Project No. 01:
Regression Analysis

By: Gu Dazhong, Zhu Yannan, DUCK-HA HWANG

Instructor: Roychowdhury Vwani

EE 239AS Winter 2016

January 31, 2016

I. INTRODUCTION

A. Purpose

This project is about Regression Analysis. In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or predictors). We will solve the regression models by using the assigned set of data, along with basic techniques to handle over-fitting; namely cross-validation, and regularization.

Cross-validation, sometimes called rotation estimation, is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set.

Regularization, in mathematics and statistics and particularly in the fields of machine learning and inverse problems, refers to a process of introducing additional information in order to solve an ill-posed problem or to prevent over fitting.

In this project, we will use sevaral regression model: linear model, randomforest model, neural network model, polynomial model, ridge model and lasso model. Two set of data will be analyzed, which are network backup dataset and housing data.

B. Equipment

There is a minimal amount of equipment to be used in this project. The few requirements are listed below:

- Rstudio Software (v3.2.3)
- Computer capable of running the software mentioned

C. Procedure

1. Find the pattern of backupsize by draw the size-time graph
2. Build linear regression model to the network data.
3. Build randomforest regression model to the network data.
4. Build neural netwok regression model to the network data.
5. Seperate the data into each workflows and build linear regession model to each workflow.
6. Build polynomial regression model to the network data.
7. Build linear regression model to the housing data.
8. Build polynomial regression model to the housing data.
9. Build ridge regression model to the housing data.
10. Build lasso regression model to the housing data.

D. Structure

This report will be seprated into 4 parts:

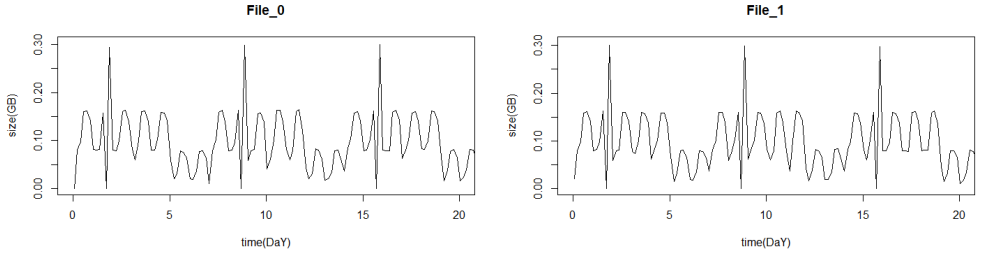
1. Introduction
2. Network backup Dataset
3. Boston Housing Dataset
4. Discussion & Conclusion

II. NETWORK BACKUP DATASET

This section we will analyze a network backup dataset. we will find the relationship between the size and other features and analyze how each feature influence the backup size. Linear regression model, randomforest model, neuralnetwork model and polynomial model will ben used in this section. We will also discuss how the parameters of each model fuction will influnce the regression result.

A. The patterns of Data

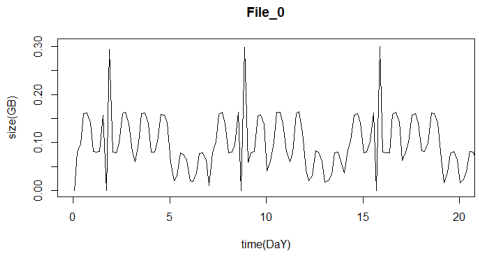
First, let us look at the size-time relationship of File0 in FIG1. We can see that the size will repeat in about 7 days. So it has a 7 day period. We also pick another file, File1, to compare. File0 and File1 are both in workflow0. We can see that the pattern of them are very similiar to each other. Actullay, all the files in the same workflow have a similiar pattern to each other. But in different workflow, the pattern is different.



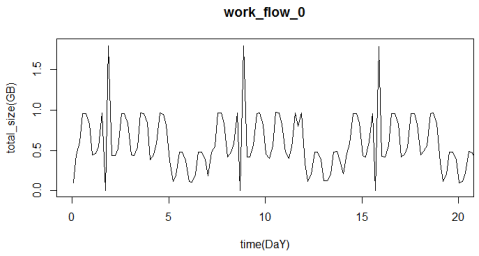
(a) File0 (b) File1

FIG. 1: Similiar pattern of file1 and file0

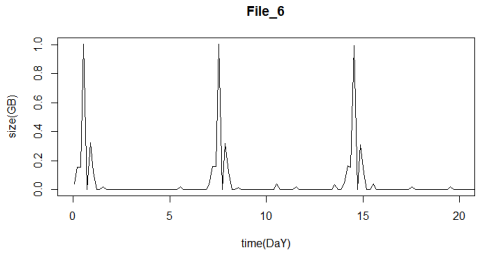
To, show the pattern of each workflow, we pick up a file from each workflow. And for each workflow, we add all the size at the same time point to get the totalsize-time relationship of each workflow, which are in FIG2 below. In FIG2, the pattern of each file is similiar to the pattern of the total size of each workflow. So a single file can represent its workflow.



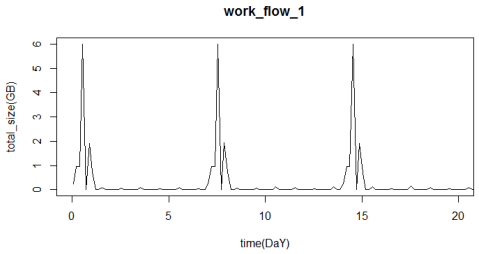
(a) File0



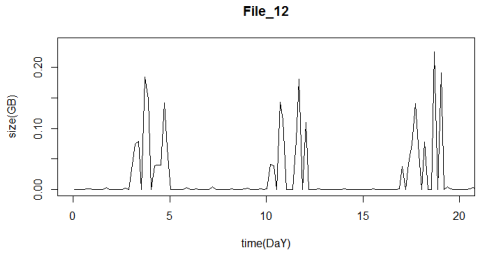
(b) workflow0



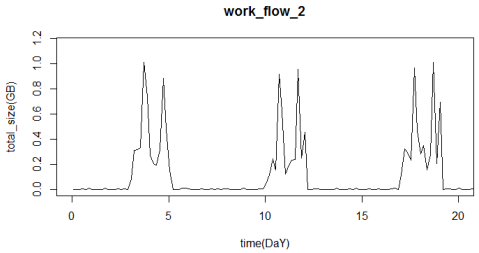
(c) File6



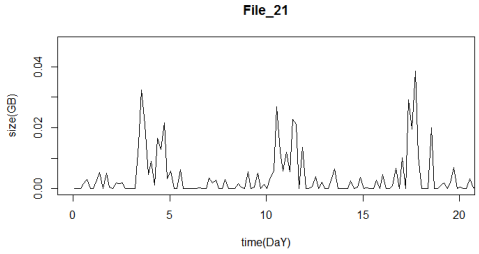
(d) workflow1



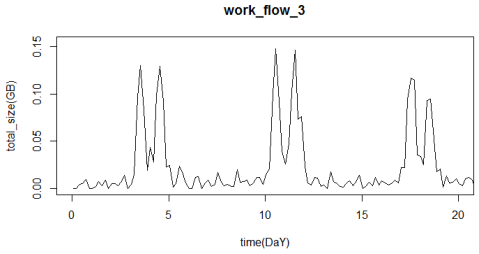
(e) File12



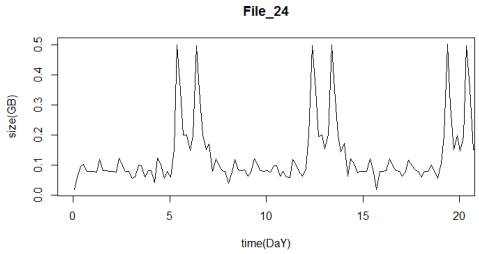
(f) workflow2



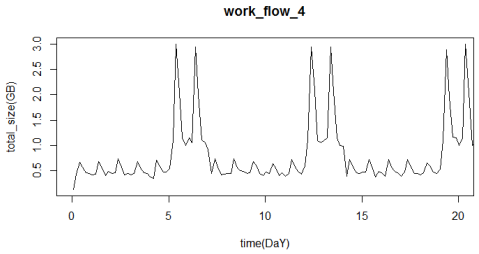
(g) File21



(h) workflow3



(i) File24



(j) workflow4

FIG. 2: Pattern of the files

B. Linear Regression Model

First, let use linear model to fit our data. linear model is the most simple regression model and is widely used. However from the section before, we can see that the pattern of our data is far from linear. So maybe we will not have a good result from linear model. But let us have a try first.

Here we will use the 10-fold cross vadilation to help us make a better model. Cross vadilation help to seperate the data into two part: the testdata and the train data. So we can test the model by using different data from what we use to train the model. If we only use the data which train the model to evaluate the performance of oue model, a over-fitting problem will show up. Over-fitting means that our model fit the train data too well that the model cannot suit other data. However, in cross vadilation, the model doesn't know any information about the testdata when it is trained. So if the model give a good result to the testdata, it means the model can also fit new data.

Here we give the result of the RMSE we got from ten test-train groups:

Group	Group_1	Group_2	Group_3	Group_4
RMSE	0.06561816	0.07896838	0.07285326	0.08021463
Group	Group_5	Group_6	Group_7	Group_8
RMSE	0.08921607	0.08206341	0.0779478	0.07534132
Group	Group_9	Group_10		
RMSE	0.08660434	0.08135932		

FIG. 3: RMSE of each group

As the figure shown above, we can obtain RMSE of 0.06561816 by using 10-fold Cross-validation, which is the RMSE of the first group. So we use the model of the first group as our result. In this case, our linear regression model is:

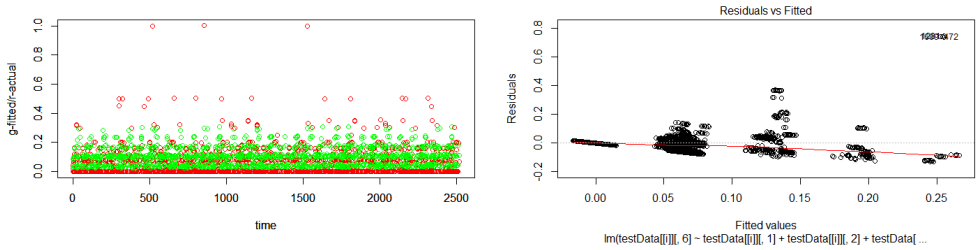
$$size = -0.0008173 * week + 0.0018172 * day + 0.0008937 * hour - 0.0005520 * filename$$
$$+0.0058346 * workflow - 0.0005520 * backuptime - 0.0133165$$

To analyze the significance of different variables, we examine the p-value between each variables and the size. p-value is a value which shows the correlation between two variables. The larger the p-value is, the closer the relationship between the two variabels.

Parameter	**Week #**	**Day of Week**	**Backup Start Time - Hour of Day**	**Work-Flow-ID**	**File Name**	**Backup Time (hour)**
p-value	0.9268	4.756e-10	2.2e-16	3.209e-07	4.902e-07	2.2e-16

FIG. 4: p-value of each features

By using function cor.test() in RStudio, we analyze the correlation between each variable and backup size. As can be seen above, the first variable Week is the most significant one. It can influence our prediction of backup size most.



(a) Fitted values and actual values scattered plot over time

(b) residuals versus fitted values plot

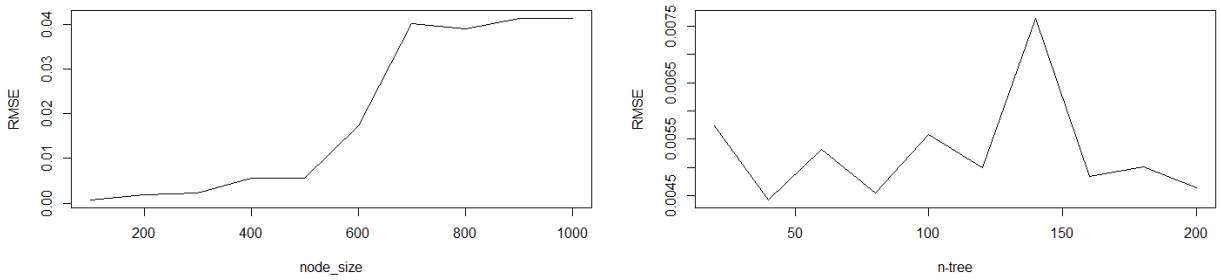
FIG. 5: Evaluation of the linear model

The left figure above shows the fitted values and actual values over time, where the green ones are fitted values while the red ones are actual values. We can see that except seldom values, which changes too rapidly, our model can fit most of the values.

As can be seen in the right figure above, most of residuals are equal to 0, which means our fitted values are the same as actual values. It consists with the dense area in the left figure.

C. Random Forest Regression Model

There are 2 major parameters that can influence RMSE in random forest model, which is number of trees and depth of each tree. We change those 2 parameters respectively and calculate corresponding RMSE as shown in following figures.



(a) Nodesize

(b) Ntree

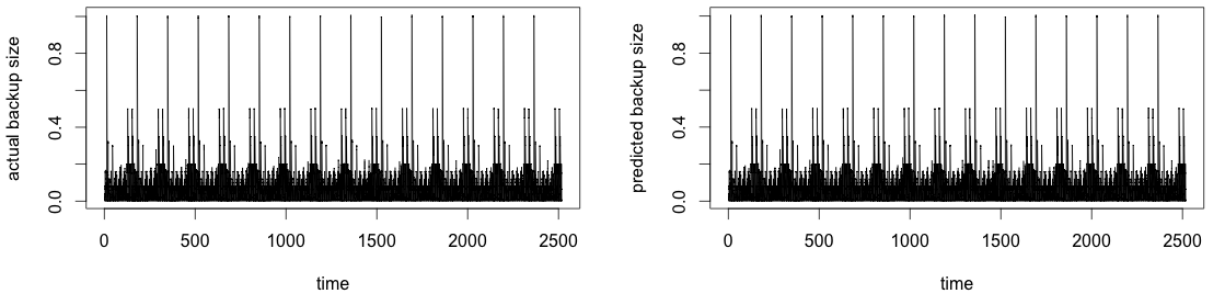
FIG. 6: Relationship between RMSE and treenum, nodesize

Nodesize can imply the depth of tree in such way that the larger nodesize is, the smaller depth becomes. The left figure shows that with the decreasing of depth, RMSE becomes larger which we do not prefer. Thus, we should use small nodesize. And we choose 200 as nodesize.

As for the number of trees, the right figure shows that RMSE fluctuates when number of trees increases.

Thus we set 50 as number of trees.

Then, we can obtain RMSE as 0.0002005608 which is much smaller than the RMSE (0.06561816) we got using linear regression model.



(a) Actual value

(b) Predict value

FIG. 7: Pattern of predict values and actual values

The figure above shows that the predicted backup size we got by using random forest model(the right figure) also presents the repeating pattern like we observed in part1(the left figure).

D. Neural Network Regression Model

We use function nnet() in RStudio to perform neural network regression model, which is a single layer neural network model. And there are 2 major parameters that can influence the performance of in RMSE, which are size and decay. Size presents the number of units in the hidden layer. And decay could limit the number of free parameters so as to avoid over-fitting, which is possible to regularize the cost function. The following figures show how they affect the performance in RMSE respectively.

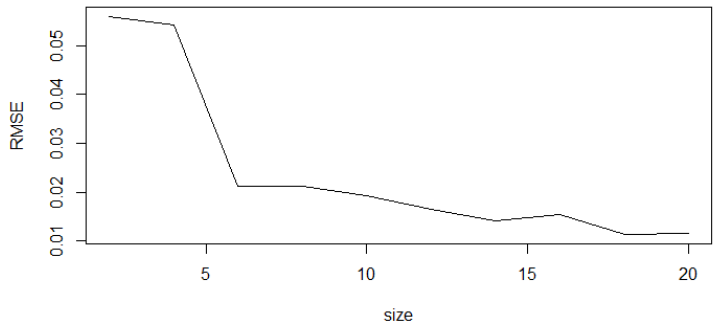


FIG. 8: Relationship between size and RMSE

Size is the most important parameter in this model. The larger size is, the smaller RMSE becomes, which means our fitted values are close to actual values.

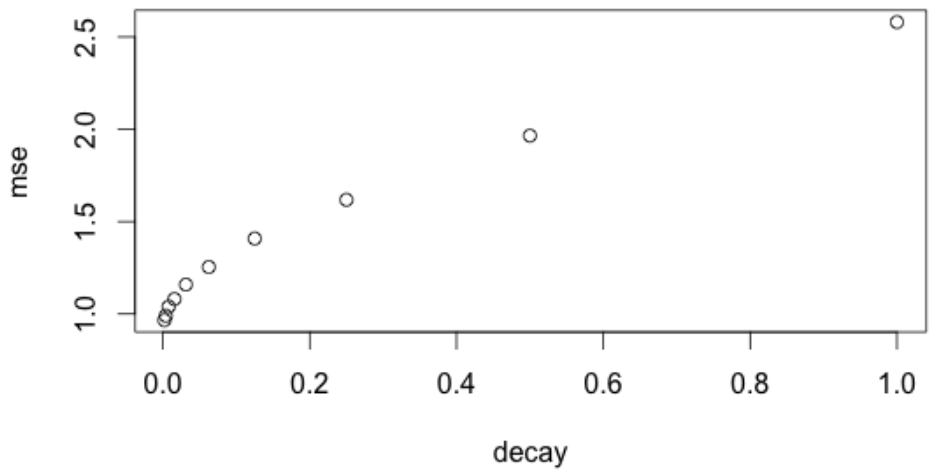


FIG. 9: Relationship between decay and MSE

As can be seen in the figure above, the smaller decay is, the smaller RMSE becomes. However, when decay is small enough, such as smaller than 0.1, it cannot improve the performance of our model anymore. Under such circumstances, we can still lower RMSE by increasing size.

E. Linear Regression and Polynomial Regression Separately by Workflow

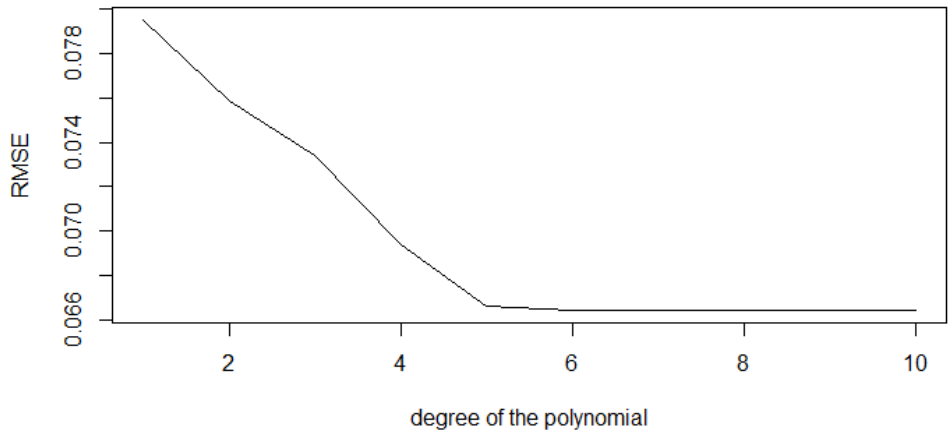
This time we analyze each workflow seperately by linear model. For each workflow we show the best RMSE and average RMSE got from 10-fold cross vadilation below:

work_flow	work_flow_0	work_flow_1	work_flow_2	work_flow_3	work_flow_4
Best RMSE	0.02780498	0.08069532	0.02301353	0.005237218	0.07797364
Average RMSE	0.02945251	0.1033918	0.0255203	0.005895825	0.08423841

FIG. 10: The RMSE of each workflow

We can see that the RMSE of each workflow is much better the the RMSE of the total data. The reason is obvious. At the beginning we show that the pattern of each is workflow is very different to each other. Workflow is a discrete variable. However, what we do before is roughly put workflow in a continuous linear fuction. So the RMSE will become better if we seperate the workflow.

Now let us the polynomial model to fit the data. What we want to study is how the degree of the polynomial will influnce our result and whether the poly model is better than linear model. Below is our result from poly model.



(a) Graph

degree of the polynomial	1	2	3	4
RMSE	0.07952918	0.07583899	0.07338765	0.06939992
degree of the polynomial	5	6	7	8
RMSE	0.0666055	0.06641248	0.06641218	0.06641201
degree of the polynomial	9	10		
RMSE	0.06641179	0.06641149		

(b) Table

FIG. 11: Relationship between the degree and RMSE

Unfortunately, we don't see a better performance of the poly model than linear model. Theoretically, poly model should be better than linear model, because linear model is a special situation of poly model when degree is one. The reason why our poly model isn't better than linear model may be that the way we group data is different. Each time we do cross vadilation, we will group the data randomly. So some way of grouping data may have a better result than others.

But we can still see the relationship between degree and RMSE. From the graph below we can see that the higher the degree is the smaller the RMSE. Because lower degree is a special condition of higher degree (we just set the coefficient of high degree component to zero). And in this data we can see that when n is greater than 5. The performance of the RMSE don't grow too muck as the degree goes high. So in this data, set n to 5 is a porper way.

III. BOSTON HOUSING DATASET

This section we will analyze a dataset of Boston Housing, which include the data of CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B:1000, LSTAT and MEDV of each houses. What we want to predict is MEDV (Median value of owner-occupied homes). We need to find out the relationship between MEDV and other features. In this section we will use linear model and polynomial model to do regression. We will also use the ridge model and lasso model which are mutation of the linear model to help us avoid the condition of ovet-fitted.

A. Linear Regression Model

First, let us do with the linear model. We used cor.test function in R to evaluate the significance of different variables. P-value means that has a strong relationship. Therefore,from the graph below we can find values of CHAS are the most important among others. However the p-value of CHAS is still very small just e-5, which means none of these features has a strong linear relationship with MEDV

Parameter	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	7.439e-05	< 2.2e-16	< 2.2e-16	< 2.2e-16
Parameter	DIS	RAD	TAX	PTRATIO	B	LSTAT	
p-value	1.253e-08	< 2.2e-16	< 2.2e-16	< 2.2e-16	1.421e-14	< 2.2e-16	

FIG. 12: P-value of each parameters

Then we use 10 fold cross validation to find a best linear model. The graph below is the RMSE for each group.In our data, we could know second group has a lowest RMSE. It indicates group2 is best group compare with other groups.

Group_house	H_Group_1	H_Group _2	H_Group _3	H_Group _4
RMSE	4.570746	4.595313	4.674507	4.68648
Group_house	H_Group _5	H_Group _6	H_Group _7	H_Group _8
RMSE	4.764529	4.726251	4.772817	4.652265
Group_house	H_Group _9	H_Group _10	Average	
RMSE	4.609983	4.633323	4.668621	

FIG. 13: RMSE of each group in housingdata

So we just pick the linear model of group2 as our result. The formula of this model is:

$$MEDV = -0.114463 * CRIM + 0.047658 * ZN + 0.017209 * INDUS + 2.244760 * CHAS$$
$$-18.132116 * NOX + 3.549907 * RM - 0.001097 * AGE - 1.508193 * DIS + 0.335655 * RAD$$
$$-0.013920 * TAX - 1.060036 * PTRATIO + 0.008559 * B - 0.498438 * LSTAT + 40.820106$$

We use the residuals vs fitted value graph to show how well our model fit the actual data.

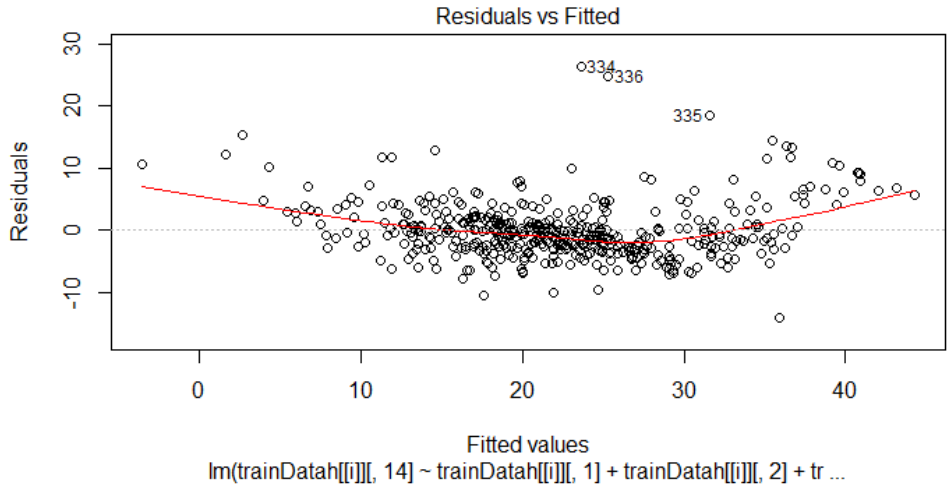


FIG. 14: Residuals vs fitted value of housing data linear model

B. Polynomial Regression Model

The RMSE of polynomial model is influnced by the highest degree of our model. The graph below show the relationship between degree and RMSE.

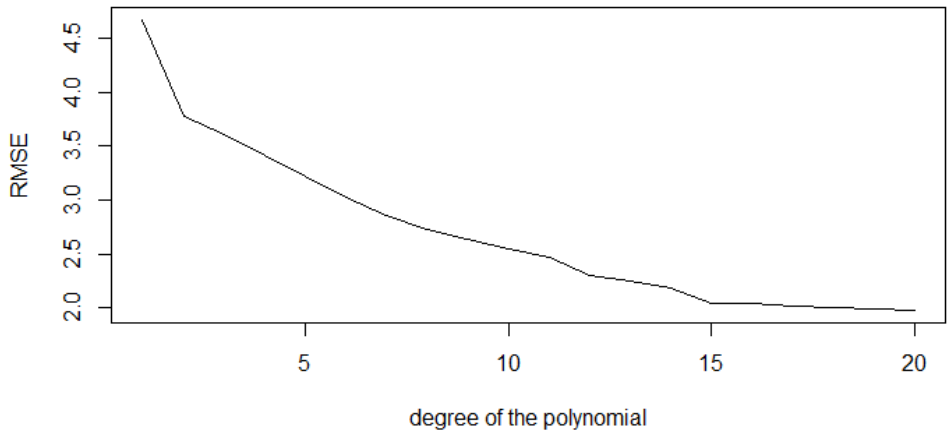


FIG. 15: The relationship between RMSE and polynomial degree

From this graph, we could analysis relationship between degree of the polynomial and RMSE. RMSE usually decrease as the degree of the polynomial increase. They are inverse proportion. When degree n is larger than 5, we can see the RMSE of poly model is much better than the linear model. And when n is

larger 15, we can see that the RMSE doesn't change too much the degree goes high. So n=15 is porper degree for our model.

C. Ridge and Lasso Regression Model

Ridge and lasso model can help us avoid ovet fitting by regulation of the parameters. the fomular of ridge and lasso is as below.

$$Ridge : min \|Y - X\beta\|_2^2 + \alpha \|\beta\|_2^2$$

$$Lasso : min \frac{1}{2n} \|Y - X\beta\| + \alpha \|\beta\|_1$$

The parameter α decide the degree that we regulate the coefficient. we will change the α to 0.1, 0.01, 0.001 to see how α influence our result. The RMSE below is the best RMSE we get from 10-fold cross validation.

α	0.1	0.01	0.001
Ridge-RMSE	3.456903	3.462289	3.46309
Lasso-RMSE	3.529446	3.485499	3.495122

FIG. 16: The influnce of α in ridge and lasso

We can see that the α does not change the RMSE too much. Actully, as long as the α isn't too close to the zero compare to the data we predict, it will not influnce the result too much. We can also see that by using ridge and lasso, a better RMSE is archived comparing to the linear model. In linear model the RMSE is about 4.5, and in ridge and lasso model the RMSE is about 3.5. So regularization of the parameters will raise the performance of the regression.

IV. DISCUSSION & CONCLUSION

In this project, we analyze two set of data: The Network backup Data and Boston Housing Data. We use linear model, random forest model, neural network model, polynomial model, ridge model and lasso model to do regression. We also use 10-fold cross vadilation to help us avoid over-fitting.

First, let us talk about the model. linear model and polynomial model is using a certain pattern of fuction to predict data. However, if the data is far from the pattern, the result will be bad. For example, the network data is a period data which is neither in linear pattern nor in polynomial pattern. So the result we got form linear and poly model is worse than the other two. In contrast, random forest model and neural network model is more general. they don't care the patter of the data too much. If we have infinte time and space they can fit any kind of data. That's why random forest model and neural network model perform better.

Second, the parameter of each model. the paramter of model will influnce the performance of each model a lot. For example, the tree number and nodesize of random forest model, the size and decay of neural network model, the degree of the polynomial model. So when we do regression, how to set the parameter is a important question.

Third, over-fitting. In this project, 10-fold cross vadilation, ridge and lasso model are ways to avoid over-fitting. Although the model we calculate should fit the traindata, it cannot be too much to lose the universality of the model. We can avoid over-fitting by seperating the train and test data, or by regulation of the coefficients.