

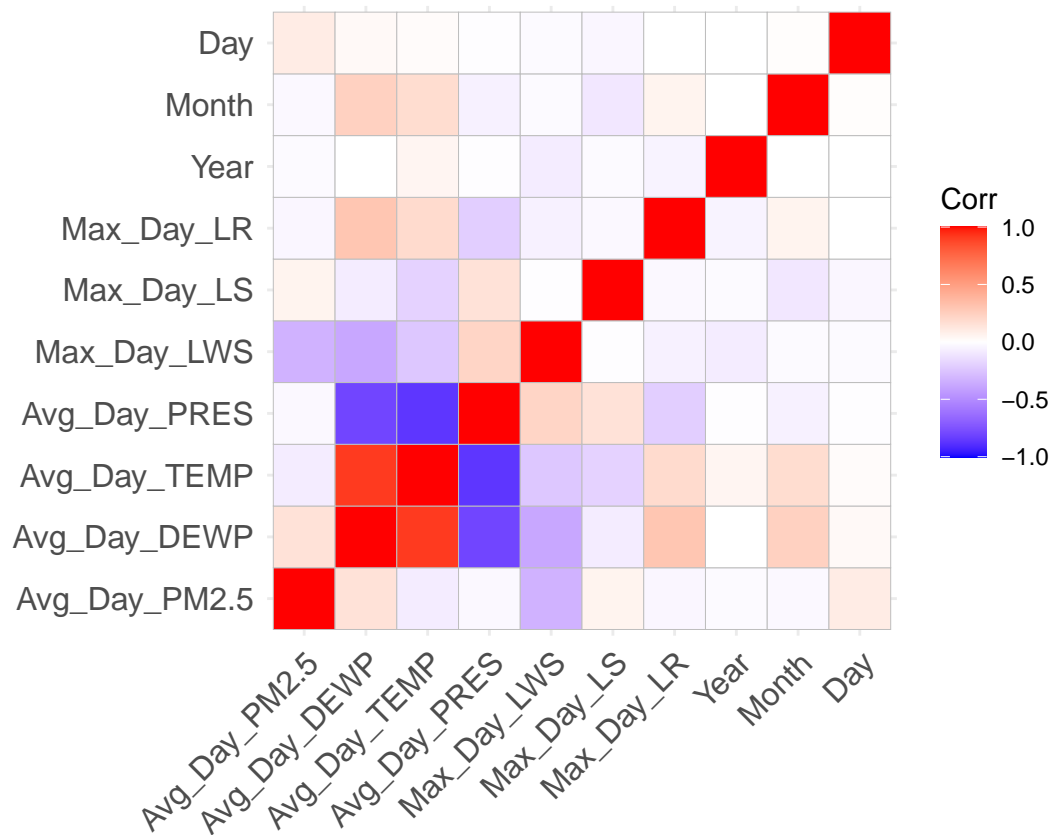
Final Project: Regression From The Mean

Zhengtao Xu, Jennings Cheng and Collin Carmichael

8/6/2021

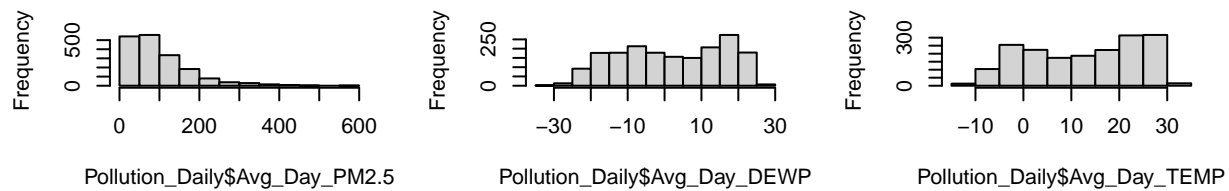
Section 1: Introduction

Section 2: Exploratory Data Analysis

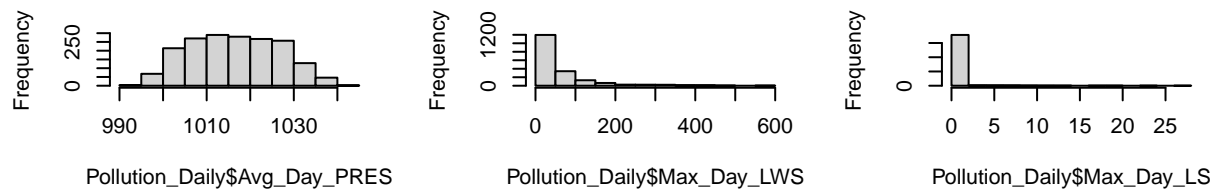


Looking at a correlation heatmap we see aside from dewpoint precipitation and temperature there are no significant correlation in other variable

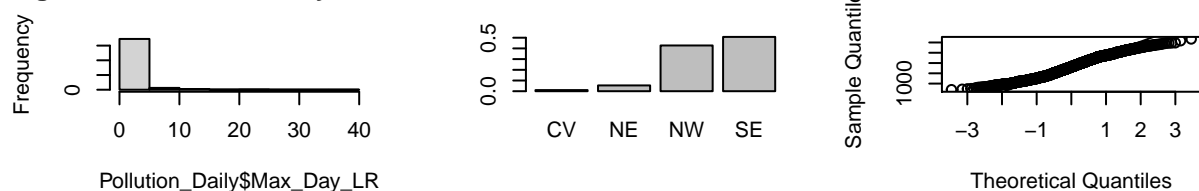
ogram of Pollution_Daily\$Avg_Daogram of Pollution_Daily\$Avg_Daogram of Pollution_Daily\$Avg_Da



ogram of Pollution_Daily\$Avg_Daogram of Pollution_Daily\$Max_D:ogram of Pollution_Daily\$Max_I



ogram of Pollution_Daily\$Max_I



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

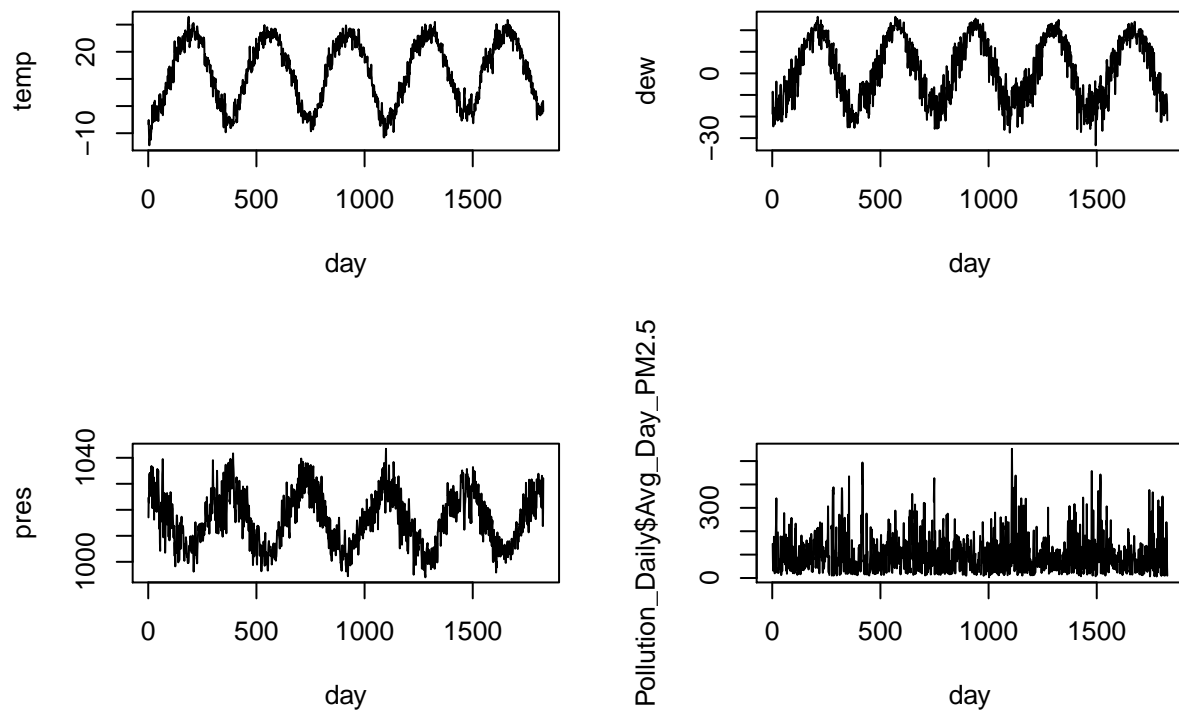
Rain is also infrequent but there are times when there is a lot of rain so we may assume that rain and snow are not the same

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

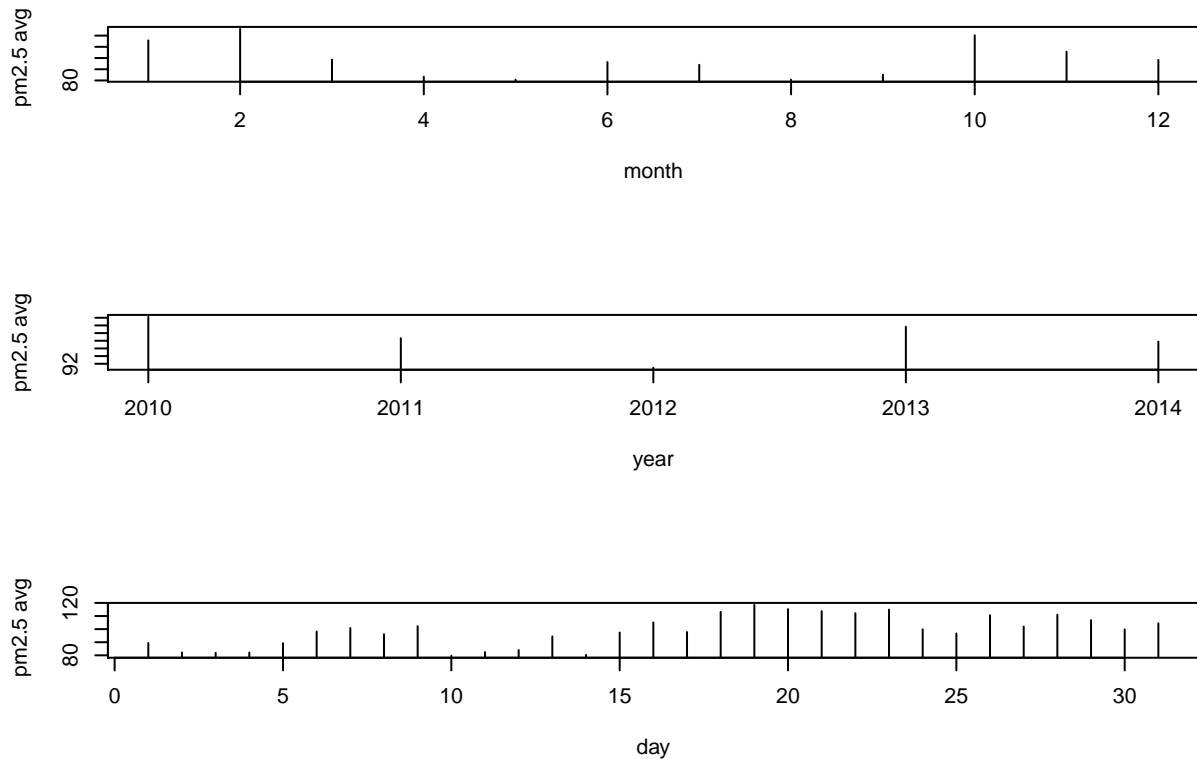
Now that we know the distributions we can take a look at some pm2.5 vs predictor

The pm and time is not a clear relationship and there are many seasonal spikes - One thing is that since there are random fluctuations 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 - and also any hypothesize increase in pm2.5 due to global warming trends in the date would be explained by the year variable and of course the other predictors so we decide not to use the column for our analysis. In the same way we would make month and day factors due to the same concern which is an increase in either of these do not necessarily mean an increase in particulate.

We can also extract that dew, temp, pres are very clearly related from the following which corroborates what was shown in heatmap.



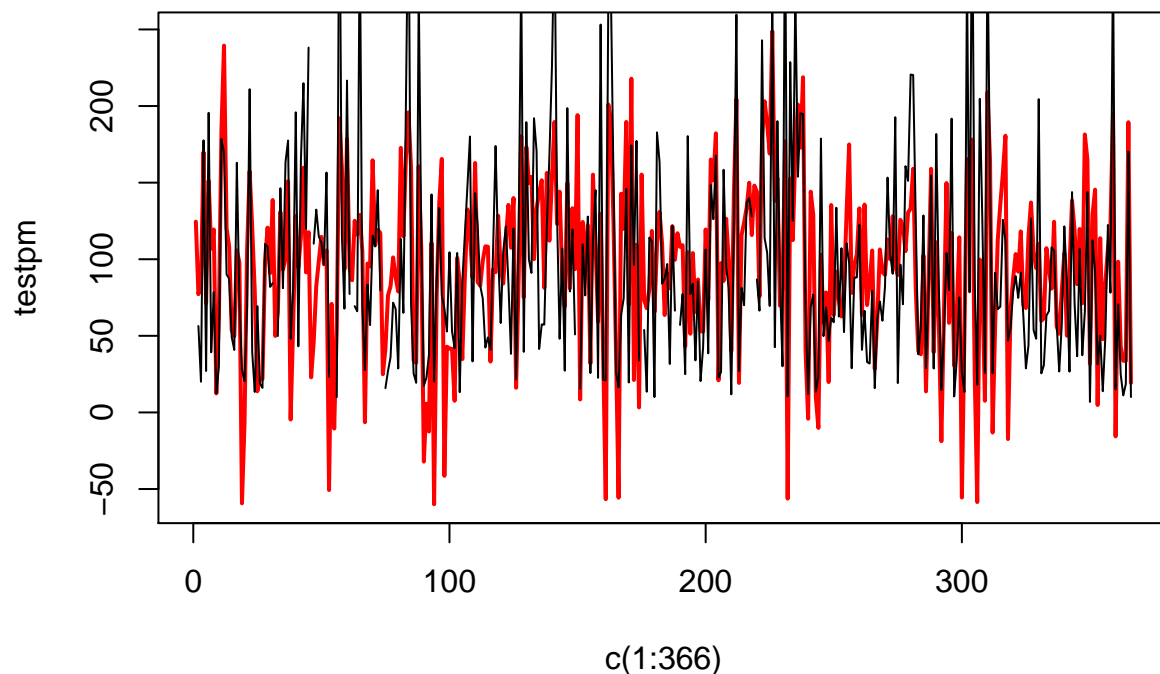
PM2.5 is obviously highly unpredictable from just looking at the time series. We will see if there is a relationship with day or month so that we can see if including these in the linear regression as factor would make sense.



As we see the variation between the month year and day is important - highly significant differences between different days, years, or months so it would make much sense to include these in MLR

3.1: Fitting model and diagnostic

Now that we know what our data looks like in accordance with the response we can fit a multiple linear regression with response Avg_Day_PM and predictor every variable besides 'Date_D', and look at the results. We can also see the predicted values of pm2.5 in in red on the testing (every 5th observation in the 5 year long data set) after fitting on the training dataset(which is the complement of test)

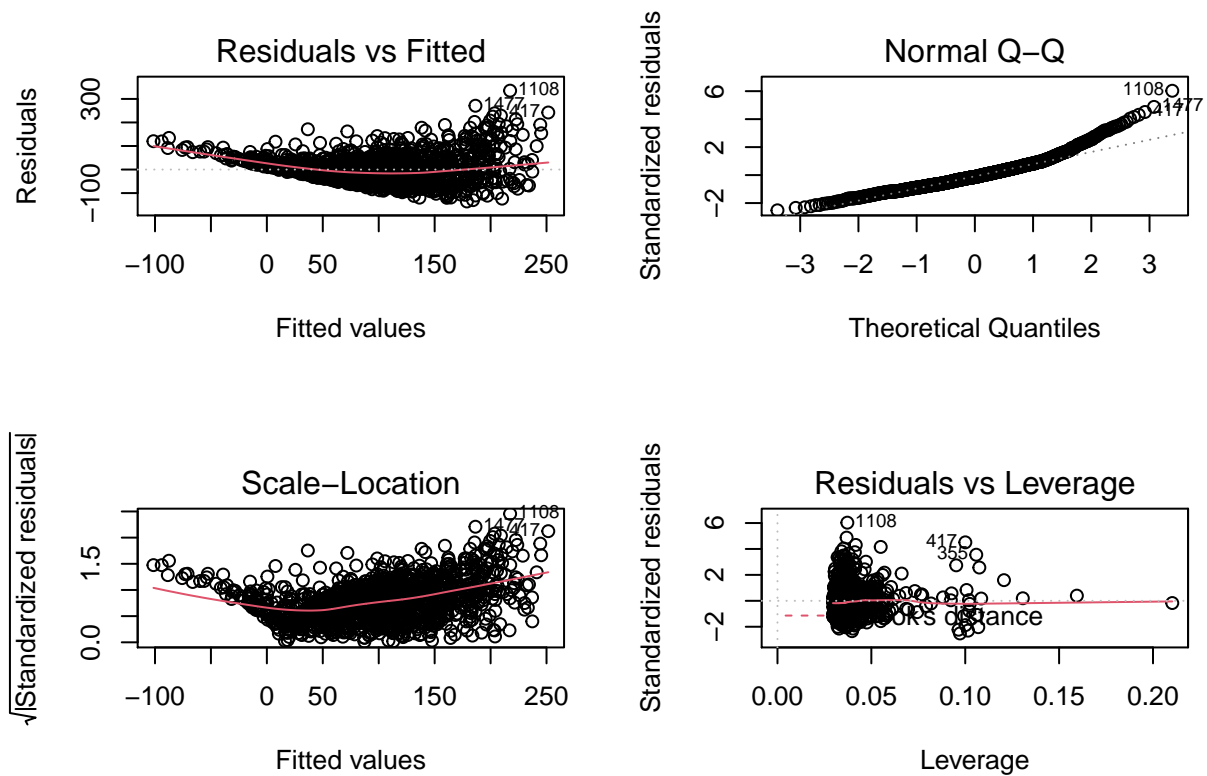


```
## $adj.r.squared
## [1] 0.4862557
```

After fitting a full model with every predictor besides 'Date_D' and Month, Year, and Day all in factor version we see that firstly there are many insignificant predictors unsurprisingly in the factors. As expected we see highly significant predictors in the factor(Day) in the 20s since that is what we saw highly in the mean particulate vs. day visualization. But the majority of days are insignificant so it makes sense to just drop the day column. Every month factor is significant besides the 2nd so this is important to keep. Year is all insignificant unless the year is 2012, which means 2012 is a lower pm anomaly. We could hypothetically exclude every non 2012 level but the simpler way is to not use the predictor.

Before considering dropping variables we need to check diagnostics. So here are some of the important assumptions that we will have to validate, namely constant variance, non correlated and normal residuals.

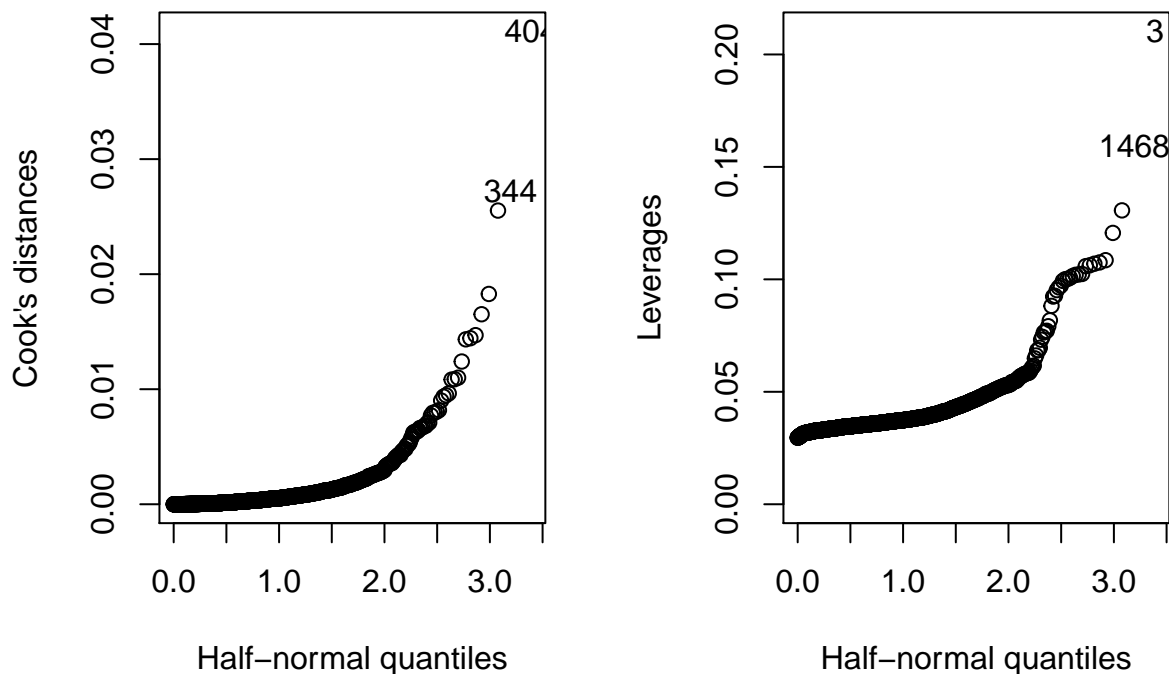
We can also check the visualizations for the regression of normality and heteroscedasticity.



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. The Residuals-Leverage plot however appears to indicate we don't have highly influential points.

```
##      bptest p shapiro test p durbin watson p
## 1 3.392827e-24  4.94394e-23  1.493388e-70
```

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 contrast, 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  4.541472 > 4.152753
## 1108 6.105989 > 4.152753
## 1125 4.181143 > 4.152753
## 1477 4.915473 > 4.152753
## 1517 4.338133 > 4.152753
```

As we can see above, we also have four outliers in training data.

All things aside, in order to remedy non linearity, heteroscedasticity, non normal residual and autocorrelation we can apply a Boxcox transformation first for non linearity and making normally distributed residuals:

This is the maximum likelihood for lambda so that will be the transformation we use

We fit the model with the boxcox transformation which involves applying the transformation $\text{responseNew} = (\text{responseOld}^{\lambda} - 1) / \lambda$

```
## R-square value      bptest p shapiro test p durbin watson p
## 1      0.5892627 0.0003424676      0.5473221      1.497302e-57
```

The shapiro wilk shows that we easily have residual normal with $p=0.54 > .05$ while we see that there is autocorrelation via dwtest and heteroscedasticity which is improve - R squared also went up by a massive ~0.1 which means the variation we explain has increased substantially

So our explanation of variation improved but there is still steep autocorrelation and heteroscedasticity - we could combine Generalised Least Squares with correlated error to our Boxcox transformed data for the first one and try a different model with OLS and the Boxcox transformation and evaluate the performance of both model:

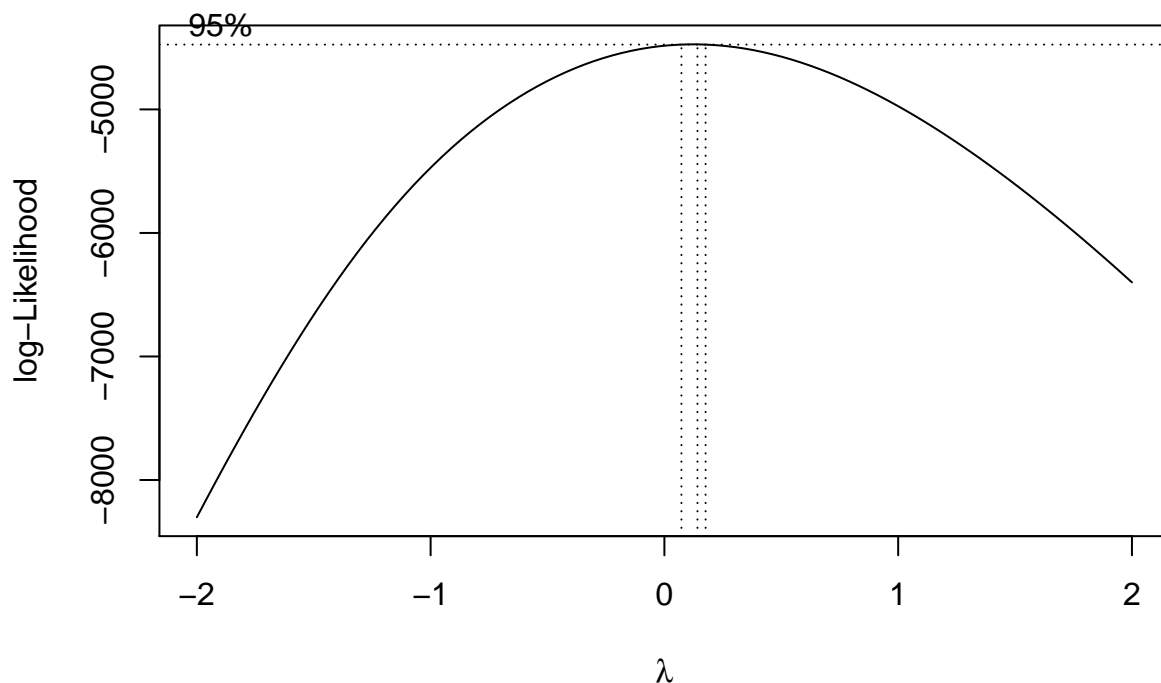
```
##   transormed only r-mse GLS Transformed r-mse original model r-mse
## 1                51.33086                52.18969                51.70159

## $corStruct
##      lower      est.      upper
## Phi 0.4397209 0.4903428 0.537859
## attr(,"label")
## [1] "Correlation structure:"

##   original aic   box aic gls with transformed aic
## 1      11631.74 -537.3763                3209.409
```

As we can see after applying a Generalised Least Square with correlated error with boxcox transformation we have an rmse of 52, which is slightly higher than the original and only transformed linear regression. Also we see that we cannot reject the null hypothesis of residual being correlated since the lower bound of Phi which is the correlation variable does not include 0 which is what we would need to reject the null hypothesis. This suggests that GLS may not be the best model for this data since there is still correlation and the r-mse is higher than OLS

Therefore, it may be a good idea to stick with OLS specifically the transformed regression for the sake of fitting a parametric model in this dataset.




```
##
## Durbin-Watson test
##
## data:  lm33
## DW = 1.029, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0

##
## Shapiro-Wilk normality test
##
## data:  residuals(lm33)
## W = 0.99765, p-value = 0.03448

##
## studentized Breusch-Pagan test
##
## data:  lm33
## BP = 2.9786, df = 8, p-value = 0.9357
```

Section 4: Conclusions

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