D5 – Model Selection

1. **Part A: MODEL SELECTION**

a) Forward Selection

Model Selection is a technique that is performed in data analysis to choose which model explains best the prediction we are trying to accomplish. Performing the Forward selection for variable selection is a technique where we start the model with one predicator, add another predicator, compute the coefficient, as well as compute the R squared and the mean squared error. The variable that has a high p-value is dropped from the model and another predicator is added to the model. This is done on all the variables of the dataset until at the end a model with the best mean squared error and R squared is returned. During our analysis, we used a function called regsubsets that would compute the forward selection and give us back the variables with the best coefficients that would give a better prediction. Using a function called which.min(), R was able to return the predictors with the lowest mean squared error. The forward selection on our dataset showed us we could use 35 predicators to build a good model.

b) Backward Selection

Backward selection starts by having all the predictors in the dataset and after performing the p value and mean squared error, the predictor with the highest p-value will be removed from the model. The model will run the calculation and remove the highest p-value until the remaining predictors are giving a good reading. In our dataset, the model returned us 35 variables that were giving a good prediction.

We pursued our analysis by performing a k-fold on our dataset and on a training and test dataset. At the end we plotted the Mean Squared errors of all the iterations and realized that the model would be good if we used 7 variables. Those variables were the following:

Predictors	**Coefficients**
Intercept	1.2822542610
Rating	-0.0809618489
Nber of adult HD pedi months w/ Ktv data	-0.0275288957
Nber of patients included in hospitalization	-0.0004973728
Patient Survival Category txt`1	0.4703799562
Offers in Center Hemodialysis	0.0000000000
Rating Code	0.0000000000
numb of patients inc in the facility serum phosp	-0.0076125335

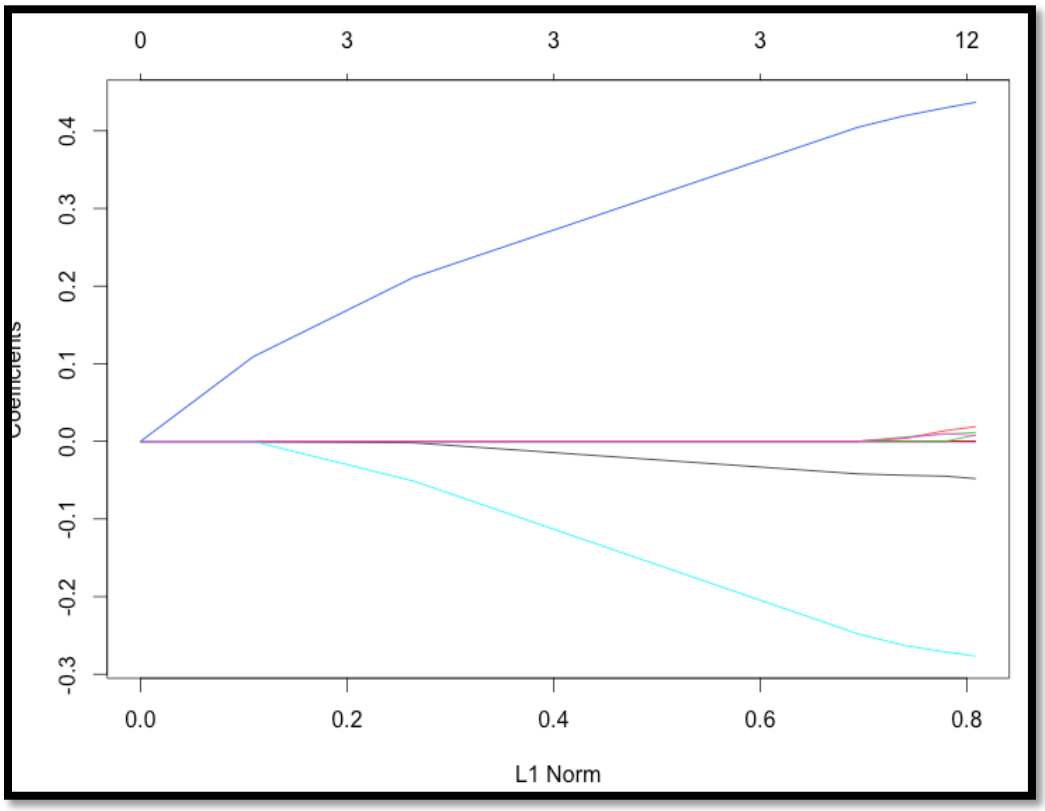
c) Ridge Model

The ridge model is important as it tells us if there is multi-colinearity on the predictors used in our model. It tries to change the intercept close to zero as the value of lambda increases… It is not very helpful for variable selection but a good

means of showing if the variables chosen are unbiased, meaning there is no correlation among the predictors that will give us a biased result on the model. Once we performed the ridge model on our dataset, we realized that not many predictors had an intercept of zero once we chose the best fit of lambda. Around 30 predictors would be used to build our models if we were to use them to build are final model. Below are the variables that would be used for a model using Ridge selection.

d) Lasso Model

The lasso Model works like the ridge model. The only difference is that the lasso model gives us a list of variables that should not be included in the final model by giving those predictors a coefficient of zero where the optimal value of lambda is met.



The picture above can show us that as the value of lambda increases, there are some predictors that are on the x-axis. Meaning they have a coefficient of zero. Looking closely we can see that around 7 or 8 predictors do not have zero. Below is a list of predictors that could be used by the lasso model to give us a good prediction of our model:

Predictors	Coefficient
Intercept	1.130964e+00
Rating	-4.614049e-02
`Profit/Non Profit`1	4.704888e-03
Facilities Transfusion Ratio	4.168200e-03
% of adult patients w/ hypercalcemia serum calcium > 10.2mg/dl	-6.361289e-06
Patient Survival Category txt`1	4.312455e-01
Standardized Hospitalization Ratio	2.873254e-02
`Patient Survival Category txt`2	-2.791367e-01
% of adult patients w/ serum phosp > 7.0mg/dl	-9.099259e-04