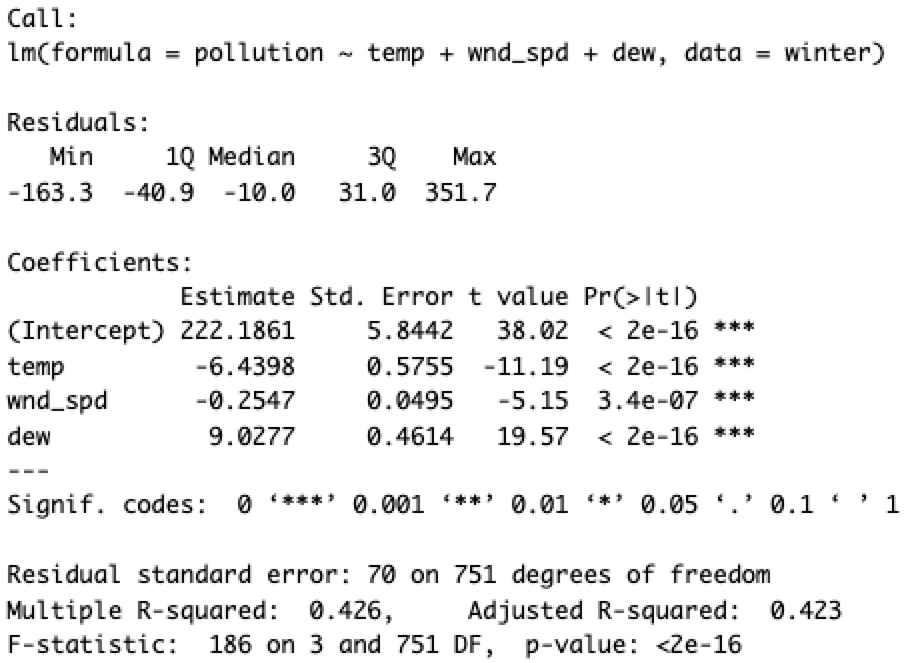
Since we see that all of our variables are statistically significant since they have a p-value below 0.05,we can use this regression model as its statistically significant. Hence the estimates for our parameters are

- $\widehat{\beta}_0 = 127.7242$
- $\widehat{\beta}_1 = -3.2278$
- $\widehat{\beta}_2 = -0.1340$
- $\widehat{\beta}_3 = 4.0631$

We see that the residual standard error is 40.1 with a degrees of freedom of 301, hence the is $SSE = (40.1)^2 \cdot (301) = 484011.01$ and that the adjusted $R^2 = 0.328$. We also see that the standard error (SE) for each parameter is

- $SE(\widehat{\beta}_0) = 10.0182$
- $SE(\widehat{\beta}_1) = 0.5734$
- $SE(\widehat{\beta}_2) = 0.0633$
- $SE(\widehat{\beta}_3) = 0.3782$

Appendix I



```
Call:
lm(formula = pollution ~ temp + wnd_spd + dew, data = winter)

Residuals:
   Min     1Q Median     3Q    Max 
-163.3  -40.9  -10.0   31.0  351.7 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 222.1861     5.8442   38.02  < 2e-16 ***
temp         -6.4398     0.5755  -11.19  < 2e-16 ***
wnd_spd      -0.2547     0.0495   -5.15  3.4e-07 ***
dew           9.0277     0.4614   19.57  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70 on 751 degrees of freedom
Multiple R-squared:  0.426,	Adjusted R-squared:  0.423 
F-statistic:  186 on 3 and 751 DF,  p-value: <2e-16
```

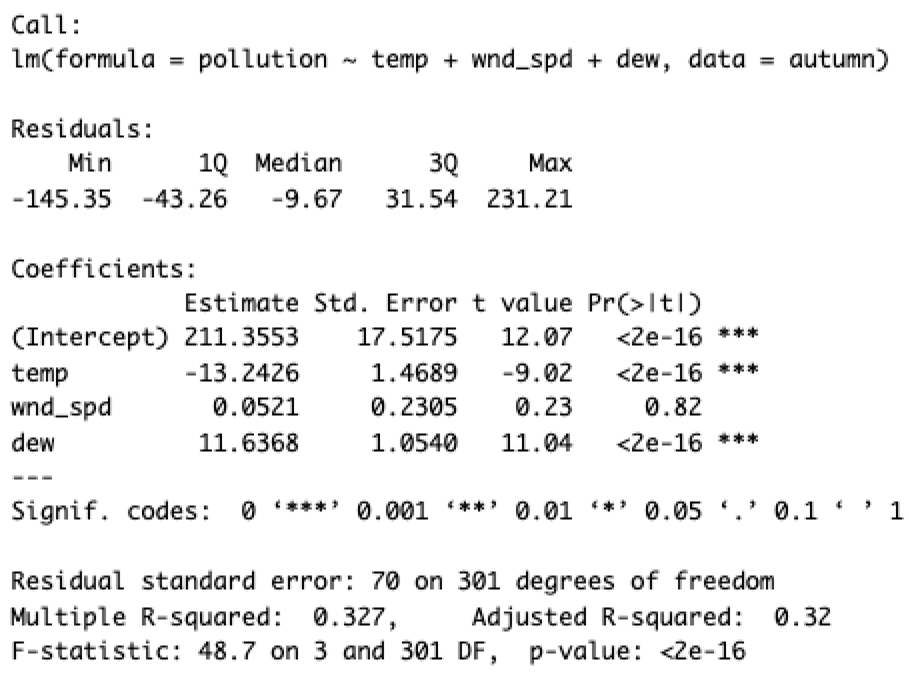
Since we see that all of our variables are statistically significant since they have a p-value below 0.05,we can use this regression model as its statistically significant. Hence the estimates for our parameters are

- $\widehat{\beta_0} = 222.1861$
- $\widehat{\beta_1} = -6.4398$
- $\widehat{\beta_2} = -0.2547$
- $\widehat{\beta_3} = 9.0277$

We see that the residual standard error is 70 with a degrees of freedom of 751, hence the is $SSE = (70)^2 \cdot (751) = 3,679,900$ and that the adjusted $R^2 = 0.423$. We also see that the standard error (SE) for each parameter is

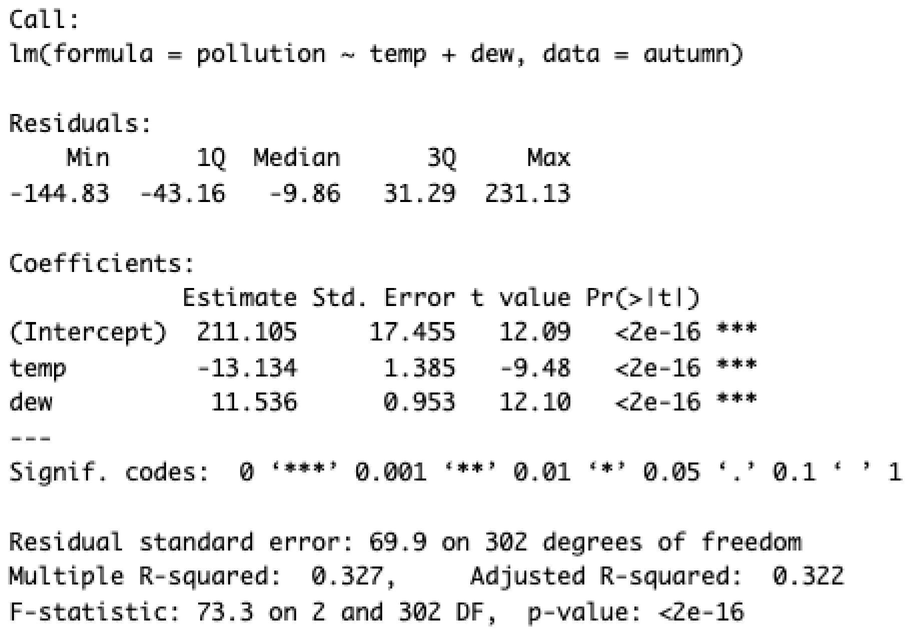
- $SE(\widehat{\beta_0}) = 5.8442$
- $SE(\widehat{\beta_1}) = 0.5755$
- $SE(\widehat{\beta_2}) = 0.0495$
- $SE(\widehat{\beta_3}) = 0.4614$

Appendix J



```
Call:
lm(formula = pollution ~ temp + wnd_spd + dew, data = autumn)

Residuals:
    Min      1Q  Median      3Q     Max 
-145.35  -43.26   -9.67   31.54  231.21 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 211.3553    17.5175   12.07   <2e-16 ***
temp        -13.2426     1.4689   -9.02   <2e-16 ***
wnd_spd       0.0521     0.2305    0.23     0.82    
dew          11.6368     1.0540   11.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70 on 301 degrees of freedom
Multiple R-squared:  0.327,	Adjusted R-squared:  0.32 
F-statistic: 48.7 on 3 and 301 DF,  p-value: <2e-16
```

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



```
Call:
lm(formula = pollution ~ temp + dew, data = autumn)

Residuals:
    Min      1Q  Median      3Q     Max 
-144.83  -43.16   -9.86   31.29  231.13 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  211.105     17.455   12.09   <2e-16 ***
temp         -13.134      1.385   -9.48   <2e-16 ***
dew           11.536      0.953   12.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69.9 on 302 degrees of freedom
Multiple R-squared:  0.327,	Adjusted R-squared:  0.322 
F-statistic: 73.3 on 2 and 302 DF,  p-value: <2e-16
```

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

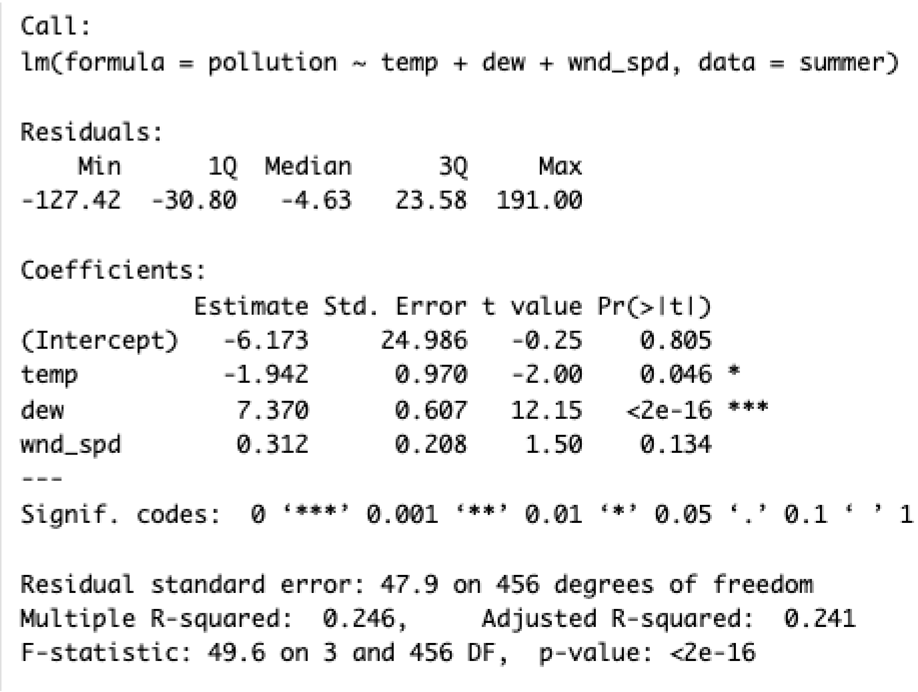
- $\widehat{\beta}_0 = 211.105$
- $\widehat{\beta}_3 = 11.536$

- $\widehat{\beta_1} = \ - 13.134$

We see that the residual standard error is 69.9 with a degrees of freedom of 302, hence the is $SSE\ =\ (69.9)^2 \cdot (302)\ =\ 1475575.02$ and that the adjusted $R^2\ =\ 0.322$. We also see that the standard error (SE) for each parameter is

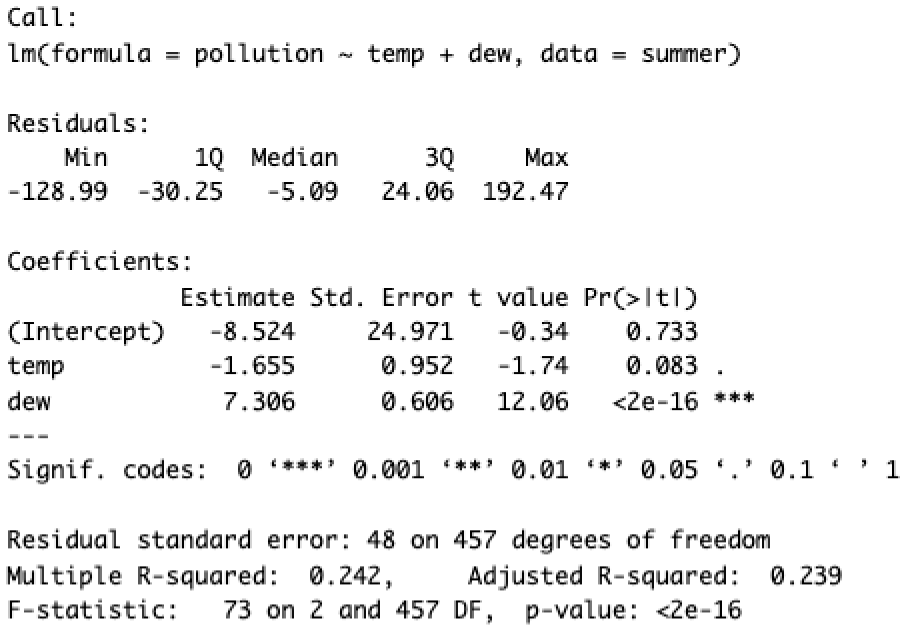
- $SE(\widehat{\beta_0})\ =\ 17.455$
- $SE(\widehat{\beta_1})\ =\ 1.385$
- $SE(\widehat{\beta_3})\ =\ \ 0.953$

Appendix K



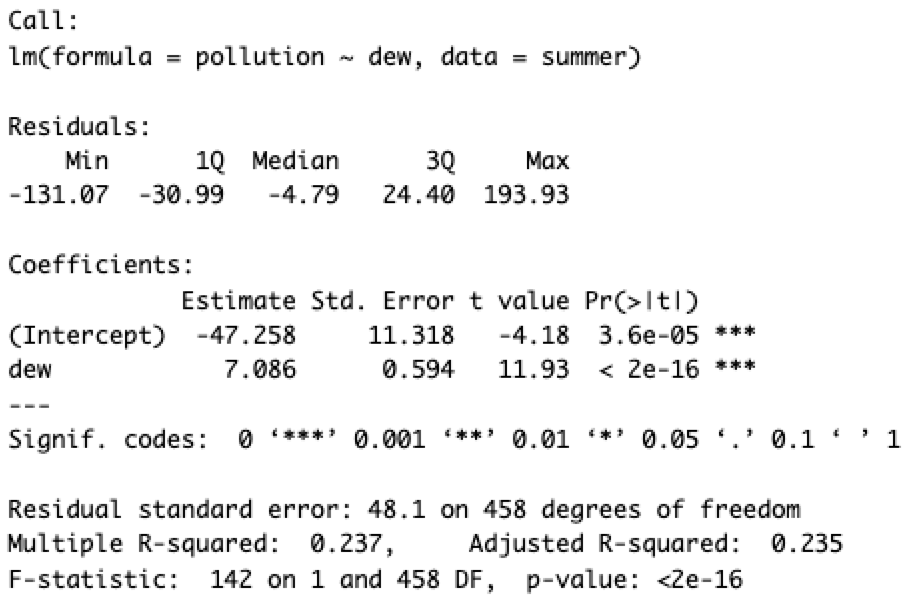
```
Call:
lm(formula = pollution ~ temp + dew + wnd_spd, data = summer)

Residuals:
    Min      1Q  Median      3Q     Max 
-127.42  -30.80   -4.63   23.58  191.00 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -6.173     24.986   -0.25    0.805    
temp          -1.942      0.970   -2.00    0.046 *  
dew            7.370      0.607   12.15   <2e-16 ***
wnd_spd        0.312      0.208    1.50    0.134    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.9 on 456 degrees of freedom
Multiple R-squared:  0.246,	Adjusted R-squared:  0.241 
F-statistic: 49.6 on 3 and 456 DF,  p-value: <2e-16
```

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



```
Call:
lm(formula = pollution ~ temp + dew, data = summer)

Residuals:
    Min      1Q  Median      3Q     Max 
-128.99  -30.25   -5.09   24.06  192.47 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -8.524     24.971   -0.34    0.733    
temp          -1.655      0.952   -1.74    0.083 .  
dew            7.306      0.606   12.06   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48 on 457 degrees of freedom
Multiple R-squared:  0.242,	Adjusted R-squared:  0.239 
F-statistic:   73 on 2 and 457 DF,  p-value: <2e-16
```

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.



```
Call:
lm(formula = pollution ~ dew, data = summer)

Residuals:
    Min      1Q  Median      3Q     Max 
-131.07  -30.99   -4.79   24.40  193.93 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -47.258     11.318   -4.18  3.6e-05 ***
dew            7.086      0.594   11.93  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.1 on 458 degrees of freedom
Multiple R-squared:  0.237,	Adjusted R-squared:  0.235 
F-statistic:  142 on 1 and 458 DF,  p-value: <2e-16
```

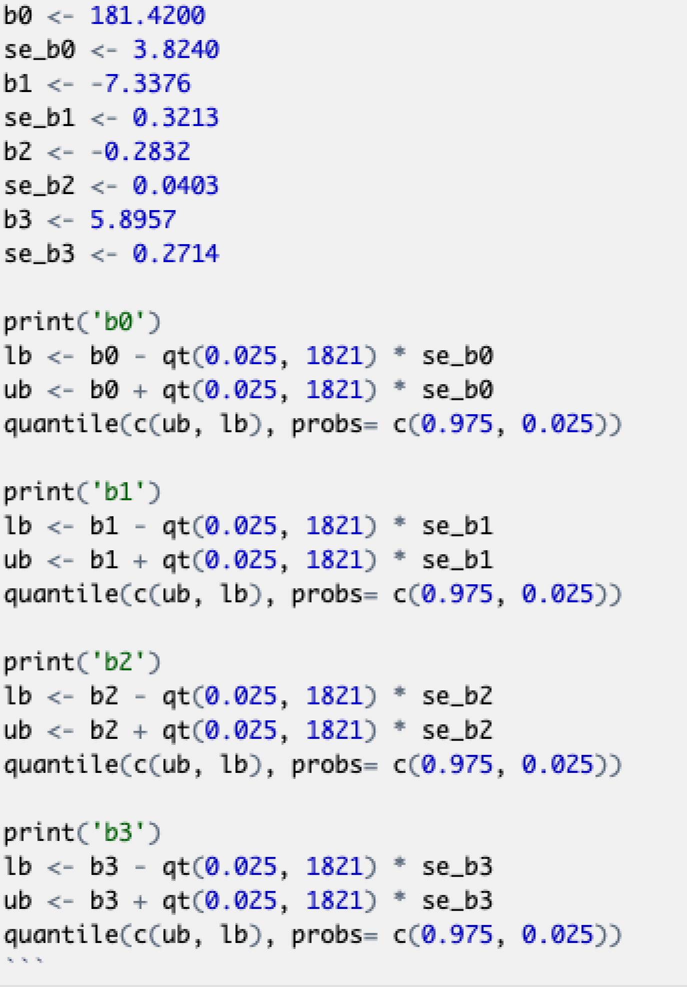
When disregarding the parameters wnd_spd and temp, we see that all of our variables are statistically significant, hence we can use this regression model as its statistically significant. The estimates for our parameters are

- $\widehat{\beta}_0 = -47.258$
- $\widehat{\beta}_3 = 7.086$

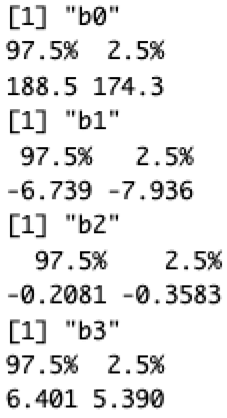
We see that the residual standard error is 69.9 with a degrees of freedom of 302, hence the is $SSE = (48.1)^2 \cdot (458) = 1,059,633.38$ and that the adjusted $R^2 = 0.235$. We also see that the standard error (SE) for each parameter is

- $SE(\widehat{\beta}_0) = 11.318$
- $SE(\widehat{\beta}_3) = 0.594$

Appendix L



~~~
b0 <- 181.4200
se_b0 <- 3.8240
b1 <- -7.3376
se_b1 <- 0.3213
b2 <- -0.2832
se_b2 <- 0.0403
b3 <- 5.8957
se_b3 <- 0.2714

print('b0')
lb <- b0 - qt(0.025, 1821) * se_b0
ub <- b0 + qt(0.025, 1821) * se_b0
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b1')
lb <- b1 - qt(0.025, 1821) * se_b1
ub <- b1 + qt(0.025, 1821) * se_b1
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b2')
lb <- b2 - qt(0.025, 1821) * se_b2
ub <- b2 + qt(0.025, 1821) * se_b2
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b3')
lb <- b3 - qt(0.025, 1821) * se_b3
ub <- b3 + qt(0.025, 1821) * se_b3
quantile(c(ub, lb), probs= c(0.975, 0.025))
```
~~~



```
[1] "b0"
97.5%  2.5%
188.5 174.3
[1] "b1"
 97.5%   2.5%
-6.739 -7.936
[1] "b2"
  97.5%    2.5%
-0.2081 -0.3583
[1] "b3"
97.5%  2.5%
6.401 5.390
```

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

$$b_x \pm t_{df,\alpha/2} \cdot SE(\beta_x)$$

to find the confidence interval where b_x is the estimated beta value $SE(\beta_x)$ is the standard error of β_x and $t_{df,\alpha/2}$ 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 $\alpha = 0.05$ our critical T value is

$$t_{df,\alpha/2} = t_{1821,\,0.025} \approx -1.961$$

as calculated in R.

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

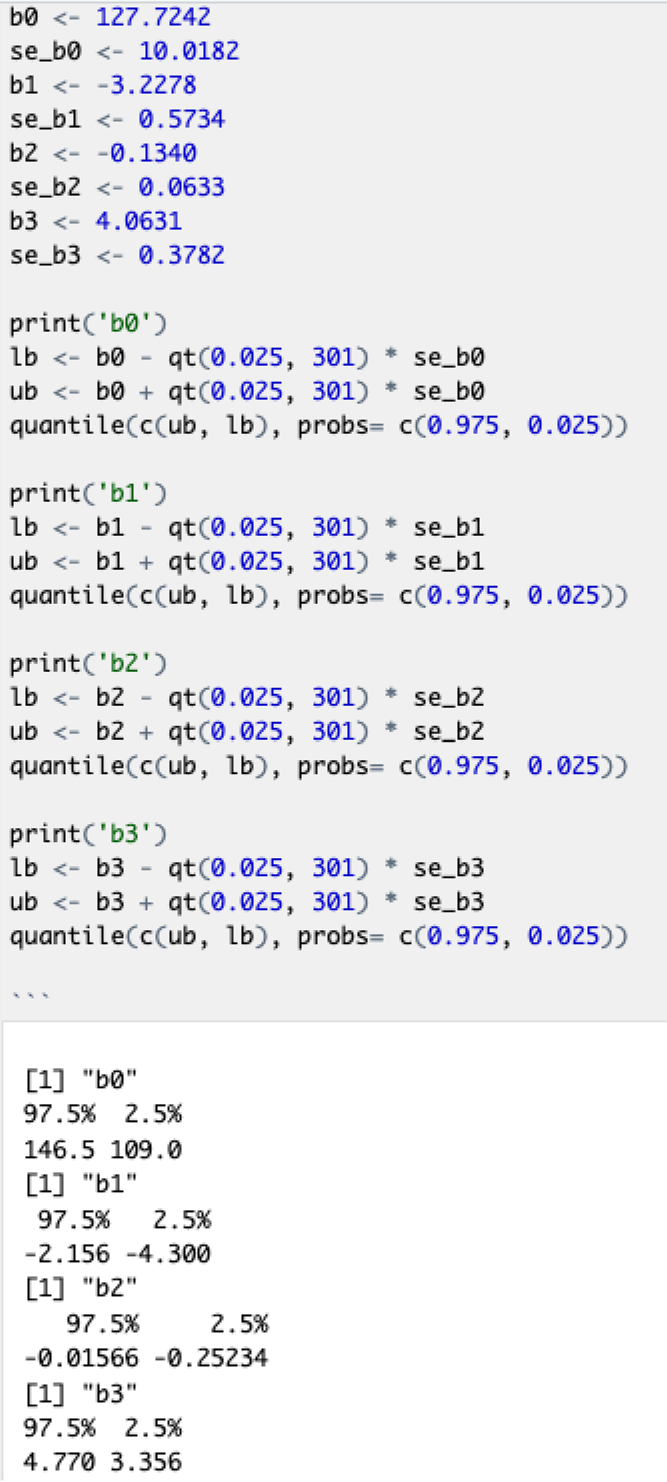
CI for b_0 is $(181.42 - [-1.961] \cdot 3.8240$, $(181.42 + 1.961 \cdot 3.8240)$ which results in an interval of $(174.3, 188.5)$.

CI for b_1 is $(-7.3376 - [-1.961] \cdot 0.3213, -7.3376 + 1.961 \cdot 0.3213)$ which results in an interval of $(-7.936, -6.739)$.

CI for b_2 is $(-0.2832 - [-1.961] \cdot 0.0403, -0.2832 + 1.961 \cdot 0.0403)$ which results in an interval of $(-0.3583, -0.2081)$.

CI for b_3 is $(5.8957 - [-1.961] \cdot 0.2714, 5.8957 + 1.961 \cdot 0.2714)$ which results in an interval of $(5.390, 6.401)$.

Appendix M



```
b0 <- 127.7242
se_b0 <- 10.0182
b1 <- -3.2278
se_b1 <- 0.5734
b2 <- -0.1340
se_b2 <- 0.0633
b3 <- 4.0631
se_b3 <- 0.3782

print('b0')
lb <- b0 - qt(0.025, 301) * se_b0
ub <- b0 + qt(0.025, 301) * se_b0
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b1')
lb <- b1 - qt(0.025, 301) * se_b1
ub <- b1 + qt(0.025, 301) * se_b1
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b2')
lb <- b2 - qt(0.025, 301) * se_b2
ub <- b2 + qt(0.025, 301) * se_b2
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b3')
lb <- b3 - qt(0.025, 301) * se_b3
ub <- b3 + qt(0.025, 301) * se_b3
quantile(c(ub, lb), probs= c(0.975, 0.025))

```
```

```
[1] "b0"
97.5%  2.5% 
146.5 109.0 
[1] "b1"
 97.5%   2.5% 
-2.156 -4.300 
[1] "b2"
   97.5%     2.5% 
-0.01566 -0.25234 
[1] "b3"
97.5%  2.5% 
4.770 3.356 
```

We will use the following the equation

$$b_x \pm t_{df,\alpha/2} \cdot SE(\beta_x)$$

to find the confidence interval where b_x is the estimated beta value $SE(\beta_x)$ is the standard error of β_x and $t_{df,\alpha/2}$ 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 $\alpha = 0.05$ our critical T value is

$$t_{df,\alpha/2} = t_{301,0.025} \approx -1.968$$

as calculated in R.

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

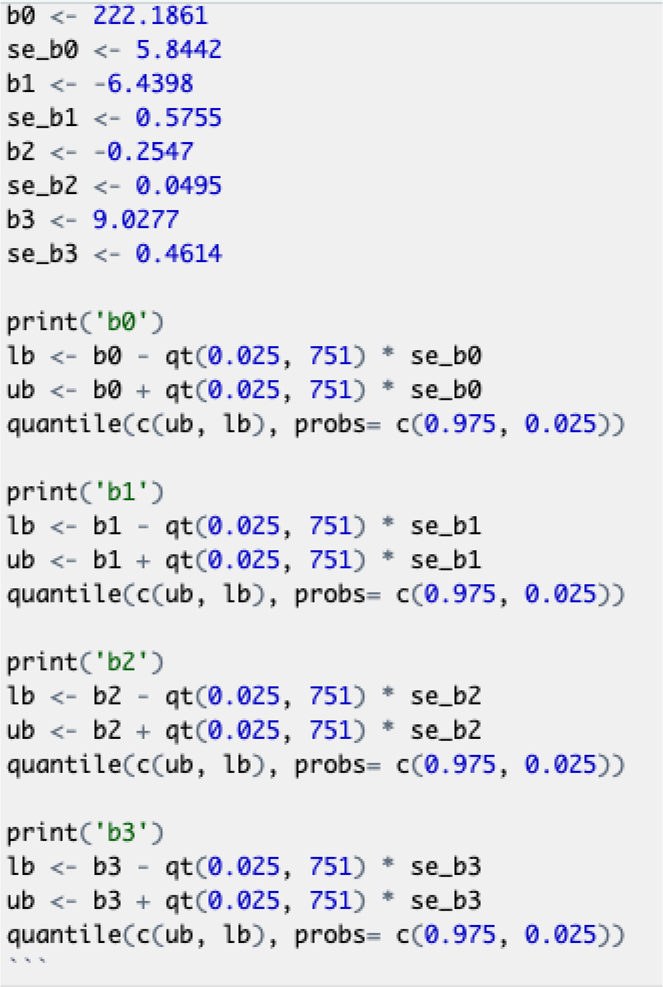
CI for $b_0$ is $(127.7242\ -\ [-\ 1.968]\ \cdot\ 10.0182\ ,\ 127.7242\ +\ 1.968\ \cdot\ 10.0182)$ which results in an interval of $(109.0,\ 146.5\ )$.

CI for $b_1$ is $(-\ 3.2278\ -\ [-\ 1.968]\ \cdot\ 0.5734\ ,\ -\ 3.2278\ +\ 1.968\ \cdot\ 0.5734)$ which results in an interval of $(-\ 4.300,\ \ -\ 2.156\ )$.

CI for $b_2$ is $(-\ 0.1340\ -\ [-\ 1.968]\ \cdot\ 0.0633,\ \ -\ 0.1340\ +\ 1.968\ \cdot\ 0.0633)$ which results in an interval of $(-\ 0.25234,\ \ -\ 0.01566\ )$.

CI for $b_3$ is $(4.0631\ -\ [-\ 1.968]\ \cdot\ 0.3782\ ,\ 4.0631\ +\ 1.968\ \cdot\ 0.3782)$ which results in an interval of $(\ 3.356,\ 4.770)$.

Appendix N

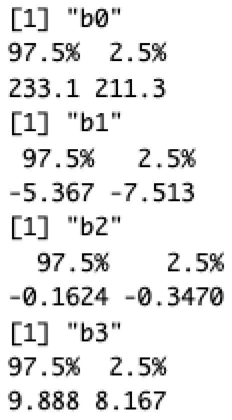
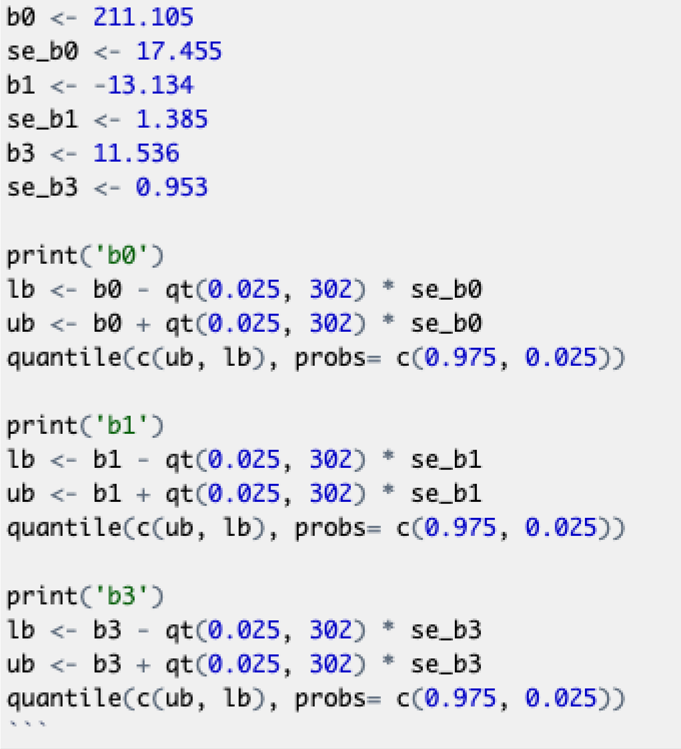
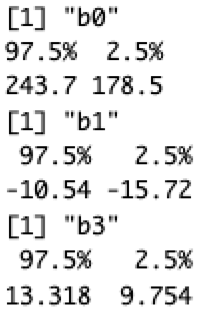
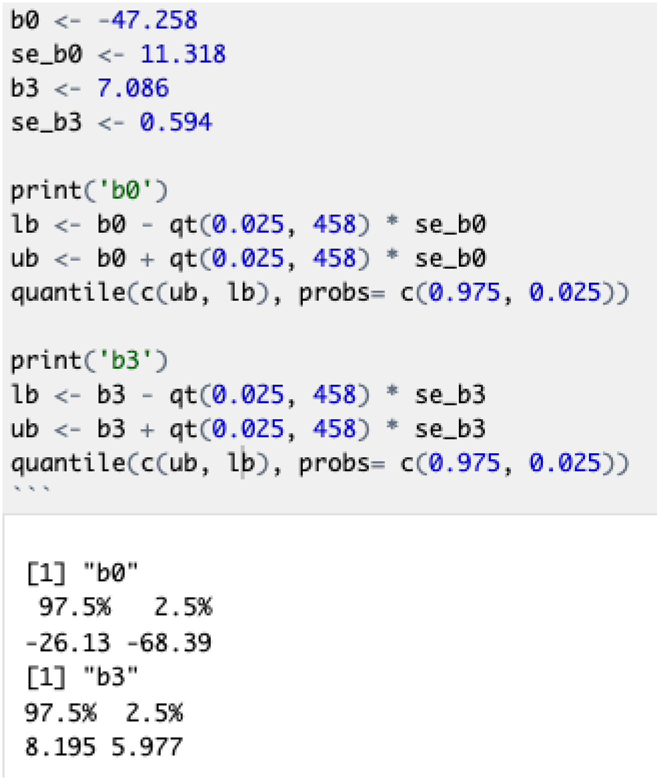
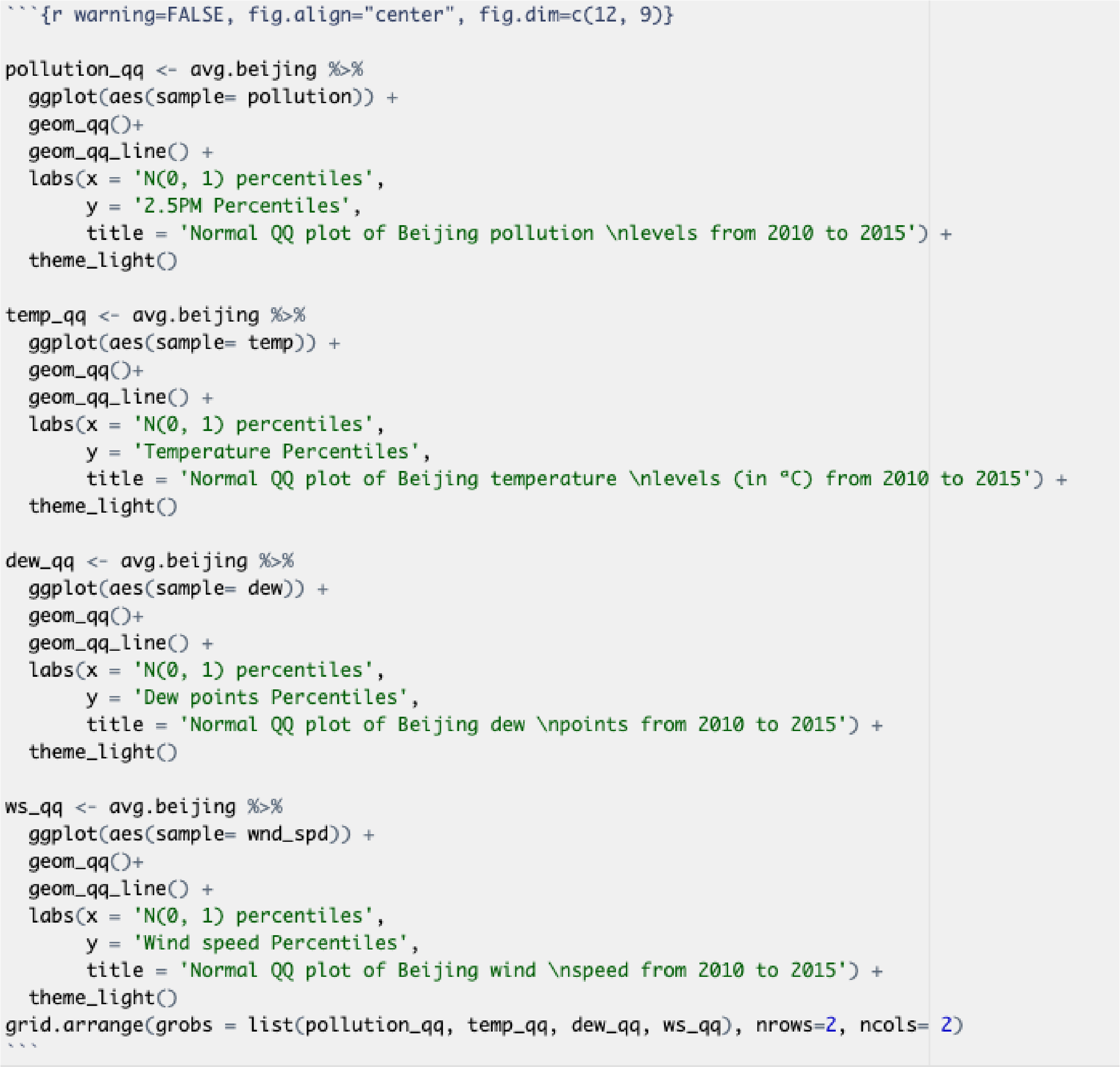


~~~~
b0 <- 222.1861
se_b0 <- 5.8442
b1 <- -6.4398
se_b1 <- 0.5755
b2 <- -0.2547
se_b2 <- 0.0495
b3 <- 9.0277
se_b3 <- 0.4614

print('b0')
lb <- b0 - qt(0.025, 751) * se_b0
ub <- b0 + qt(0.025, 751) * se_b0
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b1')
lb <- b1 - qt(0.025, 751) * se_b1
ub <- b1 + qt(0.025, 751) * se_b1
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b2')
lb <- b2 - qt(0.025, 751) * se_b2
ub <- b2 + qt(0.025, 751) * se_b2
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b3')
lb <- b3 - qt(0.025, 751) * se_b3
ub <- b3 + qt(0.025, 751) * se_b3
quantile(c(ub, lb), probs= c(0.975, 0.025))
```
~~~~



```
[1] "b0"
97.5%  2.5%
233.1 211.3
[1] "b1"
 97.5%   2.5%
-5.367 -7.513
[1] "b2"
  97.5%    2.5%
-0.1624 -0.3470
[1] "b3"
97.5%  2.5%
9.888 8.167
```

We will use the following the equation

$$b_x \pm t_{df,\alpha/2} \cdot SE(\beta_x)$$

to find the confidence interval where b_x is the estimated beta value $SE(\beta_x)$ is the standard error of β_x and $t_{df,\alpha/2}$ 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 $\alpha = 0.05$ our critical T value is

$$t_{df,\alpha/2} = t_{751,\,0.025} \approx -1.968$$

as calculated in R.

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

CI for b_0 is $(222.1861 - [-1.963] \cdot 5.8442,\ 222.1861 + 1.963 \cdot 5.8442)$ which results in an interval of $(211.3,\ 233.1)$.

CI for b_1 is $(-6.4398 - [-1.963] \cdot 0.5755,\ -6.4398 + 1.963 \cdot 0.5755)$ which results in an interval of $(-7.513,\ -5.367)$.

CI for b_2 is $(-0.2547 - [-1.963] \cdot 0.0495,\ -0.2547 + 1.963 \cdot 0.0495)$ which results in an interval of $(-0.3470,\ -0.1624)$.

CI for b_3 is $(9.0277 - [-1.963] \cdot 0.4614,\ 9.0277 + 1.963 \cdot 0.4614)$ which results in an interval of $(8.167,\ 9.888)$.

Appendix O



~~~
b0 <- 211.105
se_b0 <- 17.455
b1 <- -13.134
se_b1 <- 1.385
b3 <- 11.536
se_b3 <- 0.953

print('b0')
lb <- b0 - qt(0.025, 302) * se_b0
ub <- b0 + qt(0.025, 302) * se_b0
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b1')
lb <- b1 - qt(0.025, 302) * se_b1
ub <- b1 + qt(0.025, 302) * se_b1
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b3')
lb <- b3 - qt(0.025, 302) * se_b3
ub <- b3 + qt(0.025, 302) * se_b3
quantile(c(ub, lb), probs= c(0.975, 0.025))
```
~~~



```
[1] "b0"
97.5%  2.5%
243.7 178.5
[1] "b1"
 97.5%   2.5%
-10.54 -15.72
[1] "b3"
 97.5%   2.5%
13.318  9.754
```

We will use the following the equation

$$b_x \pm t_{df,\alpha/2} \cdot SE(\beta_x)$$

to find the confidence interval where b_x is the estimated beta value $SE(\beta_x)$ is the standard error of β_x and $t_{df,\alpha/2}$ 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 $\alpha = 0.05$ our critical T value is

$$t_{df,\alpha/2} = t_{302,\,0.025} \approx -1.968$$

as calculated in R.

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

CI for b_0 is $(211.105 - [-1.968] \cdot 17.455,\ 211.105 + 1.968 \cdot 17.455)$ which results in an interval of $(178.5,\ 243.7\ )$.

CI for b_1 is $(-13.134 - [-1.968] \cdot 1.385,\ -13.134 + 1.968 \cdot 1.385)$ which results in an interval of $(-15.72,\ -10.54\ )$.

CI for b_3 is $(11.536 - [-1.968] \cdot 0.953,\ 11.536 + 1.968 \cdot 0.953)$ which results in an interval of $(\ 9.754,\ 13.318)$.

Appendix P



~~~~
b0 <- -47.258
se_b0 <- 11.318
b3 <- 7.086
se_b3 <- 0.594

print('b0')
lb <- b0 - qt(0.025, 458) * se_b0
ub <- b0 + qt(0.025, 458) * se_b0
quantile(c(ub, lb), probs= c(0.975, 0.025))

print('b3')
lb <- b3 - qt(0.025, 458) * se_b3
ub <- b3 + qt(0.025, 458) * se_b3
quantile(c(ub, lb), probs= c(0.975, 0.025))
```
~~~~

```
[1] "b0"
 97.5%   2.5%
-26.13 -68.39
[1] "b3"
97.5%  2.5%
8.195 5.977
```

We will use the following the equation

$$b_x \pm t_{df,\alpha/2} \cdot SE(\beta_x)$$

to find the confidence interval where b_x is the estimated beta value $SE(\beta_x)$ is the standard error of β_x and $t_{df,\alpha/2}$ 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 $\alpha = 0.05$ our critical T value is

$$t_{df,\alpha/2} = t_{458,\,0.025} \approx -1.965$$

as calculated in R.

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

CI for b_0 is $(-47.258 - [-1.965] \cdot 11.318,\ -47.258 + 1.965 \cdot 11.318)$ which results in an interval of $(-68.39,\ -26.13)$.

CI for b_3 is $(7.086 - [-1.965] \cdot 0.594,\ 7.086 + 1.965 \cdot 0.594)$ which results in an interval of $(5.977,\ 8.195)$.

Appendix Q



~~~
```{r warning=FALSE, fig.align="center", fig.dim=c(12, 9)}

pollution_qq <- avg.beijing %>%
  ggplot(aes(sample= pollution)) +
  geom_qq()+
  geom_qq_line() +
  labs(x = 'N(0, 1) percentiles',
       y = '2.5PM Percentiles',
       title = 'Normal QQ plot of Beijing pollution \nlevels from 2010 to 2015') +
  theme_light()

temp_qq <- avg.beijing %>%
  ggplot(aes(sample= temp)) +
  geom_qq()+
  geom_qq_line() +
  labs(x = 'N(0, 1) percentiles',
       y = 'Temperature Percentiles',
       title = 'Normal QQ plot of Beijing temperature \nlevels (in °C) from 2010 to 2015') +
  theme_light()

dew_qq <- avg.beijing %>%
  ggplot(aes(sample= dew)) +
  geom_qq()+
  geom_qq_line() +
  labs(x = 'N(0, 1) percentiles',
       y = 'Dew points Percentiles',
       title = 'Normal QQ plot of Beijing dew \npoints from 2010 to 2015') +
  theme_light()

ws_qq <- avg.beijing %>%
  ggplot(aes(sample= wnd_spd)) +
  geom_qq()+
  geom_qq_line() +
  labs(x = 'N(0, 1) percentiles',
       y = 'Wind speed Percentiles',
       title = 'Normal QQ plot of Beijing wind \nspeed from 2010 to 2015') +
  theme_light()
grid.arrange(grobs = list(pollution_qq, temp_qq, dew_qq, ws_qq), nrows=2, ncols= 2)
```
~~~

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.