The Fine Dust Level in Seoul And Its Relation to Potential Factors: 2005-2013

Professor Jevin D. West
Data Science and Management
Date: 15/07/15
Group Name: Dust of Data Science
Group Members: Seojin Han, Seungwoo Lee, Insu Kim, Ju Ho Yoon

Abstract:

The rising issue with the fine dust level in Seoul became more and more important into the daily lives of the Seoul citizens. Having to check the daily fine dust level from the web like we check the weather forecast and having to wear masks for certain days tells how the citizens of Seoul are having to suffer everyday. Hence, our team looked into the fine dust levels in Seoul from 2005 to 2013 and studied which of the potential factors explained the most and the least to the trend. After running the subset selection function with R, we were able to eliminate some of the factors that we initially thought might affect the fine dust level and with the regression test and t-test, we were able to discover that the number of subway users in Seoul showed the highest r^2 value of 0.8521, and the following were the area of green belt in Seoul with the r^2 value of 0.6795 and the number of registered automated vehicles in Seoul with the r^2 value of 0.5602. More interestingly, we thought that the nuclear energy might be the solution to decrease the level of fine dust; however, we found out that the nuclear energy has almost no effect to the current trend of fine dust levels. After our findings, we conclude that more research needs to be done on the nuclear energy to see if it can contribute to mitigating the fine dust in Seoul.

Introduction:

Recently in Korea, there has been a rise of issue regarding the rising level of fine dusts. According to The Korea Herald, they say that at least 1 in 6 dies early from fine dust in Korea. We examine the seriousness of this rising problem and would like to conduct a study regarding this issue. Some say that the cause of such phenomenon is due to the influence of China, and others say that it is from the increase in the coal usage. There are many rising debates regarding this issue and our group intends to get on this underlying cause using the sets of data we have found. Our main set of data comprises of $\mu g/m^3$ of fine dust level in Seoul from 2005 to 2013 found from KOSIS, Korean Statistical Information Service. With that data as our benchmark, we would like to relate different suspects of factors such as the level of coal power generation from 2005 to 20013 in Seoul, nuclear energy generation level from 2005 to 2013 in Seoul, number of registered cars from 2005 to 2013 in Seoul, length of bike road facility in Seoul from 2005 to 2013, number of bicycle facilities from 2005 to 2013 in Seoul, number of subway users in Seoul from 2005 to 2013, the GRDP level in Seoul from 2005 to 2013, and the area of greenbelt in Seoul from 2005 to 2013. *From these data sets we have found, we would like to see what factors are the most relevant to this phenomenon and infer what potential actions the citizens of Seoul should implement to qualify the effects of rising level of fine dusts in Korea.*

We are going to try and answer our question by first looking into the various data regarding the potential factors that might affect the levels of fine dust level in Korea. We intend to gather the various data and conduct a regression test, and a t-test to find out the correlations to test our internal and external validities. After looking into the R values and the P-values of each potential factors, we would like to compile a simple graph showing

the diverse effects of the potential factors and examine which potential factor has the highest correlation with the fine dust level across the time period of 2005 to 2013 in Seoul, Korea.

Our potential limitations for our study are that although we would like to control for all of the other unthought-of factors, we might simply miss them. In coming up with our factors, most of them were from our common hearings and from the news articles regarding the issue. Another potential limitation is that we might not have sufficient data of the factors with identical time interval; for instance, one data is given in months and others might be given in years. We would try to minimize this drawback by summing up the months or weeks into a yearly unit and try to unify our measures as much as possible. Lastly, although we may infer correlations, we may not be able to infer causation in that testing for causation would involve another structure of study and we hope it be studied in the future.

Studies have been conducted previously regarding the fine dusts. In 1996, Tong and his colleagues conducted a study explored the effect of Asian dusts events on daily mortality in Seoul during the period of 1995-1998. Their study provided with weak association between the Asian dusts events and risk of death from all causes. In regards to this previous study, we would like to get at the 2011-2013 time frame, when the similar issue with the fine dusts in Korea was prominent.

In 2015, professors from Inha University and Ajou University have looked into the fine dust levels in regards to the death rates in Korea. They were able to find that air pollution caused 15.9% of total mortality per year in Seoul, Korea. They also took to look into what symptoms increased over the course of the higher fine dust level. Their marquee finding was that in every six people in Seoul, one would die early due to the effects of fine dusts. The study was crucial in that it looked into the effects of the fine dusts to the Korean society. In contrast to this study, our team would like to compare diverse potential factors of causes of fine dusts in regards to its statistical feasibility. After such remarkable finding by the scholars, our team would like to consider the various causes of fine dusts and develop a simple plan to warn what factors to keep in mind and evade if possible.

Methods:

In coming up with our analysis, first we have collected the various data sets from the KOSIS, the Korean Statistical Information Service. From each of the data sets we have collected, we organized all the data into an excel and have fined them into the time line of 2005 to 2013 for all of the data we collected since other years were not all present. After organizing them into a one collective data, we have set the level of fine dusts in Seoul as our dependent variable, and set the potential factors as our independent variables (x1, x2, x3…etc). In doing so, we realize that the levels of different factors might show certain trends due to other lurking variables or other factors. Hence, to control for such issue, we have generally controlled for the potential factors by dividing each of the factors to the

corresponding population of Seoul for every year in excel as displayed in *Figure A* of our appendices.

Our null hypothesis is that we expect to see no correlation in any of the factors to the fine dust level in Seoul from 2005 to 2013, whereas our alternate hypothesis will be that we expect to see a correlation in any of the factors to the fine dust level in Seoul from 2005 to 2013. We would like to build upon testing our hypothesis by first using 'R' to compare and contrast the different correlation values in each of the factors. Hence, we have exported our data into R as *Figure B* in our appendices.

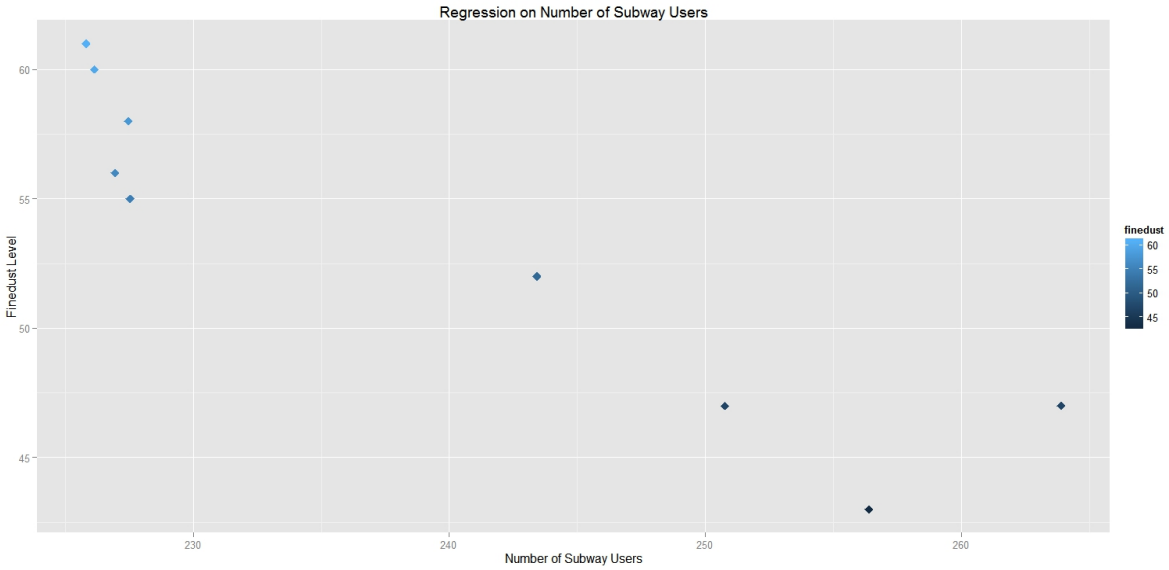
To brief, 'finedust' indicates the $\mu g/m^3$ of fine dust level, 'car_regist' indicates the number of registered cars divided by the population of Seoul, 'bikeroad_lg' indicates the km of bike roads divided by the population of Seoul, 'bikeroad_nu' indicates the number of bike facility divided by the population of Seoul, 'sub_users' indicates the number of subway users divided by the population of Seoul, 'GRDP' indicates the GRDP level divided by the population of Seoul, 'tree' indicates the number of street trees divided by the population of Seoul, 'greenbelt' indicates the m^2 of greenbelt area divided by the population of Seoul, 'coalpower' indicates the mil.kwh of coal power generation divided by the population of Seoul, 'nuclearenergy' indicates the level of nuclear energy divided by the population of Seoul, and lastly, 'construction' indicates the city's expenditure in Wons for construction divided by the population of Seoul.

In doing so, we first determined that with R's subset selection function, our data would look more fined with excluding two of our potential factors: coal power generation level in Seoul and the construction expenditure amount in Seoul as displayed in the *Figure C* of our appendices. After, we have discovered that using only 6 of the factors to run our regression test will return us a finer plot by running the regsubsets function in R and were able to run a regression test with all of the 6 independent variables to see the holistic model as shown in the *Figure D* of our appendices. Afterwards, we ran the regression test for each of our 6 independent variables individually and discovered our findings by comparing the results we found with the 2005 to 2013 trend of the fine dust levels displayed in *Figure E* and reported our results and discussions in the section below.

Results and Discussions:

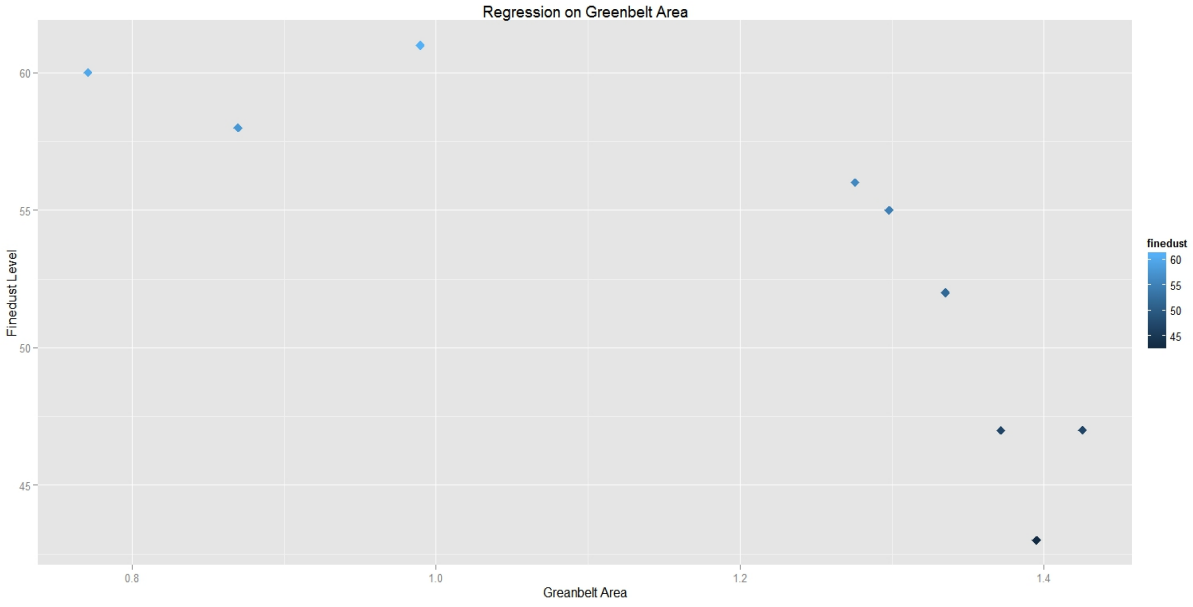
After we tested for regression values, standard errors, t-tests, and p-values on each of the 6 independent variables we sorted out using the subset selection, we were able to find out that number of subway users in Seoul showed the highest r^2 value of 0.8521, and the following were the area of green belt in Seoul with the r^2 value of 0.6795 and the number of registered automated vehicles in Seoul with the r^2 value of 0.5602. The following ggplots are displayed in the sections below for the three factors.

Figure 1:



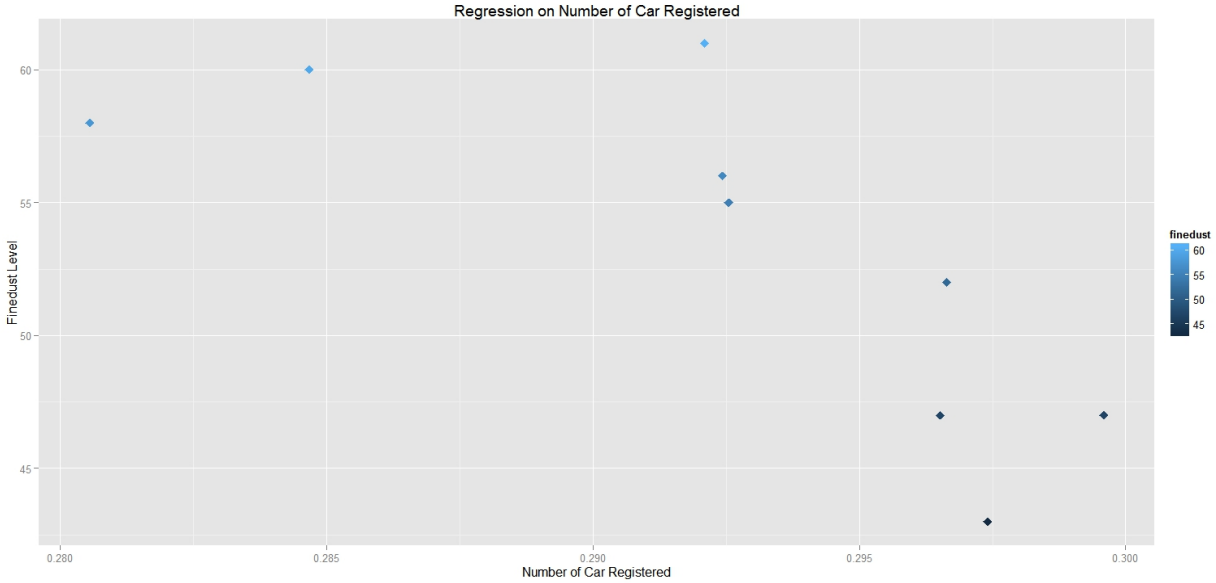
The following figure displays the effects of the number of subways users to the fine dust level in Seoul with the r^2 value of 0.8521. As shown from the values in *Figure 1A* in our appendices, we found out that the number of subway users is significant even at the near 0 alpha level, with the p-value of 0.0003855. More so, we found out that the number of subway users dramatically increased from 2,277,298,000 to 2,619,529,000. This increase in the number of subway users, however, contributes the most to the decreasing trend of fine dust level in Seoul. Given this result, we infer that the increase in subway users contribute to impeding the speed of increase in registered cars, therefore decreasing the displacement of gas. Hence, we would like to report that the upward trend of number of subway users is a significant contributor to the decreasing trend to the fine dust level in Seoul.

Figure 2:



The following figure displays the effects of the area of green belt district to the fine dust level in Seoul with the r^2 value of 0.6795. As shown from the values in *Figure 2A* in our appendices, we found out that the area of green belt is significant at the 0.01 alpha level, with the p-value of 0.006271. Given this result, we believe that as the area of green belt districts are increasing given a constant area of Seoul, the level of pollution will decrease hence act to decrease the level of fine dusts in the city. Later, according to study done by the department of rural development, we found out that approximately 2% of air pollution will be cleansed by the green belt. Thus, we would like to report that the increase in the area green belt would likely contribute to the decrease in the fine dust level in Seoul.

Figure 3:



The following figure displays the effects of the number of registered cars to the fine dust level in Seoul with the r^2 value of 0.5602. As shown from the values in *Figure 3A* in our appendices, we found out that the number of registered cars is significant at the 0.05 alpha level, with the p-value of 0.02033. It is to our knowledge that the rising number of cars will cause more cars to displace more CO2 in the atmosphere. Moreover, we found out that the friction in tires causes fine dusts level to increase through a study conducted by two professors from Doshisha University in 2013. However, the span of the increase in registered cars is seen not to have increased as much as the subway users. We infer that the citizens have begun to shift to taking the subway due to the Subway Expansion Policy, and the rising traffic in the roads. Hence, we would like to report that this shifting to taking subways contributes to this display of effect.

Conclusions and Future Direction

So far, we have found out that the fine dust level in Seoul is rather decreasing throughout the years of 2005 to 2013, and that the number of subway users, the area of green belt districts, and the number of registered cars could be contributing to such phenomenon in regards to our findings in data analysis and research. To sum, we were able to eliminate

the level of coal generation, expenditure on construction, number of street trees planted, and GRDP value to fine our independent variables better to see the effects to the fine dust levels and were able to find out that the number of subway users, area of green belt districts, and the number of registered cars contributed the most to the trend in the fine dust level in Seoul, respectively. On the other hand, length of bike road, number of bike facilities, and nuclear energy generation level showed weak r^2 values all under 0.5.
In the future, we would like to encourage more research be done on how to actually decrease the level of fine dust using our study as a stepping-stone. From our findings, we believe that more research on the nuclear energy needs to be done to see if it can contribute to mitigating the fine dust level. More so, we hope to raise the level of awareness of fine dusts locally, and also globally accounting the fact that only few studies have been conducted regarding this issue.

Appendices

Figure A:

edust(㎍/㎥)	car registration	bikeroad(len,km)	bikeroad(num)	subway users	GRDP(bil,won)	street tree	green belt(㎡)	coal power generation(mil.kwh)	nuclear energy generation	Construction Expenditure	Population
76	2691431	554	271	2230783000				118,022	119,103		10041502
69	2776536	587	281	2249226000	194891			120,277	129,672		10029787
61	2779841	616	282	2300735000	198925	276779	8185144	127,164	130,714		10036241
58	2808771	629	297	2277298000	208899	279461	8708112	133,658	146,779	11240735	10011324
60	2856857	648	313	2269410000	231223	280243	7738173	139,205	148,749	14840812	10035377
61	2933286	715	358	2267676000	249484	280499	9939148	154,674	142,937	17405539	10042096
55	2949211	728	358	2293848000	262999	279442	13085443	173,508	150,958	21242916	10081017
56	2954704	764	390	2293042000	273198	279672	12888759	193,216	147,771	18982632	10103872
52	2981400	844	422	2446519000	289718	283609	13418116	193,769	148,596	17587044	10050508
47	2977599	804	399	2518165000	303812	284305	13772994	200,124	154,723	14480078	10041783
43	2969184	666	421	2559655000	313478	284476	13926414	180,752	150,327	9980630	9983449
47	2973877	707	365	2619529000	318607	284498	14148342	200,444	138,784	10788117	9926383
46	3013541			2660907000		293389	14387077			11987392	9890661
	Index(/population)										
76	0.2680307189	0.00005517102919	0.00002698799443	222.1563069	0	0	0	0.01175342095	0.01186107417	0	
69	0.2768290094	0.00005852566959	0.00002801654711	224.2546128	0.01943122022	0	0	0.01199197949	0.01292868931	0	
61	0.2769802957	0.00006137756158	0.00002809816942	229.2427015	0.01982066792	0.02757795473	0.8155587336	0.01267048091	0.013024199	0	
58	0.2805593945	0.00006282885261	0.00002966640576	227.4722105	0.02086627103	0.02791448963	0.8698262088	0.01335068169	0.01466129755	1.122802039	
60	0.2846785925	0.00006457156517	0.00003118966034	226.1409811	0.0230407886	0.02792550793	0.771089417	0.01387142705	0.01482246257	1.478849474	
61	0.2920989801	0.00007120027532	0.00003564992806	225.8170008	0.02484381747	0.02793231612	0.9897483553	0.01540256138	0.01423378147	1.733257579	
55	0.29255094	0.00007221493625	0.00003551229008	227.5413284	0.02608853849	0.02771962392	1.298028066	0.01721135873	0.01497448125	2.107219539	
56	0.2924328416	0.00007561457627	0.0000385990638	226.9468576	0.02703894111	0.02767968557	1.275625721	0.01912296593	0.01462518528	1.878748266	
52	0.296641722	0.00008397585475	0.00004198792738	243.422422	0.02882620461	0.02821837463	1.335068436	0.01927952299	0.0147849243	1.749866176	
47	0.2965209465	0.00008006546248	0.00003973397951	250.7687131	0.03025478643	0.02831220312	1.371568575	0.01992913012	0.01540792108	1.441982763	
43	0.2974106444	0.0000667104124	0.00004216979523	256.3898508	0.03139976976	0.02849476168	1.394950182	0.01810516586	0.01505762187	0.9997176327	
47	0.2995932154	0.00007122433217	0.00003677069482	263.8956204	0.0320969884	0.02866079215	1.42532703	0.02019305521	0.01398132633	1.086812487	
46	0.3046855008	0	0	269.032272	0	0.02966323484	1.454612285	0	0	1.211990988	

Figure B:

```
> data <- read.csv(file.choose(),header=T)
> data
   finedust car_regist bikeroad_lg bikeroad_nu sub_users        GRDP
1        58   0.2805594    6.28e-05   2.9700e-05  227.4722 0.02086627
2        60   0.2846786    6.46e-05   3.1200e-05  226.1410 0.02304079
3        61   0.2920990    7.12e-05   3.5600e-05  225.8170 0.02484382
4        55   0.2925509    7.22e-05   3.5500e-05  227.5413 0.02608854
5        56   0.2924328    7.56e-05   3.8600e-05  226.9469 0.02703894
6        52   0.2966417    8.40e-05   4.2000e-05  243.4224 0.02882621
7        47   0.2965209    8.01e-05   3.9734e-05  250.7687 0.03025479
8        43   0.2974106    6.67e-05   4.2200e-05  256.3899 0.03139977
9        47   0.2995932    7.12e-05   3.6800e-05  263.8956 0.03209699
         tree greenbelt  coalpower nuclearenergy construction
1 0.02791449 0.8698262 0.01335068    0.01466130    1.1228020
2 0.02792551 0.7710894 0.01387143    0.01482246    1.4788495
3 0.02793232 0.9897484 0.01540256    0.01423378    1.7332576
4 0.02771962 1.2980281 0.01721136    0.01497448    2.1072195
5 0.02767969 1.2756257 0.01912297    0.01462519    1.8787483
6 0.02821838 1.3350684 0.01927952    0.01478492    1.7498662
7 0.02831220 1.3715686 0.01992913    0.01540792    1.4419828
8 0.02849476 1.3949502 0.01810517    0.01505762    0.9997176
9 0.02866079 1.4253270 0.02019306    0.01398133    1.0868125
```

Figure C:

```
> fit.full = regsubsets(data$finedust~., data=data, nvmax=11, nbest=3)
경고메시지:
1: In leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in = force.in,  :
  2  linear dependencies found
2: In leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in = force.in,  :
  nvmax reduced to  8
> summary(fit.full)
Subset selection object
Call: regsubsets.formula(data$finedust ~ ., data = data, nvmax = 11,
    nbest = 3)
10 Variables  (and intercept)
             Forced in Forced out
car_regist       FALSE      FALSE
bikeroad_lg      FALSE      FALSE
bikeroad_nu      FALSE      FALSE
sub_users        FALSE      FALSE
GRDP             FALSE      FALSE
tree             FALSE      FALSE
greenbelt        FALSE      FALSE
coalpower        FALSE      FALSE
nuclearenergy    FALSE      FALSE
construction     FALSE      FALSE
3 subsets of each size up to 8
Selection Algorithm: exhaustive
         car_regist bikeroad_lg bikeroad_nu sub_users GRDP tree greenbelt
1  ( 1 ) " "        " "         " "         "*"       " "  " "  " "
1  ( 2 ) " "        " "         " "         " "       "*"  " "  " "
1  ( 3 ) " "        " "         " "         " "       " "  " "  "*"
2  ( 1 ) " "        " "         " "         "*"       " "  " "  " "
2  ( 2 ) " "        " "         " "         " "       " "  " "  "*"
2  ( 3 ) " "        " "         " "         "*"       " "  "*"  " "
3  ( 1 ) " "        " "         " "         " "       " "  " "  "*"
3  ( 2 ) " "        " "         " "         "*"       " "  "*"  " "
3  ( 3 ) " "        " "         " "         "*"       " "  " "  "*"
4  ( 1 ) " "        " "         " "         " "       "*"  " "  "*"
4  ( 2 ) " "        "*"         " "         "*"       " "  " "  "*"
4  ( 3 ) " "        " "         " "         "*"       " "  " "  "*"
5  ( 1 ) " "        " "         "*"         "*"       " "  "*"  " "
5  ( 2 ) " "        "*"         " "         " "       " "  "*"  "*"
5  ( 3 ) " "        "*"         " "         "*"       " "  " "  "*"
6  ( 1 ) "*"        "*"         "*"         "*"       " "  " "  "*"
```

Figure D:

```
> summary(fit)

Call:
lm(formula = data$finedust ~ data$car_regist + data$bikeroad_lg +
    data$bikeroad_nu + data$sub_users + data$greenbelt + data$nuclearenergy)

Residuals:
        1         2         3         4         5         6
-0.011439  0.049317 -0.066126  0.128692 -0.141158  0.164696
        7         8         9
-0.145659  0.019987  0.001691

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.108e+02  1.260e+01   8.798 0.012675 *
data$car_regist     2.363e+02  4.399e+01   5.372 0.032944 *
data$bikeroad_lg    2.096e+05  1.713e+04  12.237 0.006612 **
data$bikeroad_nu   -3.115e+05  4.499e+04  -6.923 0.020235 *
data$sub_users     -2.904e-01  8.994e-03 -32.281 0.000958 ***
data$greenbelt     -1.196e+01  8.051e-01 -14.860 0.004498 **
data$nuclearenergy -3.178e+03  2.312e+02 -13.743 0.005253 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2147 on 2 degrees of freedom
Multiple R-squared:  0.9997,    Adjusted R-squared:  0.9989
F-statistic:  1170 on 6 and 2 DF,  p-value: 0.0008543
```

Figure E:

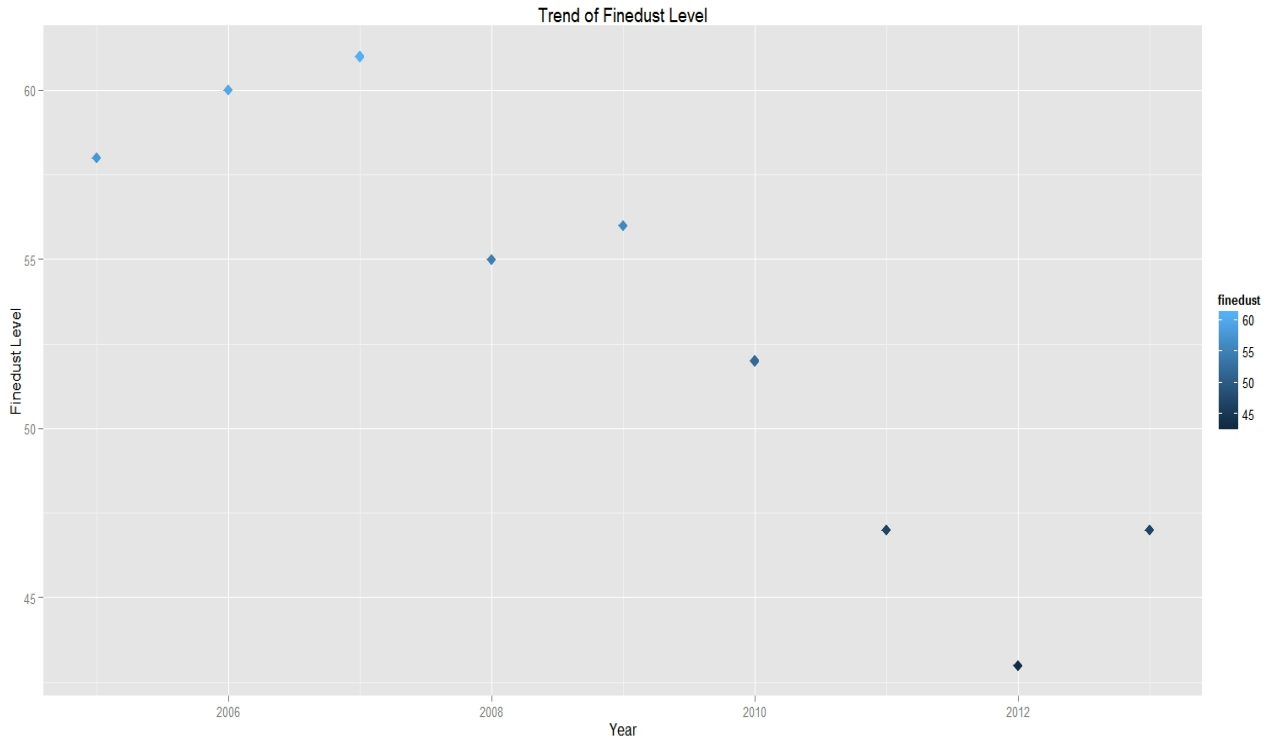


Figure 1A:

```
> summary(lm(data$finedust~data$sub_users))

Call:
lm(formula = data$finedust ~ data$sub_users)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3568 -1.7905  0.4135  1.8966  3.5580

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    145.92159   14.62586   9.977 2.17e-05 ***
data$sub_users  -0.38833    0.06116  -6.349 0.000385 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.615 on 7 degrees of freedom
Multiple R-squared:  0.8521,    Adjusted R-squared:  0.8309
F-statistic: 40.31 on 1 and 7 DF,  p-value: 0.0003855
```

Figure 2A:

```
> summary(lm(data$finedust~data$greenbelt))

Call:
lm(formula = data$finedust ~ data$greenbelt)

Residuals:
   Min     1Q Median     3Q    Max
-5.924 -2.160 -1.280  3.479  4.544

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      78.520      6.690  11.736 7.38e-06 ***
data$greenbelt  -21.216      5.507  -3.853  0.00627 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.849 on 7 degrees of freedom
Multiple R-squared:  0.6795,    Adjusted R-squared:  0.6337
F-statistic: 14.84 on 1 and 7 DF,  p-value: 0.006271
```

Figure 3A:

```
> summary(lm(data$finedust~data$car_regist))

Call:
lm(formula = data$finedust ~ data$car_regist)

Residuals:
   Min     1Q Median     3Q    Max
-6.476 -3.155  0.814  1.938  7.473

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       276.29      74.71   3.698  0.00767 **
data$car_regist  -762.64     255.38  -2.986  0.02033 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.508 on 7 degrees of freedom
Multiple R-squared:  0.5602,    Adjusted R-squared:  0.4974
F-statistic: 8.918 on 1 and 7 DF,  p-value: 0.02033
```

References

Ning, Da-Tong. Zhong, Liang-Xi. Chung, Yong-Seung. (1996), Aerosol size distribution and elemental composition in urban areas of Northern China. *Atmospheric Environment*, Vol 30, Issue 13, pp. 2355-2362.

Leem, Jong Han. Kim, Soon Tae. Kim, Hwan Cheol. (2015), Public-health impact of outdoor air pollution for 2nd air pollution management policy in Seoul metropolitan area, Korea. *Annals of Occupational and Environmental Medicine,* 2015, 27:7.

Lee, Hyun-jeong. (2015), At least 1 in 6 dies early from fine dust: study. *The Korean Herald*. http://www.koreaherald.com/view.php?ud=20150420001188.

Yamashita, Masakazu. Yamanaka, Shohei. (2013), Dust Resulting from Tire Wear and the Risk of Health Hazards. *Journal of Environmental Protection,* 2013, 4, 509-515.