Ridge Regression Analysis on Correlates of Cars' Miles Per Gallon

Jiannan (Jeffrey) Zhang

Department of Mathematics

University of Texas at Austin, Austin TX 78712

Email: jiannanzhang@utexas.edu

Abstract

Nowadays, fuel economy is catching more and more attention. Therefore, it is essential to have good estimates of fuel consumption for different types of cars. MPG, which stands for mile per gallon, is one of the most standard measures of fuel consumption. In this project, I used the ridge regression method to generate the most stable regression model with the best ridge parameter. The data set I used is from the StatLib library (Bache and Lichman 2013. UCI Machine Learning Repository), which is maintained at Carnegie Mellon University. As my predictor variables I used number of cylinders, engine displacement, horsepower, weight of the vehicle, acceleration time, origin country and the year of the car. I used cross-validation to train and test the data. I obtained three different models by including in one all predictor variables and in the other two only categorical and continuous variables. I found HP (Horse Power) and weight as the most important factors affecting MPG in a negative way. Origin country and the year of the car were the only two variables positively related with MPG. Ridge regression model has smaller error than Ordinary Least Squares (OLS) regression model and it can reduce multicollinearity between predictor variables. Hence it is better to consider using ridge regression than OLS regression to build models with multiple potentially correlated variables.

Introduction

Concerns about the fuel supply and fuel efficiency have been growing a lot in recent years especially for cars. Fuel economy standards have long been a federal policy instrument to reduce gasoline use and are currently slated for an increase in stringency of 35 percent by 2020 (Jacobsen 2010). There are many variables associated with improvements in fuel efficiency such as horsepower and weight of the car. When choosing between energy-using durable goods such as autos and air conditioners, consumers are assumed to form beliefs about the energy costs of different models (Allcott 2011). Therefore, it is essential to have good estimates of fuel consumption for different types of cars. The standard for measuring fuel efficiency in the U.S. has been miles per gallon (MPG) (Bartkovich 2013).

Many different statistical studies have been done trying to determine the best model explaining fuel efficiency in cars. Some researchers have used Ordinary Least Squares (OLS) regression to study the correlation between car's MPG and its associated features such as HP and the weight of the car. Gautam (2010) used an OLS regression model to determine the effects of these explanatory variables on fuel economy.

In this study, there are seven predictor variables. Because of a big number of predictor variables, there is a big chance of multicolinearity between variables. In regression analysis, researchers often encounter the problem of multicollinearity. Multicollinearity leads to high variance and instable parameter estimates when estimating linear regression models using OLS (Wu and Liu 2014). To handle this problem, ridge estimator is used by many researchers (Wu and Liu 2014).

My study focuses on using ridge regression to analyze the correlation between car's MPG and a series of possible predictor variables. There has been increasing interest in using penalized regression in the analysis of high dimensional data (Cule et al. 2011). Ridge regression is one such penalized regression technique, which does not perform variable selection, instead estimating a regression coefficient for each predictor variable (Cule et al. 2011). Moreover, I also want to exam if ridge regression model is better than OLS regression model in this study. The data set I will use to achieve these goals is from the StatLib library (Bache and Lichman 2013. UCI Machine Learning Repository), which is maintained at Carnegie Mellon University.

Methods

The data set has 398 instances; each instance consists of the MPG value and values of seven attributes of the car (number of cylinders, engine displacement, horsepower, weight of the vehicle, acceleration time, origin country and the year of the car). There are eight missing values in some of the predictor variables. I decided not too include these cases in the analysis due to the incompleteness of the data.

I used the statistical software MATLAB (MathWorks 2014) to do my study. The main methods I used are the ridge regression and cross-validation. Ridge regression penalizes the size of the regression coefficients by adding a degree of bias to the regression estimates. In the ridge regression analysis, the estimation of ridge parameter *k* is an important problem (Kibria 2003). Cross-validation is a very robust method to train and test data. Next I explain the implementation of these techniques.

Suppose Y is the response variable we want to predict, X is the matrix containing all the features, *W* is the coefficients vector for the ridge regression model, lambda (λ) is the ridge parameter. As a result, we can derive *W* by the following formula:

$$\boldsymbol{W = (X^T X + \lambda I)^{-1}\ X^T Y}$$

The goal is to find the best ridge parameter that minimize the following expression:

$$\Sigma(Y_i - X_i^T W)^2 + \lambda\Sigma\, W_j^{\ 2}$$

Let us call the result of the above expression **"error"**. Note that if the *lambda* is zero in the expression above, the result is the error for the OLS model.

I used 10–fold Cross-Validation to train and test the data. Therefore I divided the data evenly into ten folds and each fold has 39 instances of the data. For a specific lambda, I used nine folds to train the data and get the regularization parameter '*W*'. After that, I used that '*W*' to test the "test data". Therefore there will be 39 predicted values for that corresponding test fold each time for a corresponding lambda. First, I calculated the square of every difference between the predicted value and the true value. Second, I calculated the mean of those differences for the test fold and then add the regularization term, which is *lambda*||W||^2*. The result of this is the mean error for one fold out of ten folds. Third, I calculated all the mean errors for ten folds and I took the mean again of them, denoting that mean by Lambda_error. Here, Lambda_error is the error I want to minimize. Fourth, I stored the lamba_error for the corresponding lambda into an array for each corresponding lambda. Finally, I plotted the scatter plot for Lambda_error against corresponding lambda and found the best lambda that gives me the smallest error.

I implemented three different models. In the first model (Model I) used the data set with six features (number of cylinders, engine displacement, horsepower, weight of the vehicle, acceleration time, origin country). In the second model (Model II) I used the data set with all continuous variables (displacement, HP, weight, acceleration time). Finally, in the third model (Model III) I used the data set with all discrete variables (number of cylinders, origin, years).

Results

Model I.

First I ran the model from lambda = 0 to 10 with step = 1 and plotted the Error VS Lambda (see figure 1). So the best lambda is between 5 to 7, then I ran it from 5 to 7 with step 0.05 (see figure 2). It shows that the best lambda is between 5.5 to 6.5. Then I ran it with step = 0.1 (see figure 3). Now I found the best lambda to fit this model with lambda = 5.9 and the corresponding error = 1.0e-04 * 0.8482. The error is pretty small. This means this model is an excellent model. There are some sample weights from model I (see table I). There samples weights are used to analyze the correlations between the MPG and cars and how big the effect is for a specific predictor variable.

Model II.

First I ran the model from lambda = 0 to 20 with step = 1 (see figure 4). It shows the best lambda is between 8 and 12, then I ran it again from 8 to 12 with step = 0.5 (see figure 5). So the best lambda for this model is 10 and error is 1.0e-04 * 0.8756. Notice that this error is a little bit higher than the error of the previous model. There are some sample weights from model II (see Table II).

Model III.

First I ran the model from lambda = 0 to 20 with step = 1 (see figure 6). This shows the best lambda is between 10 and 16. Then ran it again from 10 to 16 (see figure 7). Now we find the best lambda is 13 and the error is 1.0e-04 * 0.7281. Notice that this error is the smallest one in these three models. There are some sample weights for model III. (Table III)

From the graphs (Error VS Lambda) (see Figure1 to Figure 7), when lambda = 0, the corresponding error is the error for OLS regression model. From the graphs, ridge regression model has a smaller error than the OLS regression model, which indicates the ridge regression model is better than the OLS model in this study.

Discussion

In this project, I used the ridge regression method to analyze which variables best predict a car's MPG using the StatLib data set. In addition I also examined whether the ridge regression method is better than the OLS regression model in this study. My results shows weight of the car and horsepower are the best two predictors for the MPG. My study indicates the advantages of using ridge regression on a large set of predictor variables. The ridge regression showed smaller error than the OLS regression.

Some researchers found the similar results as mine. Kibria (2003) indicates that under certain conditions the proposed estimators perform well compared to least squares estimators (LSE) and other popular existing estimators (LSE is obtained in the OLS regression model). Gautam (2010) suggests that large gains in fuel economy are associated with technological factors - vehicle's weight and horsepower.

Additionally, I found there are five variables that are negative correlated with MPG. They are number of cylinders, engine displacement, horsepower, weight of the vehicle and acceleration time. Among these variables, weight of the vehicle has the biggest effect on MPG, followed by horsepower and the other three. Gautam (2010) concludes that the weight of a vehicle is a significant factor affecting fuel economy.

On the other hand, there are two variables that are positively correlated with MPG. They are year of the car and the country origin. For the country origin, 1 = America, 2 = Germany and 3 = Japan. Therefore Japanese cars usually have the highest MPG, followed by German cars and American cars.

Further work could be done to test if logistic regression is better for this part of the study using ridge regression. Cessie and van Houwelingen (1992) show how ridge regression can be used to improve the parameter estimates in logistic regression when the number of predictors is relatively large or highly correlated. Ridge regression penalizes the size of the regression coefficients by adding a degree of bias to the regression estimates. By controlling the size of coefficients, ridge regression is a good way to battle multicollinearly for a large set of predictor variables in a regression model. My study is important because it introduces a new way to analyze the correlation between car's attributes and car's MPG. Moreover, this study provided some more accurate and stable models than models using OLS.

Literature Cited

Allcott, H. 2011. Consumers' Perceptions and Misperceptions of Energy Costs. American Economic Review 101(3): 98-104.

Bartkovich, G.K. 2013. Modeling Fuel Efficiency: MPG or GPHM? The Mathematics Teacher 107(1): 20-27.

Cule, E., P. Vineis, and M. De Iori. 2011. Significance testing in ridge regression for genetic data. BMC Bioinformatics 12:372.

Cessie, S.L. and JCV. Houwelingen. 1992. Ridge Estimators in Logistic Regression. Applied Statistics 41(1): 191-201.

Gautam, S. 2010. What Factors Affect Average Fuel Economy of US Passenger Vehicles? Honors Projects. Paper 104.

Jacobsen, M.R. 2011. Fuel Economy, Car Class Mix, and Safety. American Economic Review 101(3): 105-109.

Kibria, B.M.G. 2003. Performance of some new ridge regression estimators. Communications in Statistics-Simulation and Computation 32: 419-435.

MATLAB version 8.3.0. The MathWorks Inc., 2014.

Wu, J. and C. Liu. 2014. Performance of Some Stochastic Restricted Ridge Estimator in Linear Regression Model. Journal of Applied Mathematics.

Table 1. Sample weights for model I (K = Intercept, Num = Number)

Variables	K	Num of Cylinders	Engine displacem	Horsepower	Weight	Acceleration time	Origin
Weights (best lambda)	0.0873	-0.1021	0.0474	-0.2682	-0.5452	-0.0365	0.1033
Weights	0.0842	-0.0576	-0.0204	-0.2365	-0.4796	-0.0675	0.1021
Weights	0.0843	-0.0565	-0.0160	-0.2383	-0.4864	-0.0676	0.1025

Table 2. Sample weights for model II (K = Intercept)

Variables	**K**	**Engine displacement**	**Horsepower**	**Weight**	**Acceleration time**
Weights (best lambda)	0.0919	-0.0965	-0.2072	-0.5733	-0.0369
Weights	0.0923	-0.1243	-0.2068	-0.5246	-0.0657
Weights	0.0936	-0.0841	-0.2312	-0.5807	-0.0408

Table 3. Sample weights for model III (K = intercept)

Variables	K	Number of Cylinders	Origin	Years
Weights (best lambda)	-0.0468	-0.5582	2.2883	0.1355
Weights	-0.0342	-0.5868	2.0849	0.1027
Weights	-0.0460	-0.5780	2.2930	0.1314

Figure Legends

Figure 1. Plot of model error VS lambda (ridge parameter) for Model I. Lambda is from 0 to 10 with step = 1

Figure 2. Plot of model error VS lambda (ridge parameter) for Model I. Lambda is from 5 to 7 with step = 0.05

Figure 3. Plot of model error VS lambda (ridge parameter) for Model I. Lambda is from 5.5 to 6.5 with step = 0.1

Figure 4. Plot of model error VS lambda (ridge parameter) for Model II. Lambda is from 0 to 20 with step = 1

Figure 5. Plot of model error VS lambda (ridge parameter) for Model II. Lambda is from 8 to 12 with step = 0.5

Figure 6. Plot of model error VS lambda (ridge parameter) for Model III. Lambda is from 0 to 20 with step = 1

Figure 7. Plot of model error VS lambda (ridge parameter) for Model III. Lambda is from 10 to 16 with step = 1

Figure 1

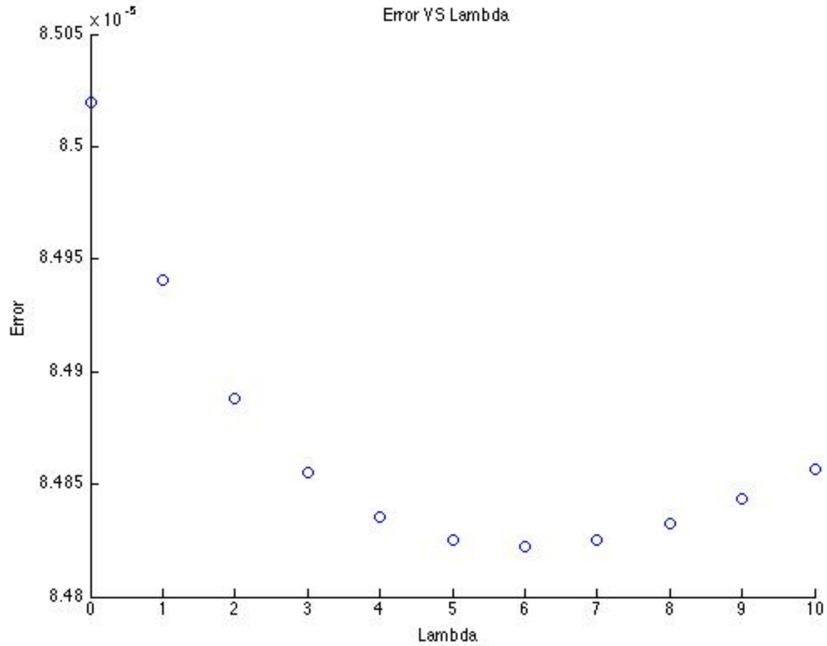

Error VS Lambda
x 10^-5
8.505
8.5
8.495
8.49
8.485
8.48
Error
0
1
2
3
4
5
6
7
8
9
10
Lambda

Figure 2

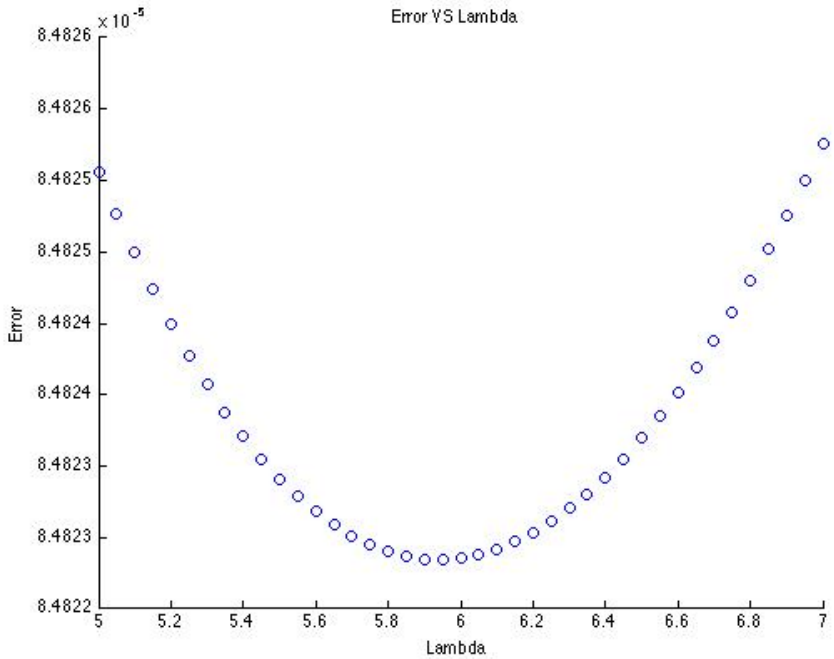

Error VS Lambda
$\times 10^{-5}$
8.4826
8.4826
8.4825
8.4825
8.4824
8.4824
8.4823
8.4823
8.4822
Error
5
5.2
5.4
5.6
5.8
6
6.2
6.4
6.6
6.8
7
Lambda

Figure 3

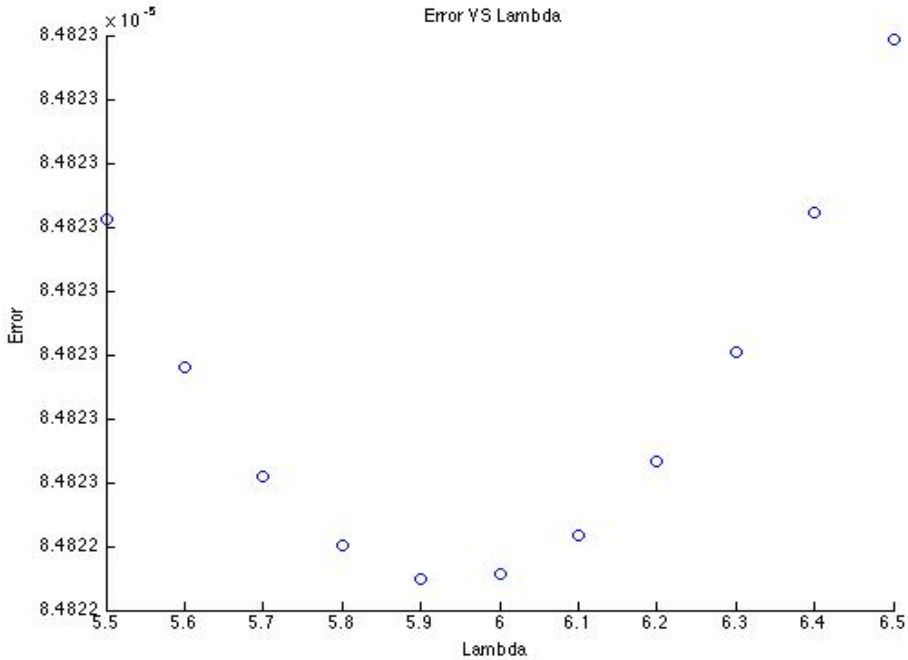

Error VS Lambda
× 10^-5
8.4823
8.4823
8.4823
8.4823
8.4823
8.4823
8.4823
8.4823
8.4822
8.4822
Error
5.5
5.6
5.7
5.8
5.9
6
6.1
6.2
6.3
6.4
6.5
Lambda

Figure 4

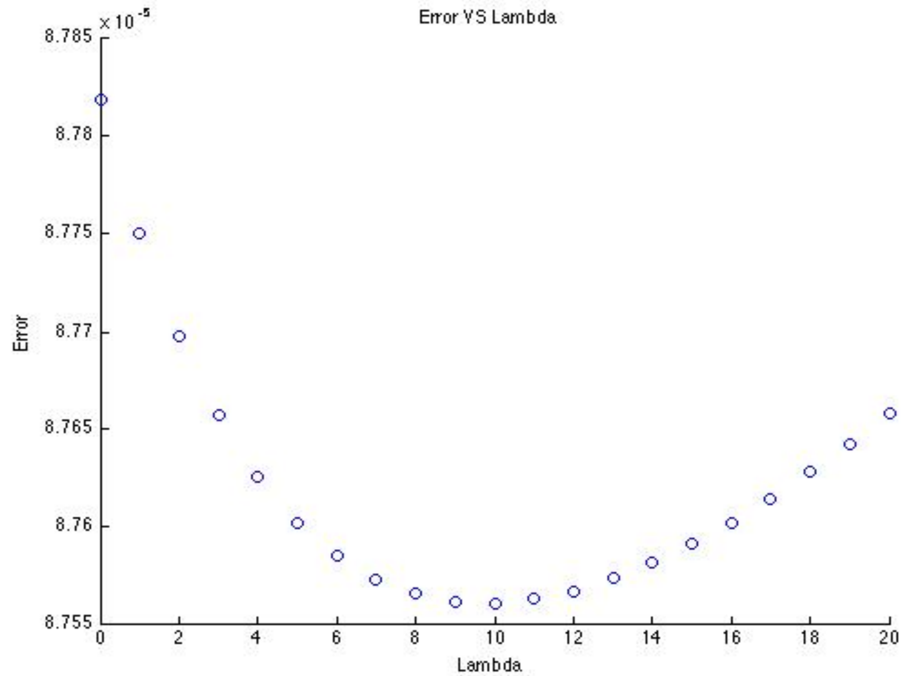

Error VS Lambda
× 10^-5
8.785
8.78
8.775
8.77
8.765
8.76
8.755
Error
0
2
4
6
8
10
12
14
16
18
20
Lambda

Figure 5

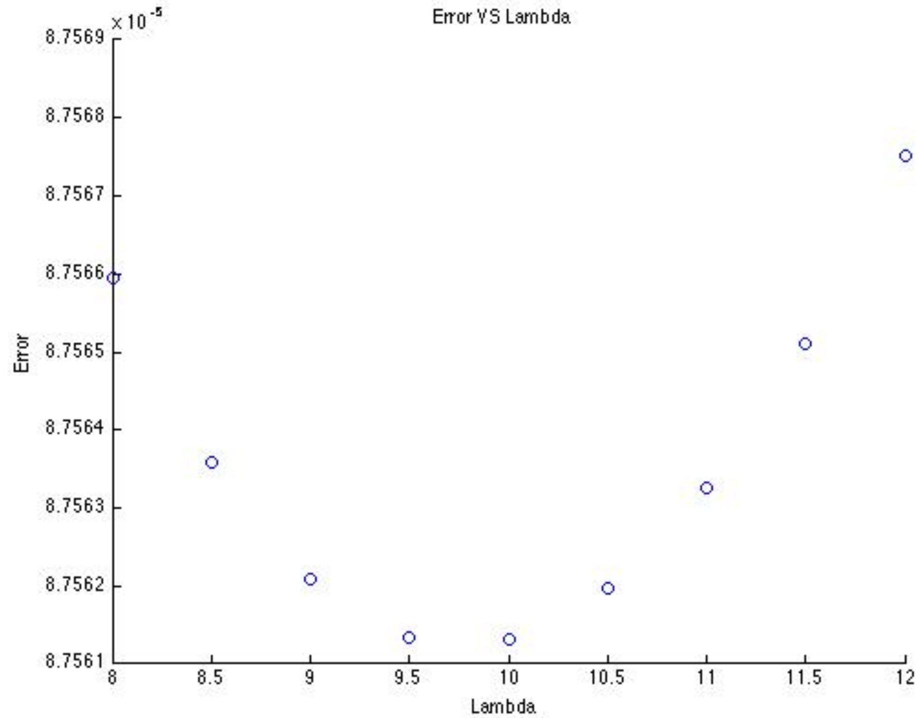

Error VS Lambda
x 10^-5
8.7569
8.7568
8.7567
8.7566
8.7565
8.7564
8.7563
8.7562
8.7561
Error
8
8.5
9
9.5
10
10.5
11
11.5
12
Lambda

Figure 6

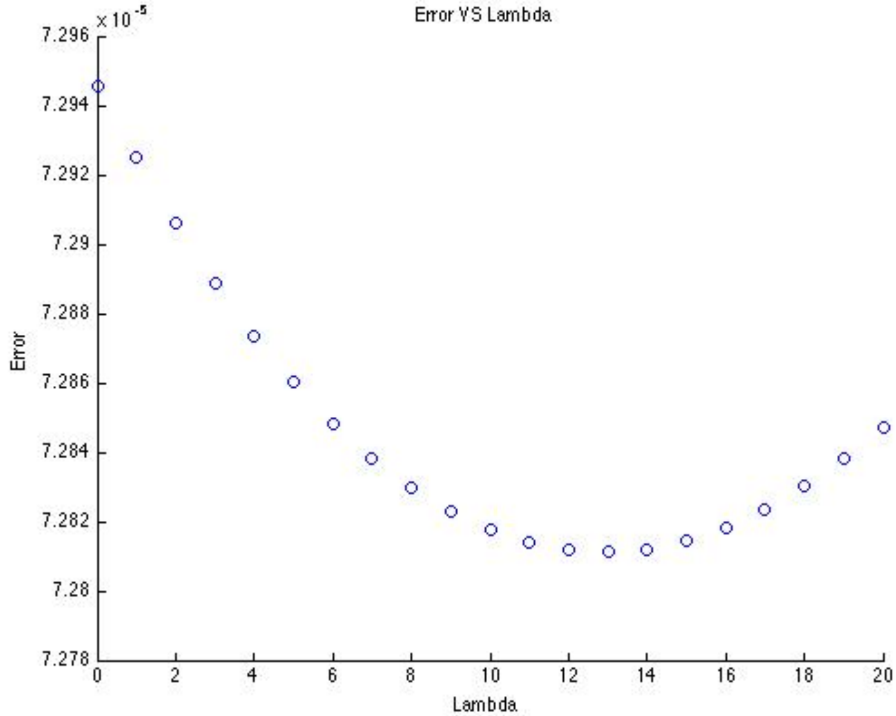
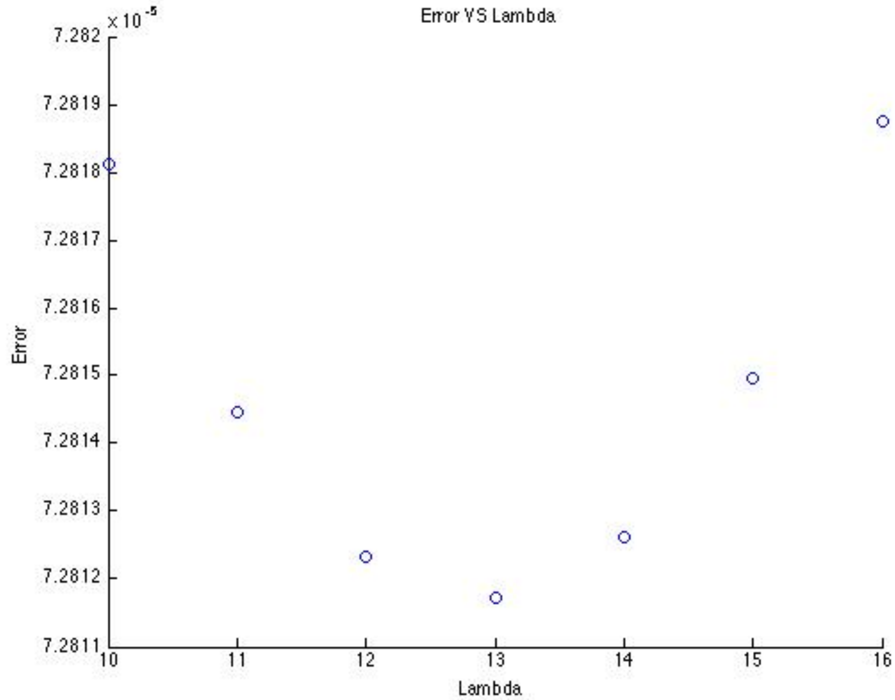

Error VS Lambda
× 10^-5
7.296
7.294
7.292
7.29
7.288
7.286
7.284
7.282
7.28
7.278
Error
0
2
4
6
8
10
12
14
16
18
20
Lambda

Figure 7


Error VS Lambda
x 10^-5
7.282
7.2819
7.2818
7.2817
7.2816
7.2815
7.2814
7.2813
7.2812
7.2811
Error
10
11
12
13
14
15
16
Lambda