P8106 Midterm Project Report

Sabrina Lin stl2137

Introduction

The National Health and Nutrition Examination Survey (NHANES) conducted by the CDC collects health and nutrition data among Americans going back to 1999. Our group is primarily interested in finding a model to predict LDL cholesterol levels in adults, as LDL cholesterol is associated with cardiovascular disease. We are also curious to see how different cross validation methods would affect which model was the best to predict LDL cholesterol.

The NHANES data is categorized into the following six categories: demographics, dietary, examination, laboratory, questionnaire, and limited access data. When building our dataset, we decided to look through and extract variables from the demographics, dietary, laboratory, and questionnaire data, utilizing the "nhanesA" package. Since the risk factors of having high LDL according to the Mayo Clinic are poor diet, obesity, lack of exercise, smoking, age, and diabetes, we included the variables that capture these risk factors in our dataset, in addition to other variables that could potentially capture social or genetic factors when predicting LDL cholesterol. Our final dataset consists of 661 observations with 63 predictors, 21 of them being categorical and the remaining 41 as numeric variables.

Exploratory Analysis

Upon coming to our final dataset, we looked for near-zero variance predictors and found the variable, hiq011, representing whether the subject had healthcare, had only 1 subject in our training dataset who did not have healthcare. We thus took out hiq011, resulting in our eventual 63 predictor dataset. After properly factoring the categorical variables, removing missing values and observations that were imputed as unknown or missing, and excluding subjects under age 18, we joined our datasets to complete our final dataset. Because we are looking to predict LDL cholesterol using a model, we performed a 80/20 split into training and test datasets respectively.

Models

To predict LDL cholesterol, we used the 63 variables we pulled from the NHANES data (Appendix 1). Our group had decided that we would do a linear model, lasso model, ridge model, and elastic net model utilizing different cross validation techniques. I utilized 10-fold and leave one out cross validation (LOOCV). To do so, we utilized the `caret` package.

10-fold CV

Linear Model with 10-fold CV

The linear model with 10-fold CV has a predicted MSE of 1353.7600907.

Lasso Model with 10-fold CV

The lambda that gives the lowest RMSE value is 3.2602966. The lasso model for this lambda has 6 predictors in the final model, including the various factor levels to predict LDL cholesterol levels.

The lasso model gives a cross-validated MSE of 1441.3547424.

Ridge Model with 10-fold CV

The lambda that gives the lowest RMSE value is 32.6174847. The lasso model for this lambda has 85 predictors in the final model, including the various factor levels to predict LDL cholesterol levels.

The lasso model gives a cross-validated MSE of 1357.8259902.

Elastic Net Model (alpha = 0.75) with 10-fold CV

The lambda that gives the lowest RMSE value is 4.2825364. The lasso model for this lambda has 6 predictors in the final model, including the various factor levels to predict LDL cholesterol levels.

The lasso model gives a cross-validated MSE of 1441.167912.

Elastic Net Model (alpha = 0.25) with 10-fold CV

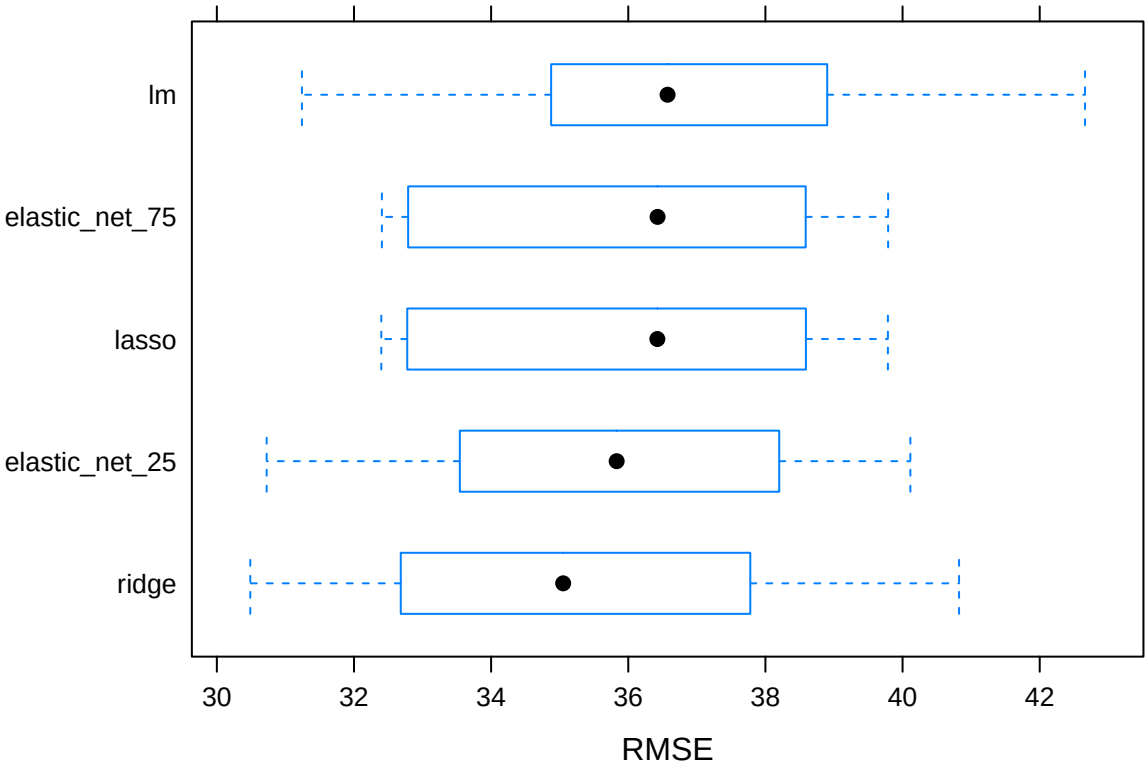
The lambda that gives the lowest RMSE value is 4.3260136. The lasso model for this lambda has 44 predictors in the final model, including the various factor levels to predict LDL cholesterol levels.

The lasso model gives a cross-validated MSE of 1358.9095435.

Comparing models

```
##
## Call:
## summary.resamples(object = resamp)
##
## Models: lm, lasso, ridge, elastic_net_75, elastic_net_25
## Number of resamples: 10
##
## MAE
##                     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## lm             23.95382 26.47768 27.85571 28.32138 29.93487 34.00952    0
## lasso          24.53510 26.11822 28.25318 27.90220 29.47330 31.07430    0
## ridge          23.78106 25.02946 27.52867 27.26863 29.40698 30.62777    0
## elastic_net_75 24.53098 26.11064 28.25292 27.90734 29.49693 31.08315    0
## elastic_net_25 24.51908 25.38334 28.13838 27.74933 29.59878 31.55338    0
##
## RMSE
##                     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
## lm             31.24079 34.97338 36.57335 36.93192 38.76880 42.66092    0
## lasso          32.39855 32.85593 36.42427 35.89154 38.24164 39.78603    0
## ridge          30.48774 32.88673 35.05073 35.48087 37.71310 40.82420    0
## elastic_net_75 32.40706 32.86989 36.42745 35.90334 38.24992 39.78962    0
## elastic_net_25 30.72614 33.78430 35.83073 35.91662 38.12628 40.11398    0
##
## Rsquared
##                       Min.    1st Qu.     Median       Mean    3rd Qu.
## lm             5.277924e-03 0.05354253 0.08796727 0.08485882 0.10224052
## lasso          1.830910e-05 0.01157490 0.04826540 0.06515193 0.08063767
## ridge          3.341437e-03 0.02735938 0.06586788 0.07375424 0.07289065
```

```
## elastic_net_75 8.085229e-05 0.01159923 0.04675473 0.06419567 0.07994471
## elastic_net_25 3.810353e-03 0.02367574 0.05642958 0.06212303 0.07427410
##                      Max. NA's
## lm              0.1908614    0
## lasso           0.2039932    0
## ridge           0.2322703    0
## elastic_net_75 0.2004992    0
## elastic_net_25 0.2052825    0
```



From the boxplot and confirmed by `summary(resamp)`, we can see that the ridge regression gives the lowest RMSE among the five models when predicting cholesterol level when using 10-fold cross-validation.

LOOCV

Linear Model with LOOCV

The linear model with LOOCV has a predicted MSE of 1353.7600907.

Lasso Model with LOOCV

The lambda that gives the lowest RMSE value is 0.4832251. The lasso model for this lambda has 59 predictors in the final model, including the various factor levels to predict LDL cholesterol levels.

The lasso model gives a cross-validated MSE of 1349.3138277.

Ridge Model with LOOCV

The lambda that gives the lowest RMSE value is 25.3384558. The lasso model for this lambda has 85 predictors in the final model, including the various factor levels to predict LDL cholesterol levels.

The lasso model gives a cross-validated MSE of 1352.2229609.

Elastic Net Model (alpha = 0.75) with LOOCV

The lambda that gives the lowest RMSE value is 0.6542653. The lasso model for this lambda has 59 predictors in the final model, including the various factor levels to predict LDL cholesterol levels.

The lasso model gives a cross-validated MSE of 1348.6063904.

Elastic Net Model (alpha = 0.25) with LOOCV

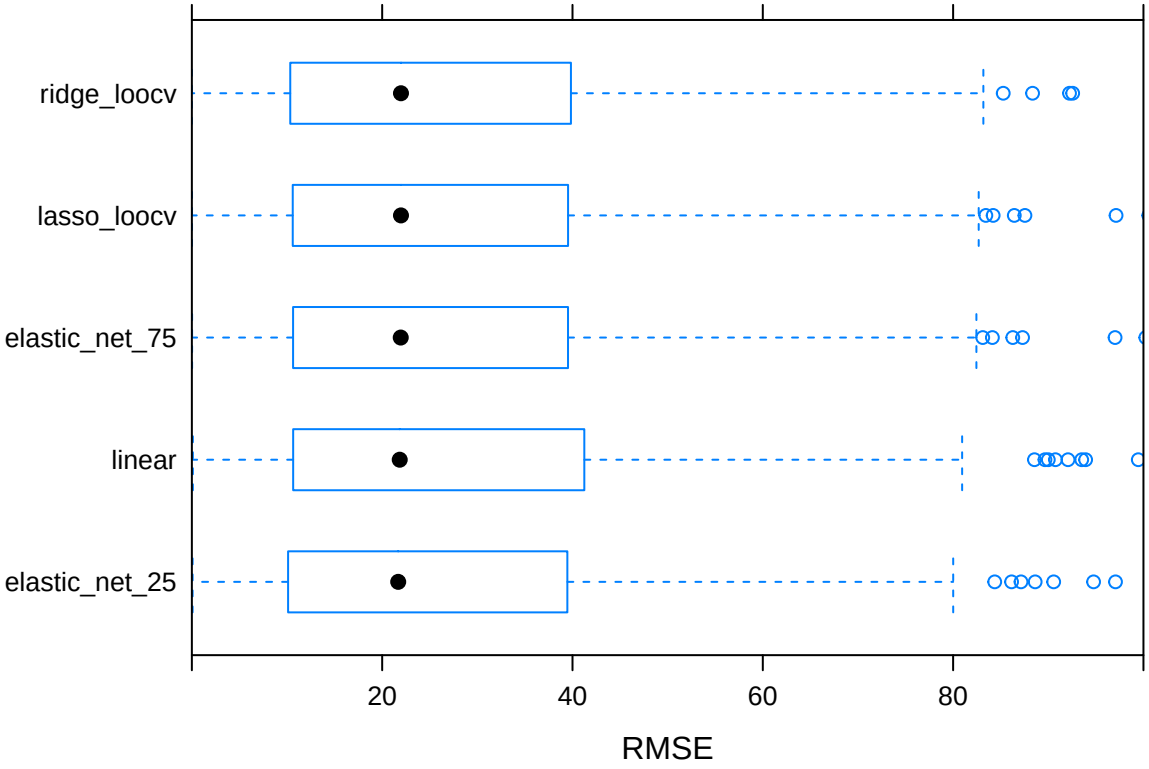
The lambda that gives the lowest RMSE value is 0.7613004. The lasso model for this lambda has 69 predictors in the final model, including the various factor levels to predict LDL cholesterol levels.

The lasso model gives a cross-validated MSE of 1358.9627052.

Comparing models

```
##
## Call:
## summary.resamples(object = resamp_loocv)
##
## Models: linear, lasso_loocv, ridge_loocv, elastic_net_75, elastic_net_25
## Number of resamples: 529
##
## MAE
##                       Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## linear         0.107376801 10.64663 21.84473 28.12841 41.24396 159.7836
## lasso_loocv    0.018165172 10.60262 21.97510 27.60123 39.53496 161.4819
## ridge_loocv    0.001812031 10.34089 21.97708 27.34720 39.83712 146.5593
## elastic_net_75 0.025835343 10.63873 21.95237 27.59410 39.53006 161.2231
## elastic_net_25 0.079298262 10.11322 21.68390 27.58897 39.46063 160.8559
##                NA's
## linear            0
## lasso_loocv       0
## ridge_loocv       0
## elastic_net_75    0
## elastic_net_25    0
##
## RMSE
##                       Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## linear         0.107376801 10.64663 21.84473 28.12841 41.24396 159.7836
## lasso_loocv    0.018165172 10.60262 21.97510 27.60123 39.53496 161.4819
## ridge_loocv    0.001812031 10.34089 21.97708 27.34720 39.83712 146.5593
## elastic_net_75 0.025835343 10.63873 21.95237 27.59410 39.53006 161.2231
## elastic_net_25 0.079298262 10.11322 21.68390 27.58897 39.46063 160.8559
##                NA's
## linear            0
```

```
## lasso_loocv       0
## ridge_loocv       0
## elastic_net_75    0
## elastic_net_25    0
##
## Rsquared
##                Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## linear           NA      NA     NA  NaN      NA   NA  529
## lasso_loocv      NA      NA     NA  NaN      NA   NA  529
## ridge_loocv      NA      NA     NA  NaN      NA   NA  529
## elastic_net_75   NA      NA     NA  NaN      NA   NA  529
## elastic_net_25   NA      NA     NA  NaN      NA   NA  529
```



By `summary(resamp)` since the boxplot makes it hard to see which RMSE is the lowest, the ridge regression gives the lowest RMSE among the five models when predicting cholesterol level when using 10-fold cross-validation.

Conclusions

Looking at both the 10-fold cross validation and leave one out cross validation, the ridge model had the lowest RMSE when comparing the various models we tested. We did not expect the cross validation methods to give us drastically conclusions, but the run time required 10-fold and LOOCV is very different, as LOOCV took much longer. The range of RMSE when comparing the models when using LOOCV (given the bias-variance tradeoff with LOOCV generally having less bias) was also much wider than 10-fold. For the purposes of our dataset and object, I believe 10-fold CV is the more appropriate cross-validation method.

Since the ridge regression accounts for correlated predictors and ultimately shrinks correlated predictors together, it makes sense that the ridge model performed the best. A person's health, and in our case, LDL

cholesterol depends on a plethora of health and diet choices; these choices are often correlated with each other.

That being said, all the models still have high RMSE's and low R-squared values. All the regression techniques we set out using are linear regression techniques, allowing for not much model flexibility. This suggests that the models created in this project still can be improved to predict LDL cholesterol with either more flexible polynomial models or techniques we will learn later in the course.

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)
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

variable	definition
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