EE239AS Special Topics in Signals and Systems

Project 1 Report:

Regression Analysis

Group Members:

Jiayu Guo(504513188),Yitian Hu(904516321),
Peidong Chen(204432674), Yang Yang(804522285)

Winter 2016

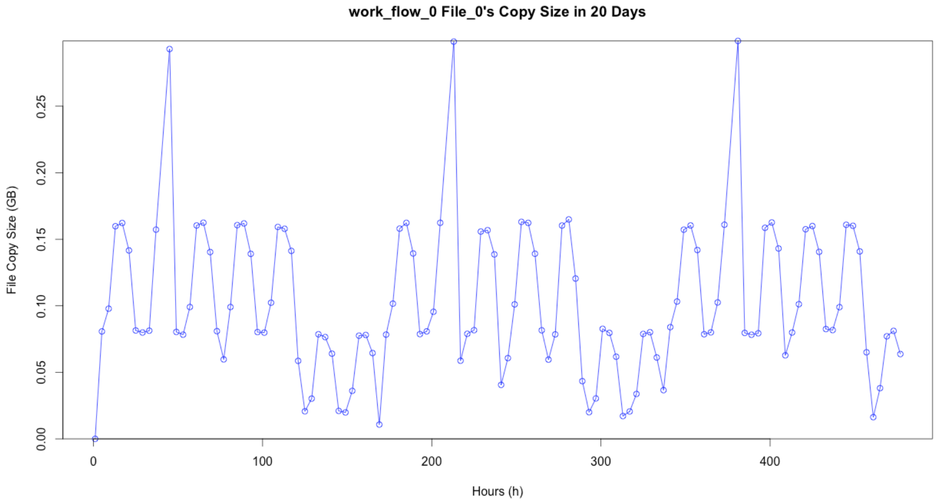
01/30/2016

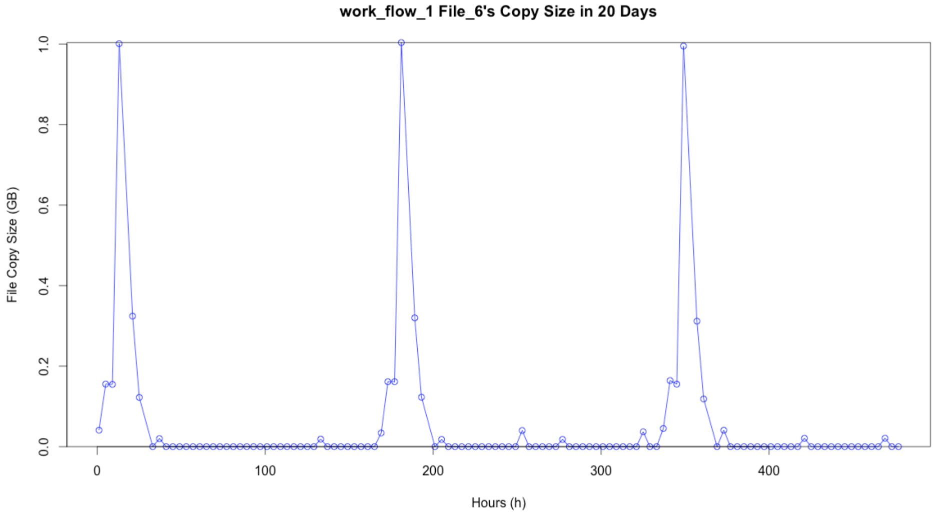
Question 1

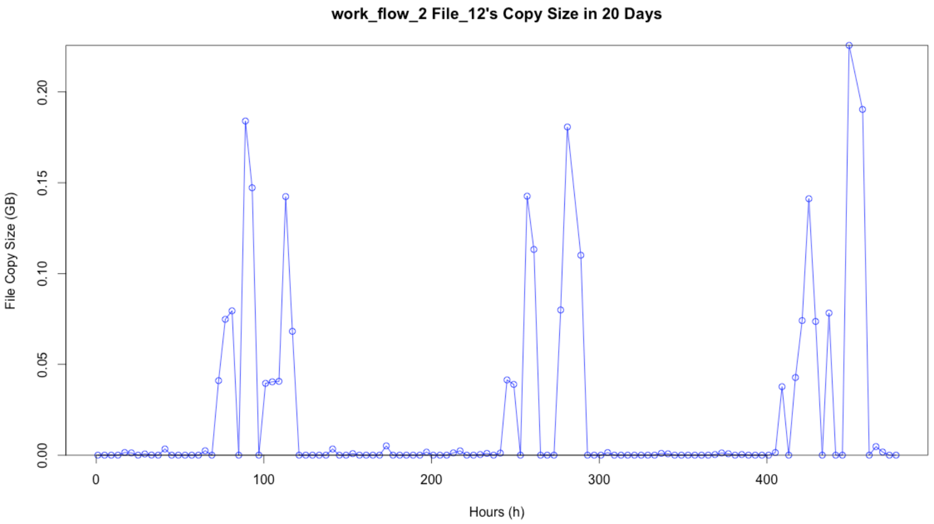
For each workflow, we can use one file to represent the upload condition. Thus we choose one file in each workflow and plotted the actual copy sizes of all the files on a time period of 20 days. The plot is shown below.

In workflow 0 we choose file 0; In workflow 1 we choose file 6; In workflow 2 we choose file 12; In workflow 3 we choose file 21; In workflow 4 we choose file 24. In each workflow, we can see a rough repeat pattern in the pictures. The repeat cycle is about 1 week for workflow 0-2. For workflow 3, although the repeat is not as obvious as the others, we can still figure out a repeat trend. The peak of the copy size repeatedly in a cycle of 1 week. For workflow 4, the cycle is also about 1 week.

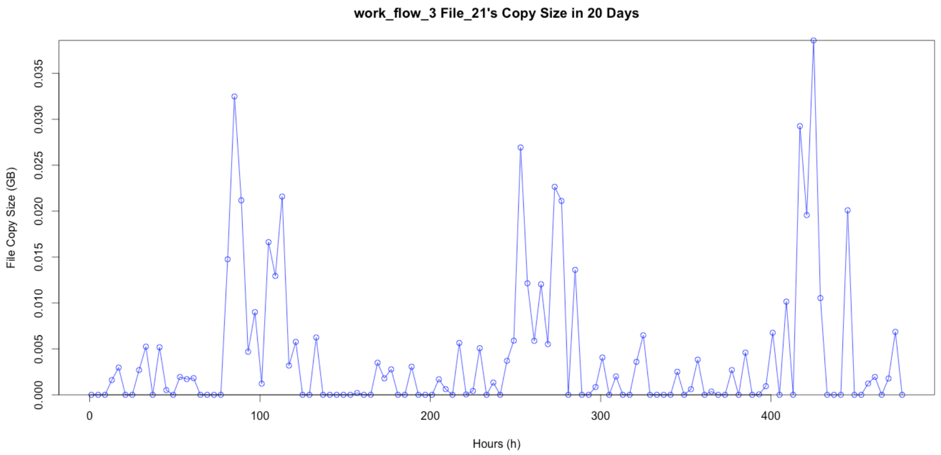
Figure 1 #nfl tweets per hour over time



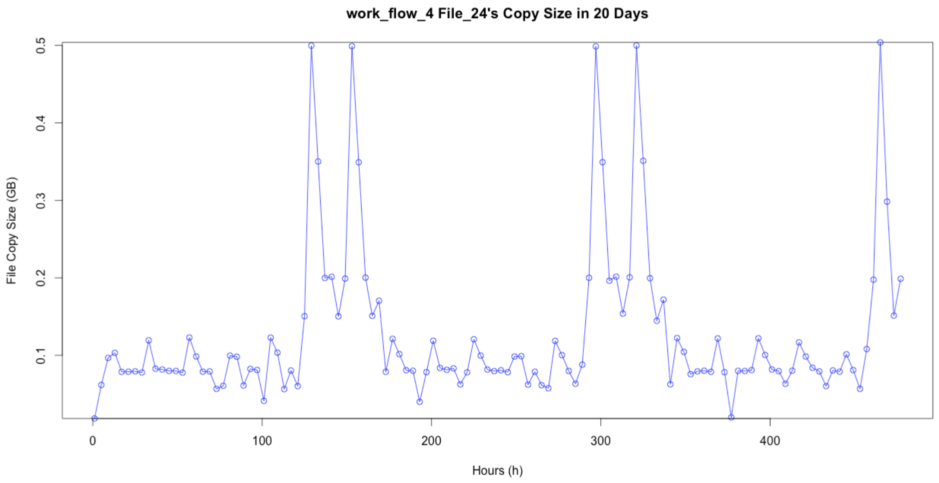




work_flow_2 File_12's Copy Size in 20 Days



work_flow_3 File_21's Copy Size in 20 Days



work_flow_4 File_24's Copy Size in 20 Days

Question 2

a) Fit linear regression model:

We do 10-fold Cross-validation using linear model.

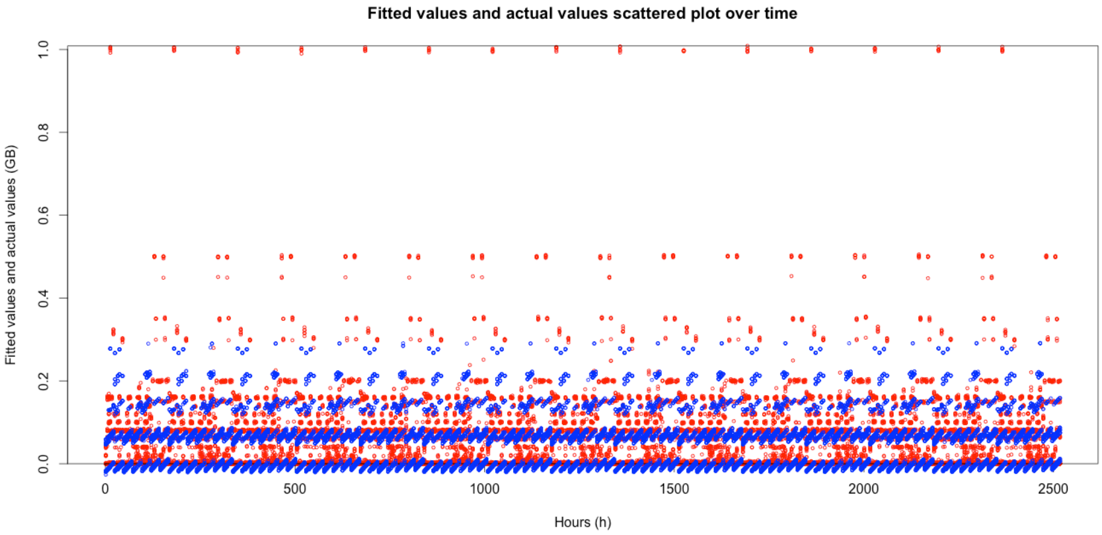
The significance of different variables with the statistics obtained from the model are shown below:

Variables	Significance
Week	0.0001212423713
DayOfWeek	0.0013189671131
StartTime	0.0009633665451
WorkFlowName	0.0021885715069
FileName	0.0001190829549
TimeBackup	0.0712918216619

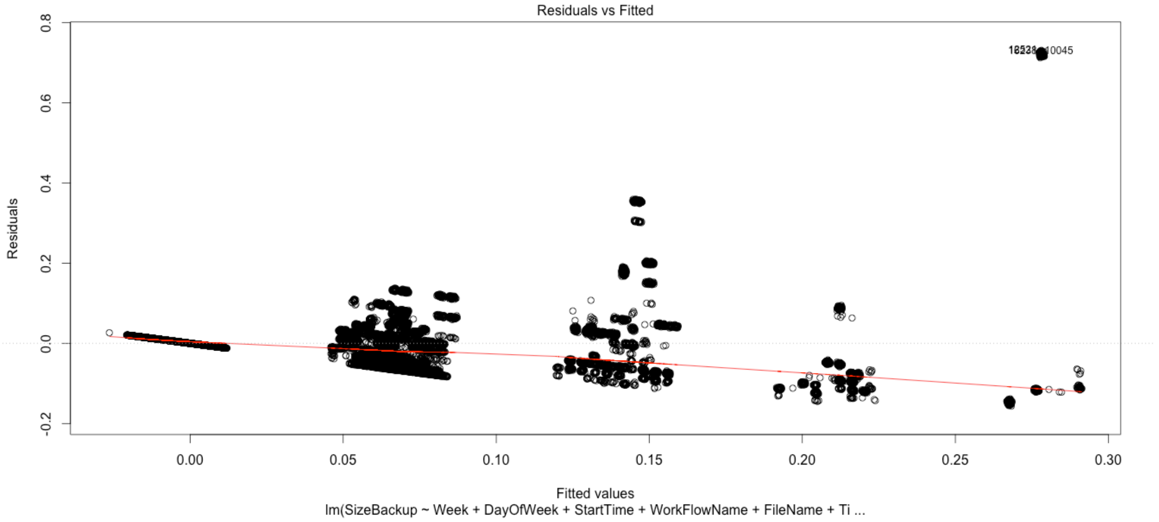
We can see that TimeBackup is the most significant variable.

The average RMSE is 0.0792344.

The following figure shows "Fitted values and actual values scattered plot over time". The red points show the actual values, and the blue points show the fitted values.



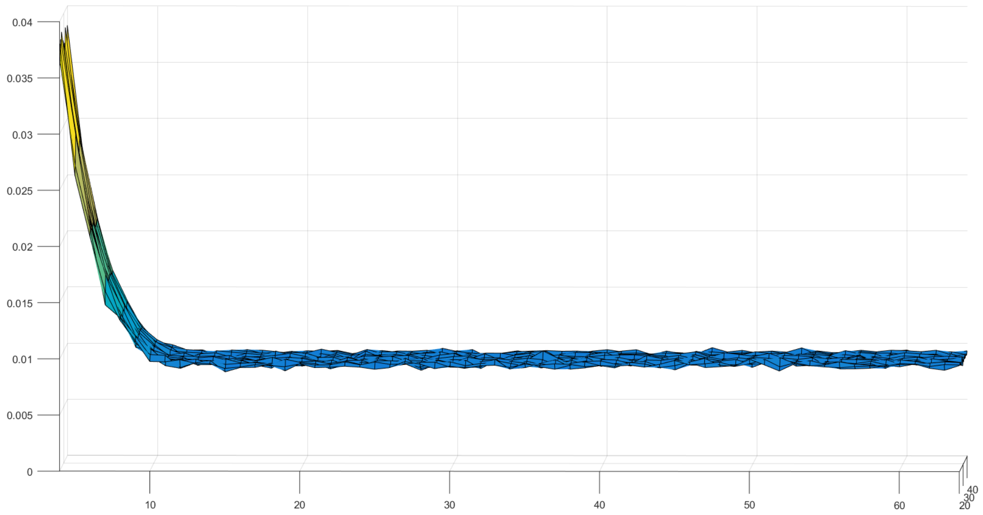
The "residuals versus fitted values plot" is shown below.

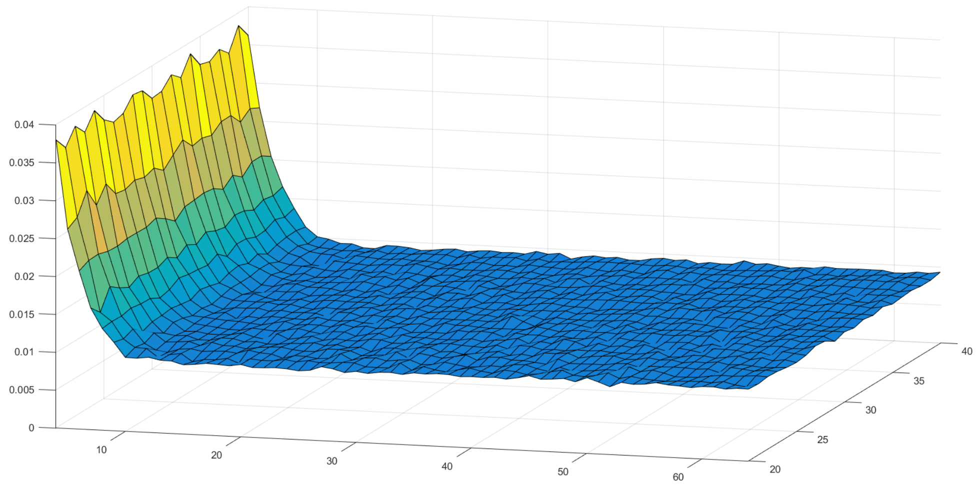


From the 'Fitted values and actual values scattered plot over time', we can see there are many value our model can not fit. From the 'residuals versus fitted values plot', we can see the residual do not "bounce randomly" around the 0 line. There are also some outliers in the basic residual pattern. This all indicates the relationship of our data is not linear. Linear model is not a ideal model for our data.

b) Use a random forest regression model:

We plot 3D figures to see the relationship between "Number of trees", "Depth of each tree" and "RMSE values".

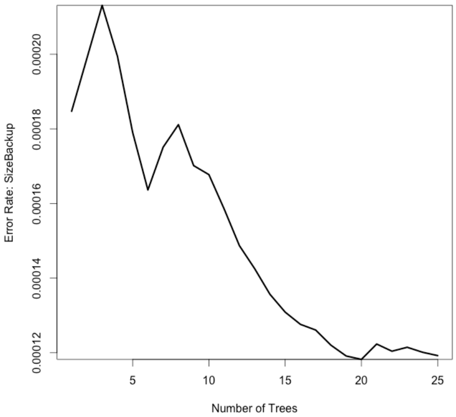


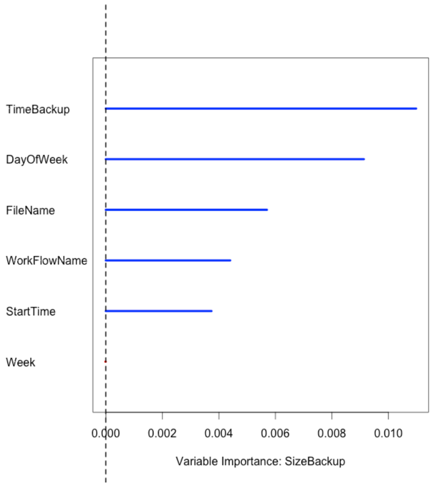


After analyzing the figures, we tune the "Number of trees" to be 25 and "Depth of each tree" to be 15. And the best RMSE we get is 0.008703.

The patterns we observed are that when the depth is larger than 10, it contributes little to the RMSE value. And also the number of trees doesn't affect the value of RMSE.

Here is the detail of our fitted model. From the variable importance we can see that, the variable week will not affect the result. This indicates that sizeBackup has a repeat cycle of 1 week. Which is the same as our pattern found in part 1.



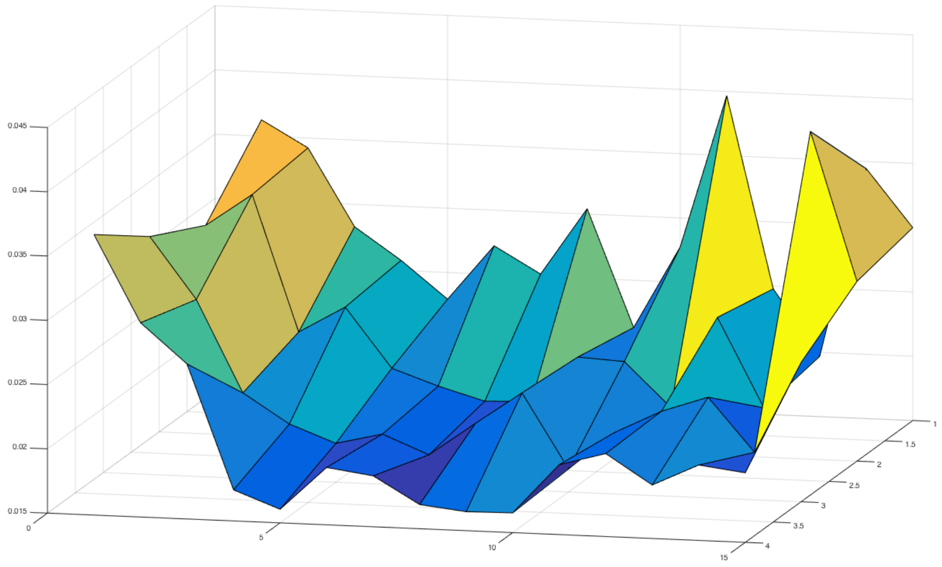


c) Use neural regression model:

We choose a single file "File_16" for the neural network regression model. We use two parameters for the model. The parameters are the number of hidden nodes each layer and the number of hidden layers.

We use "neuralnet" library in R to implement this neural regression model. We mainly tune two parameters, which is the "number of hidden nodes each layer" and "the number of hidden layers." The hidden layer(s) makes it possible for the modle to perform non-linear regression, however, if the number of the hidden layers if too high, it is prone to overfitting. In order to secure the ability of the model to generalize and predict, the number of hidden nodes has to be kept small. If not, it will do worse in the validation data.

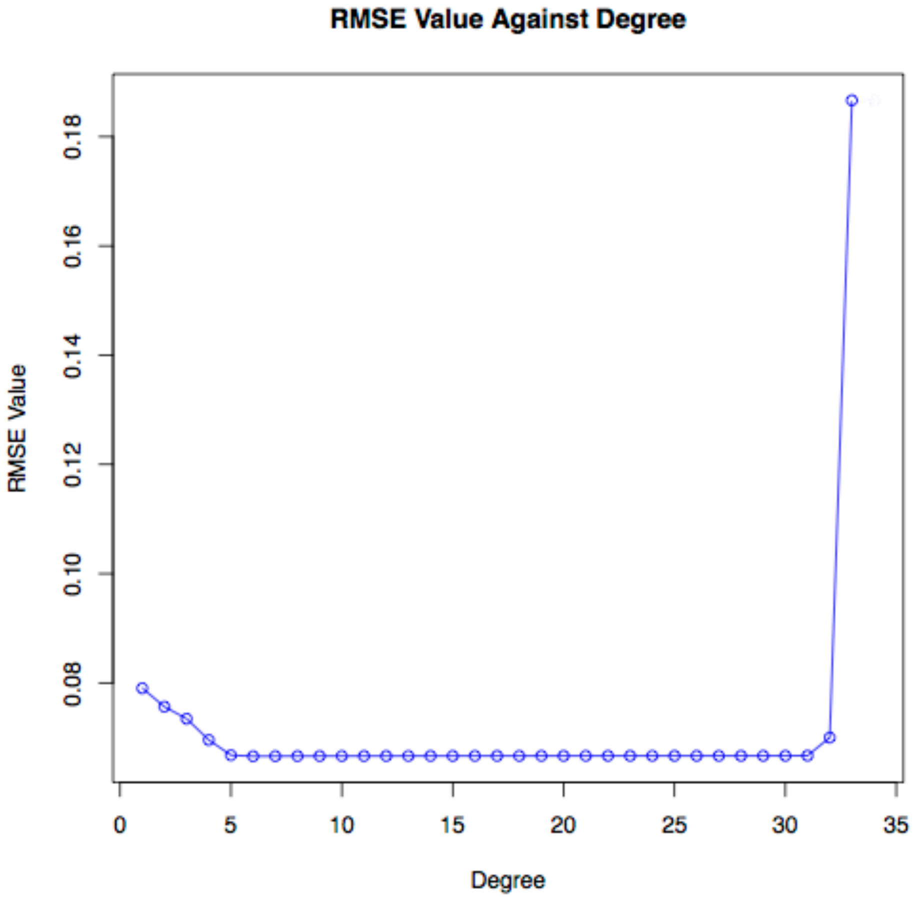
The following 3D figure shows the relationship between "number of hidden nodes each layer", "number of hidden layers" and RMSE value.



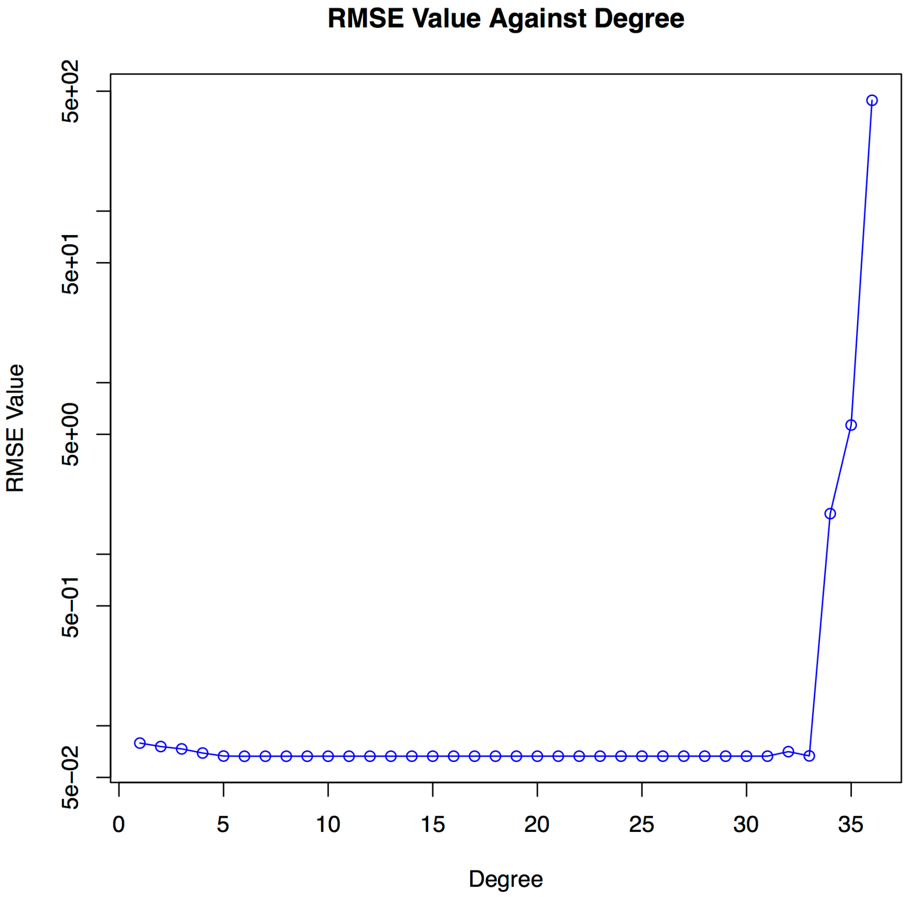
From the figure above, we can see that when the "number of hidden nodes for each layer" is equal to 10 and "number of hidden layers" is equal to 2, the RMSE is the best. And the best RMSE is 0.016157.

Question 3

The following figure shows the relationship between degree and RMSE value.



And the following figure shows the RMSE in log scale.



We can see that the best degree is 6.
We find the threshold degree is 33, and when the degree is greater than 33, the RMSE value is getting worse.

In general, the more complex our model is, the more success will our model 'learning' the training data. However, With the improve of our model complexity, when it is excessively complex, overfitting will happen in our model. This will cause our model to "memorize" training data rather than "learning" to generalize from trend. So the model will do well in training data, but when use this model to do prediction, the result will be bad. By using cross validation, we will divide our data into training set and validation set. This two set are separate, thus when overfitting happen, the model will not 'memorize' the validation data and will have much errors when doing validation. For this reasons, cross validation will help us evaluate the model objectively. It can help us choose the most suitable complex of our model.

Question 4

We do 10-fold Cross-validation using linear model.
The significance of different variables with the statistics obtained from the model are shown

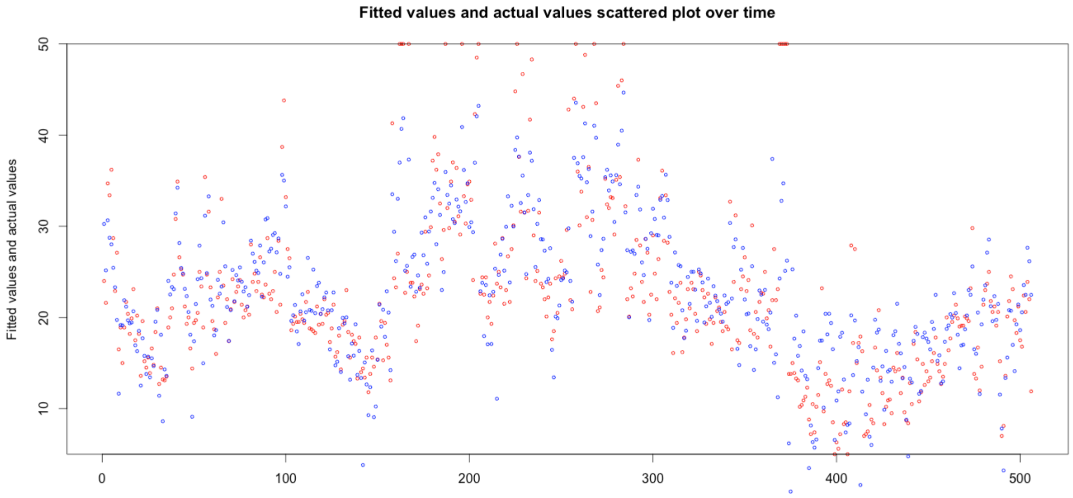
below:

Variables	Significance
CRIM	-0.10151449937
ZN	0.05373695243
INDUS	0.01887539615
CHAS	2.82064623774
NOX	-17.16498921368
RM	3.62595999313
AGE	0.01262356371
DIS	-1.48548234670
RAD	0.29722579772
TAX	-0.01136922489
PTRATIO	-0.96120709560
B	0.01025428746
LSTAT	-0.58130370238

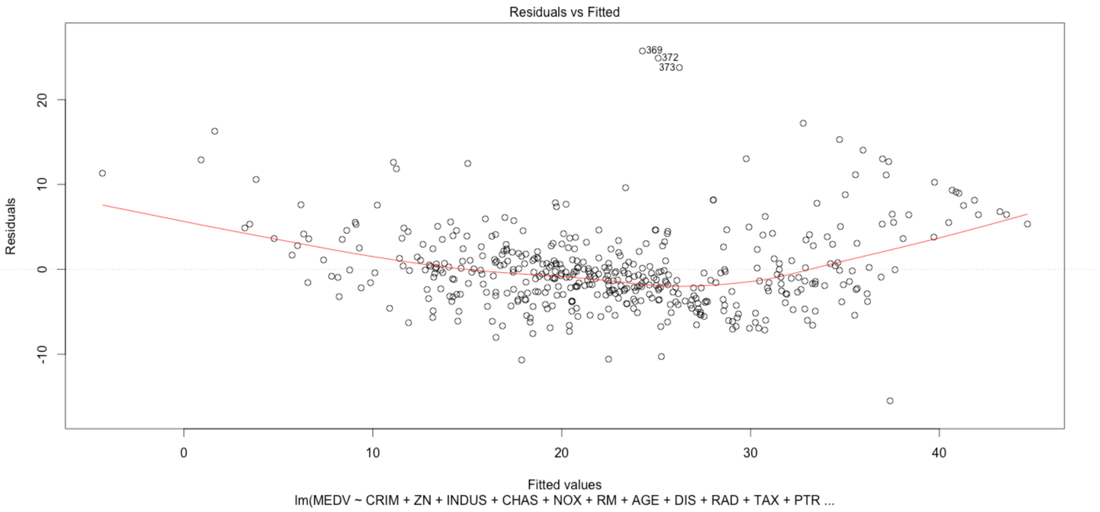
We can see that NOX is the most significant variable.

The average RMSE is 4.8137655.

The following figure shows "Fitted values and actual values scattered plot over time". The red points show the actual values, and the blue points show the fitted values.

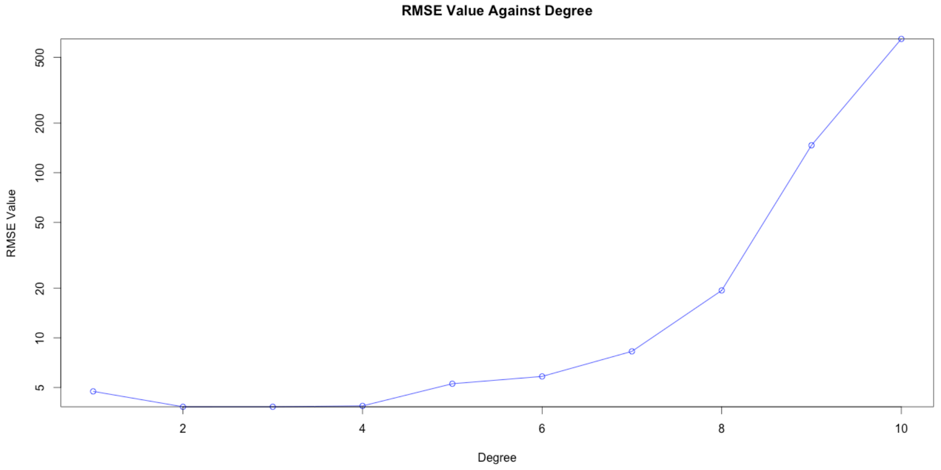


The "residuals versus fitted values plot" is shown below.



From the 'Fitted values and actual values scattered plot over time', we can see our model has a better result in question 2a, though there are also some value our model can not fit. From the 'residuals versus fitted values plot', we can see the residual do not "bounce randomly" around the 0 line, which indicates linear is also a ideal model for this data.

The following figure shows the relationship between degree and RMSE value in log scale.



And the optimal degree is 3.

Question 5

a) When using ridge regression

The relationship between α and RMSE is shown in the following table.

α	RMSE Value
0.1	4.807664
0.01	4.768524
0.001	4.770637

The best RMSE is 4.807664. And we choose "α" to be 0.01.

b) When using Lasso regularization

The relationship between α and RMSE is shown in the following table.

α	RMSE Value
0.1	7.676292
0.01	8.990613
0.001	9.135838

The best RMSE is 7.676292. And we choose "α" to be 0.1.