Project Step 4

Minu Pabbathi, Ziqian Zhao, Paul Zhang

2024-06-11

Introduction

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains information about patients and risk factors for heart disease as well as whether or not they eventually end up developing heart disease. For the sake of convenience, we have randomly chosen 500 observations from the original data set. In this step, we are going to execute LASSO and ridge regression to find alternative models, and finally use logistic regression to determine the probability of a patient having heart disease.

Shrinkage Methods

LASSO Regression

LASSO regression is a type of linear regression that is used for shrinkage and variable selection. A penalty term equivalent to the absolute value of the coefficients times a tuning parameter λ is added to discourage large values of β. As a result, some coefficients are shrunk to 0, making LASSO a useful tool for variable selection. Thus, finding an optimal λ is crucial to maintaining a balance between shrinkage and a usable model. Additionally, by introducing bias through the penalty term, LASSO regression can decrease the variance of the coefficient estimates, leading to a lower out-of-sample MSE and better generalization.

Here, λ is selected by testing a grid of potential values and then conducting 10-fold cross validation to determine which value minimizes the test MSE. Fig. 1 plots the test MSE against potential λ values

```
## [1] "Best lambda: 0.310161341566227"
```

Final LASSO Model

Using the calculated optimal λ, we fit the final model:

$$\begin{aligned}\text{max heart rate}_i = 143.61 &- 6.99 \cdot age_i + 8.89 \cdot 1\{cp_i = 1\} + 3.87 \cdot 1\{cp_i = 2\} + 13.43 \cdot 1\{cp_i = 3\} + 1.67 \cdot trestbps_i \\ &- 2.14 \cdot 1\{restecg_i = 1\} - 7.70 \cdot 1\{restecg_i = 2\} - 8.47 \cdot 1\{exang_i = 1\} \\ &- 0.81 \cdot oldpeak_i - 1.78 \cdot 1\{slope_i = 1\} + 9.98 \cdot 1\{slope_i = 2\} - 0.36 \cdot ca_i \\ &+ 0.22 \cdot chol_i + 0.59 \cdot thal_i + 6.29 \cdot 1\{target_i = 1\}\end{aligned}$$

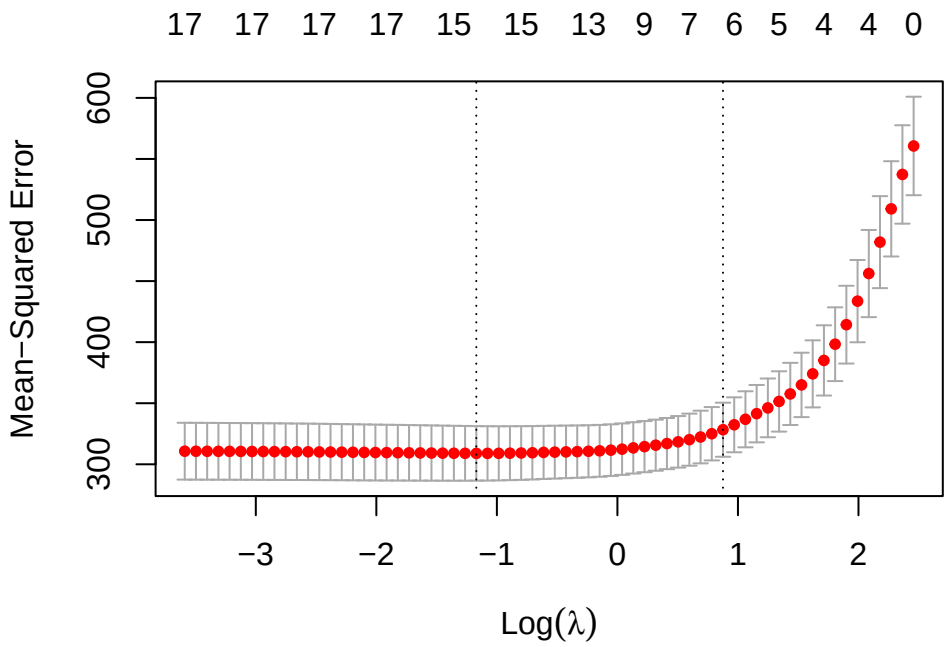

Figure 1: MSE vs. Lambda

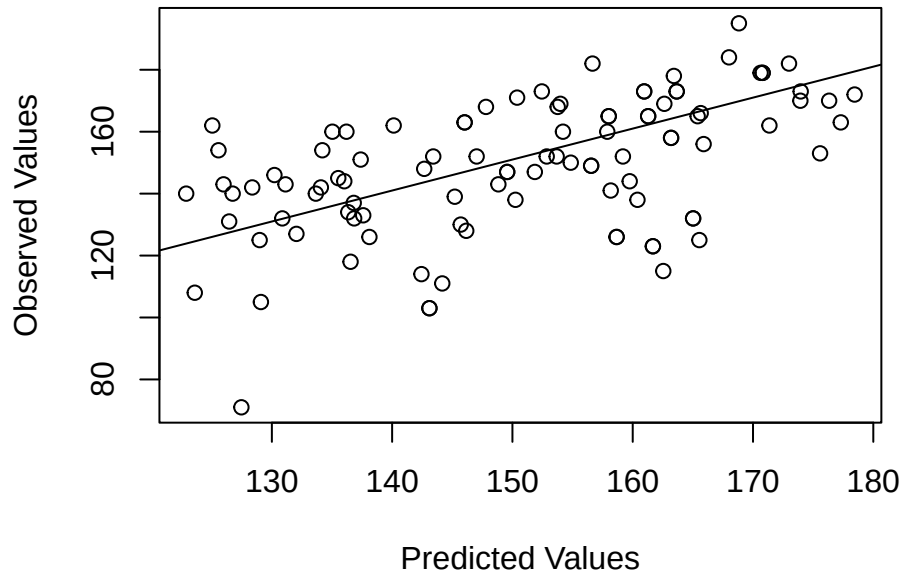

Figure 2: Plotting predicted values using LASSO regression against observed values

Predictions

Using the test data, the predicted values plotted against the observed values follow a near linear trend, indicating the model is helpful for prediction.

Ridge Regression

Ridge model is a method of estimating the coefficients of multi-variable linear regression.Different from Ordinary Least Square(OLS) method, Ridge regression, including a penalty term, provides a trade-off between variance and bias, which are the two criteria for MSE what we want to minimize.

Here, we fit a Ridge regression with the optimal lambda found through cross validation

```
## [1] "By cross validation between 10 folds, the optimal lambda is 3.2493"
```

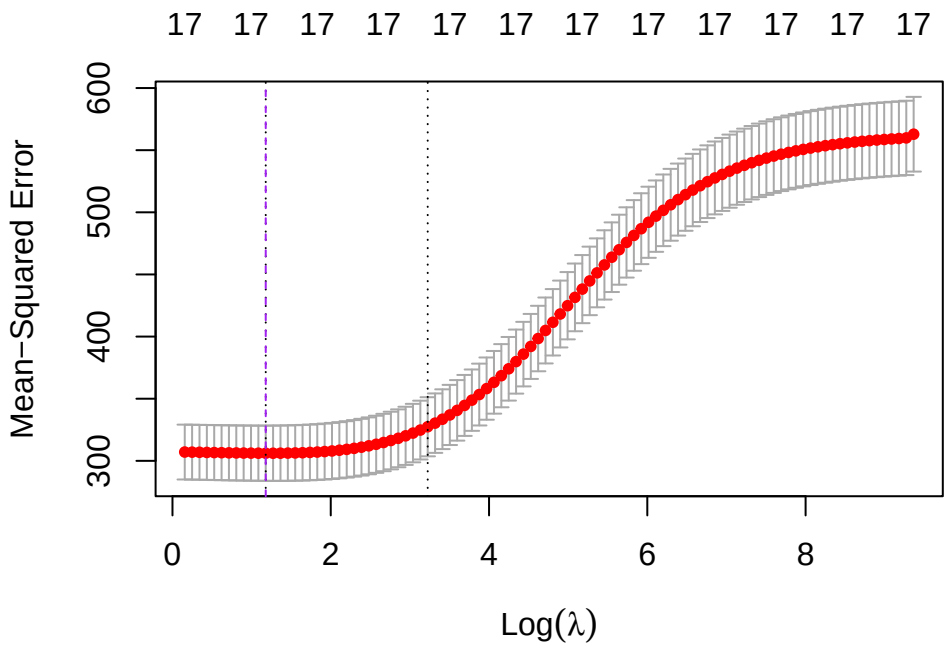

Figure 3: MSE vs. Lambda

Fig. 3 visualized how MSE change with the regularization parameter λ. As our goal is to minimize MSE, we locate the $log(\lambda)$ correspond to the smallest MSE, which shown by the purple line and select the lambda as the regularization factor.

We then fit the model with the penalty term and obtain the estimated coefficients. In terms of variable selection, more variables are considered significant than forward step-wise selection (the method used previous). Specifically, `oldpeak`,`ca`, and `thal` are also considered as significant (the |coefficient| > 0.5). The final model is

$$\begin{aligned}\text{max heart rate}_i = 146.21 &- 6.34 \cdot age_i + 8.66 \cdot 1\{cp_i = 1\} + 4.03 \cdot 1\{cp_i = 2\} + 13.03 \cdot 1\{cp_i = 3\} + 1.60 \cdot trestbps_i \\ &- 2.22 \cdot 1\{restecg_i = 1\} - 10.13 \cdot 1\{restecg_i = 2\} - 8.12 \cdot 1\{exang_i = 1\} \\ &- 1.39 \cdot oldpeak_i - 4.33 \cdot 1\{slope_i = 1\} + 7.37 \cdot 1\{slope_i = 2\} - 0.68 \cdot ca_i \\ &+ 0.83 \cdot thal_i + 6.12 \cdot 1\{target_i = 1\}\end{aligned}$$

Making Predictions

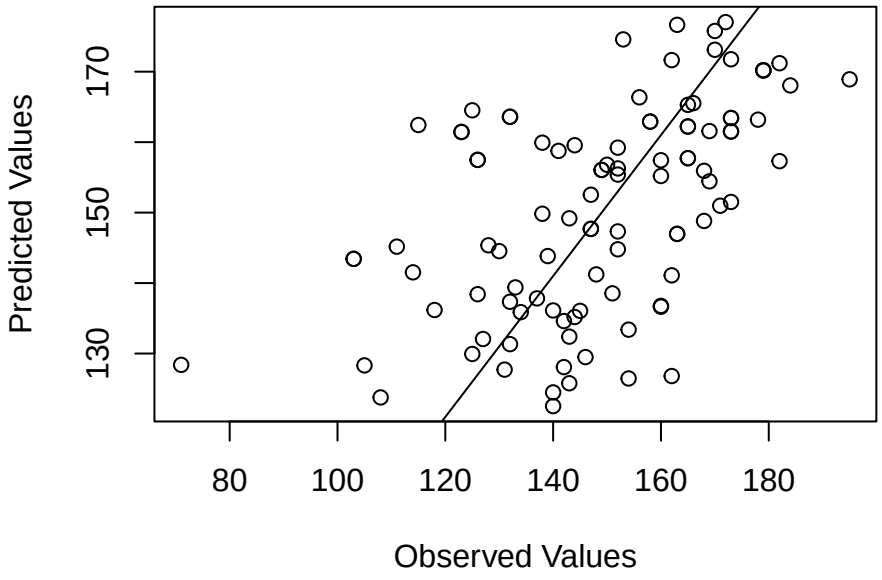

Figure 4: Ridge Predicted Values vs. Observed Values

Fig. 4 is predicted value vs. observed value, which visualizes the accuracy of the prediction with ridge regression. The dots, as expected, are spread around y=x, which indicates that the ridge regression, though with some outliers, provided a relatively good prediction.

Making Combined Plot

Model Comparison

The plot shows that the predictions between three models do not differentiate much. Specifically, the observations of three models, represented by three different colors, located close to each other along the line $y = x$.

Specifically, dots of LASSO (red) and Ridge(blue) are closely located together and the small deviation could be caused by differences between the penalty term (i.e. λ).

The dots of OLS(green) deviated relatively more from the others which could be a result of different predictor selection. Specifically, lasso and ridge regression consider three more variables significant. However, the deviation is not significant, so it is reasonable to conclude that predictor are consistently selected.

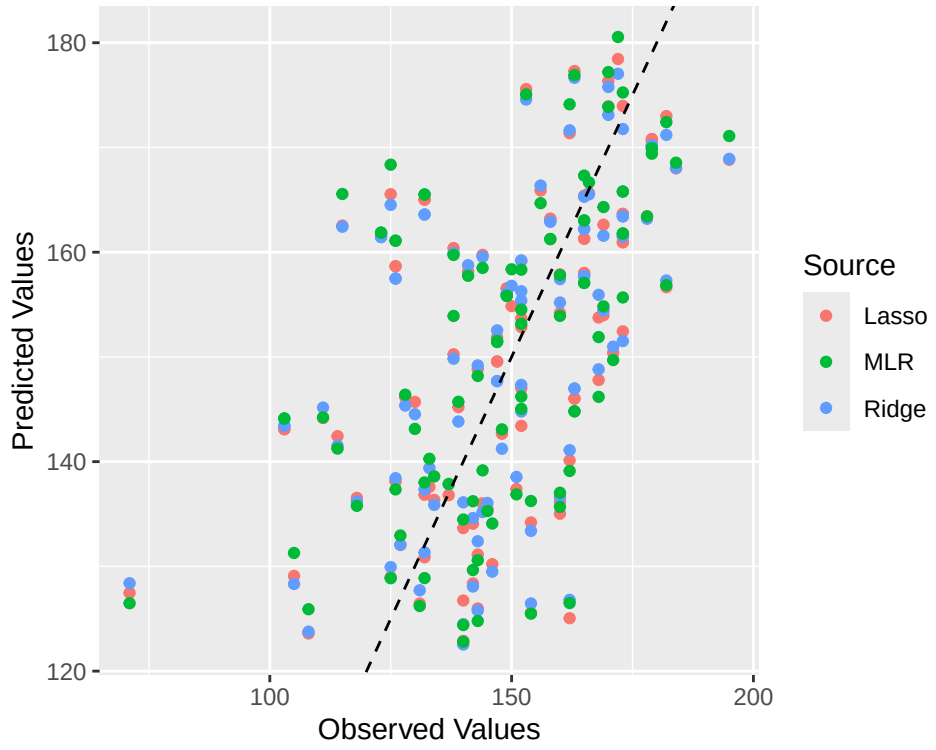

Figure 5: Combined Plot

Predictors	Lasso	Ridge	MLR
Age	-6.99	-6.34	0
Sex = 1	0	-0.29	0
Chest Pain type =1	8.89	8.66	10.56
Chest Pain type=2	3.87	4.03	5.17
Chest Pain type=3	13.43	13.03	15.11
Rest bps	1.67	1.60	0.12
Serum Cholesterol	0.22	0.31	0
Fasting blood sugar =1	0	0.31	0
Rest ecg=1	-2.14	-2.22	-3.18
Rest ecg=2	-7.69	-10.12	-12.76
Exercise angina=1	-8.47	-8.12	-8.35
oldpeak	-0.81	-1.39	0
ST slpoe =1	-7.18	-4.33	-1.16
ST slope =2	9.98	7.37	11.27
Target =1	6.29	6.12	6.43

From the table above, which shows the estimated coefficients from three regression methods, it further convinced that the coefficients are inherently and similarly estimated.

Conclusion

After applying RR and LASSO to our model, we have found similarly selected predictors, and also the predictions are generally promising compared to observed data. When compared with the model we had

from MLR, several predictors are reconsidered and we still have relatively consistent predicted result. It is also noteworthy that when looking at the superimposed graph, we had outliners that distributing similarly across three methods, which indicating that further steps are needed to take care of them.

Innovation

Logistic regression is used to predict the outcome of a binary variable, for example 0 or 1. Suppose the probability of 1 occurs is p and the probability of 1 does not occur is 1-p, then the ratio of 1 occurring vs. 1 not-occurring is simply $\frac{1}{1-p}$.

Since we then need a linear regression to estimate the coefficient, we then apply log map the probability to real number for linear regression, that is

$$Logit(p) = log(\frac{p}{1-p})$$

Apply $logit(p)$ as the new response and fit a linear model with the predictor variables.

$$Logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Then we inverse the logit to get p, namely,

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

Technical Conditions

1. The response variable is binary. This fits our model that the response is the target variable, which indicates whether the patient have heart disease.
2. The model should have minimal multicollinearity. The LASSO regression is removing predictors that are having high correlations.
3. Sufficient sample size and independent observation. Our dataset is large and the patients conditions are independent.
4. There shouldn't be prefect separation. We don't have some predictors that can precisely predict the outcome.
5. There shouldn't be highly influential outliners. For this one we didn't particularly remove the outliners.

Method Justification

To explore further models, we decided to use logistic regression since we were interested in seeing how the variables can be used to predict heart disease, or the `target` variable. Since heart disease is a binary variable, OLS, ridge, and LASSO cannot be used since the response variable in these cases must be continuous.

Variable Selection

First, we used LASSO to determine which variables to include in the model.

```
## [1] "Best lambda: 0.00487039373579772"
```

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept)  2.46342446
## (Intercept)  .
## age         -0.01001151
## sex1        -1.66750818
## cp1          0.62465074
## cp2          2.27538242
## cp3          2.67834570
## trestbps    -0.01441694
## chol         .
## fbs1        -0.28912612
## restecg1     0.42219897
## restecg2     .
## thalach      0.01867867
## exang1      -0.83167092
## slope1       .
## slope2       1.78187914
## ca          -0.95402650
## thal        -1.10788050
```

Fitting Logistic Model

Based on the results from the LASSO regression, we fit a logistic model using the non-zero coefficients.

```
##
## Call:
## glm(formula = target ~ age + sex + cp + trestbps + fbs + restecg +
##     thalach + exang + slope + ca + thal, family = binomial, data = train_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.215911   2.691304   1.195  0.23212
## age         -0.013983   0.023098  -0.605  0.54493
## sex1        -2.011873   0.445663  -4.514 6.35e-06 ***
## cp1          0.838787   0.482932   1.737  0.08241 .
## cp2          2.817552   0.506933   5.558 2.73e-08 ***
## cp3          3.350009   0.717658   4.668 3.04e-06 ***
## trestbps    -0.017784   0.009636  -1.846  0.06496 .
## fbs1        -0.539926   0.491566  -1.098  0.27204
## restecg1     0.540507   0.351903   1.536  0.12455
## restecg2    -0.380370   2.182778  -0.174  0.86166
## thalach      0.019662   0.011372   1.729  0.08381 .
## exang1      -0.853594   0.419316  -2.036  0.04178 *
## slope1       0.173169   0.746474   0.232  0.81655
## slope2       2.248040   0.765454   2.937  0.00332 **
```

```
## ca          -1.116049   0.193442  -5.769 7.95e-09 ***
## thal        -1.307612   0.284819  -4.591 4.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 552.83  on 399  degrees of freedom
## Residual deviance: 235.42  on 384  degrees of freedom
## AIC: 267.42
##
## Number of Fisher Scoring iterations: 6
```

ROC Curve

A tool to visualize logistic regression is a ROC (Receiver Operating Characteristic) curve, in which the true positive rate (sensitivity) is plotted against the false positive rate (1 - specificity) at different threshold values. The closer the ROC curve is to the top-left corner, the better the model is at distinguishing between positive and negative classes. A model with no discriminative power will have an ROC curve along the diagonal line, essentially equivalent to random guessing. As can be seen in Fig. 6, the ROC curve is relatively close to the top-left corner, indicating that the model has high sensitivity and high specificity.

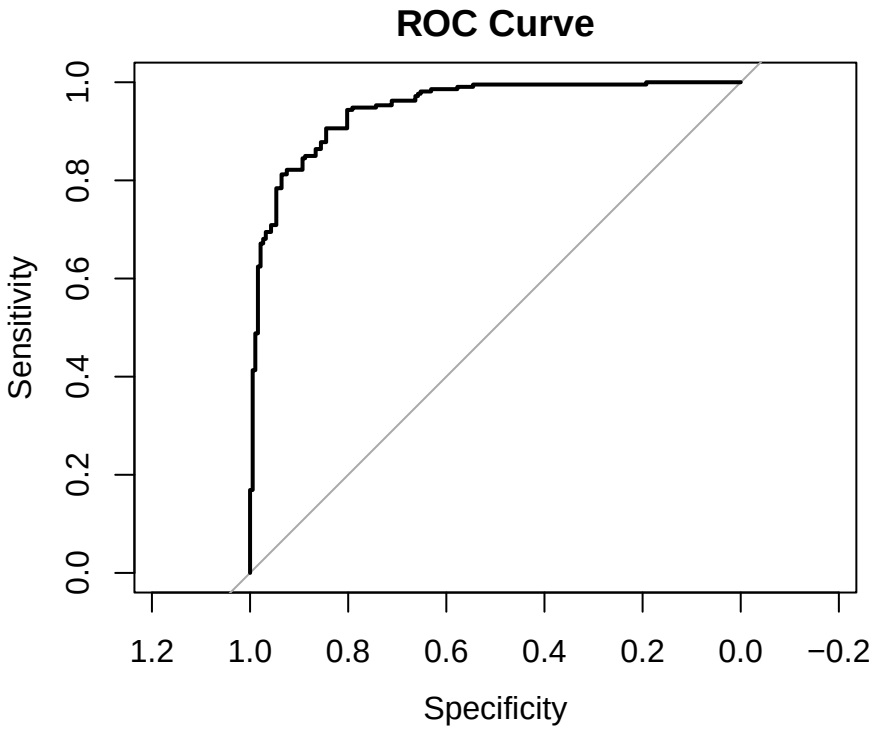

Figure 6: ROC Curve for logisitic model

AUC

The AUC, or area under the curve, is a measure of how well the model is able to discriminate between positive and negative cases. The AUC ranges from 0 to 1, with the following interpretations:

- < 0.5: No discriminative power (equivalent to random guessing).
- 0.5 - 0.7: Poor discrimination.
- 0.7 - 0.8: Acceptable discrimination.
- 0.8 - 0.9: Excellent discrimination.
- > 0.9: Outstanding discrimination.

```
## [1] "AUC: 0.947176822073259"
```

Since the AUC is above 0.9, we can conclude that our model is highly accurate at distinguishing between cases.

Optimal Threshold Value

Based on the ROC curve, we can determine the optimal threshold value, which is the value that maximizes the distance from the diagonal line (representing random chance) or maximizes the area under the ROC curve (AUC). If the predicted probability is less than the threshold, then the response can be classified as 0, or no heart disease. If the predicted probability is greater than threshold, then the response can be classified as 1, or having heart disease.

```
## [1] "Optimal threshold value: 0.524617014252765"
```