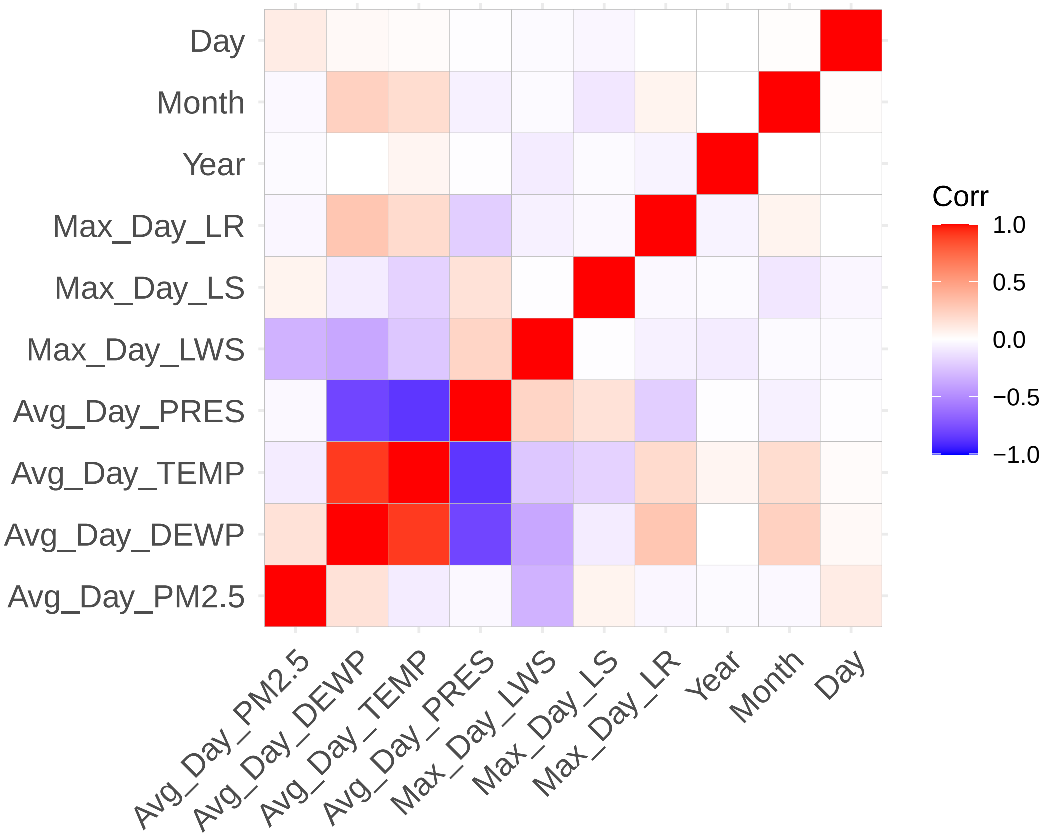
Final Project: Regression From The Mean

Zhengtao Xu, Jennings Cheng and Collin Carmichael

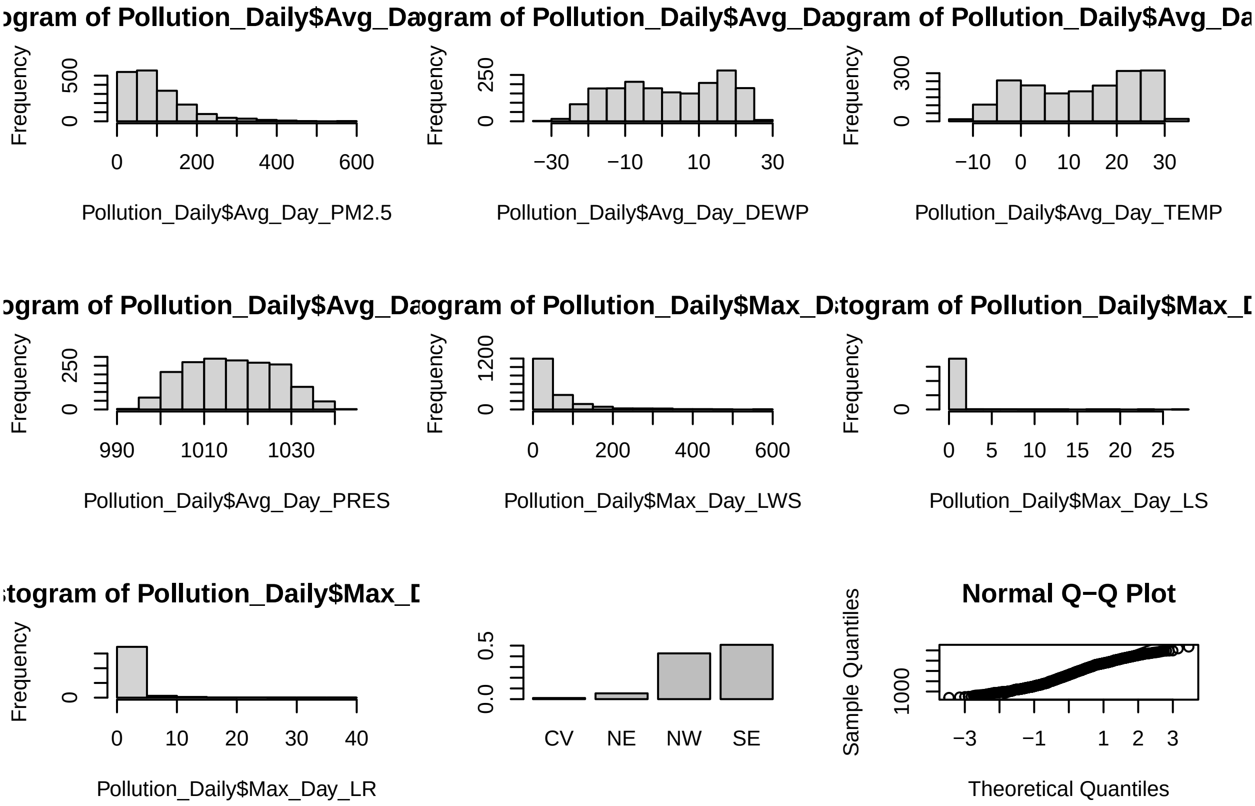
8/6/2021

Section 1: Introduction

Section 2: Exploratory Data Analysis



Looking at a correlation heatmap we see aside from dewpoint precipitation and temperature there are no significant correlation in other variable



We can see that PM2.5 has an inverse distribution with values skewed towards low PM2.5 but many high values that go beyond the median

Temp and Dew point are highly related and so may not need to be included in the same model

Precipitation looks like a bell curve but not normal as seen in the plot in 3,3

Wind speed is similar but not identical to PM2.5 in that it has an inverse distribution so it may be highly important in prediction

There is almost always not any snow in Beijing and the max is 27 for a day

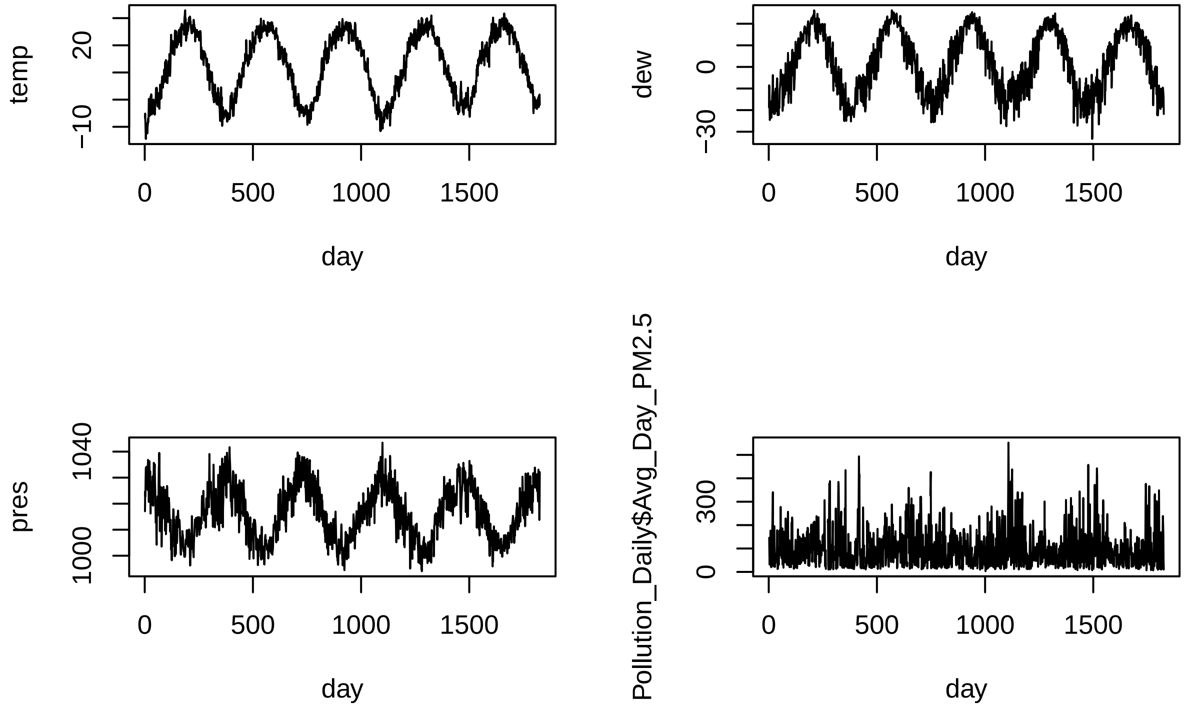
Rain is also infrequent but there are times when there is a lot of rain so we may assume that rain and snow are not the same

The most common direction is SE and NW while sometimes NE or CV

Now that we know the distributions we can take a look at some pm2.5 vs predictor

The pm and time is not a clear relationship and there are many seasonal spikes - One thing is that since there are random fluctations in our final model we will not use date since in a regression model the advance of one day would theoretically advance pm2.5 which is not the relationship we see

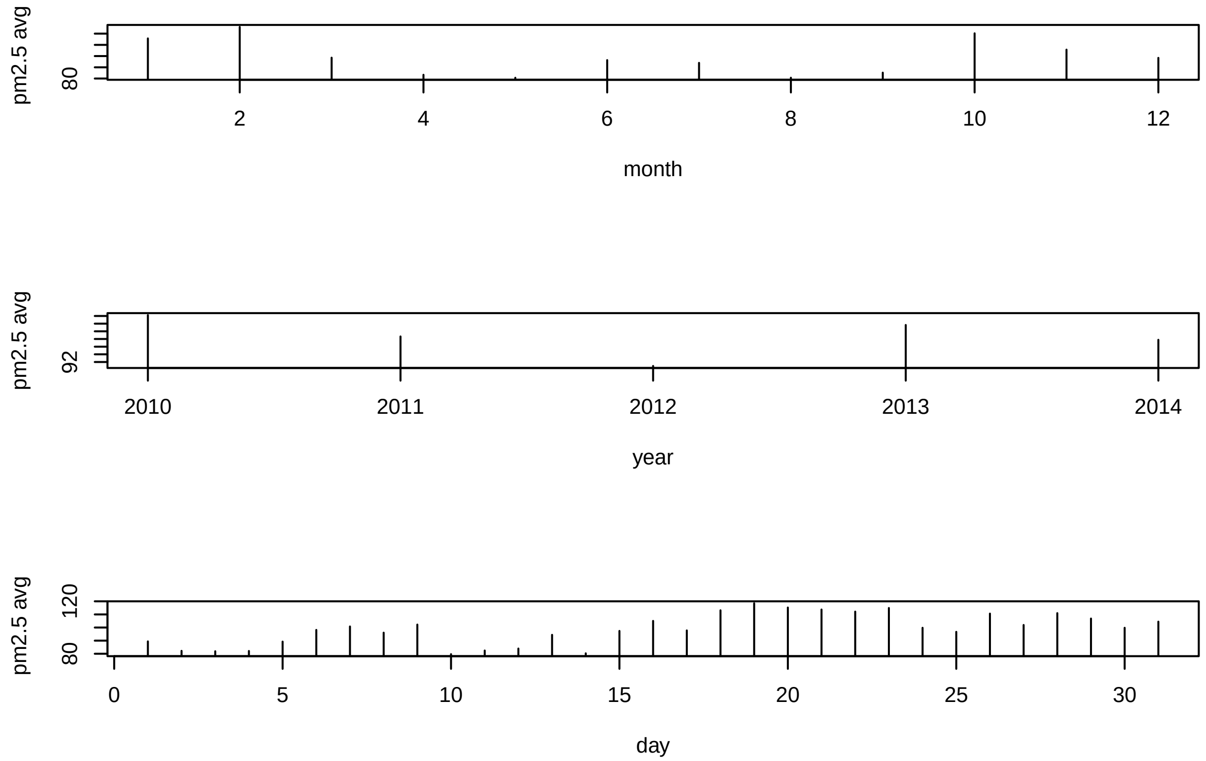
We can also extract that dew, temp, pres are very clearly related from the following:



Dew and temp are very similar and pres is the opposite so we may not necessarily need all three in the final model

Max wind is also related to Avg__Day__PM2.5 since more wind produces less 2.5

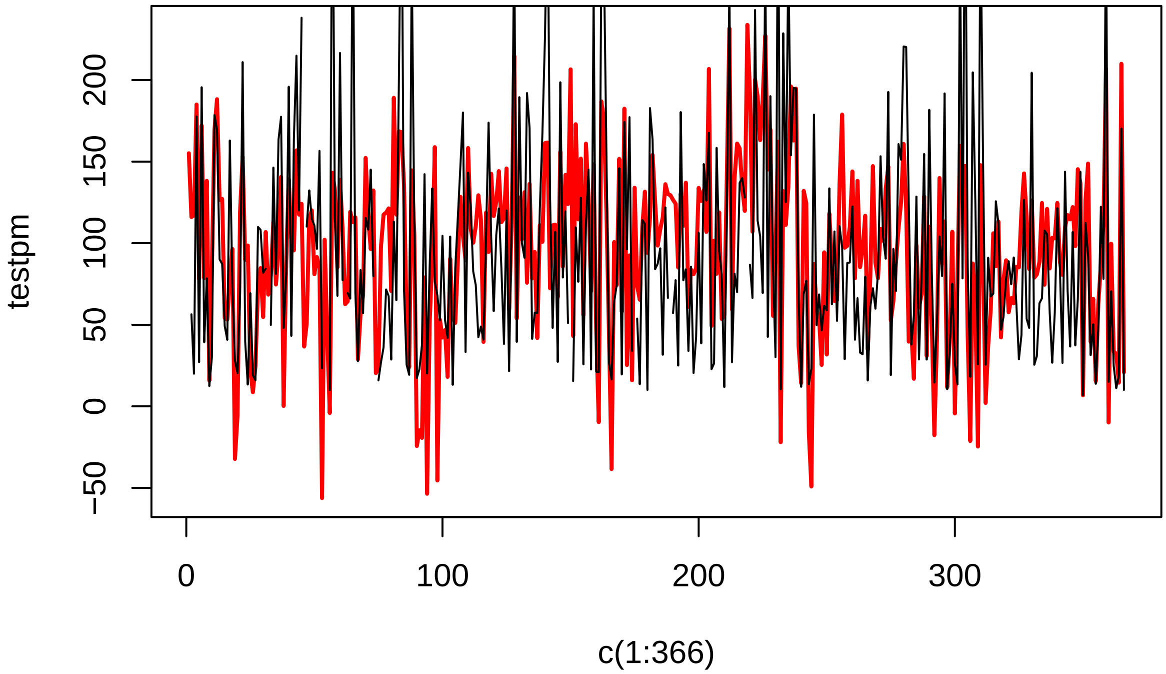
At first glance it may seem that pm2.5 may not vary with month day or year since there are a lot of random spike but looking at a mean of all pm2.5 with each factor shows us that it indeed is influenced highly by these predictor



As we see the variation between the month year and day is important and vary so therefore for the sake of fitting a model between these years every variable will be important beside Date_D

3.1: Fitting model and diagnostic

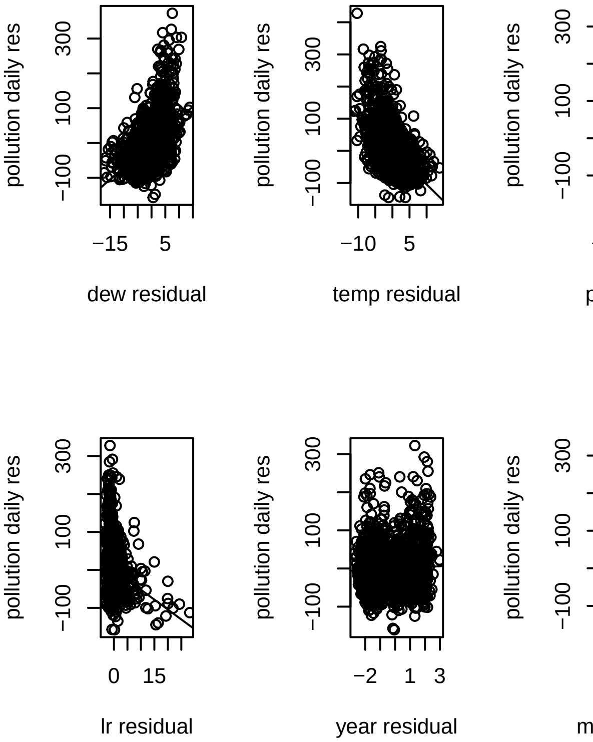
We can fit a multiple linear regression with response Avg_Day_PM and predictor being every variable besides date and below we see the prediction in red on the testing (every 5th observation in the 5 year long data set) after fitting train



```
##    (Intercept)    Avg_Day_DEWP    Avg_Day_TEMP    Avg_Day_PRES     Max_Day_LWS
##   6.284491e-01    4.717073e-78   3.203974e-104    2.643099e-14    2.759269e-09
##     Max_Day_LS      Max_Day_LR Max_Day_CBWDNE Max_Day_CBWDNW Max_Day_CBWDSE
##   4.680786e-05    3.341618e-16    8.172423e-06    6.685417e-04    2.509930e-03
##           Year           Month             Day
##   8.690987e-02    3.685244e-04    7.488178e-05
```

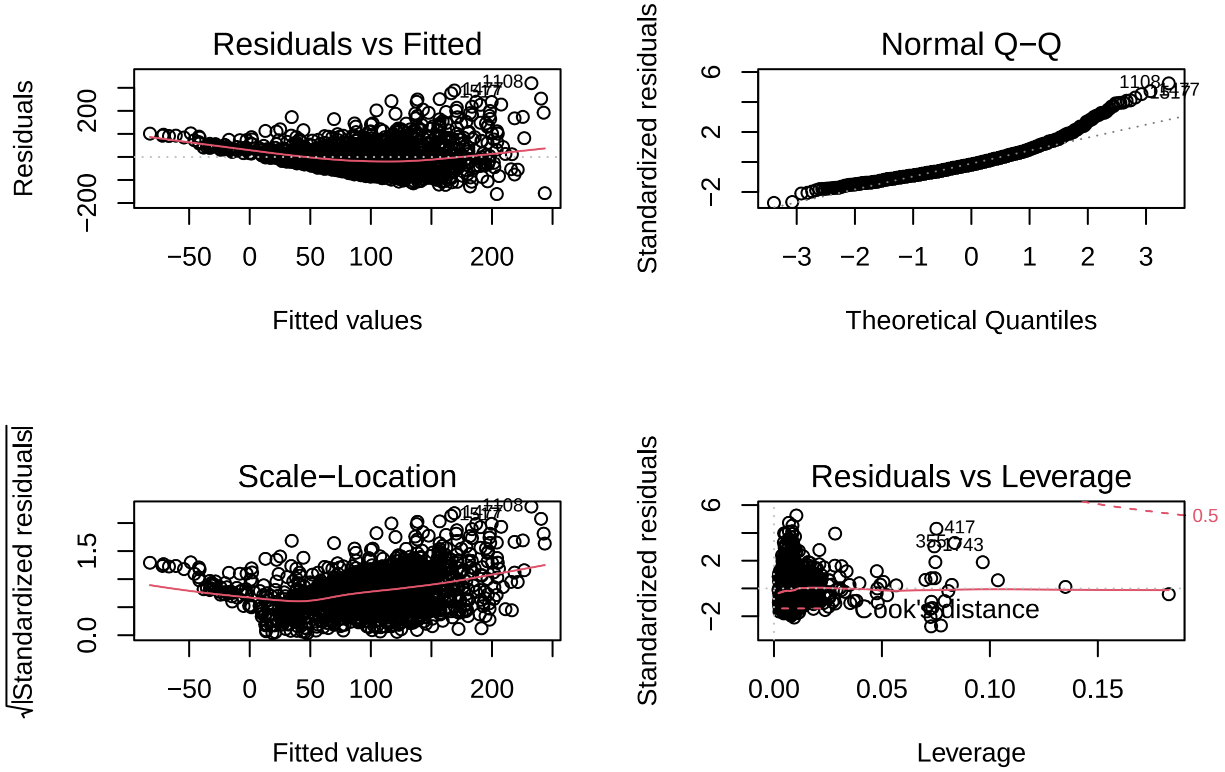
```
## $r.squared
## [1] 0.4041776
```

After fitting a full model we see every single predictor besides year is significant however R square is only 0.4 which suggests even though each coefficient is directionally correct the variation of response is not accurate



We can see some diagnostics as to why our model may not be that effective
We see that there are a lot of high influence points in partial regression for each numerical variable

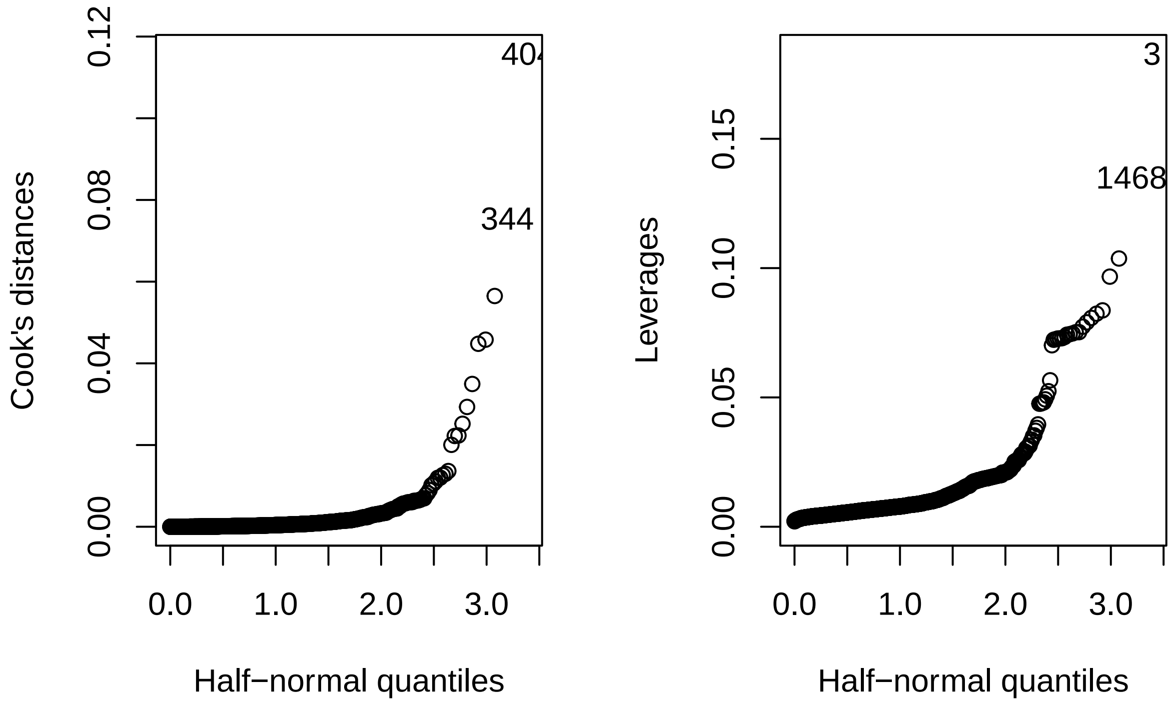
Now we can check some of the plot of diagnostic



Taking a glance at our linear model, we see that it does not appear to meet all error assumptions. Particularly, the Q-Q plot suggests non-normality and the Fitted-Residual plot suggests heteroscedasticity and some degree of non-linearity. However, the Residuals-Leverage plot appears to indicate that we don't have any high influence points.

```
##       bptest p shapiro test p durbin watson p
## 1 3.188052e-31   3.601797e-25     6.90063e-80
```

As we can tell through testing more formally, none of the error assumptions of homocedasticity, normality, and uncorrelated errors are accurate.

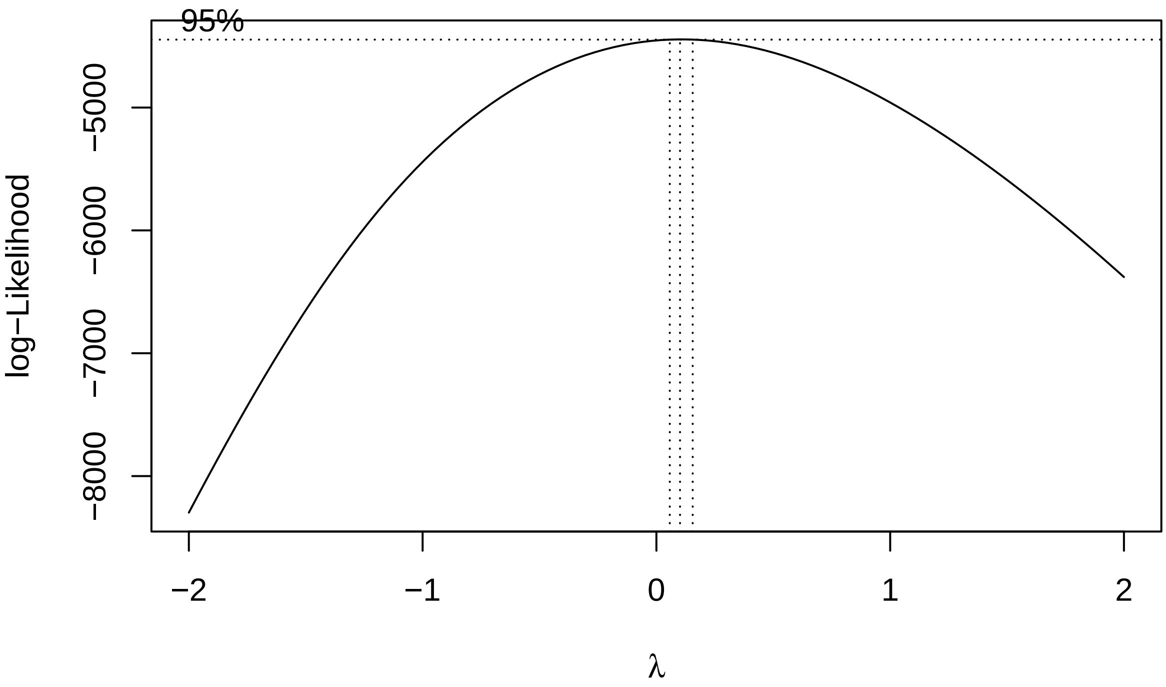


In regards to unusual observations, our maximum cooks distance as seen above indicates we have no datapoints that would be classified as highly influential, since it is much less than one. In constrast, however, we seem to have an abundance of high-leverage points. However, it is unclear at first glance how many of these are "good" or "bad" leverage points, although we can assume some points like observations 3 and 1468 with wildly different y-values are likely "bad."

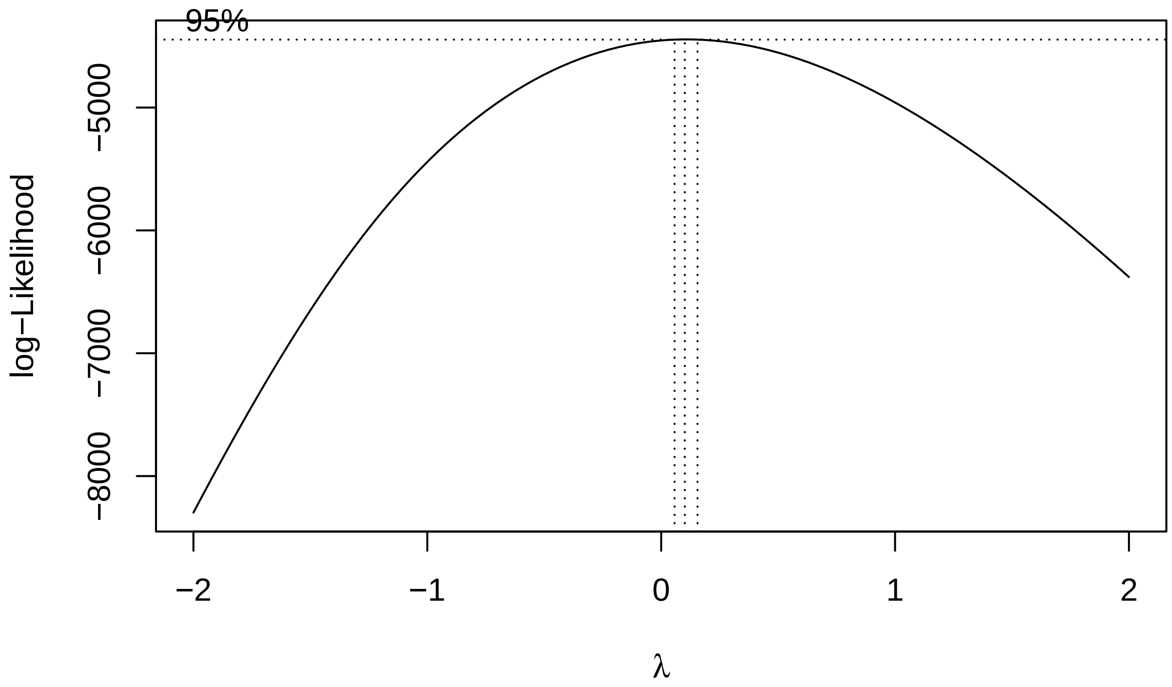
```
##             b d        c
## 417  4.327428 > 4.152348
## 1108 5.299076 > 4.152348
## 1477 4.762509 > 4.152348
## 1517 4.561424 > 4.152348
```

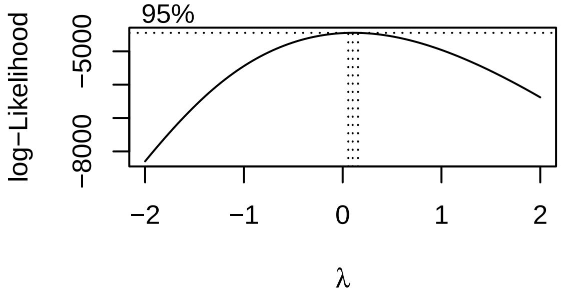
As we can see above, we also have four outliers in the dataset.

So in order to remedy non linearity, heteroscedasticity, non normal residual and autocorrelation we would try to



```
##               b d        c
## 417  4.327428 > 4.152348
## 1108 5.299076 > 4.152348
## 1477 4.762509 > 4.152348
## 1517 4.561424 > 4.152348
```


95%
log−Likelihood
−5000
−6000
−7000
−8000
−2
−1
0
1
2
λ



Section 4: Conclusions

Through our analysis we have learned many things, such as: