Final Project Report

Yifeng Lan, Gloria Wu, Yunwei Lu

The dataset our group used for final project is the hourly data set contains the PM2.5 record in Beijing from 2010 to 2015. Meanwhile, meteorological data for Beijing are also included.

Here is the attribute information:
No: row number
year: year of data in this row
month: month of data in this row
day: day of data in this row
hour: hour of data in this row
season: season of data in this row
PM: PM2.5 concentration (ug/m^3)
DEWP: Dew Point (Celsius Degree)
TEMP: Temperature (Celsius Degree)
HUMI: Humidity (%)
PRES: Pressure (hPa)
cbwd: Combined wind direction
Iws: Cumulated wind speed (m/s)
precipitation: hourly precipitation (mm)
Iprec: Cumulated precipitation (mm)
Recourse:
https://archive.ics.uci.edu/ml/datasets/PM2.5+Data+of+Five+Chinese+Cities

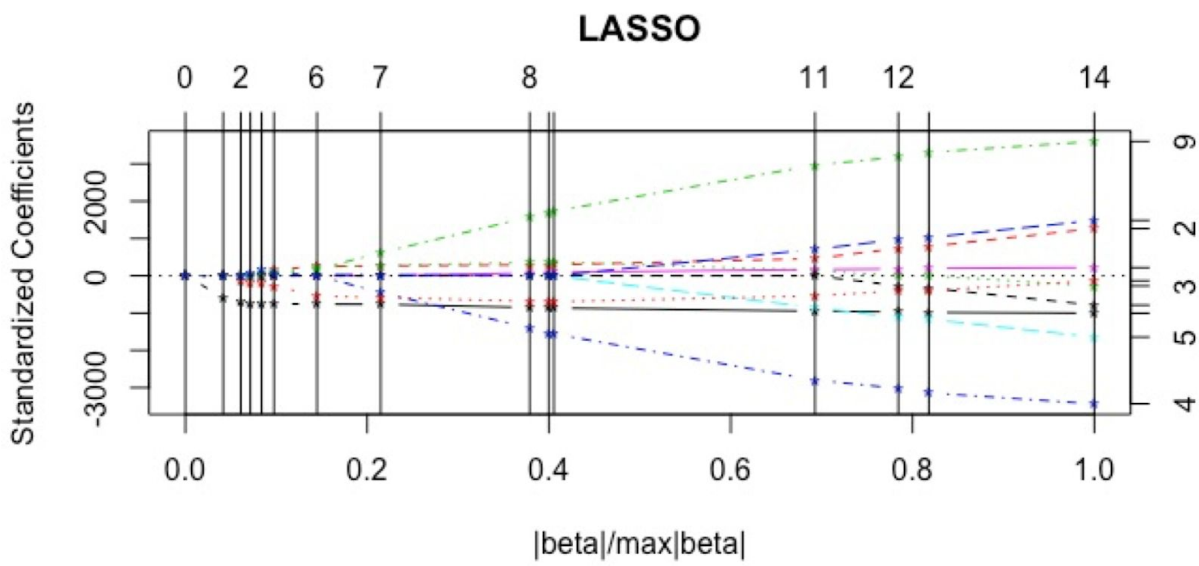
The goal of our final project is that based on the simple stochastic time series model simulated from data from 2010 to 2014, we will try to predict the change of PM2.5 in the next hour by the hourly data several hours ago. Therefore, we will observe the testing result and conclude the most appropriate number of hours to step behind.

After several test, we decided to only include the most influential variables among those 14 variables, which are PM_US.Post, DEWP, HUMI, PRES, TEMP.
By running the Least Square Fit for 2 hours, we have the following result:

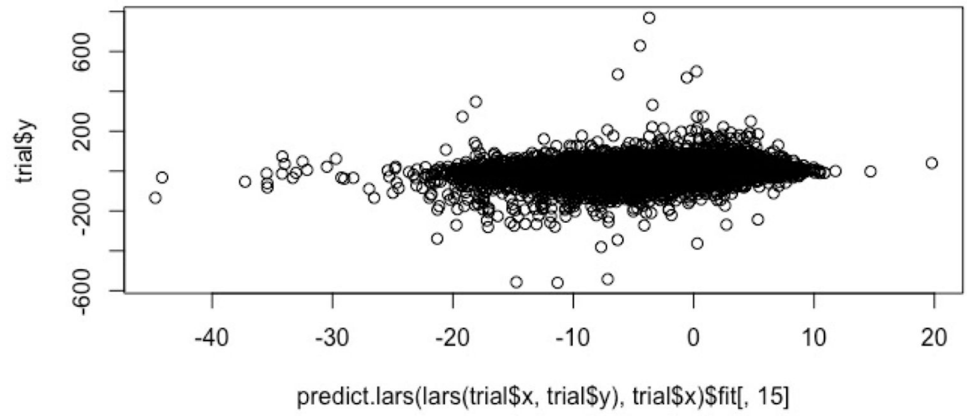
```
> trial<-delay.map.make(mat.pred,ycol,xcol,3)
[1] 41731     5
[1] 41728
[1] 41728    10
> ls.print(lsfit(trial$x,trial$y))
Residual Standard Error=24.1111
R-Square=0.035
F-statistic (df=10, 41716)=151.4707
p-value=0

           Estimate Std.Err  t-value Pr(>|t|)
Intercept  -73.0249 22.3926  -3.2611   0.0011
PM_US.Post  -0.0533  0.0049 -10.8215   0.0000
DEWP         0.4292  0.1402   3.0603   0.0022
HUMI        -0.0510  0.0430  -1.1876   0.2350
PRES        -1.6321  0.1713  -9.5267   0.0000
TEMP        -0.6609  0.1587  -4.1654   0.0000
PM_US.Post   0.0116  0.0049   2.3509   0.0187
DEWP        -0.2667  0.1404  -1.8993   0.0575
HUMI        -0.0252  0.0429  -0.5874   0.5569
PRES         1.7125  0.1701  10.0677   0.0000
TEMP         0.5974  0.1578   3.7855   0.0002
```

```
> summary(lars(trial$x,trial$y))
LARS/LASSO
Call: lars(x = trial$x, y = trial$y)
   Df      Rss       Cp
0   1 25131982 1505.707
1   2 24469655  368.406
2   3 24366626  193.182
3   4 24346886  161.226
4   5 24336571  145.483
5   6 24329105  134.641
6   7 24313412  109.645
7   8 24300452   89.352
8   9 24277485   51.846
9   8 24275273   46.040
10  9 24274716   47.083
11 10 24255755   16.467
12 11 24253295   14.236
13 10 24252700   11.212
14 11 24251414   11.000
```

From this output, we can see that R-square equal to 0.035, which is not very high, but Pr(>|t|) of variables are all very close to 0, except HUMI, which shows the correlation of these variables to our diff(PM2.5).



By running Lasso, we can tell that there are two variables that enter the model in the last, and has the least correlation, and we believe these two variables are the HUMI from the first hour and from the second hour.




```
> cor(predict.lars(lars(trial$x,trial$y),trial$x)$fit[,15],trial$y)
[1] 0.1871837
```

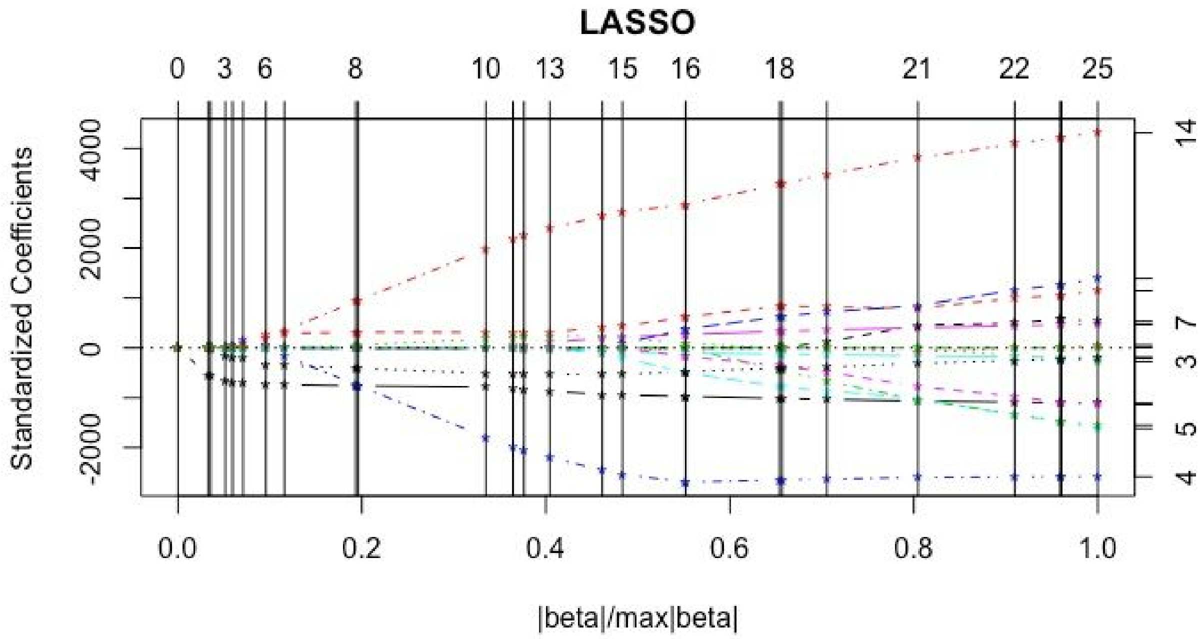
While lars() produces the entire path of solutions, predict.lars allows one to extract a prediction at a particular point along the path.

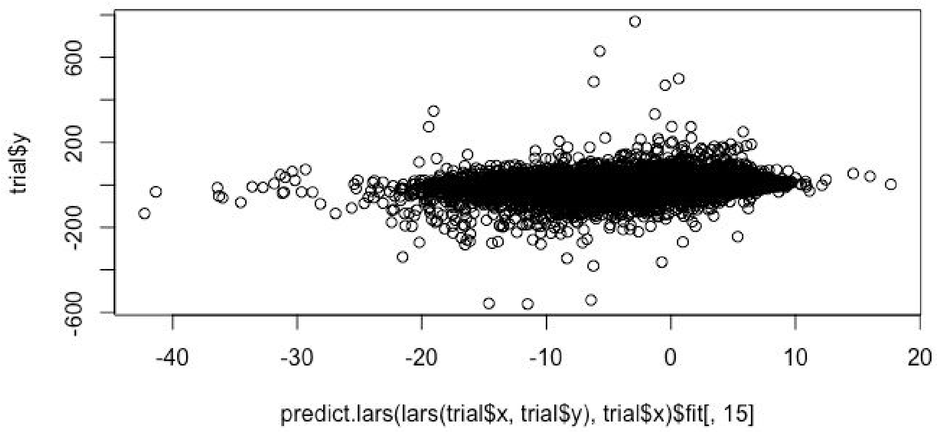
Thus, after making a prediction from the fitted lars model, we can see that the correlation is the close to flat but having too many outliers, leading to the correlation value not very high.

This is the result from the 2 hour behind.

```
> trial<-delay.map.make(mat.pred,ycol,xcol,4)
[1] 41731     5
[1] 41727
[1] 41727    15
> ls.print(lsfit(trial$x,trial$y))
Residual Standard Error=24.0574
R-Square=0.0394
F-statistic (df=15, 41710)=114.0591
p-value=0

           Estimate Std.Err  t-value Pr(>|t|)
Intercept  -77.8192 22.4741  -3.4626   0.0005
PM_US.Post  -0.0587  0.0050 -11.7341   0.0000
DEWP         0.3888  0.1424   2.7298   0.0063
HUMI        -0.0518  0.0436  -1.1874   0.2351
PRES        -1.2277  0.1770  -6.9369   0.0000
TEMP        -0.6510  0.1621  -4.0155   0.0001
PM_US.Post   0.0250  0.0075   3.3501   0.0008
DEWP         0.1811  0.2079   0.8711   0.3837
HUMI         0.0112  0.0617   0.1812   0.8562
PRES        -0.7363  0.2681  -2.7463   0.0060
TEMP         0.5596  0.2388   2.3432   0.0191
PM_US.Post  -0.0099  0.0050  -1.9701   0.0488
DEWP        -0.3841  0.1428  -2.6889   0.0072
HUMI        -0.0391  0.0434  -0.9017   0.3672
PRES         2.0498  0.1732  11.8372   0.0000
TEMP         0.0038  0.1642   0.0234   0.9814
```

By running 3 hours behind, we can see that R-square increase slightly. Three hours of HUMI are still bad, but the third hour or the earliest hour of HUMI always behaves better than the rest of HUMI based on the observation.
What worth mentioning is that second year of DEWP and first year of TEMP are very unstable.






```
> cor(predict.lars(lars(trial$x,trial$y),trial$x)$fit[,15],trial$y)
[1] 0.1953856
```

By running 4 and 5 hours behind:

```
> ls.print(lsfit(trial$x,trial$y))
Residual Standard Error=24.0315
R-Square=0.0416
F-statistic (df=20, 41704)=90.4536
p-value=0
```

	Estimate	Std.Err	t-value	Pr(>|t|)
Intercept	-87.4561	22.5941	-3.8707	0.0001
PM_US.Post	-0.0601	0.0050	-12.0086	0.0000
DEWP	0.3195	0.1437	2.2230	0.0262
HUMI	-0.0296	0.0441	-0.6709	0.5023
PRES	-0.8684	0.1833	-4.7366	0.0000
TEMP	-0.5430	0.1651	-3.2886	0.0010
PM_US.Post	0.0272	0.0076	3.6050	0.0003
DEWP	0.1965	0.2079	0.9454	0.3444
HUMI	0.0096	0.0618	0.1546	0.8771
PRES	-0.8633	0.2682	-3.2185	0.0013
TEMP	0.5123	0.2396	2.1385	0.0325
PM_US.Post	-0.0223	0.0076	-2.9469	0.0032
DEWP	0.2109	0.2078	1.0150	0.3101
HUMI	-0.1517	0.0617	-2.4582	0.0140
PRES	0.2875	0.2686	1.0703	0.2845
TEMP	-0.3543	0.2391	-1.4819	0.1384
PM_US.Post	0.0112	0.0051	2.2152	0.0268
DEWP	-0.5598	0.1441	-3.8846	0.0001
HUMI	0.1004	0.0438	2.2914	0.0219
PRES	1.5388	0.1777	8.6606	0.0000
TEMP	0.3161	0.1684	1.8764	0.0606

```
> ls.print(lsfit(trial$x,trial$y))
Residual Standard Error=24.0025
R-Square=0.044
F-statistic (df=25, 41698)=76.799
p-value=0
```

	Estimate	Std.Err	t-value	Pr(>|t|)
Intercept	-96.2944	22.7306	-4.2363	0.0000
PM_US.Post	-0.0613	0.0050	-12.2406	0.0000
DEWP	0.2574	0.1443	1.7841	0.0744
HUMI	-0.0095	0.0442	-0.2154	0.8295
PRES	-0.7266	0.1847	-3.9343	0.0001
TEMP	-0.4609	0.1666	-2.7667	0.0057
PM_US.Post	0.0269	0.0075	3.5702	0.0004
DEWP	0.2368	0.2078	1.1392	0.2546
HUMI	0.0092	0.0619	0.1485	0.8820
PRES	-0.6291	0.2693	-2.3362	0.0195
TEMP	0.4966	0.2395	2.0735	0.0381
PM_US.Post	-0.0167	0.0076	-2.1847	0.0289
DEWP	0.2513	0.2078	1.2092	0.2266
HUMI	-0.1598	0.0618	-2.5842	0.0098
PRES	0.1516	0.2689	0.5639	0.5728
TEMP	-0.4194	0.2398	-1.7492	0.0803
PM_US.Post	-0.0181	0.0076	-2.3935	0.0167
DEWP	-0.2593	0.2079	-1.2473	0.2123
HUMI	-0.0058	0.0617	-0.0945	0.9247
PRES	-0.2080	0.2692	-0.7724	0.4399
TEMP	0.1348	0.2392	0.5637	0.5730
PM_US.Post	0.0253	0.0050	5.0088	0.0000
DEWP	-0.3203	0.1443	-2.2194	0.0265
HUMI	0.0994	0.0439	2.2648	0.0235
PRES	1.5151	0.1786	8.4815	0.0000
TEMP	0.1804	0.1700	1.0613	0.2885

As the number of hours increase, we can see that the Pr(>|t|) increase dramatically for many variables in the hours except the first and last hours.
And after 5 hours, as the hour increase, the R-square and correlation don't have jumps but the behavior of variables become more and more bad and random.
Therefore, we suggest that we choose between 3 hours and 5 hours.