Assignment 1 - Report

STAT497

Presented to

Dr. Frederic Godin

By

Nam Phuong Nguyen (40011040)

Tony Yuan (40029336)

Spyros Orfanos (40033280)

February 16, 2018

Question 1

**Part (b)**

Created a design matrix that is a 291x11. First column is a column of 1s and the next 10 columns is for our Fourier expansion.

**Part (c)**

The function takes the xdesignTrain matrix and the yTrain observed response and produces the optimal betas in an OLS Regression using matrix multiplication. Using the betas calculated from the xdesignTrain and the yTrain a prediction is made for XdesignValid. The prediction and yValid are then used to calculate the sum of squared residuals and is returned by the function.

**Part (d)**

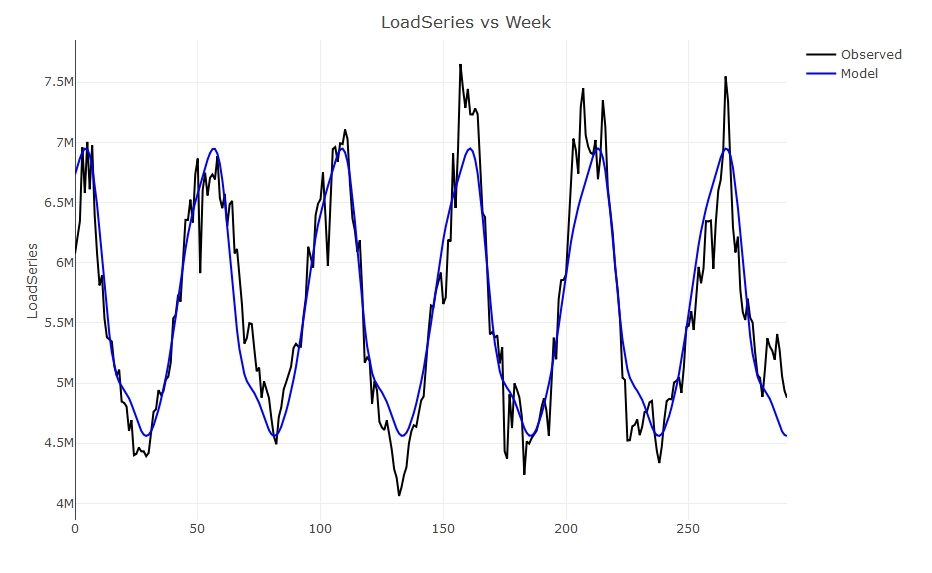
The SSETot was calculated for many 5 different values of p.

SSETot:



P Chosen = 4

**Part (e)**



**Part (f)**

* Calculated week to be t = 293
* Prediction = 4652320

Question 2

**Part (a)**

* RMSE = 100.7576

**Part (b)**

* 6 significant predictor variables: Income, Limit, Rating, Cards, Age, and StudentYes
* RMSE = 99.78596

**Part (c)**

**Minimizing BIC using:**

1. **Forward Approach**

* Best model has 5 predictor variables: Income, Limit, Rating, Cards, and StudentYes
* BIC = -1197.096
* The predictor variables have the following coefficients:



1. **Exhaustive Approach**

* Best model has 4 predictor variables: Income, Limit, Cards, and StudentYes
* BIC = -1198.053
* The predictor variables have the following coefficients:



**Comparison:**

* The exhaustive approach provides a better model since it has a lower BIC. Note that the predictors used are different, as the exhaustive approach model does not use Rating as a predictor.

**Part (d)**

* The optimal value of lambda is 0.04037
* The associated out-of-sample RMSE is 102.6047

Some comments regarding code for this question:

* Lambdagrid is an arithmetic sequence, so its terms can be explicitly represented:

seq(10,-2,length=100) = (10, A2, A3,…, A99, -2)

⬄ A1 + d(99)= A100

⬄ 10 +99d = -2

⬄ d = -12/99

⬄ Aj = 10 -12/99\*(j-1) for j = 1,2,…,99,100

By expressing lambdagrid this way, we can easily use a for loop to calculate the MSE for different values of lambda and store the results in a matrix.

* The matrix we create is 5x100, where each column gives the SSEi for a value of lambda. By taking the sum of the elements in a column, we obtain the SSETOT for each lambda. Dividing by 400 and taking the square root, we obtain the RMSETOT for a given value of lambda. We find the column which has the smallest RMSE to find the optimal value of lambda.

**Part (e)**

* The optimal value of lambda is 0.01
* The associated out-of-sample RMSE is 102.6022

**Part (f)**

We suggest to use the model trained in Part (b), which uses predictor variables that are significant at the 5% level. Since both prediction and inference are important for our model, having 6 predictor variables allows for accurate predictions and also provides insight to which predictor variables contribute to credit payment defaults. The following are the predictors with their coefficients for the suggested model:



Question 3

**Part (a)**

The dmvn function created takes in 3 components

* a 2x192 matrix “xmat” with the predictor values of the sample
* a 2x1 column vector “muvec” with the means of the Maturity and the Strike predictors
* and a 2x2 predetermined covariance matrix “Sigmamat”.

It then uses the probability density function of the multivariate Gaussian distribution

to return “pdfvec”, a vector of 192 components.

**Part (b)**

Using the dmvn function, we associate a weight to each of the 192 sample points to predict the implicit volatility for each of the three covariance matrices Sigma. First, we find the betas for each of the 336 points in grid X and also for each sigma using the formula:

**,**

where from the sample points,

from the weights calculated previously

and y is the vector of 192 implicit volatility values taken from the sample.

Second, we calculate the predicted implicit volatility by calculating the matrix , where gridX represents all the possible combinations of the predictors gridMaturity and gridStrike for which we want the prediction. A is a 336x336 matrix, and we notice that the (i,i)th component A is the ith prediction calculated using the beta coefficients for the ith prediction, which is . We then calculate , which we store in fhatmatlist, a list of three 16x21 matrices.

**Part (c)**

The values are the plotted against their respective maturity and strike.



**Part (d)**

We create a function PredictNearestNeighbors (PNN) which takes in 4 components:

* “yval” which is a 1x192 vector of the sample implicit volatility;
* “xval”, a 2x192 matrix with the corresponding maturity and strike values of each yval value;
* “xpred” which are the maturity and strike that we want to predict the implicit volatility for
* and NNnumber, a scalar which represents the number of nearest neighbors we want to consider.

The function calculates the Euclidean distance between each sample point and the predictions we want, selects the NNnumber closest sample points to the predictions and returns the mean of the corresponding y values as “NNpreds” vector.

**Part (e)**

We rescale the X and the gridX matrices by changing each Strike values to Strike/400. Then, we use the PNN function for NNnumber = 5, 15 and 30 for the previous data set to predict the implicit volatility values that we store in fhatmatlistKNN, a list with three 16x21 matrices.

**Part (f)**

We plot the values against their respective maturity and strike.

