Analysis

Ngoc Duong

4/2/2020

Introduction

Low-density lipoprotein (LDL) is a group of lipoprotein that transports fat molecules around the body. Levels of low-density lipoprotein cholesterol (LDL-C) has historically been used as future clinical predictors of cardiovascular events, in particular the risk of myocardial infarction. Studies have found that reductions in the number of myocardial infarction and coronary heart diseases were concurrent with a reduction in LDL-C concentration [1]. LDL-C is also associated with other cardiovascular diseases. A 2018 prospective cohort study by Zhao et al. linked levels of oxidized-low density lipoprotein to premature coronary artery disease [5]. A 2018 retrospective cohort study by Pokharel et al. found that small dense LDL-C predicts incident diabetes mellitus [6]. Therefore, understanding what might contribute to high levels of LDL among people living in America is important as it may offer leads to preventive measures for cardiovascular diseases through controlling LDL-C levels. In this project, I aim to use data obtained from NHANES 2015-2016 and apply statistical learning techniques to obtain a LDL-C predictive model with low error as well as understanding the underlying relationships between certain predictors and LDL-C.

Data

Data were obtained from NHANES 2015-2016, a public survey data source provided by the CDC. The response variable is low-density lipoprotein (LDL-C). The design matrix is made up of three smaller components. These represent the following categories: demographics and questionnaire (income- and health-related), diets and nutrients, and laboratory and physical examination. There are 63 total predictor variables, among which 21 are categorical predictors, all from the former two categories (socio-economic determinants of health). The remaining 42 variables are numerical, which are mostly diet, nutrients, lab and physical examination data. The inclusion of these variables in the final data was based on some preliminary knowledge or assumption of their relation to the outcome variable.

When cleaning the data, we made categorical variables factors, and treated answers such as "Don't know" and "Refused to answer" as NA variables. For the sake of simplicity, we did not impute but only looked at complete cases (i.e., people who have data on all 63 predictors of interest), although this means we already made a big assumption that all the missing data were completely at random. The final dataset contains 661 subjects. I then split the data into a training set (80% of original data) and a held-out/validation set (the remaining 20%).

Exploratory data analysis

First, I wanted to look at some key variables in the dataset. A barplot of the response variable - low-density lipoprotein (LDL) shows that there are some outliers in the upper tail but otherwise the distribution looks normal. From Figure 1, we can see that LDL does not seem to differ much by self-perception of health, sex, or race (although other/mixed race tend to have high LDL on average). Other interesting observations could be: immigrants and people who do not have hypertension tend to have slightly higher LDL, while people on special diet have lower LDL on average.

Models

I did a linear model regressing LDL on all predictors as the standard model. Then I looked at regularized linear models to either determine the importance of some predictor variables/select features. The shrinkage methods I looked at were ridge regression, the LASSO, and elastic net (considering four different mixing percentages: 0.2, 0.4, 0.6, and 0.8). I also included a dimension-reduction technique (PCR) to compare their predictive ability.

To select the best tuning parameter/linear combinations of predictors for each model, I used 10-fold repeated cross-validation and Monte-Carlo cross-valuation (Leave-Group-Out-Cross-Validation). The 10-fold cross-validation was repeated 5 times, whereas Monte-Carlo resampled 50 times at p = 0.9 (the train:test split being 9:1) to make the two cross-validation conditions as similar as possible (50 train/test splits each at 9:1 train:test ratio). However, the workings of each CV method is different. Repeated 10-fold CV randomly divides the data into 10 equal-sized blocks of data, then each blocks is left out one by one and serve to test the prediction ability of the trained model (trained by the remaining 9 blocks). This process is repeated a specified number of times. Monte-Carlo CV randomly samples from the data without replacement until there are enough observations (by pre-specified percentage) in the training set.

- 10-fold CV (repeated 5 times) We can see that the elastic net tuned by repeated 10-fold CV has very similar mean cross-validated RMSE as the LASSO (36.08 vs. 36.15). The validation RMSE on the (independent) held-out data was also very similar 1245.50 for LASSO and 1244.05 for elastic net. Ridge regression's performance is also very similar, followed by PCR and linear model has the highest CV RMSE.
- Monte Carlo CV (repeated 50 times, train/test split = 9/1)

We can see that the elastic net tuned by MCCV has slightly lower mean and median CV RMSE than the LASSO and Ridge regression (36.05 vs. 36.09 vs. 36.11, and 35.83 vs. 35.9 vs. 36.16, respectively).

Looking at Figure 4, we can see both cross-validation methods give relatively consistent results for the "optimal" regularized linear model. We can also see that elastic net model tuned by 10-fold CV method has slightly smaller median CV RMSE and more consistent performance in getting lower RMSE (signified by fatter bottom part of violin plot). In addition, the test error (MSE) on the held-out data for this elastic net model tuned by 10-fold CV is 1244, slightly smaller than that tuned by Monte-Carlo CV (validation MSE = 1246). Therefore, I choose to go with the elastic net model tuned by 10-fold repeated CV for variable selection (mixing percentage alpha = 0.2, and selected lambda = 8.216). That said, the models from both CV methods are very comparable and Monte-Carlo CV even selects elastic net models with a lower range of CV RMSE than 10-fold CV.

Caveats of regularized linear regression

The LASSO/elastic net model was not flexible enough to capture the underlying truth since they still assume linear relationships between predictors and response. From the scatterplot, there seem to be some non-linear trends. Therefore, this leads to the use of GAM and MARS model to explore the non-linear relationships between certain predictors and the response variable.

GAM

Using the function `varImp` on the elastic net model, I obtained 17 most important variables (both categorical and numerical). These variables include: age (ridageyr), gender (riagendr), being told to have diabetes by doctor (diq010), daily alcohol consumption (dr1talco), diastolic blood pressure (bpxdi), lymphocyte percentage (lbxlypct), neutrophil percentage (lbxnepct), high-sensitivity C-reactive protein level (lbxhscrp), globulin level (lbxgb), alkaline phosphate level (lbxsapsi), being told to have hypertension by doctor (bpq020), hemoglobin count (lbxhgb), hematocrit level (lbxhct), self-perceived healthiness of diet (dbq700), past-year household food security (fsdhh), uric acid level (lbxsua), and lactate dehydrogenase ldc (U/L). I then used

pairwise-scatter plots (Figure 4) to explore the relationship between the selected numerical variables and the response variable.

For GAM, I fitted two models, one excluded variables that seemed not strongly associated with the response variable (determined from box plots in EDA step), and the other one was the full model. The ANOVA F-test was conducted to determine which model is better. Since F-test p-value = $0.13 > 0.05$, I decided to go with the more parsimonious GAM, whose expression is below:

$\hat{LDL} = 89.6 - 0.47 \times age + 18.5 \times I(diabetes = No) + 14.32 \times I(diabetes = Borderline) - 0.2 \times alcohol + 9.87 \times Glucose + 7.23 \times I(Hypertension = No) + 11.04 \times I(diet = verygood) + 7.57 \times I(diet = good) + 9.04 \times I(diet = fair) + 16.62 \times I(diet = poor) + \hat{f}_1(HSCRP) + \hat{f}_2(Hematocrit) + \hat{f}_3(AlkalinePhosphate) + \hat{f}_4(LDH)$

The individual relationships between the predictors and LDL can be observed in figure 5. A few observations are; holding other covariates fixed, predicted LDL tends to: 1. fluctuate a bit across levels of high-sensitivity C-reactive protein (HSCRP), hitting the lowest point when HSCRP is around 20 mg/L, and highest point when HSCRP is 40 mg/L. 2. increase and peaks when alkaline phosphate reaches 75U/L and the effect decreases slightly beyond that point.

Multivariate adaptive regression spline (MARS)

I let the number of degrees vary from 1 to 3 to explore any potential interactions between the hinge functions. Number of prunes varies from 1 to 25 to give a wide enough range for the number of terms to search for lowest cross-validated RMSE.

MARS selects the number of basis functions h_i as 6 that minimize the cross-validated RMSE. The product degree selected is 1 so there is no interaction between the hinge functions. I then created partial dependence plots, which are used to examine the marginal effect of the predictors on the response. MARS model selected HSCRP, Glucose level, and lymphocyte percentage.

Looking at the partial dependence plots in figure 6, we can see the non-linear marginal effect of each predictor on the predicted LDL. Both GAM and MARS suggested there was a non-linear relationship between high-sensitivity C-reactive protein/HSCRP (an inflammation signal) and LDL-C, with the LDL-C maximum being in the middle range of HSCRP. However, due to the difference between hinge function and smoothing spline, the PDP describes a simpler relationship between HSCRP and LDL than the smoothing spline in GAM.

The validation RMSE of the MARS model is 1345.45. The GAM model's validation RMSE is 1256, a little higher than that of tuned-LASSO (1244) and tuned-elastic net (1246), despite having fewer variables, offering more flexibility and taking into account non-linearity between the predictors and outcome.

Conclusion

Given the distribution of out response variable (mean = 114 and median = 112), the best cross-validated RMSE (by the elastic net tuned by repeated 10-fold CV) is at about 30% of the mean value of the response variable, which is not small, but it does not signify extreme variance either. That said, the prediction ability given the selected variables is limited at best. Small cross-validated R-squared value, averaged at 7.5% signifies there is still a large share of variance in LDL not captured by the regularized model. This is likely due to the selection of variables in the data collection process. We did not dig deep into the pre-exisitng literature but rather assumed associations between some predictors and response. A better job at selecting predictors is desired. Tuned GAM and MARS models, despite capturing non-linear associations, did not perform as well with respect to validation RMSE. Further literature search and more domain knowledge is needed in deciphering the truthfulness of the non-linearity captured by these models. All in all, the models offer some insights into the relationships between certain variables and offer some predictive ability. However, more work needs done in order to improve the predictive ability of these models.

Figure 1. Distribution of LDL across some demographic and socioeconomic determinants of health

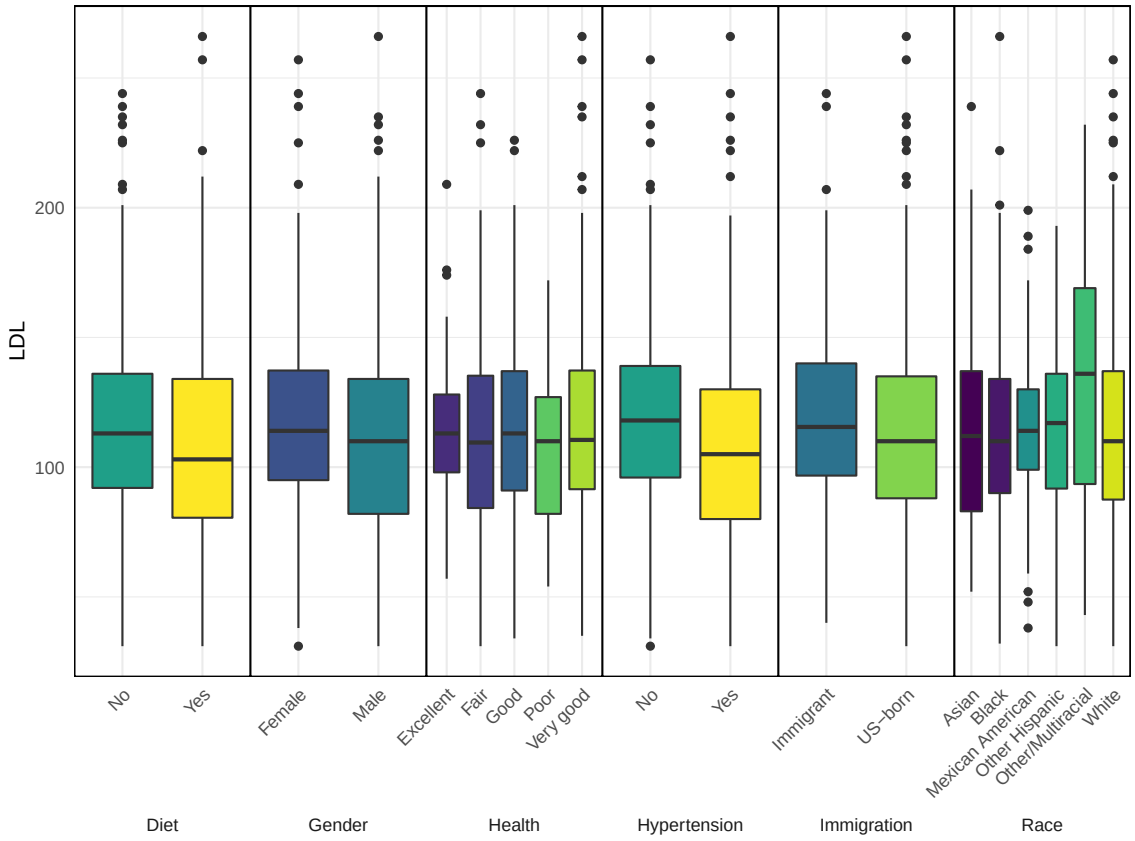

Figure 2. 10-fold repeated CV hyperparameter tuning plots

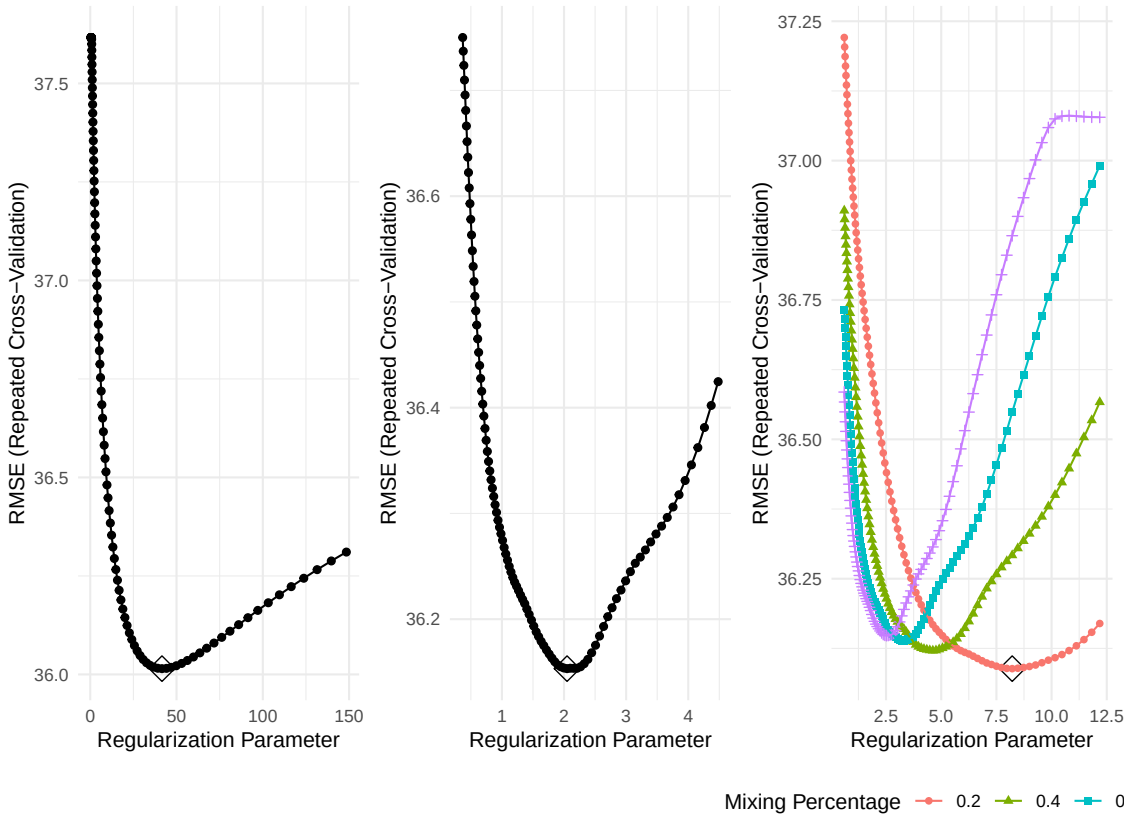

Figure 3. Distribution of Cross−validated RMSE across models by two CV methods

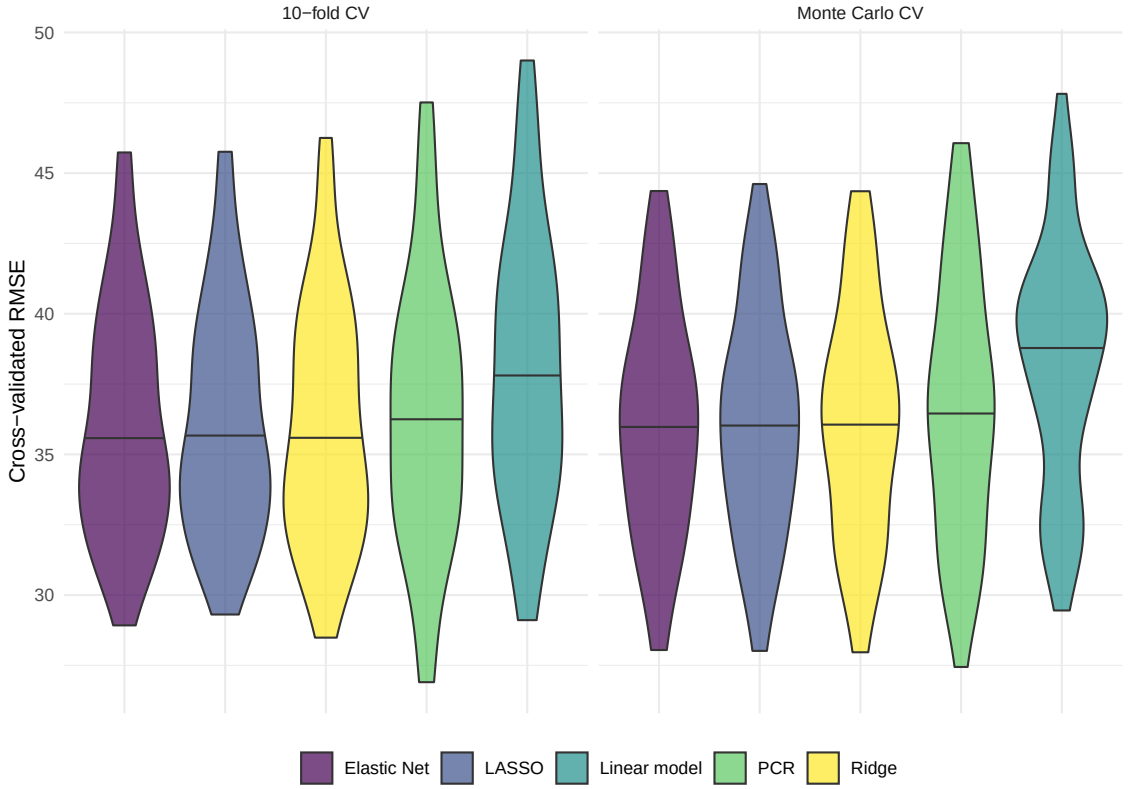

Figure 4. Pairwise scatterplots between important numerical predictors and LDL-C

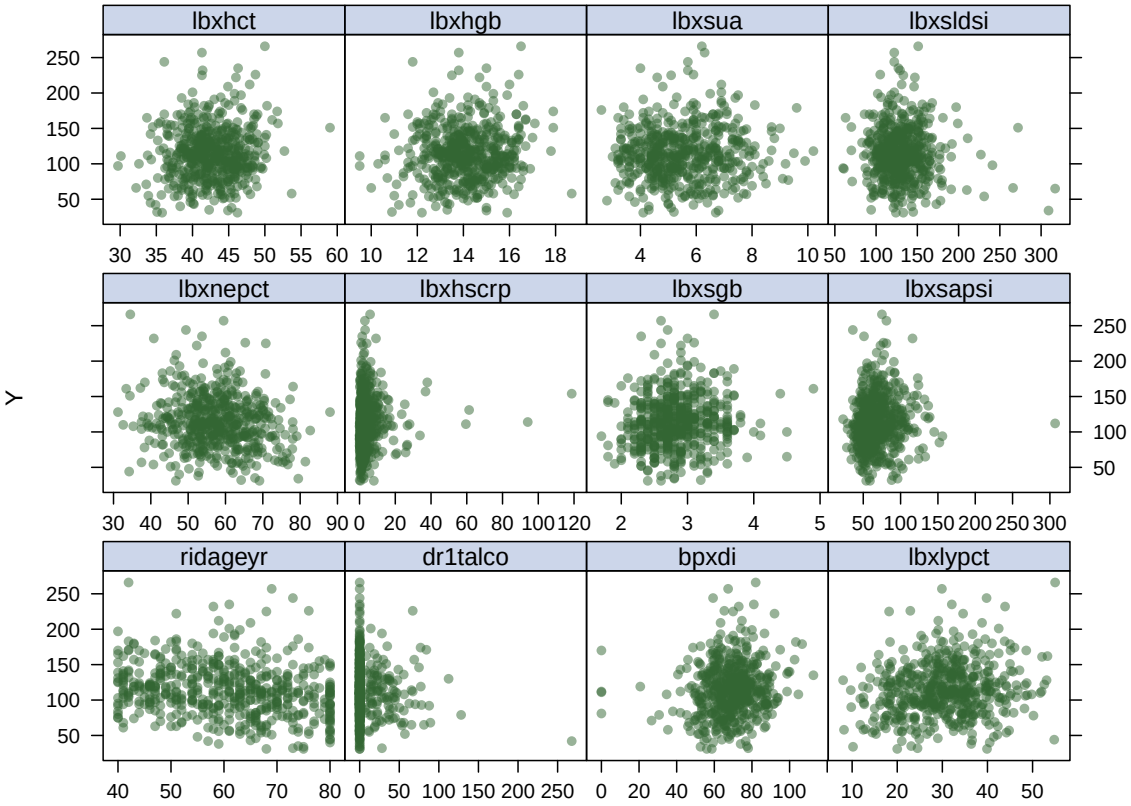

Figure 5. Visualizations of smoothing splines in GAM

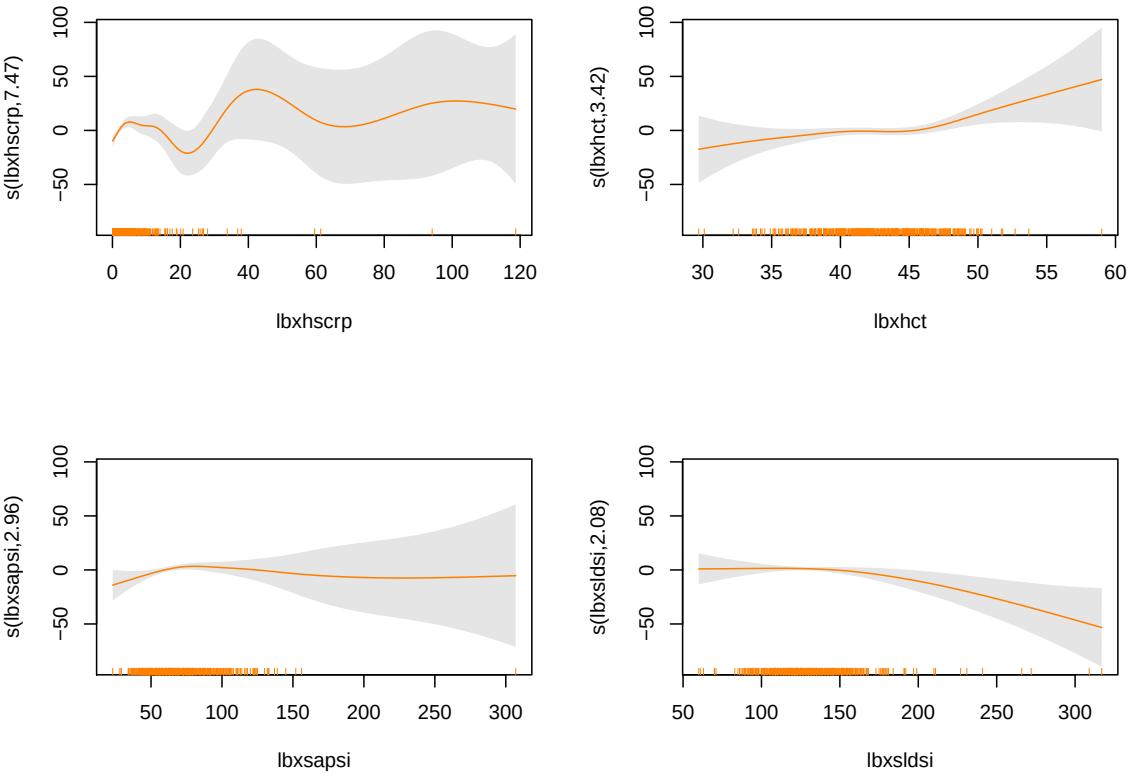

Figure 6. Partial Dependence Plots