22294500 C7081_Assignment_ATC_2022_11_22_v20

A T Chamberlain

2022-11-21

Title

An appraisal of the ability of a range of statistical learning algorithms to model air quality (PM10) in Sweden from meteorological and traffic data.

GitHub link

https://github.com/TomChamberlain61/Assessment-C7081

Background

Particulate matter pollution is a growing global problem resulting in rising levels of ill health and mortality. There are two major sources of particulate pollution. Firstly open fires and burning wood creates small particles (typically under 2.5 microns and referred to as PM2.5). Problems are seen when wood fuel and open fires are used for heating and cooking and are recorded predominately in developing countries. The second source is from vehicle traffic - predominantly road. Pollution particles are larger (2.5 to 10 microns)and do not disperse so widely in the environment.

Particulate pollutants are typically measured as PM2.5 which includes are particles in the range 0.3 - 2.5 microns and as PM10 which refers to particles in the range 0.3 to 10 micron. Devices to measure particulate pollutant concentrations (microgram per cubic meter) are expensive and require regular attention to maintain consistent performance and to collect data. There is therefore an interest in predicting particulate pollution concentrations from meteorological and traffic data as these are easier and cheaper to obtain.

Aldrin and Haff (2005) used Generalised Additive Modelling to model PM10 concentrates from meteorological and traffic data. Other researchers have used a rage of traditional and more modern modelling algorithms - see Aldrin and Haff (2005) but they have not been formally compared and some more modern techniques (eg BART - Sparapani et al, 2021) have not been considered.

Objective

The objective of this study was to apply a range of statsical learning techniques as described by James et al (2021) to a publicly available subset of the data used by Aldrin and

Haff (2005). Performance will be formally assessed using RMSE and classification error estimations as well as for transparency, parsimony and simplicity.

Data

A publically available subset of the data used by Aldrin and Haff (2005) were obtained from http://lib.stat.cmu.edu/datasets/PM10.dat and consisted of 500 observations. The original data were collected by the Norwegian Public Roads Administration over 4 sites in Oslo but only one is site (Alnabru) is present in the data subset available. The original data were collected between October 2001 and August 2003.

The data available are as follows:

Response variable

- hourly values of the logarithm of the concentration of PM10 particles

Predictor variables

- logarithm of the number of cars per hour
- temperature 2 meter above ground (degree C)
- wind speed (meters/second)
- the temperature difference between 25 and 2 meters above ground (degree C)
- wind direction (degrees between 0 and 360)
- hour of day
- day number from October 1. 2001.

The original variable is continuously distributed which limits the types of analysis techniques that can be used. WHO (2021)defines a high concentration as being over 45 micrograms per cubic meter and this threshold will be used to define a categorical variable for pollution risk which allows a range of categorical analysis techniques to be explored.

Initial Data frame structure

```
names(PM10raw)
```

```
## [1] "logPM10"       "logCarNumbers" "temp2m"        "windSpeed"
## [5] "temp2mcf25m"   "windDirection" "hourofDay"     "dayNumber"
```

MATERIALS AND METHODS

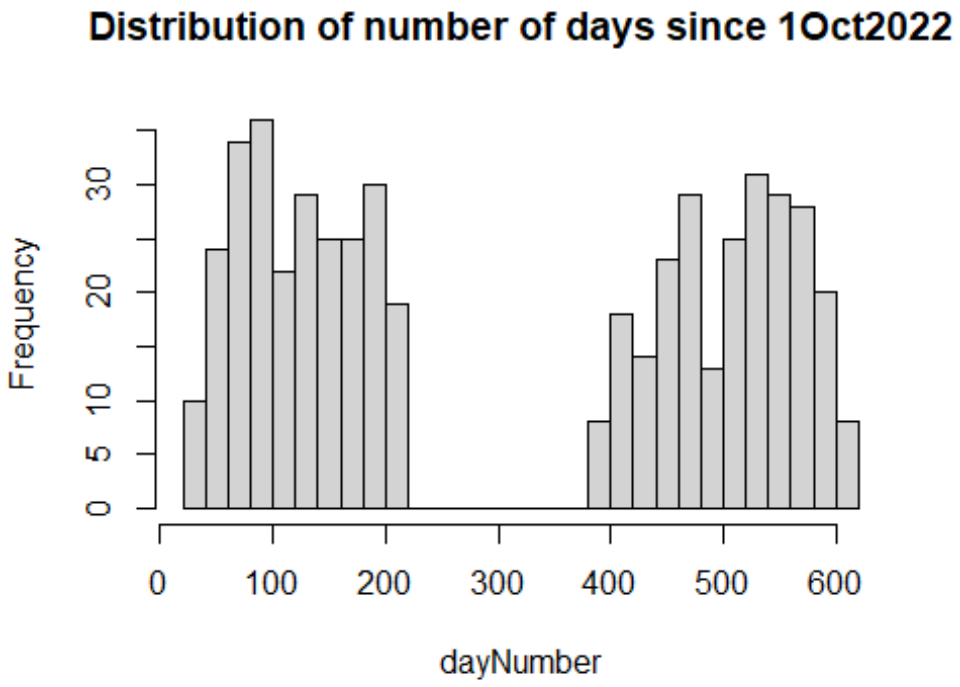
EDA

Data distributions, transformations and derivations

Data distributions were assessed using histograms and Gaussian distribution was formally assessed using the Shapiro-Wilks test. logCarNumbers looked skewed but simple power transformations did not notable improve the Shapiro-Wilks statistic. By contrast the

distribution of windSpeed was improved by taking the square root and this transformed variable was added to the data-frame.

Distribution by dayNumber (days since 1 Oct 2001) was unusual and worthy of note (see plot below)



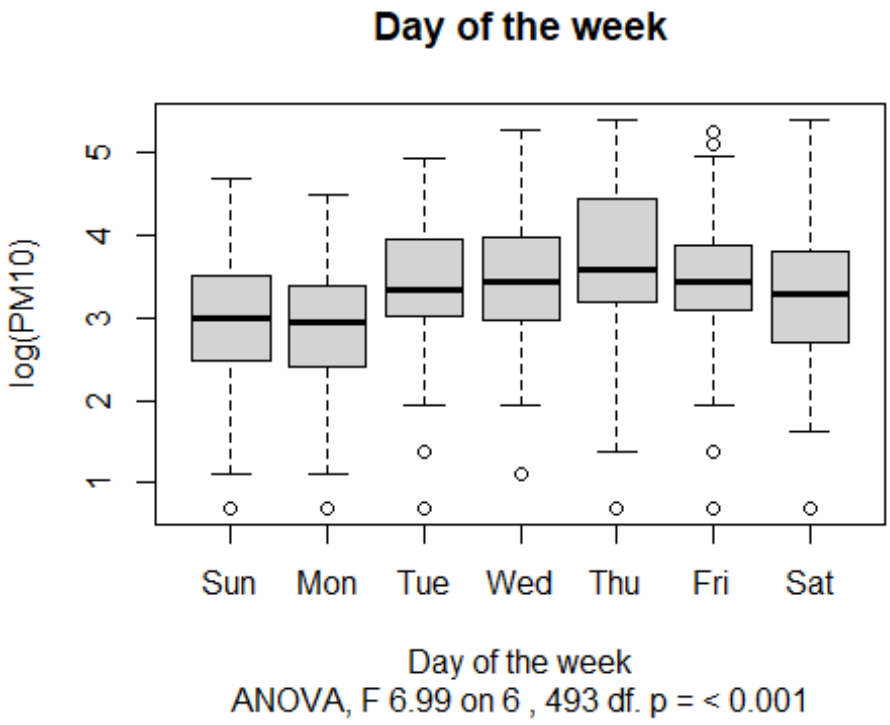
The data subset seems to cover the winter periods of two successive years. This may create some problems in further analysis as the data are a time-series set as they are on successive days. There is a risk that some data variables will be highly correlated over time such that the best predictor for day(t+1) is the data for day(t). If training and test data sets are created by selecting every other value they will be very similar and the 'test' subset will not be a true, independent set. For this reason the training set was taken as the first 250 (approx) values and the test set the second 250-ish values.

The variable windDirection required attention. In its raw state direction was represented as degrees on the compass. Although it seems continuous it has the odd characteristic that the difference between 1 and 2 degrees is the same as the difference between 360 and 1 degree. To overcome this a new factorial variable call windRose was created with windDirection ranging from 45 to 135 degrees classed as E and so forth.

Similarly dayNumber can be improved. Initially the dayNumber was transformed into a calendar date and from this the month and season (Dec,Jan,Feb = Winter etc) was derived. Day of the week (Monday, etc) was derived from the data and this was classified as Weekend (Sat, Sun) or weekday.

Further investigation of how logPM10 varied with day of the week showed that PM10 was lowest on Mondays rather than at the weekend suggesting there may be a possible lag. Successive ANOVA analyses showed that splitting out Sunday and Monday from the other days has the highest significance of any split (F = 34.25, df = 1,498) and a new categorical variable was set up to reflect this.

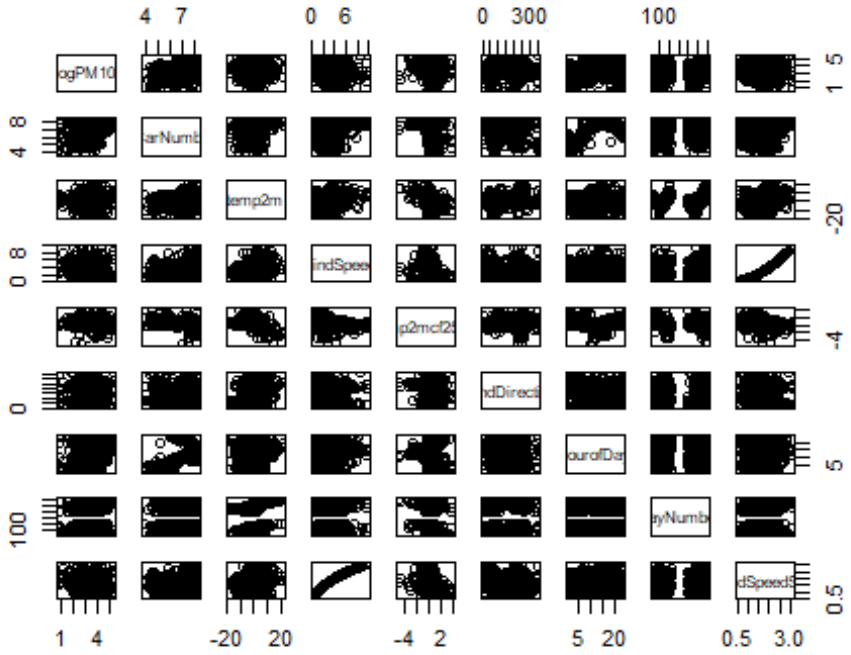
```
##   treatment        x
## 1       Sun 2.976900
## 2       Mon 2.853041
## 3       Tue 3.384632
## 4       Wed 3.394945
## 5       Thu 3.611331
## 6       Fri 3.418153
## 7       Sat 3.243621
```



The original dataset contained variables for temperature at 2m and the difference between temperatures at 2m and 25m but not the raw data for temperature at 25m. This seems rather inconsistent so the 25m temperature was derived and added to the data frame.

Data visualisation

Pairs analysis is too large and complex across all the variables that we now have, but can be visualised for the first nine variables.



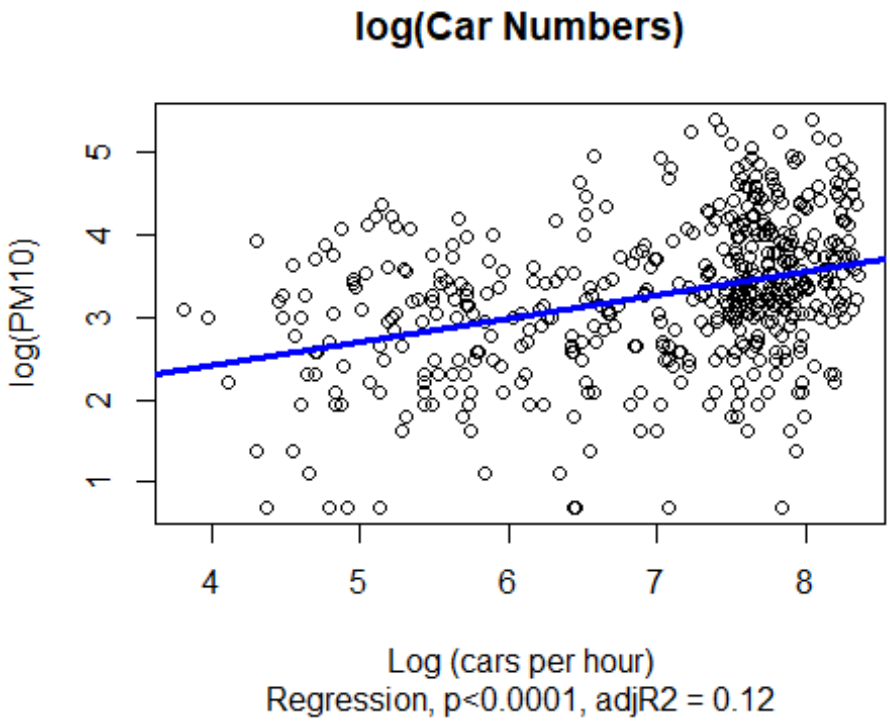
However even is visualisation is difficult to interpret so will look at numerical correlations to 3 dp.

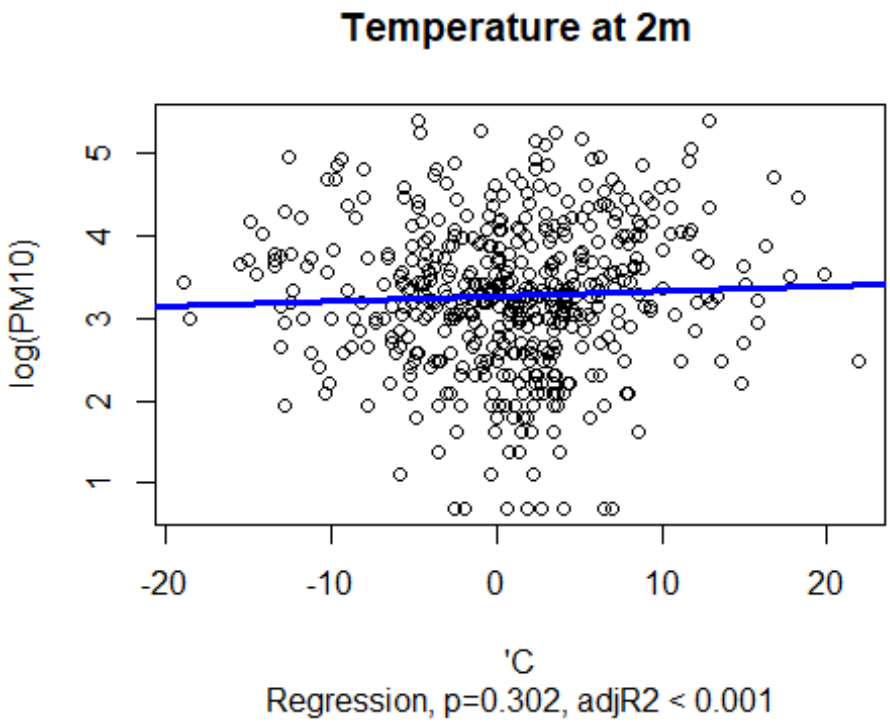
```
##                logPM10 logCarNumbers temp2m windSpeed temp2mcf25m
windDirection
## logPM10          1.00          0.36   0.05     -0.15       -0.07
0.02
## logCarNumbers    0.36          1.00   0.26      0.19       -0.35
0.01
## temp2m           0.05          0.26   1.00      0.21       -0.36
0.32
## windSpeed       -0.15          0.19   0.21      1.00       -0.27         -
0.09
## temp2mcf25m     -0.07         -0.35  -0.36     -0.27        1.00         -
0.04
## windDirection    0.02          0.01   0.32     -0.09       -0.04
1.00
## hourofDay        0.21          0.56   0.12      0.03       -0.07
0.04
## dayNumber        0.05          0.00   0.16      0.04       -0.14
0.04
## windSpeedSqrt   -0.17          0.17   0.21      0.98       -0.29         -
0.12
## year             0.10          0.01   0.20      0.03       -0.14
0.15
## monthNum        -0.13         -0.03  -0.16      0.00        0.06         -
0.24
```

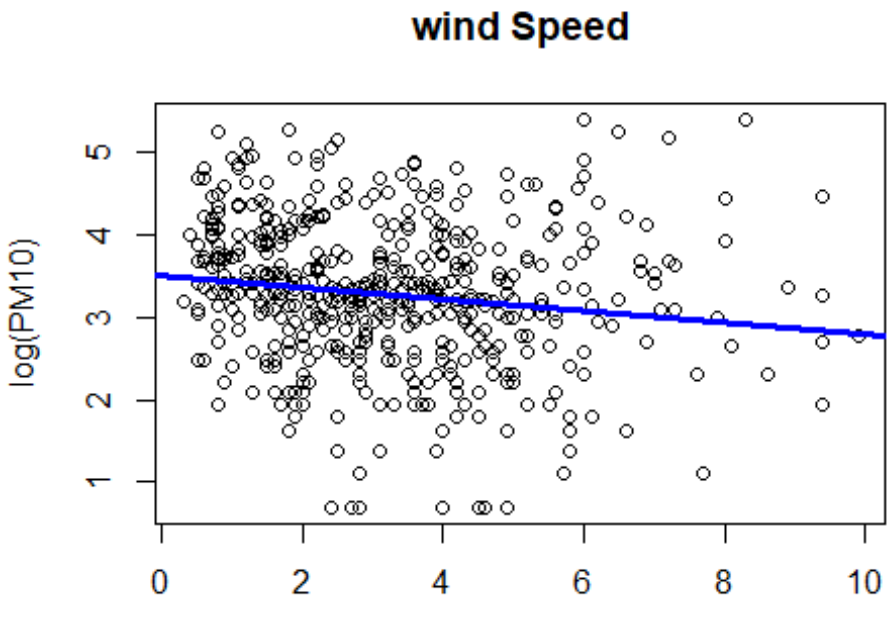
```
## dayOfWeek        0.17          0.08  -0.02      0.03        0.02         -
0.03
## trialMonth       0.05          0.00   0.16      0.04       -0.14
0.05
## temp25m          0.04          0.21   0.99      0.18       -0.21
0.33
##                hourofDay dayNumber windSpeedSqrt  year monthNum dayOfWeek
## logPM10             0.21      0.05         -0.17  0.10    -0.13      0.17
## logCarNumbers       0.56      0.00          0.17  0.01    -0.03      0.08
## temp2m              0.12      0.16          0.21  0.20    -0.16     -0.02
## windSpeed           0.03      0.04          0.98  0.03     0.00      0.03
## temp2mcf25m        -0.07     -0.14         -0.29 -0.14     0.06      0.02
## windDirection       0.04      0.04         -0.12  0.15    -0.24     -0.03
## hourofDay           1.00     -0.05          0.01 -0.05     0.02     -0.02
## dayNumber          -0.05      1.00          0.04  0.89    -0.27      0.02
## windSpeedSqrt       0.01      0.04          1.00  0.02     0.01      0.03
## year               -0.05      0.89          0.02  1.00    -0.68      0.03
## monthNum            0.02     -0.27          0.01 -0.68     1.00     -0.03
## dayOfWeek          -0.02      0.02          0.03  0.03    -0.03      1.00
## trialMonth         -0.05      1.00          0.04  0.89    -0.27      0.02
## temp25m             0.11      0.15          0.17  0.19    -0.16     -0.02
##                trialMonth temp25m
## logPM10              0.05    0.04
## logCarNumbers        0.00    0.21
## temp2m               0.16    0.99
## windSpeed            0.04    0.18
## temp2mcf25m         -0.14   -0.21
## windDirection        0.05    0.33
## hourofDay           -0.05    0.11
## dayNumber            1.00    0.15
## windSpeedSqrt        0.04    0.17
## year                 0.89    0.19
## monthNum            -0.27   -0.16
## dayOfWeek            0.02   -0.02
## trialMonth           1.00    0.15
## temp25m              0.15    1.00
```

Several variables are correlated. logCarNumbers is correlated with time of day (r=0.56) because at some repeatable times of the day the roads are busier. Temperature at 2m and at 25m are highly correlated (r=0.99) indicating that we do not need both in any future model. The temperature difference is correlated with logCarNumbers (r=-0.35) and temperature at 2m (r=-0.36). Other correlations are artifacts of the derivation of new variables such as dayNumber being correlated with year (r=0.89)

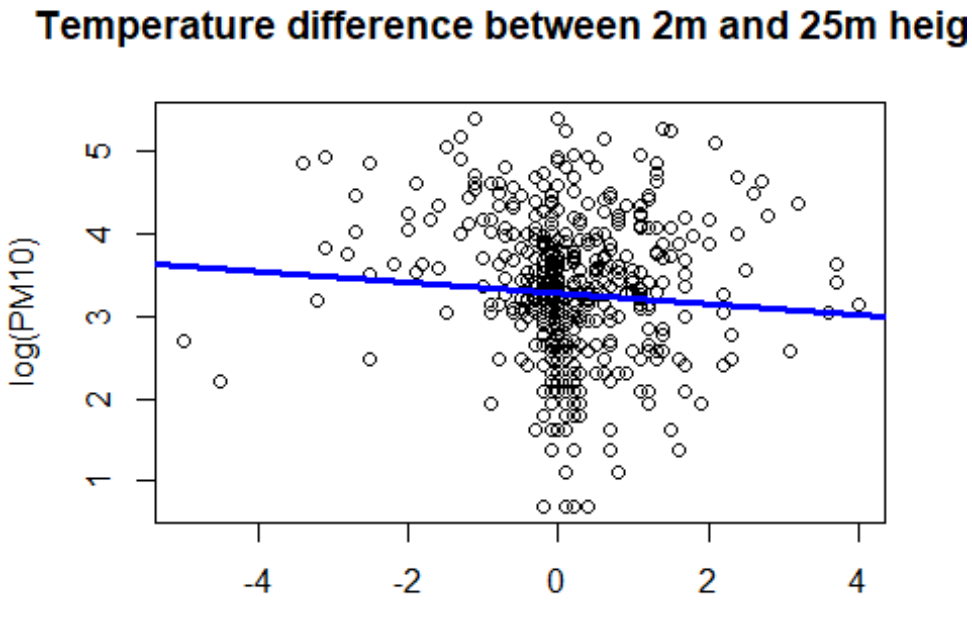
Finally look at some of the regression plots for logPM10 against possible explanatory variables.


log(Car Numbers)
log(PM10)
1
2
3
4
5
4
5
6
7
8
Log (cars per hour)
Regression, p<0.0001, adjR2 = 0.12

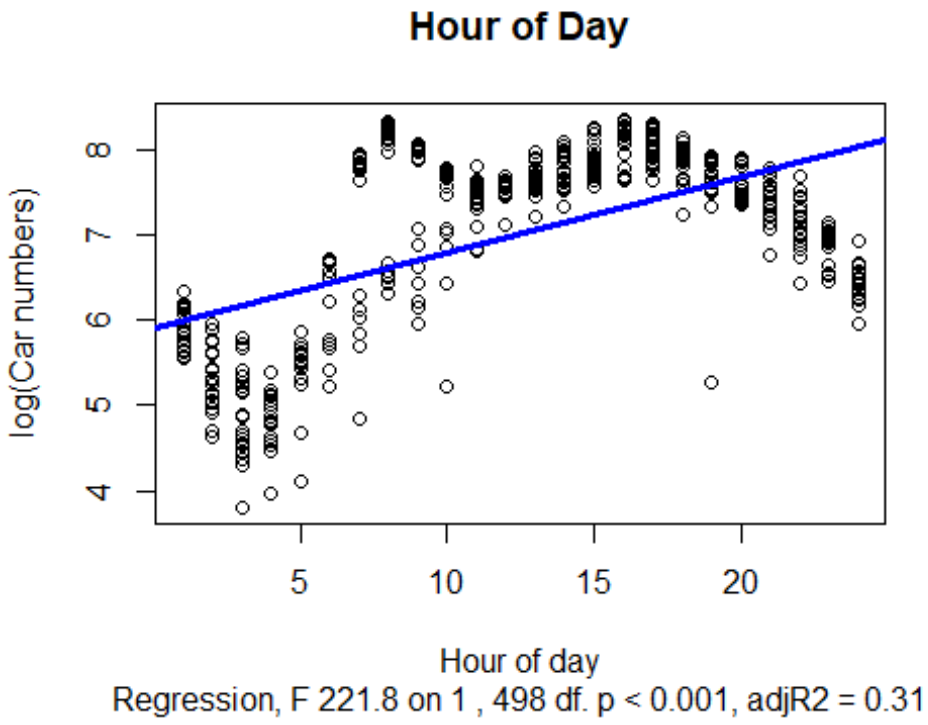

Temperature at 2m
log(PM10)
1
2
3
4
5
-20
-10
0
10
20
'C
Regression, p=0.302, adjR2 < 0.001



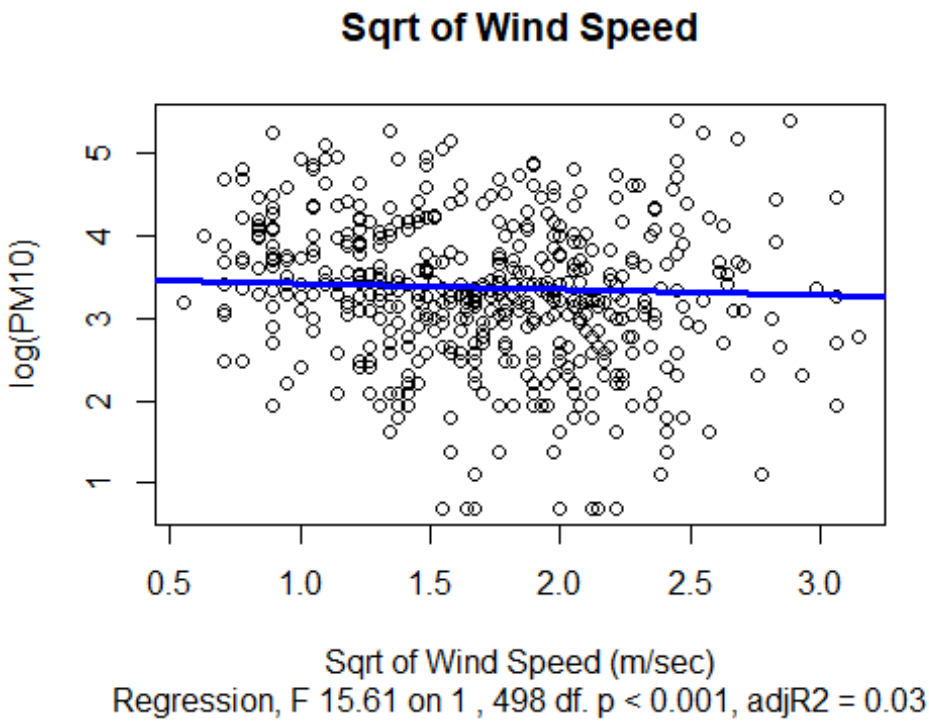
Regression, F 10.7 on 1 , 498 df. p < 0.001, adjR2 = 0.02



Regression, F 2.6 on 1 , 498 df. p < 0.11, adjR2 = 0.003



Regression, F 221.8 on 1 , 498 df. p < 0.001, adjR2 = 0.31



Regression, F 15.61 on 1 , 498 df. p < 0.001, adjR2 = 0.03

Conclusions from initial EDA and visualisation

- regression plots identify very few outliers - suggesting that data has been cleaned by the original authors.

- There are a small number of moderate correlations between variables ($p<0.05$) but none appear strong enough to indicate that a simple model will suffice.
- Single component regressions are generally poor with very small amounts of the variance accounted for in the adjRsqr.

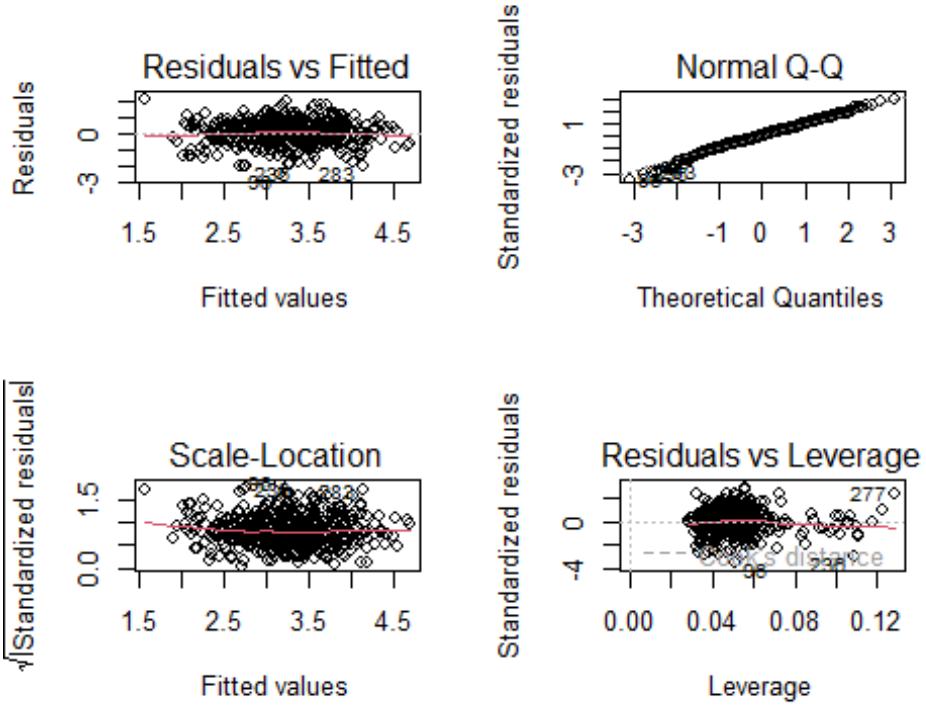
Assessment of a variety of Statistical Learning techniques

This section will explore the use of traditional and newer statistical techniques to model the particulate pollution concentration both as a continuous and as a categorical variable.

Simple linear regressions

Fit all explantory variables

```
## [1] 0.7100922

## Warning: not plotting observations with leverage one:
##   432
```



The regression has an adj R2 of 0.324 and is a highly significant fit (F = 10.55 on 25,474 df p < 0.0001). The residuals plot shows they are tightly clustered around the fitted line which is approx horizontal and they look to be normally distributed and consistent across the range of observations. One value has a very high leverage; this is related to a Temp2m value of was 21.9 which is high but not ridiculous so do not reject.

The RMSE for this regression is 0.710 log units. The IQR for logPM10 is 1.17 so the error is 60% of the IQR indicating that this is not a very useful model. However it does give a baseline RMSE against which other techniques can be compared.

The next section will look at several other simple regressions and determine the RMSE for each so that comparisons can start to be made. The subsequent section will look at the risk of the models being over fitted by using at training and test dataset.

Original variables

The variables used were logCarNumbers, temp2m, windSpeed, temp2mcf25m, windDirection, hourofDay and dayNumber.

Although the model's fit was highly significant ($F=15.06$, $df = 492,7$, $p<0.001$) the adj the model adj R2 was lower at 0.1765, and the RMSE had risen to 0.803. In general the models are highly significant even when they do NOT have good RMSE values -the latter will be reported more in subsequent analyses.

Transformed variables

The variables used were logCarNumbers, temp2m, temp25m, windSpeedSqrt, temp2mcf25m, windDirection, hourofDay, dayNumber.

The model's adj R2 is 0.1876, and the RMSE 0.798 which is still below the model that included every variable.

Manually selected variables - based on correlations and scientific background of the required models

The variables used were logCarNumbers, temp2m, temp25m, windSpeedSqrt, temp2mcf25m, windRose, hourofDay, season and SplitDaySunMon.

The model's adj R2 is 0.3089, and the RMSE 0.736 which is better but still below the model that included every variable.

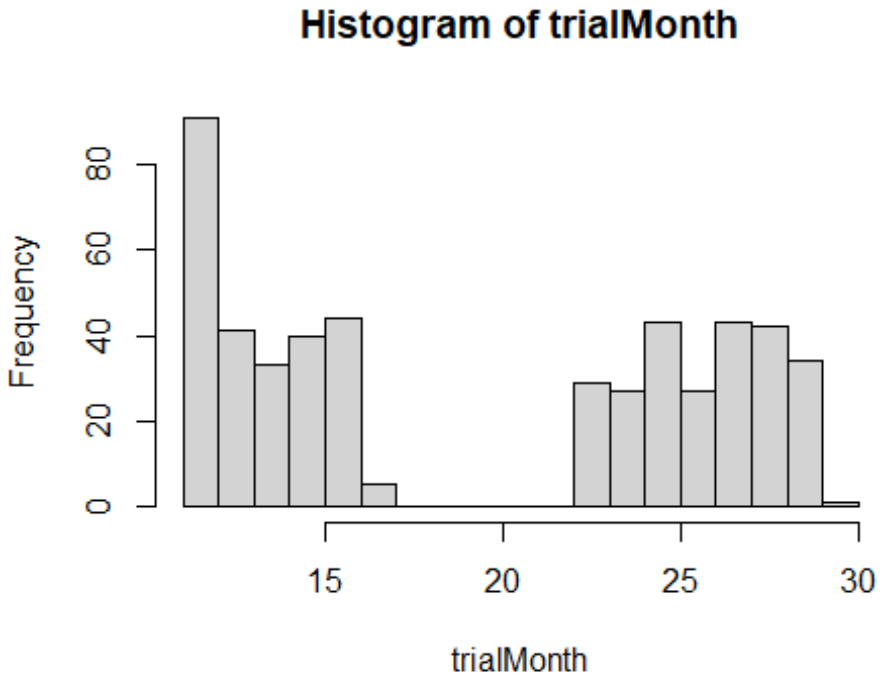
ANOVA was used to compare the five models developed so far
* logCarNumbers * Original * TransformedVars * ChosenVars * Everything

The Original model is better than CarNumbers alone ($p<0.001$), adding the transformed variables does not improve the model. Moving to chosenVars is an improvement ($p<0.0001$)and is using everything ($p=0.0008$). Initial data visualizations does not suggest that higher powers (quad, cubic etc) of any of the variables will improve the models so none were explored.

Model training

The above models were fitted to the entire data set and there is a risk that they are dataset-specific and are over fitted. The are various methods to avoid this such as Kfold and test-train splits. Some algorithms include a Kfold option; where this is not a feature a Test:train approach was used. As can be seen in the accompanying figure the trialMonth splits into

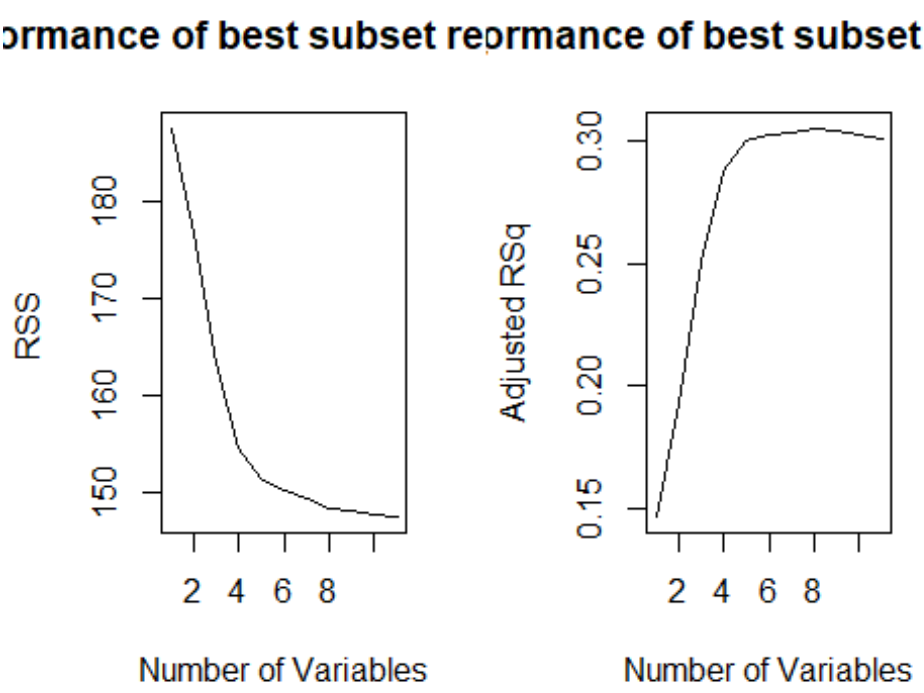
two clusters. Data collected before month 20 were used as the training dataset, after month 20 were the test set. This results in 254 observations in the training data set and 246 in the test set. The test dataset contained one observation where season was classified as summer, there were no summer classifications in the training set. The 'summer' observation was for 1 June when Spring ended 31 May; this observation was recoded as Spring and 31May.



Best subset linear regression

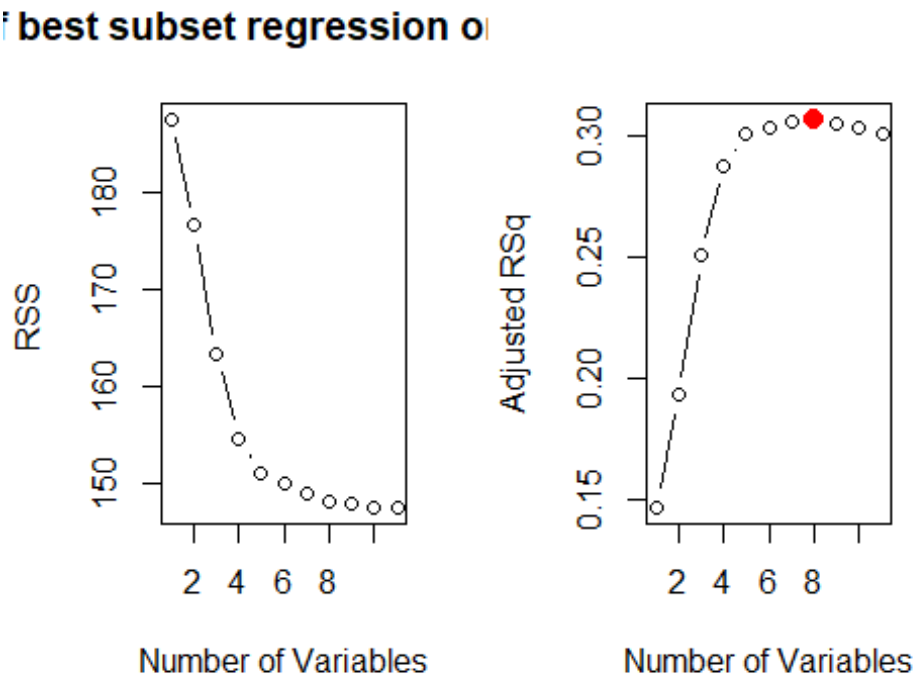
With 21 variables in any possible model there are over four million possible combinations which took too long on my computer (over 10 hours). A smaller set of 9 variables was therefore chosen.

```
## Reordering variables and trying again:
```



There is no maximal adj R2 but gains tend to slow after about 7 or 8 components. Model selected (p=8) is logCarNumbers, temp2m, windSpeedSqrt, temp2mcf25m, windRose, hourofDay, season and weekDay. This is a similar composition to fitChosenVars above. When running a train/test procedure need to do ALL Of the analyses on the Train set and then apply to the test set rather than use the model fitted to ALL of the data.

```
## Reordering variables and trying again:

## [1] 8

## [1] 0.3069068
```

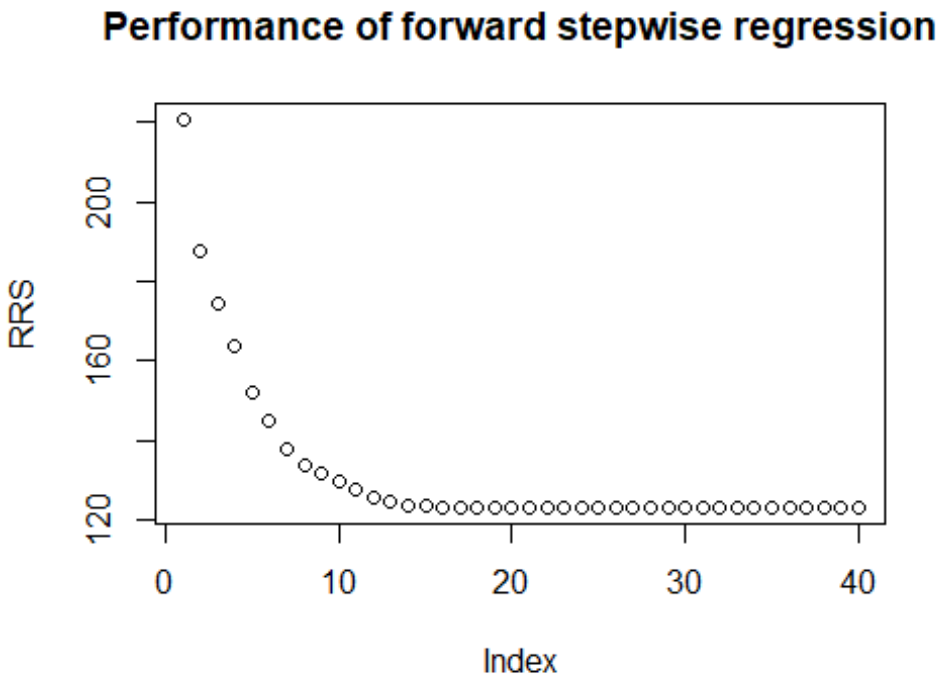


Rate of gain in R2 falls off after 8 variables (adj R2 at p=8 is 0.307 which is highest value) and suggests an 8 component model will be useful. Eight component model is logCarNumbers, hour of day, windSpeedSqrt, windRose and season. 2 are continuous, 2 are categorical (hourofDay is ordinal).

Forward stepwise regression

Compare with forward and backward stepwise regression with entire set. Use all of data set as still looking to find best LM. The plot of RSS against model size shows the change in fall in RSS starts to diminish after about p=13. This model is logCarNumbers+ temp2m + windSpeed + windSpeedSqrt + windRose + day + month + year season. The RMSE for this model when applied all of the data was 0.711. The model was then applied to the Training set only and assessed on the Test set and the RMSE rose to 0.781.

```
## Reordering variables and trying again:
```



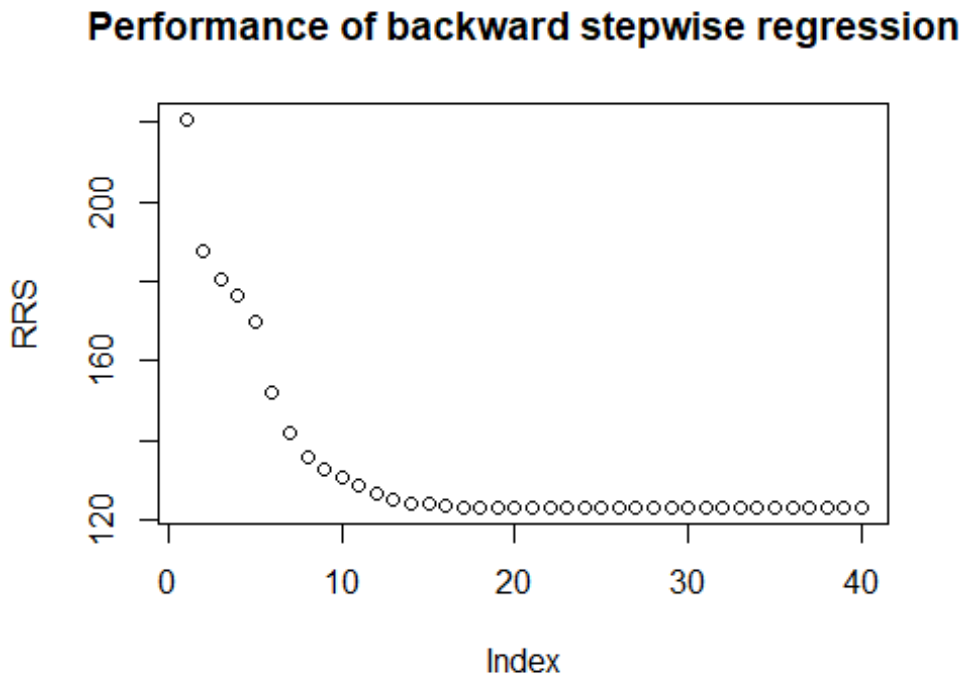

```
## [1] 0.7109654
```

```
## [1] 0.8184697
```

Backward stepwise regression

Backward stepwise regression was run on all variables available and RSS plotted against number of variables. The pattern of decline in RSS had a slightly different pattern but again improvement falls off after 13 components. The chosen model is logCarNumbers, temp2m, windDirection, hourofDay, WindSpeedSqrt, windRose, day and month. The RMSE when this model was applied to the entire data set was 0.716, when the model was trained then tested the RMSE rose to 0.8184.

```
## Reordering variables and trying again:
```



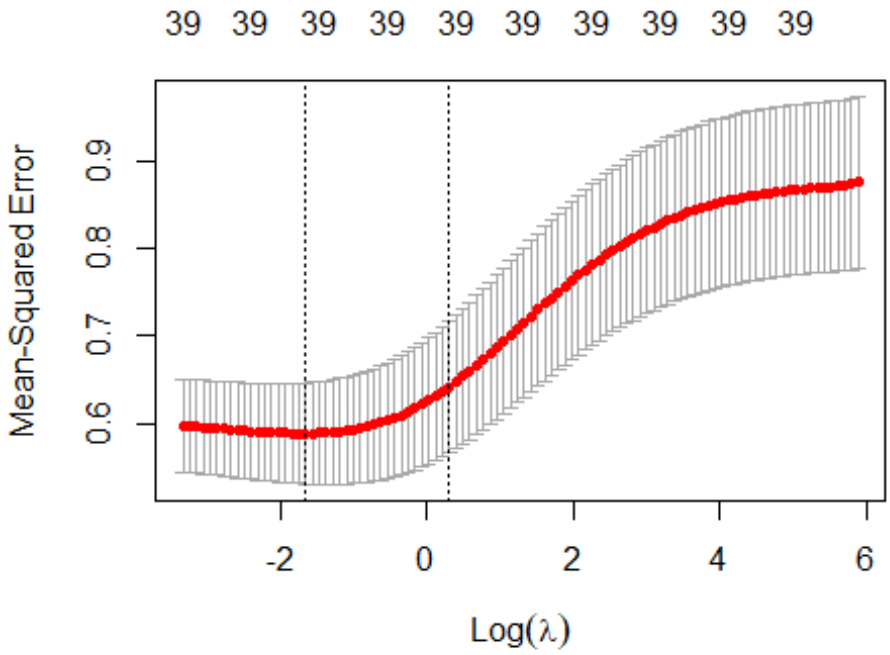

```
## [1] 1.023187
```

```
## [1] 0.8184697
```

Ridge regression

All variables

Ridge regression was applied with a grid of Lambda values. The model was developed with all variables using the Training set and validated with ten fold cross validation over the range of Lambda values. The best lambda value was 0.192 and a model using this value was used to predict values in the Test data set. Performance was assessed using the RMSE which was 0.7032. The model derived used all predictors so should be compared to lm.fitEverything where RMSE was 0.710 so this was a marginal improvement.

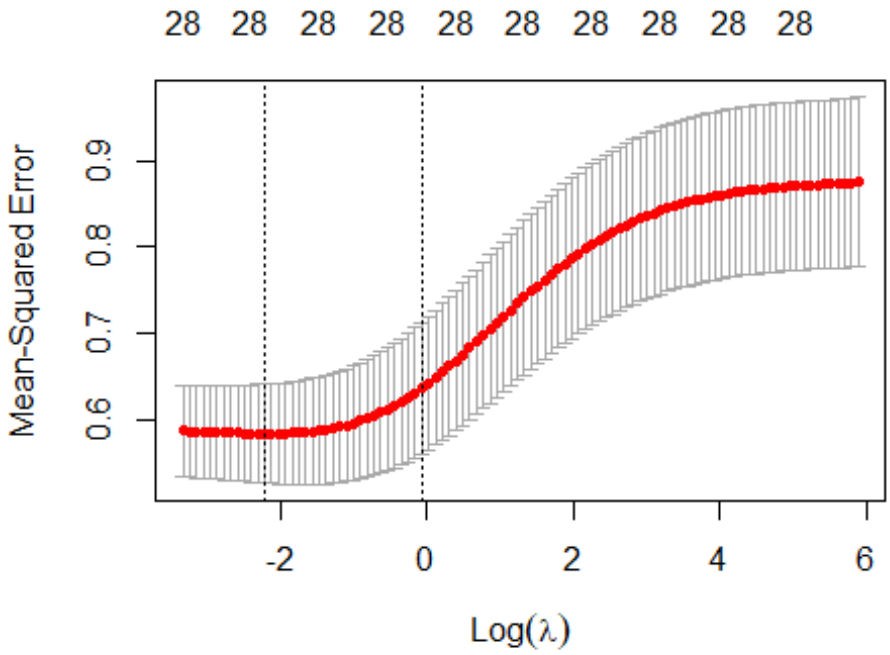



```
## [1] 0.1927846
```

```
## [1] 1.103101
```

Chosen variables

Ridge regression does not select and rank variables so repeat with chosen and SW.forward variables sets in case the full set of variables caused too much over-fitting. The optimal lambda value was 0.1103 and the RMSE when applied to the test set was 0.6994 which is a marginal improvement.

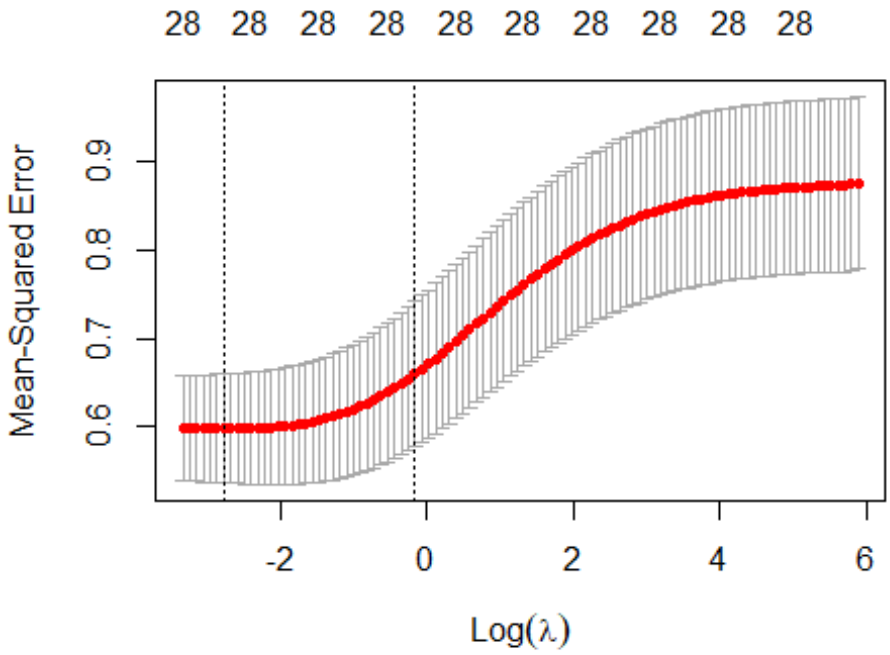



```
## [1] 0.1103184
```

```
## [1] 0.9962832
```

Stepwise Forward variables

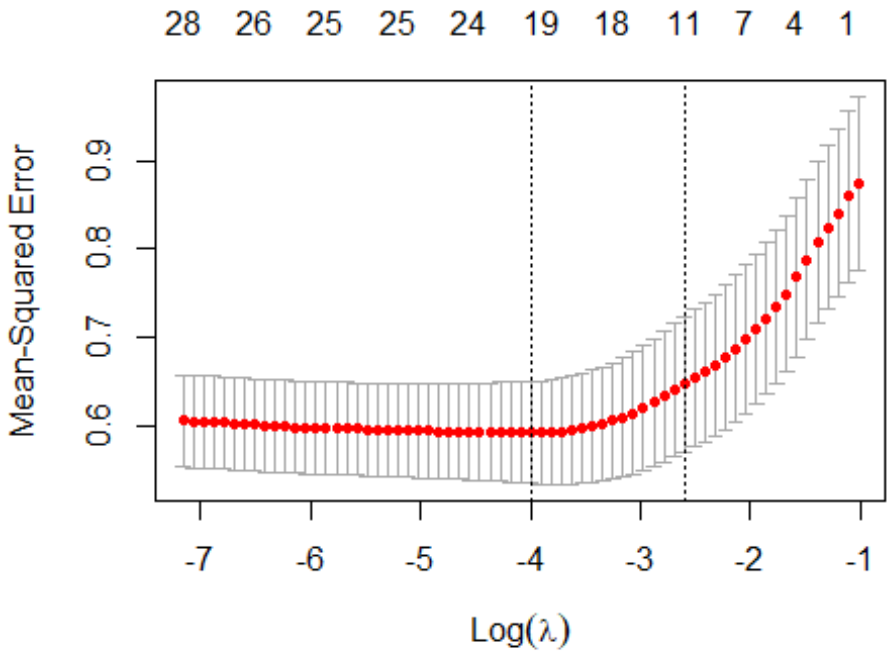
The RMSE for the test data set was 0.6977 which is again a (very) marginal improvement. With such marginal gains models should be selected for the quality of their input variables; what is the error in obtaining the data - for example is measuring temperature and windspeed.




```
## [1] 0.06312827

## [1] 1.006802
```

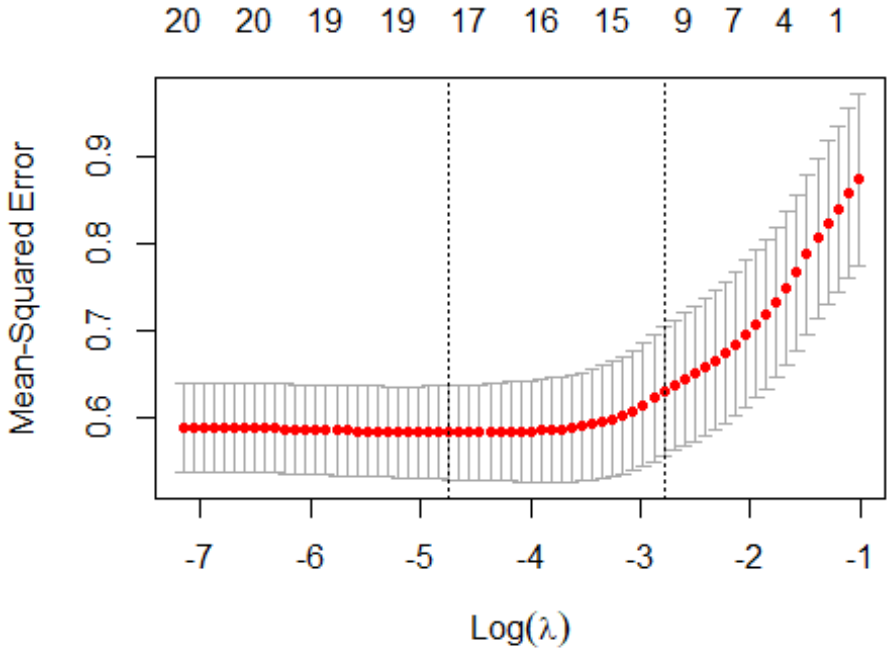
Lasso regression

Again the first model uses all of the variables and is trained and then tested. The resultant RMSE was 0.8219, for the chosen variables the RMSE was 1.103 and for the SWForward variable set it was also 1.103.




```
## [1] 0.01840223
```

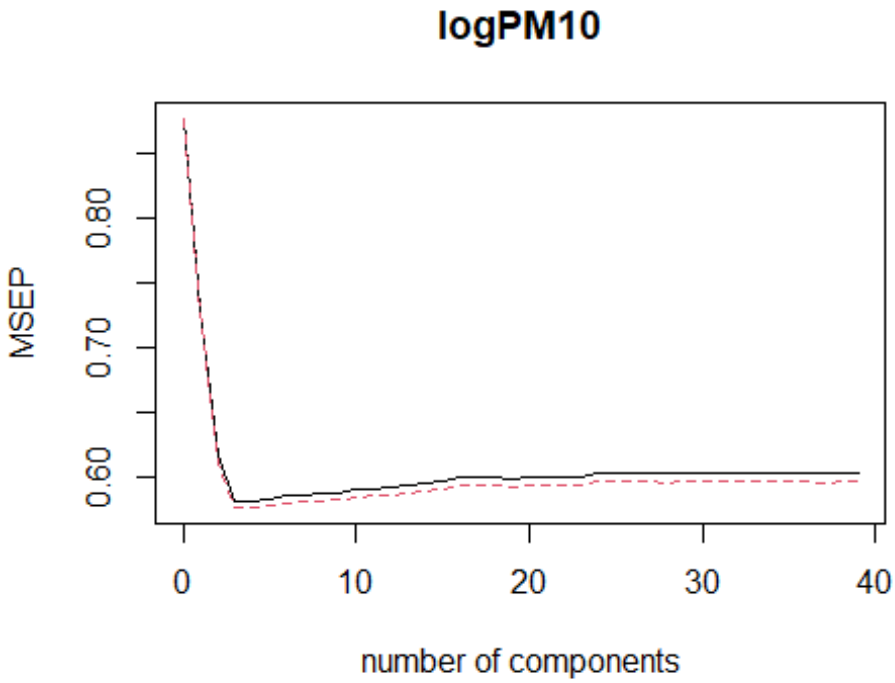
```
## [1] 0.8218905
```



```
## [1] 0.008742548
```

```
## [1] 1.006802
```

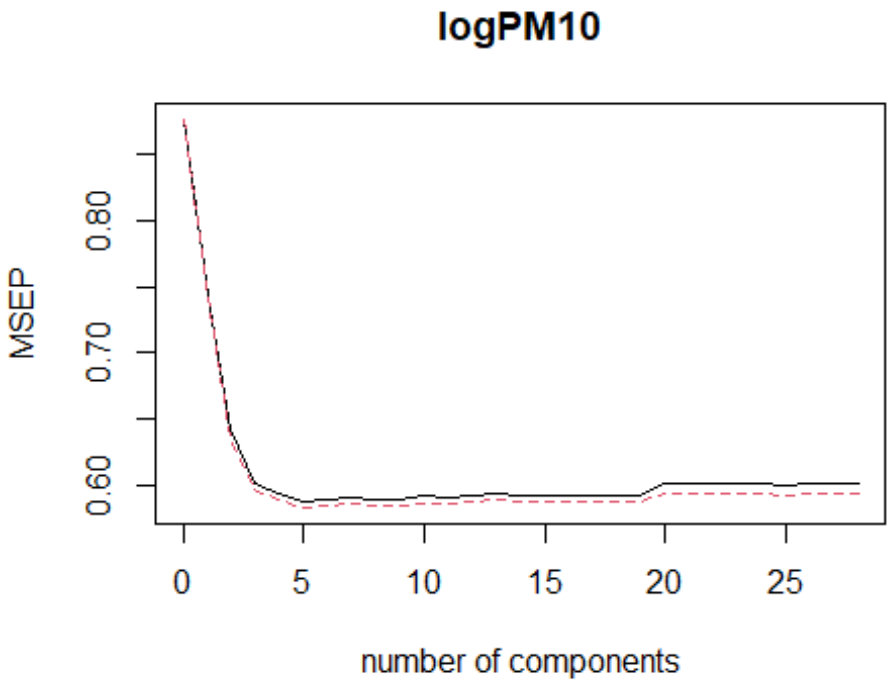
Initially run the model on all the variables and applied to the training data set. This carries out ten fold cross-validation and a plot of the MSEP against number of components shows a clear low point with three components. Applying this model to the Test dataset gives a RMSE of 0.7074.




```
## [1] 0.7074268
```

Variable set identified with stepwise forward algorithm

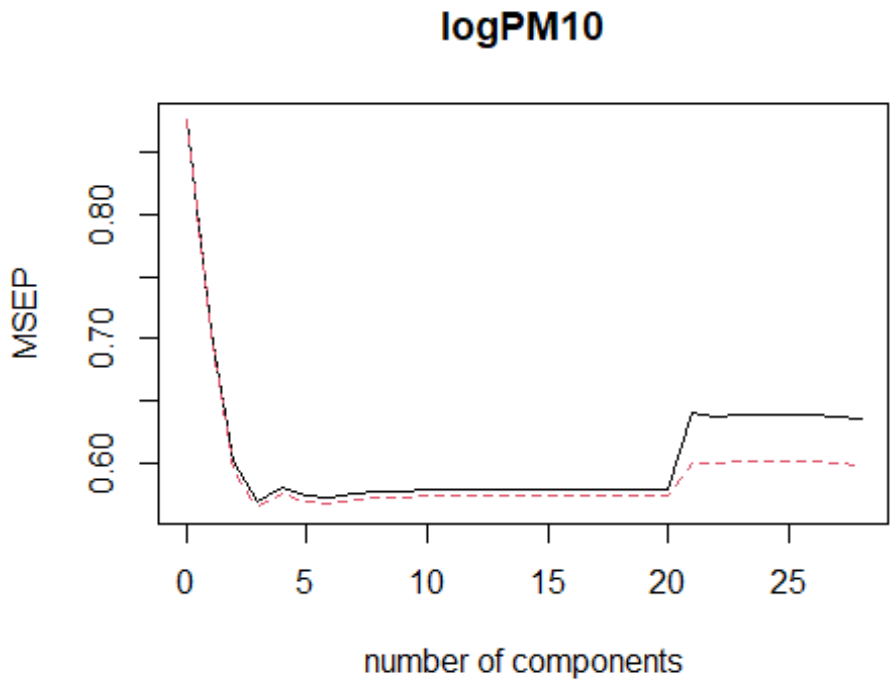
With this model the minimum MSEP is not so clear but assume it is at M = 5. When this model is applied to the Test data set the RMSE is 0.7106 so very similar and fewer variables involved.




```
## [1] 0.710651
```

Variable set chosen from earlier assessments.

The variables offered to the PLS algorithm were logCarNumbers, temp2m, windDirection, hourofDay, windSpeedSqrt, windRose, day, month and season. The PLSR model with three components had the lowest MSEP so this was applied to the test data set an gave a RMSE of 0.7025.




```
## [1] 0.7025516
```

Generalised Additive Models - GAM

The above PLS models identified very little gain in RMSE from the different variable combinations. As the GAM routine requires a lot of manual fitting to find the best model will only use the Chosen9 dataset in this section. Initially apply without smoothing splines and then with smoothing splines. The RMSE for these two variants were 0.8015 and 0.8185 respectively.

```
## [1] 0.8015062

## [1] 0.8184697
```

Could redo the natural splines with differing numbers of variables and compare predictions using ANOVA but the 9 component model did not have a very good RMSE on the test set so little point and removing variables will not improve the RMSE.

That completes the modelling of PM10 counts as a continuous (log transformed) variable Ridge, Lasso and PLS regression gave an improvement in RMSE over simple regressions but GAM did not give much improvement.

Categorical / dichotomous variable analysis

If logPM10 is treated as a dichotomous variable can then investigate the value of investigate algorithms such as logistic regression, clustering, SVM. Split will be if PM10 is

above or below FAO(2020) guideline of 45 ug/cubic meter.First step required is to create a Category variable in all three data sets and convert this to a factor.

```
   High  Low
```

raw 130 370

Train 61 185

Test 69 185

All three datasets have a 3:1 bias in favour of Low so may end up being better at predicting Low. Furthermore just classifying all predictions as Low will be correct in 75% of cases overall - but will have zero accuracy in predicting High observations which are probably of greater practical interest.

```
##
## High  Low
##  130  370

##
## High  Low
##   61  185

##
## High  Low
##   69  185
```

Logistic Regression

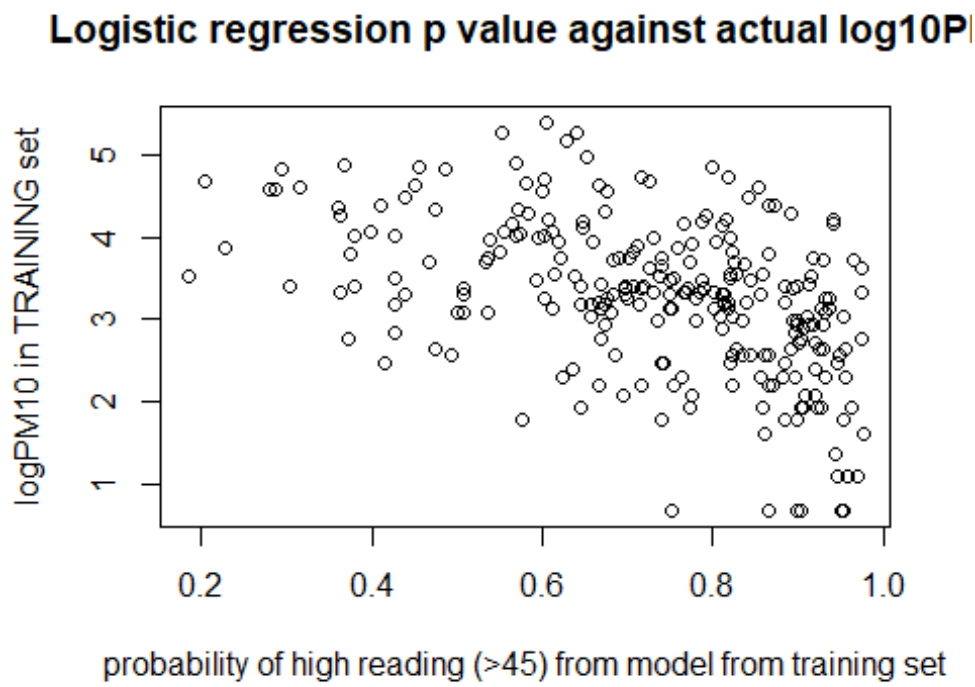
Explore some logistic regression models as initial EDA. Initially offer up all the possible variables. From this large model just pick out those variables with interesting p values (ie p<0.2) - avoid Type II errors - FOMO and then cross validate using the Test and training data sets. The smaller model looks a better model as no variables with high p-values that just add noise. The plotted results for the initial model do not look impressive - even on the training set! The confusion matrix is:

```
        High Low
```

HighPred 44 156

LowPred 17 29

Only 29% of observations are predicted correctly but 44/(44+17) but 72% of High observations were correctly predicted.




```
##
## glm.pred   High Low
##   HighPred   51 171
##   LowPred    18  14

##
## glm.pred   High Low
##   HighPred   44 156
##   LowPred    17  29

## [1] 0.296748

## [1] 0.7213115
```

Before we move on from Logistic regression will look at different probability cut-off points for defining 'High'

Cut off

0.5 (58 + 0) / 246 23.5 %

0.75 (11 + 72) / 246 33.7 %

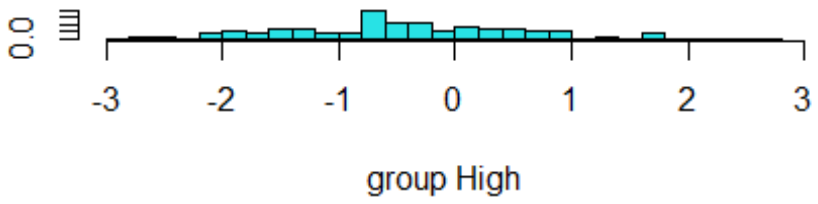
0.85 (5 + 72) / 246 31%

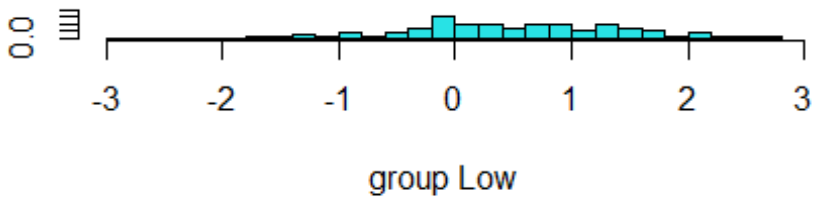
None of these look materially better than just assuming it will always be Low which has a 100-26 = 74% success rate.

```
##
## glm.pred   High Low
##   HighPred   58 185
##   LowPred     3   0

## [1] 0.2357724

##
## glm.pred   High Low
##   HighPred   11 113
##   LowPred    50  72

## [1] 0.3373984

##
## glm.pred   High Low
##   HighPred    5  72
##   LowPred    56 113

## [1] 0.3130081
```

Linear Discriminant Analysis (LDA)

A model offered all the possible variables showed too much co-linearity so have used a reduced variable set. Some separation can be seen and the predictive accuracy is 71% overall which is almost as good as assuming all are low! 29.5% (18 out of 61) High observations were predicted correctly.





```
## [1] "class"     "posterior" "x"

##
## lda.class High Low
##      High   18  28
##      Low    43 157

## [1] 0.7113821
```

Quadratic Discriminant Analysis

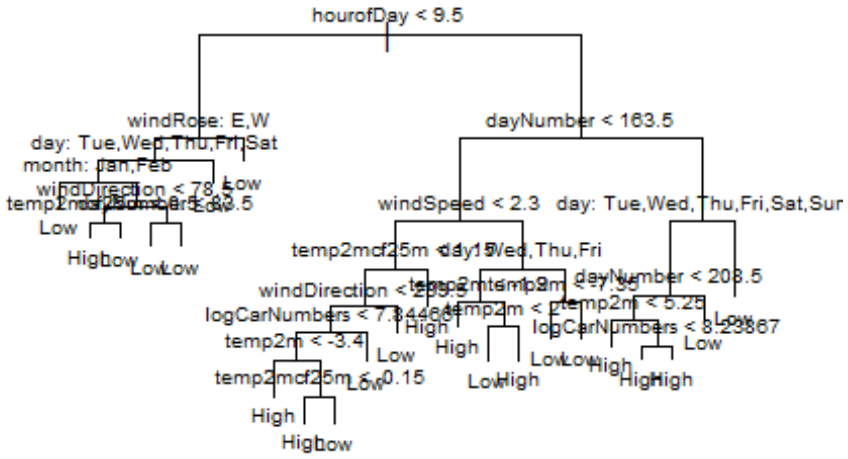
Used the same, chosen data set as in LDA. Again can see some separation - maybe a bit better but the classification is not as good - 65%

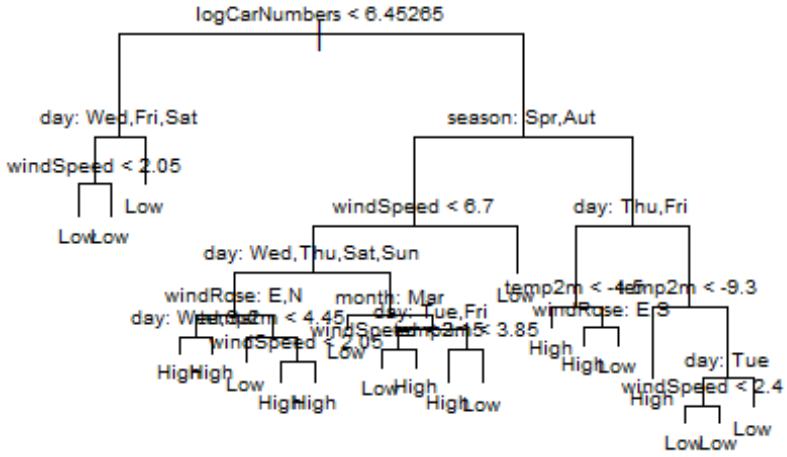
```
##
## qda.class High Low
##      High   29  53
##      Low    32 132

## [1] 0.6544715
```

K-Nearest neighbours

This section will use the smaller 'chosen' variable set and will iterate through a range of values for K. With the raw data the optimal k looks to be 5 where 71% of observations are correctly classified. Where the variables are standardised k rises to 10 and 75.6% of observations are correctly classified.

Decision trees



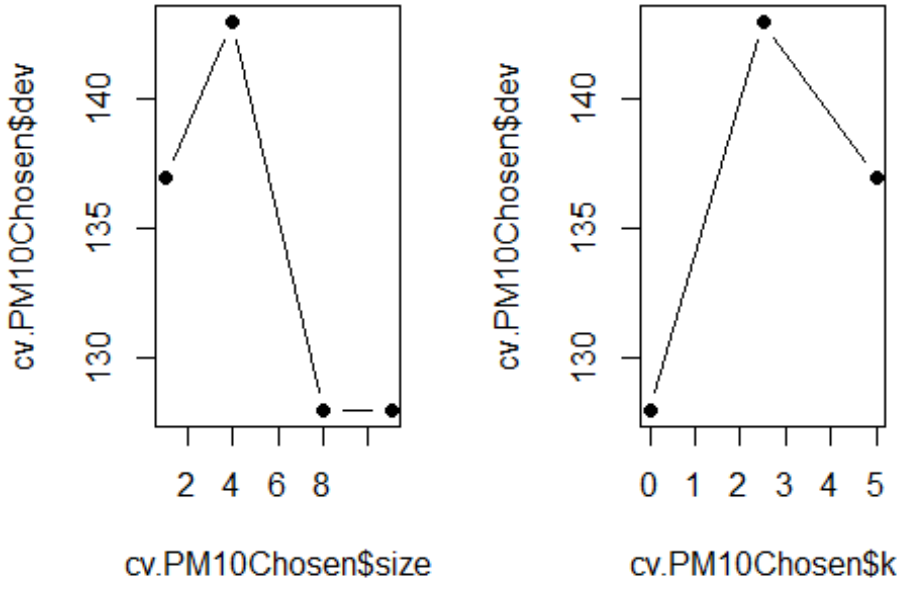



```
##
## tree.PM10TestChosen High Low
```

```
##                High   24  49
##                Low    37 136

## [1] 0.6585366
```

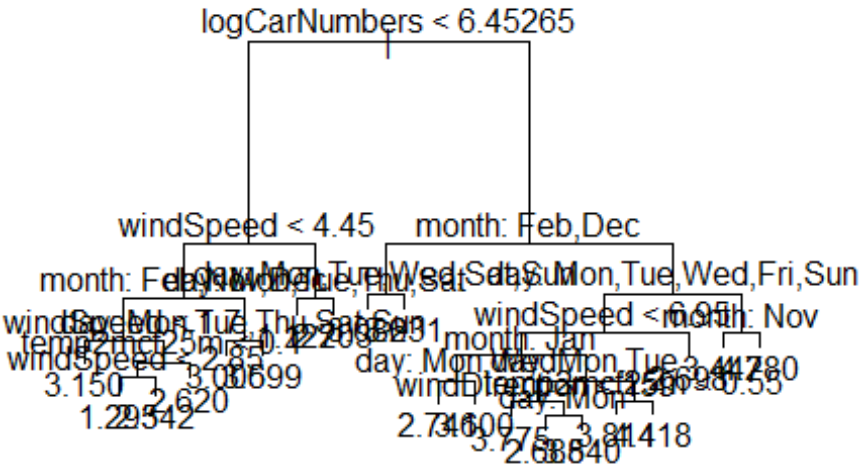
The initial whole-set model uses 10 variables and mis-classifies 8% of observations. The Chosen variable set had a mis-classification rate of 11% and the SWForward set 13%. The latter looks a simpler tree and many of the branches are the same (Low and Low) so could be pruned. When applied to the Test dataset correct classifications falls to 65%.

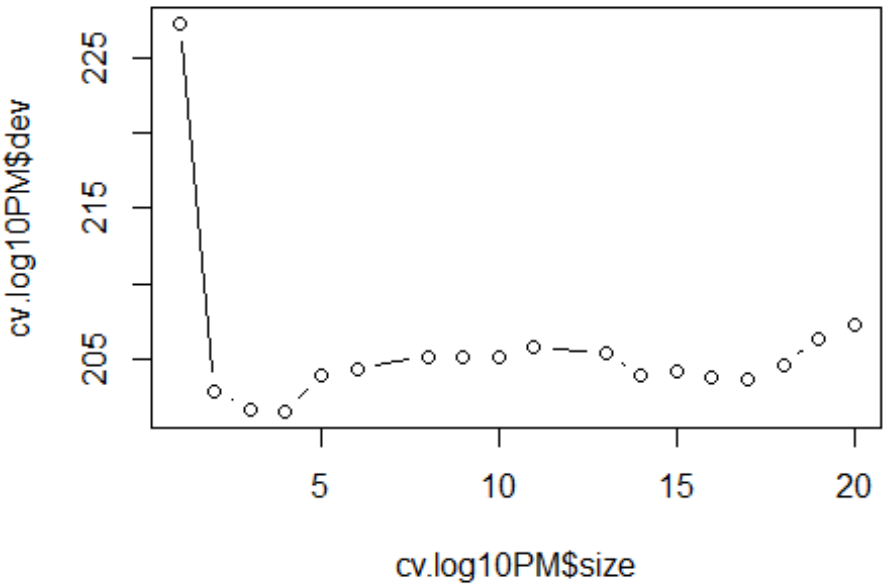
Look at the effect of pruning on the 'Chosen' variable set. 8 or 11 looks to be optimal node count - parsimony suggests 8 as the gains are small. With 8 nodes 79.2% of observations are correctly classified, with 11 nodes this drops marginally to 78.8% on the Test datasets. The latter classifies more highs correctly (51 cf 49) so might be preferred.

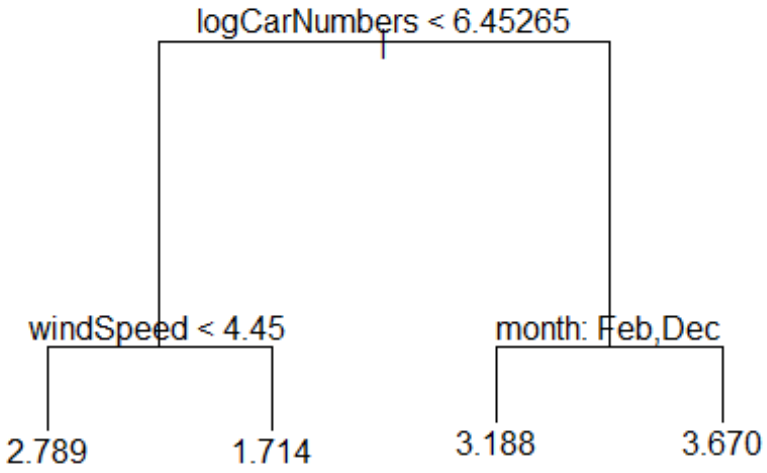


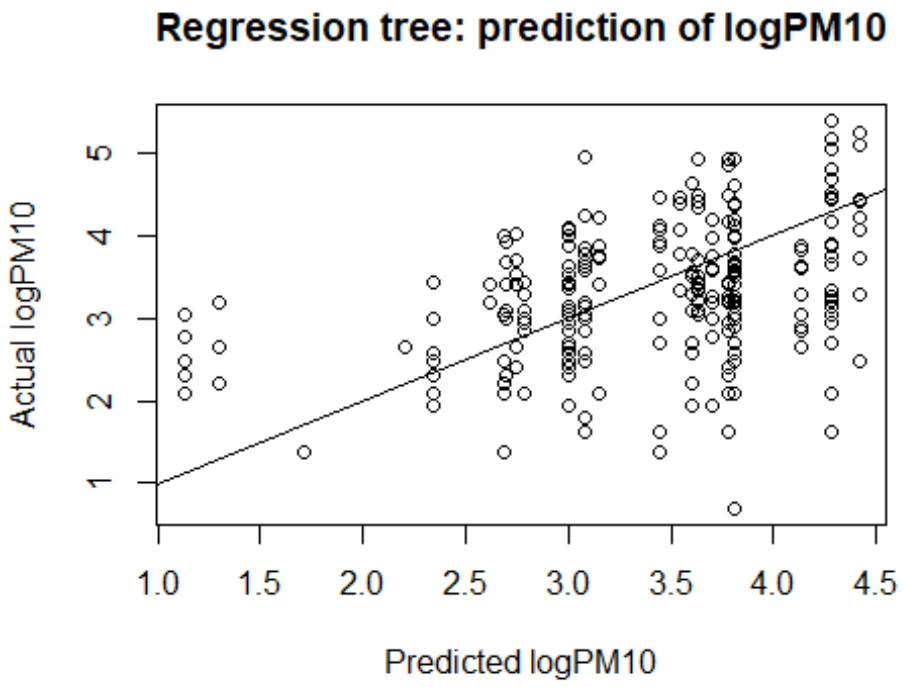

```
##
## tree.pred High Low
##      High   49  39
##      Low    12 146

## [1] 0.7926829

##
## tree.pred High Low
##      High   49  42
##      Low    12 143

## [1] 0.7886179
```

Can try creating a regression tree - bit hybrid as just chops up log10PM into sections. Using all variables gives an ugly tree and NA's have been introduced by coercion. Re-running the model with just those variables used (see tree figure) avoids this warning. A plot of the gain in performance against tree structure shows the gains come early in the branching and a 'best' of 4 nodes was selected and the predicted logPM10 plotted against actual. This had a RMSE of 0.893 which is poorer than previous methods.


logCarNumbers < 6.45265
windSpeed < 4.45
month: Feb,Dec
day: Mon,Tue,Wed,Fri,Sun
windSpeed < 6.95
month: Nov
month: Jan
day: Mon,Tue
3.150
2.620
2.734
3.418


cv.log10PM$dev
225
215
205
5
10
15
20
cv.log10PM$size


logCarNumbers < 6.45265
windSpeed < 4.45
month: Feb,Dec
2.789
1.714
3.188
3.670

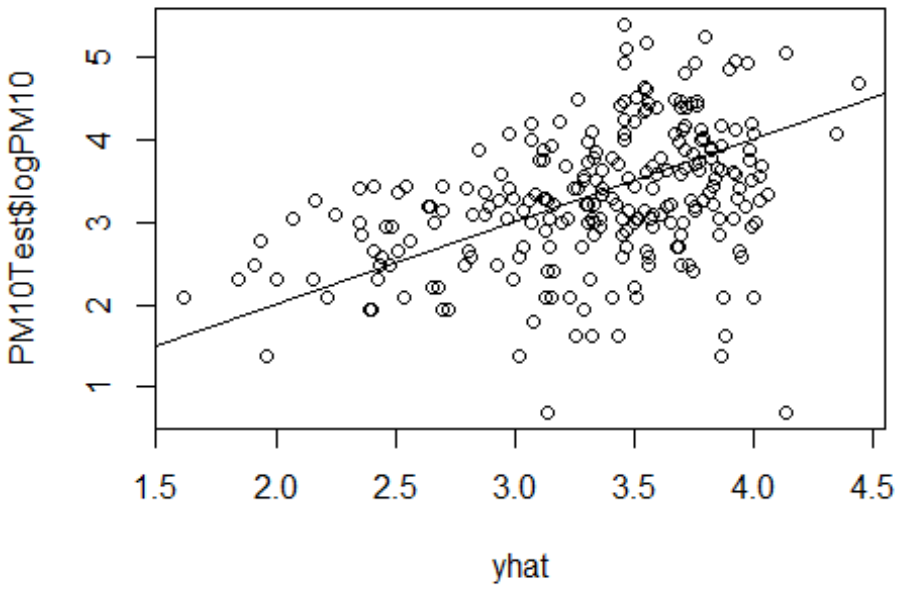

Regression tree: prediction of logPM10
Actual logPM10
Predicted logPM10
1
2
3
4
5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

```
## [1] 0.8932166
```

Bagging

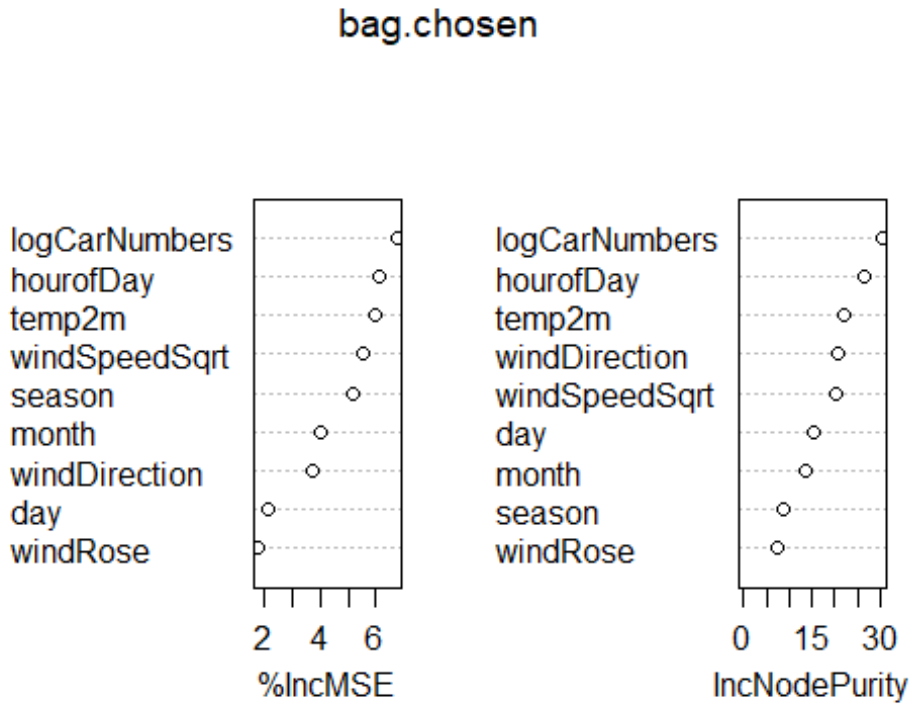
Again as there is some manual steering of parameters so will limit modelling to the Chosen9 and SWForward variable set. Initially, as EDA, just work with fairly standard forest size and mtry values. A plot of predicted against actual shows a decent relationship (which as usual gets clearer once the regression line is added) and an encouraging RMSE value of 0.7945. Next steps are to manually optimise tree numbers and mtry value. Model does not improve very much beyond a forest size of 100 trees. and this value is taken forward into an optimisation of mtry. The final optimised model has an RMSE of 0.768 which is a notable improvement over the initial RMSE obtained with bagging. The relative importance of each variable is ranked as expected with a fairly linear decline in importance with no obvious break point or 'elbow'.

- Trees = 25 RMSE = 0.840
- Trees = 100 RMSE = 0.794
- Trees = 500 RMSE = 0.7944
- Trees = 1000 RMSE = 0.790
- mtry = 1 RMSE = 0.768
- mtry = 2 RMSE = 0.787
- mtry = 3 RMSE = 0.780
- mtry = 5 RMSE = 0.791
- mtry = 8 RMSE = 0.800

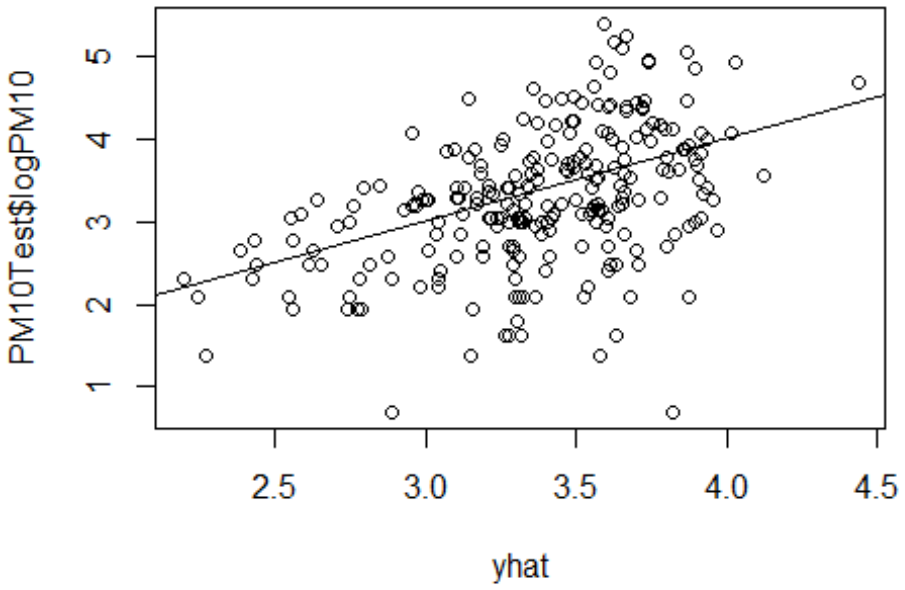



```
## [1] 0.7944662
```

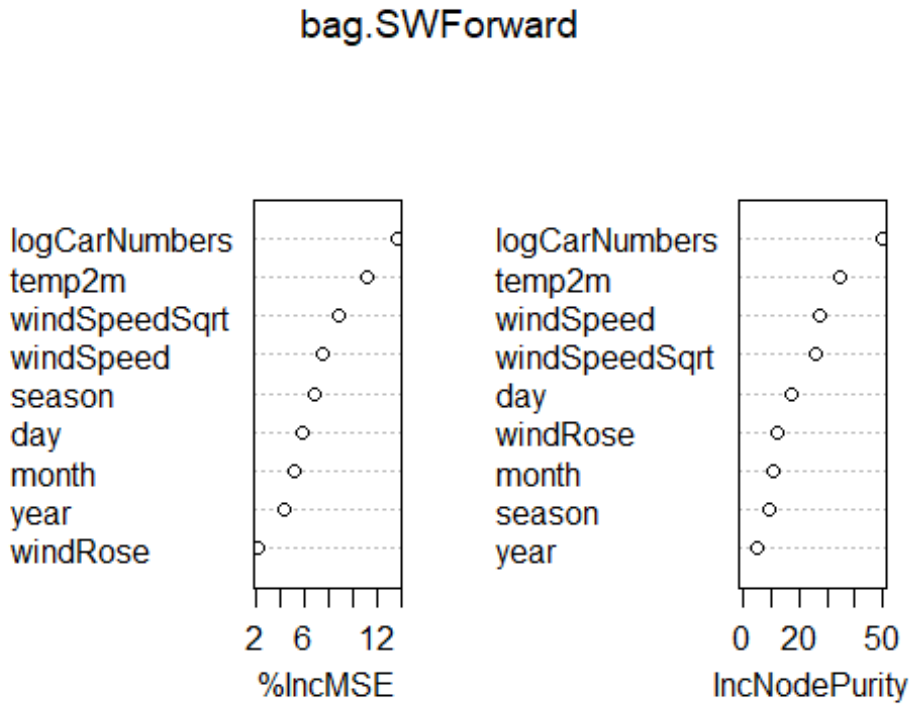
```
## [1] 0.7683198
```



Repeating the modelling with the SWForward variable set yields a model with an optimal forest size of 150 trees and an mtry of 2. The best model has an similar RMSE 0.7498 and a very similar pattern of variable importance.



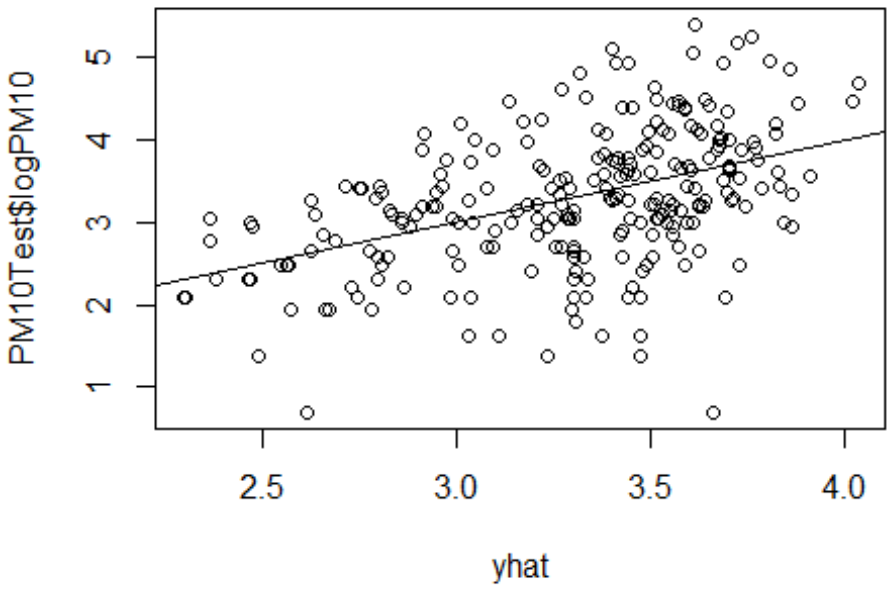

```
## [1] 0.7497742
```



Boosting

Will run a similar optimisation and comparison. The optimum looks to be 20 tress and an interaction depth of 4. The ranking of the importance of the variables is similar to previous results as is the plot of actual and predicted values as is the final RMSE at 0.7453.

```
## Using 20 trees...
```




```
## [1] 0.7453301
```

Bayesian additive reg trees

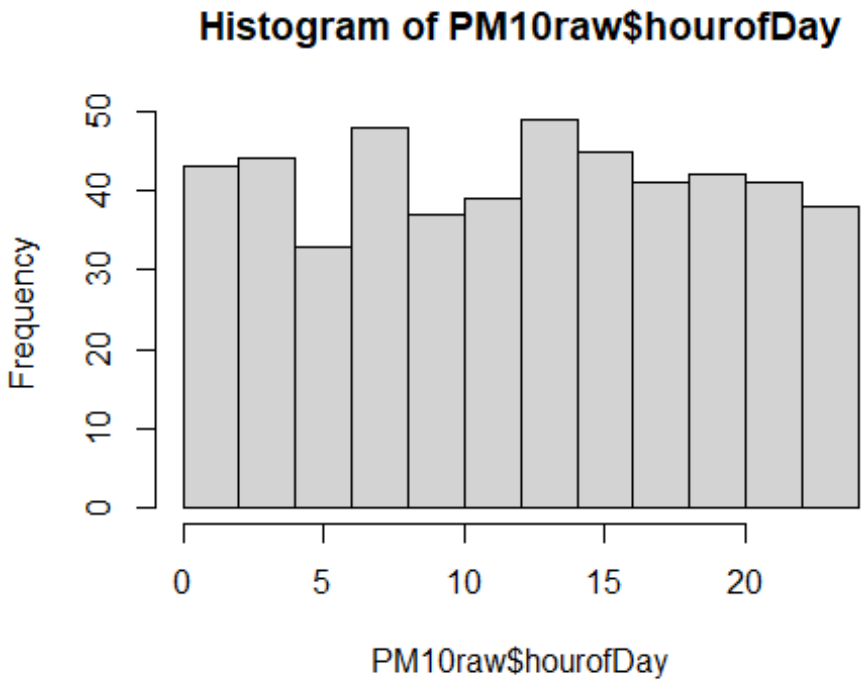
BART was run on the Chosen9 variable set and the best model identified. The plot of actual and predicted was similar as was the RMSE 0.7453.

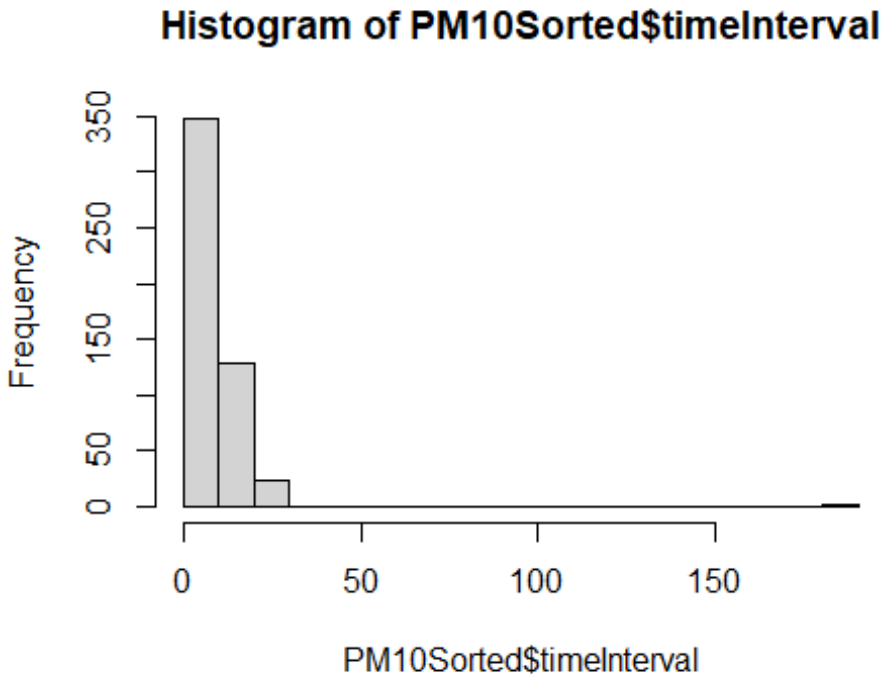
```
## *****Calling gbart: type=1
## *****Data:
## data:n,p,np: 254, 28, 246
## y1,yn: 0.393460, 0.418780
## x1,x[n*p]: 7.744140, 0.670820
## xp1,xp[np*p]: 8.033980, 0.670820
## *****Number of Trees: 200
## *****Number of Cut Points: 100 ... 2
## *****burn,nd,thin: 100,1000,1
## *****Prior:beta,alpha,tau,nu,lambda,offset:
2,0.95,0.0829321,3,0.103006,3.2701
## *****sigma: 0.727187
## *****w (weights): 1.000000 ... 1.000000
## *****Dirichlet:sparse,theta,omega,a,b,rho,augment: 0,0,1,0.5,1,28,0
## *****printevery: 100
##
## MCMC
## done 0 (out of 1100)
## done 100 (out of 1100)
## done 200 (out of 1100)
```

```
## done 300 (out of 1100)
## done 400 (out of 1100)
## done 500 (out of 1100)
## done 600 (out of 1100)
## done 700 (out of 1100)
## done 800 (out of 1100)
## done 900 (out of 1100)
## done 1000 (out of 1100)
## time: 4s
## trcnt,tecnt: 1000,1000

## [1] 0.770531
```

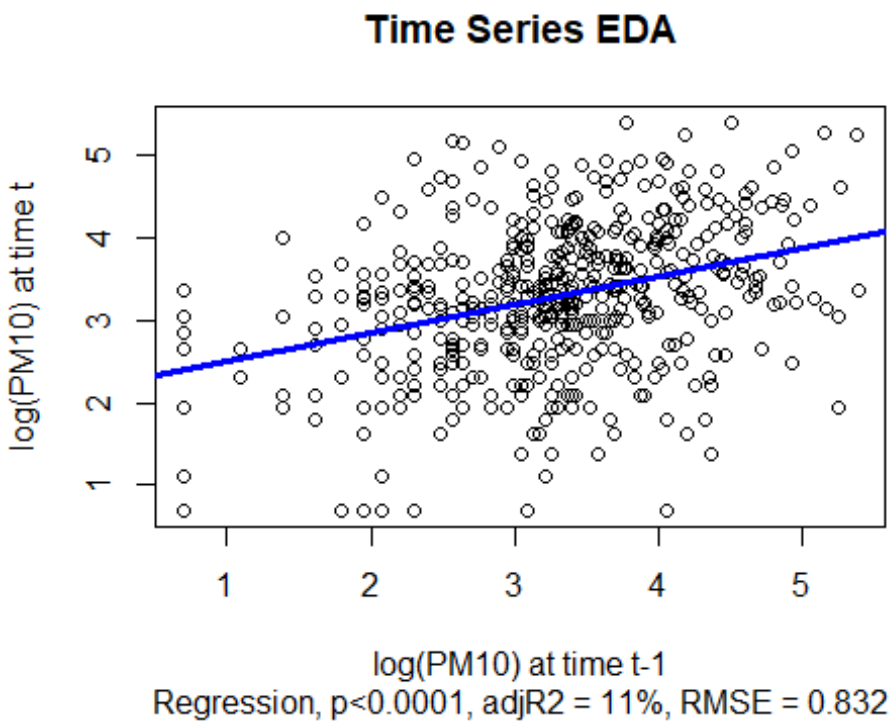
Is Time Series Analysis a valid approach?

As mentioned earlier the data are extracted from a time series - if the intervals between observations are approx constant then could look at simple Time series analyses: so some EDA was carried out to explore this option. Data are available over two winters and look to have data across the entire 24 hour day. First step is to construct a time stamp for each observation then look at the sampling intervals we have. The median sampling interval is 5 with an IQR from 2 to 12. This is a narrow enough range that time series analysis may be worth exploring. A new variable was created being the difference in logPM10 at time(t) and time(t-1). When logPM10 at time(t-1) is regressed against time(t) a relationship can be seen which is significant ($p<0.001$) but only accounts for 11% of the variation. The RMSE is 0.832.






```
## [1] 0.340706

## [1] 0.8329164
```



Results

The table below below summerises the various analyses and tests performed detailing the statistical method and the data and variable sets used in each analysis. Performance is assessed using the RMSE when logPM10 is treated as a continuous variable and overall percent correct when logPM10 is treated as a dichotomous variable with the break at 45 micrograms.

method	dataset	variableset	assessment	value
Simple multiple regression	all data	Original variables	RMSE	0.80300
Simple multiple regression	all data	Transformed variables	RMSE	0.79800
Simple multiple regression	all data	Manually selected variables	RMSE	0.73600
Forward stepwise multiple regression	Test data	All variables	RMSE	0.78100
Backward stepwise multiple regression	Test data	All variables	RMSE	0.81840
Ridge regression	Test data	All variables	RMSE	0.69940
Ridge regression	Test data	Chosen variables	RMSE	0.99630

method	dataset	variableset	assessment	value
Ridge regression	Test data	Stepwise forward variables	RMSE	0.99630
Lasso regression	Test data	All variables	RMSE	0.82190
Lasso regression	Test data	chosen 9 stepwise forward variables	RMSE	1.00680
Partial least squares regression	Test data	All variables	RMSE	0.70740
Partial least squares regression	Test data	SWForward variables	RMSE	0.71065
Partial least squares regression	Test data	Chosen variables	RMSE	0.70255
GAM regression - normal Splines	Test data	Chosen 9 variables	RMSE	0.81840
GAM regression - normal Splines	Test data	Chosen 9 variables	RMSE	0.81850
DICHOTOMOUS VARIABLE ANALYSIS				0.00000
Logistic regression - 0.5 cutoff	Test data	Chosen 9 variables	% Correct	29.70000
Logistic regression - 0.75 cutoff	Test data	Chosen 9 variables	% Correct	33.70000
Logistic regression - 0.85 cutoff	Test data	Chosen 9 variables	% Correct	31.00000
Linear Discriminant Analysis (LDA)	Test data	Chosen 6 variables	% Correct	71.00000
Quadratic Discriminant Analysis (QDA)	Test data	Chosen 5 variables	% Correct	65.00000
Best K-NN k = 5, raw data	Test data	Chosen 5 variables	% Correct	70.70000
Best K-NN k = 10, standardised variables	Test data	Chosen 5 variables	% Correct	75.60000
Decision Tree	Test data	Chosen 9 variables	% Correct	65.00000
Decision Tree 8 nodes	Test data	Chosen 9 variables	% Correct	79.20000
Decision Tree 8 nodes	Test data	Chosen 9 variables	% Correct	78.80000
Decision Tree 8 nodes -	Test	Chosen 9 variables	RMSE	0.89300

method	dataset	variableset	assessment	value
linear regression	data			
Bagging - optimised	Test data	Chosen 9 variables	RMSE	0.76830
Bagging - optimised	Test data	SW Forward variables	RMSE	0.77960
Boosting - optimised	Test data	Chosen 9 variables	RMSE	0.76870
BART - optimised	Test data	Chosen 9 variables	RMSE	0.76870
Time series - one step ahead	All data	logPM10 at t-1	RMSE	0.83290

Discussion

Data interpretation.

The PM10 data is expressed on the natural log scale such that a one unit change is a 2.7 fold increase.

The IQR for logPM10 is 1.17 and the average RMSE across all the tests performed was 0.810 such that the average RSME is 69% of the IQR or more than a doubling of the PM10 value. Such large errors will make any predictive test of limited value.

When considering the dichotomous predictors it should be noted that 75% of the observations were for a low reading (under 45) so just assuming all predictions would be for a 'Low' concentration would be correct 75% of the time. Against this the percent correctly assigned is not impressive. However there was some evidence that the ability to predict 'High' was better than just random chance.

The original paper (Aldrin and Haff, 2005) used Generalised Additive Modelling (GAM) but they explored the techniques performance more than its predictive ability. It is interesting to note that the 'best' test method (Ridge Regression) performed considerably better than GAM (RMSE 0.6994 cf 0.8184) and many of the tests (12/21 57%) assessed performed better than the GAM method.

Why did the tests have low predictive ability

PM10 is made up of smaller particles (PM25) and larger particles (Segersson et al, 2017). The smaller particles tend to come from fuel burning whereas the larger ones from vehicle traffic. The parameters available to include in these models largely related to roads and vehicle traffic and did not pick up variables that might have been associated with fuel burning (house density, use of open fires etc). This inability to capture all the relevant variables may account for the poor predictive ability. The original authors also noted that snow cover could act as a buffer for PM material trapping particles as it fell and releasing them as it melted - we did not have any snow cover data available.

The data set available was from one of four sites around Oslo that were originally surveyed. Looking at Google Maps the site for which we had data was very industrial and bounded to the NW by a large rail marshaling yard. It is unlikely that predictions made from data for the one site would extrapolate successfully to the other, more residential, sites.

Conclusion

A wide range of modern statistical data learning techniques have been applied to a sample of data from an original study on particulate matter pollution (Aldrin and Haff, 2005). For a range of reasons the predictive capacity of these models is such that they will be of very limited applied use. The full data set were collected over sequential days and hours so incorporating and element of time series analysis (such as the logPM10 count at same time of previous day) may well improve the predictive ability.

References

Aldrin M. and Haff I.H. (2005) Generalised Additive Modelling of air pollution, traffic volume and meteorology. Atmospheric Environment 39, 2145-2155

James G, Witten D, Hastie T. and Tibshirani R (2021) An introduction to Statistical Learning. 2nd Ed. 607 pp. Pubs Springer.

Segersson D, Eneroth K, Gidhagen L, Johansson C, Omstedt G, Nylén A Forsberg B. (2017) Health Impact of PM10, PM2.5 and Black Carbon Exposure Due to Different Source Sectors in Stockholm, Gothenburg and Umea, Sweden. Int. J. Environ. Res. Public Health 2017, 14, 742

Sparapani, R., Spanbauer, C., & McCulloch, R. (2021). Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package. Journal of Statistical Software, 97(1), 1–66. https://doi.org/10.18637/jss.v097.i01 Accessed 15Oct22

WHO (2021) https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health. Accessed 24Oct22