Model Comparison in Predicting Cholesterol Levels

P8106 Midterm Project

Adeline Shin, Group Members: Sabrina Lin and Ngoc Duong

Introduction

In the US, high cholesterol is a common health problem, affecting more than 12% of adults over the age of 20, according to the CDC. This data on cholesterol was collected as a part of the NHANES database, which consists of answers to a national survey conducted on nutrition and health behavior among Americans.

Our group used the NHANES data to predict cholesterol based on demographic, dietary, laboratory, and questionnaire data from the surveys conducted during 2015-2016. We picked a combination of 63 potential predictors from these categories in order to cover potential social, behavioral, and genetic determinants of the outcome variable, LDL cholesterol levels. We were interested in determining which model type was the most effective in predicting cholesterol levels given values for all other predictors.

The 63 variables as potential predictors were chosen by looking at the entirety of the NHANES dataset. Based on our research on causes of high cholesterol levels, we decided to pick variables across all sections of the NHANES dataset. The chosen variables are listed in Appendix A1, along with their variables names and categories.

Using these 63 variables, we first separated the data into training and test data using an 80/20 split. The training data was used to generate the models, and the test data was used to compare to the predicted values from the models. Using the RMSE calculated between the test data and the predicted data, we were able to compare which method had the lowest RMSE value, and therefore the best predictive abilities.

Exploratory Analysis and Visualization

In order to conduct a preliminary exploratory analysis, the dataset was loaded using the nhanesA package, then the summary() and table() functions were used to find potential outliers or data that was coded differently than expected.

During this procedure, we found many missing values or unknown values that were coded with the values "7, 9, 777, 999, 777777, 999999, and ." in particular cases. These values were all converted to NAs for the purpose of this project, since we were not attempting to recover or predict missing data. We also noticed that some of the variables were separated into two categories for youth and adult, so we decided to just focus on adults for the scope of the project.

After filtering out all of the rows with NAs and using only rows with data from adults, a dataframe of 661 observations was left.

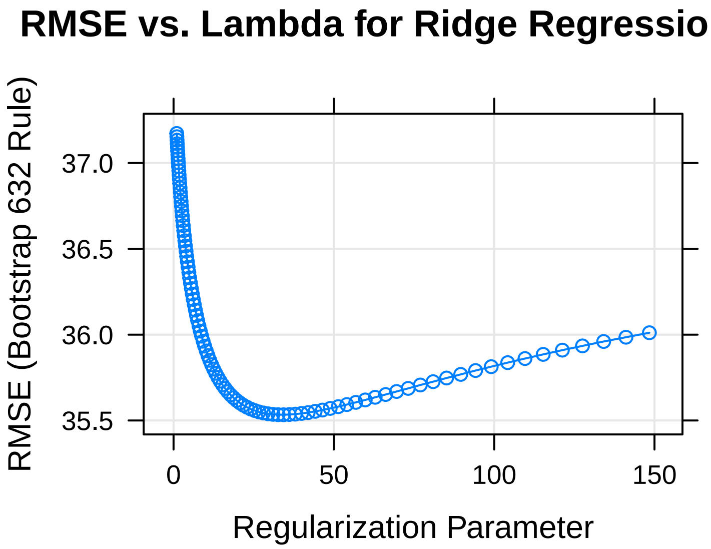
Models

In this project, the caret package in R was used to train all the models, and thus, the models compared were a linear model, a ridge regression model, a lasso model, and an elastic net model with an alpha value of 0.75. Our group members decided to use different cross-validation methods in order to see whether that would make a difference in terms of the final chosen model. In the models below, the 632 bootstrap cross-validation method was used, while other team members chose to use Monte Carlo cross-validation or leave one out cross-validation.

Linear Model

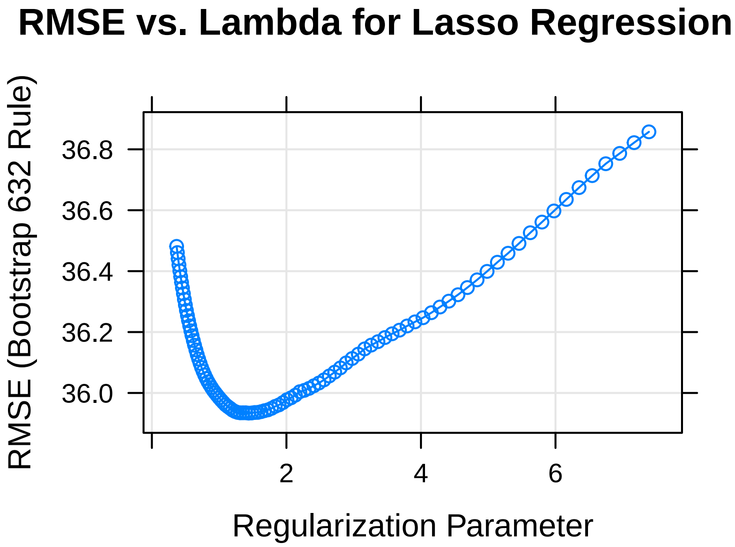
At a 95% significance level, only 6 of the 63 predictors are significant, which likely means that the model does not fit the data well. In addition, from the model summary, the adjusted R-squared value is only 0.1146, which confirms that the model is not a good fit for the data and therefore will likely not predict well. The linear model has a cross-validated MSE of 904.0828578, which is high, but expected, as the model itself did not have many significant variables.

Ridge Regression Model



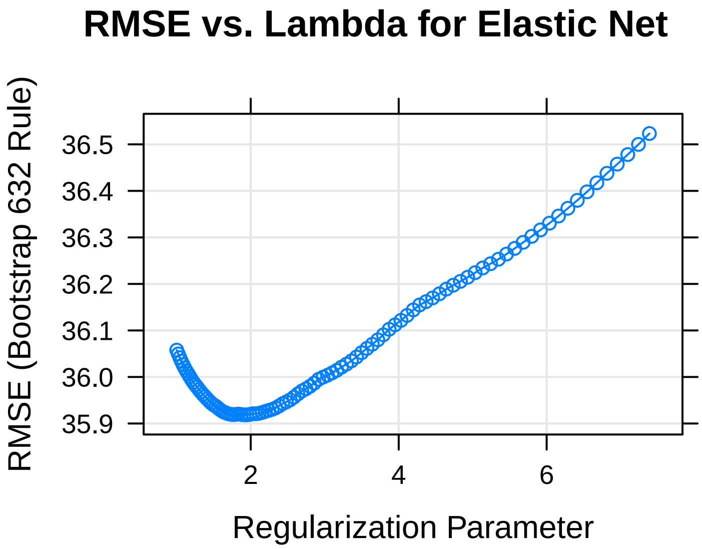
As shown on the graph above, the value of lambda that gives the lowest RMSE value is 34.3071414. The ridge model at this value of lambda gives 85 predictors in the final model, including the different factor levels, which can then be used to predict cholesterol levels. The ridge model gives a cross-validated MSE of 969.8954725, which is lower than that of the linear model.

Lasso Model



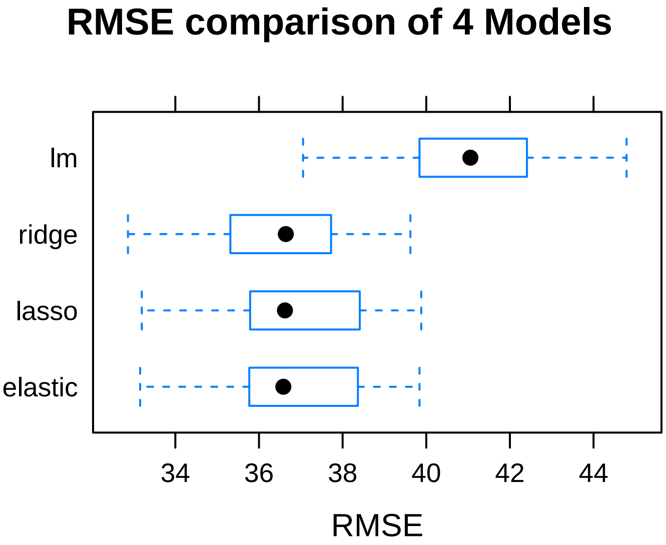
As shown on the graph above, the value of lambda that gives the lowest RMSE value is 1.438551. The lasso model at this value of lambda gives 33 variables in the final model, which can then be used to predict cholesterol levels. With this value of lambda, the lasso model gives a cross-validated MSE of 984.8223326.

Elastic Net Model



As shown in the graph above, the elastic net model with an alpha of 0.75 has the lowest RMSE at a lambda value of 1.9087807. This combination of alpha and beta gives a model that contains 33 predictors in the final model. This model is then used to predict the cholesterol level and compared to the test data. Prediction of cholesterol levels using the elastic net model gives a cross-validated MSE of 986.4715369.

Model Comparison



The box plot shows that the model created using the elastic net gives the lowest RMSE among the four models created for the purpose of predicting cholesterol levels. Therefore, the elastic model should be chosen when using 632 bootstrap as the cross validation method to predict cholesterol levels.

Conclusion

All three of our group members used different forms of cross-validation for our model training, which resulted in different conclusions of which model did the best in predicting cholesterol levels. For 632 bootstrap and Monte Carlo cross-validation methods, elastic net proved to have the best model, but for LOOCV abd 10-fold cross-validation, the ridge model was best at prediction. Therefore, it is extremely important to specify model parameters, including cross-validation technique, when determining which model will work best in predicting outcomes.

While the elastic net model worked best with the 632 bootstrap cross-validation method, none of the models gave great results. The R-squared values for the models were all under 0.2, suggesting that none of the models created fit the data well, and therefore were not expected to predict cholesterol levels well. This is likely because we were dealing with real data with limited predictors, and the NHANES dataset did not necessarily capture all the predictors of high cholesterol level. In addition, given more time, there could have been other model methods used to capture the data better, but those were not in the scope of this class so far.

Appendices

A1

Table of Predictors and Corresponding Variable Names

variable	definition
seqn	Respondent Sequence Number
lbdldl	LDL/Triglyceride Levels (Outcome)
urxuma	Albumin (ug/mL)
urxucr	Creatinine (mg/dL)
lbxsapsi	Alkaline Phosphotase (IU/L)
lbxsc3si	Bicarbonate (mmol/L)
lbxsgl	Glucose (mg/dL)
lbxsgb	Gamma Glutamyl Transferase (U/L)
lbxsgtsi	Lactate Dehydrogenase (U/L)
lbxsldsi	Phosphorus (mg/dL)
lbxsph	Potassium (mmol/L)
lbxsksi	Total Bilirubin (mg/dL)
lbxstb	Uric acid (mg/dL)
lbxsua	Lymphocyte percent (%)
lbxlypct	Monocyte percent (%)
lbxmopct	Segmented neutrophils percent (%)
lbxnepct	Hemoglobin (g/dL)
lbxhgb	Hematocrit (%)
lbxhct	High-Sensitivity C-Reactive Protein (hs-CRP) (mg/L)
lbxhscrp	Body Mass Index (kg/m**2)
bmxbmi	Waist Circumference (cm)
bmxwaist	Systolic: Blood pressure
bpxsy	Diastolic: Blood pressure
bpxdi	How often do you add ordinary salt to your food at the table?
dbd100	How often is ordinary salt or seasoned salt added in cooking or preparing foods in your household?
drqsprep	Are you currently on any kind of diet?
drqsdiet	Total number of foods/beverages reported in the individual foods file
dr1tnumf	Energy (kcal)
dr1tkcal	Protein (gm)

variable	definition
dr1tprot	Carbohydrate (gm)
dr1tcarb	Total sugars (gm)
dr1tsugr	Dietary fiber (gm)
dr1tfibe	Total fat (gm)
dr1ttfat	Caffeine (mg)
dr1tcaff	Alcohol (gm)
dr1talco	Was the amount of food that you ate yesterday much more than usual, usual, or much less than usual?
dr1_300	Total plain water drank yesterday
dr1_320z	During the past 30 days did you eat any types of shellfish?
drd340	During the past 30 days did you eat any types of fish?
drd360	Total # of Dietary Supplements Taken
ds1dscnt	Total # of Antacids Taken
ds1ancnt	Money Spent at Grocery Stores in the Last 30 Days
cbd071	Money Spent on Eating Out in the Last 30 Days
cbd121	Money Spent on Takeout in the Last 30 Days
cbd131	Number of Meals Ordered
dbd895	Have you ever been told by a doctor or other health professional that you had hypertension?
bpq020	Ever had any pain or discomfort in chest
cdq001	General Health Level
hsd010	Ever told that you had Diabetes
diq010	How healthy is overall diet?
dbq700	Worried food would run out
fsd032a	Can't afford to eat balanced meals
fsd032c	Household food security
fsdhh	Does the insurance plan cover prescription medicine?
hiq270	Do you get regular physical activity?
paq605	Smoked 100 cigarettes in lifetime
smq020	Gender
riagendr	Age
ridageyr	Education
ridreth1	Recorded Race Hispanic Origin
ridreth3	Recorded Race with Non-Hispanic Asian Category
dmdborn4	Born in the US or not
indfmpir	Ratio of Family Income to Poverty