Tashdeed Kader Faruk & Ryan Burden

Professor Eran Mukamel

COGS 109

06/04/21

Predicting Presence of Cardiovascular Disease and Ranking Predictive Risk Factors

1. Introduction and Hypothesis

Introduction:

Cardiovascular disease (often shortened to CVD) is a classification for diseases that involve the heart and/or blood vessels. CVD primarily concerns coronary artery diseases like angina (chest pain caused by inadequate blood supply to the heart), but CVD also encompasses strokes, heart failure, and abnormal heart rhythms, among others (*CDC - Cardiovascular Disease*). In 2015 alone, 17.9 million deaths were caused by some form of cardiovascular disease (*CDC - Cardiovascular Disease*). These 17.9 million deaths accounted for 32.1% of total deaths worldwide that year making CVDs the leading cause of death worldwide. By comparison, CVDs were the cause of 12.3 million deaths worldwide (25.8% of all deaths) in 1990 (*CDC - Cardiovascular Disease*). Therefore, the severity of cardiovascular diseases is understood not only through it being the leading cause of death for some time, but by their share as a cause of total deaths worldwide growing rapidly each year.

The Center for Disease Control (CDC) estimates that up to 90% of CVDs may be preventable given appropriate care and healthy living habits to improve relevant risk factors (*CDC - Cardiovascular Disease*). Although the causes of CVDs vary depending on the specific disease, the most common risk factors that are studied in relation to cardiovascular disease include a subject's age, sex, resting blood pressure, fasting blood sugar levels, and maximum heart rate achieved, among others.

For this report, data from a public health dataset on heart disease was analyzed. The data set comprises data from four separate datasets: Cleveland, Hungary, Switzerland, and Long Beach (*Lapp, David*). The dataset consists of 13 predictors (many of which are similar to the aforementioned risk factors) – *age*, *sex*, *chest pain type* (from 0 to 3), *resting blood pressure* (in mmHg), *serum cholesterol* (in mg/dl), *resting electrocardiographic results* (from 0 to 2), *maximum heart rate achieved* (in bpm), *exercise induced angina* (binary: 0 or 1), *oldpeak = ST depression induced by exercise relative to rest*, *the* *slope of the peak exercise ST segment*, *number of majors vessels observable by fluoroscopy* (0 to 3), and *condition of thalassemia* (0 = normal; 1 = fixed defect; 2 = reversable defect) (*Lapp, David*). The dataset then has 1 outcome variable, *target*, which is a binary value that indicates the presence of heart disease in the patient (0 = no disease present; 1 = disease present) (*Lapp, David*). The dataset held data for the 13 predictors and 1 outcome variable for 1024 patients (n = 1024 number of observations).

Hypothesis:

The most basic logistic regression model that can be fit to the dataset is a Multiple Linear Regression model. This model takes into consideration all 13 of the predictors and an intercept to predict the target outcome variable for any given patient. Although predictions from a multiple linear regression model would ensure the lowest possible *Mean Squared Error* (MSE) compared to other available models for this dataset, a model dependent on 13 different independent variables is considered too complex. This is because a patient trying to use the model to predict their own chances of experiencing a heart disease would need 13 separate values, which can be a burden on them in terms of costs, time, and energy. Our hypothesis is: we can develop a sparse model, using *Forward Subset Selection* or *LASSO*, that is more interpretable and less complex than the *Multiple Linear Regression* model and the increase in average *Mean Squared Error* (MSE) will not be greater than 0.05.

2. Methods - Cross Validation

The decisions made during the model selection phase of this analysis are supported by *k-Fold Cross-Validation*, with k = 10 folds. The choice of k in k-Fold cross validation impacts the performance of the process. In particular, the number of folds is directly related to the proportion of total data used for training the model. The size of the training set in cross validation dictates the amount of bias and variance that the model will exhibit. When the training set consists of almost the entire dataset, such as in *LOOCV*, the cross-validated *MSE* of the model will increase in variance, which is an indication of overfitting. This is because the model is being trained by the majority of the dataset, and only being tested by an extremely small amount of randomly sampled observations. Conversely, if the training dataset is too small, the model will have high bias due to being trained on a small subset of randomly sampled observations, and tested against a large amount of the entire dataset. This relationship is important to consider when choosing the value of k in k-Fold validation. The value of k = 10 is a generally accepted and commonly used choice in data analysis, and does not have a significant impact on classification error *(Marcot & Hanea, 2020).*

This process was performed using the *KFold* library from *scikit-learn*. First, the library shuffles the observations within the dataset to take a random order. Then the observations are separated into 10 subsets, or ‘folds’, of equal size. These folds are then iterated over, with each fold serving as the validation dataset once, while the remaining nine folds are used as training data. The test-set error is calculated by fitting the model with the training dataset. Each validation-set observation is used by the model to compute an estimate for the dependent variable. The test set error is generated by comparing these estimates to the true values of the dependent variable in the validation-set observations. This iteration continues until each fold has been used as validation data, and 10 test-set mean squared errors have been collected. The model’s performance is evaluated by the average of the test-set mean squared errors across all 10 folds. Cross-validated MSE is an important factor when selecting an appropriate model, but not the only factor to be considered.

3. Methods - Model Analysis

The objective of this project is to develop a sparse model that is more interpretable and less complex than the *Multiple Linear Regression* model that incorporates all 13 predictors. We hypothesized that the MSE of our sparse model will not be greater than 0.05 from the MSE of our *Multiple Linear Regression* model. A *sparse* model is one where only a few parameters are non-zero, thus improving the interpretability and reducing the complexity of the model by reducing the number of independent variables. *Forward Subset Selection* and *Least Absolute Shrinkages and Selection Operator* (*LASSO*) are both examples of sparse models because they remove “unneeded” predictors (i.e., predictors that are less consequential to the outcome compared to other predictors). LASSO is similar to subset selection, but it uses a continuous regularization parameter (lambda) rather than trying out all of the subsets.

This project will find two models, one using *Forward Subset Selection* and one using *LASSO*, and the effectiveness of the two models in predicting heart disease in patients will be evaluated and compared to each other as well as to the more complex multiple linear regression model. In our project we are focusing primarily on prediction, as we evaluate the effectiveness of our models by calculating the Mean Squared Error in predicting the *target* outcome variable. However, the two models we compare are being developed using inference, as the predictors that are less consequential to the outcome compared to other predictors are removed in both *Forward Subset Selection* and *LASSO* models. Therefore, our project focuses on both prediction and inference. The two models we have chosen are appropriate for the hypothesis we are testing because they both produce sparse models.

4. Results - Model Selection

Model Selection - Forward Subset Selection:

Forward Subset Selection (*FSS*) is a modeling technique that is more efficient than the closely related *Best Subset Selection*. In FSS, the formula used to establish a relationship between predictors and

the outcome variable is built upon in each iteration as a new predictor is added in each iteration. FSS begins with the null model, which contains no predictors and results in modeling the following formula: *target* ~ 1 (where, *target* = outcome variable; 1= the intercept). Then in the first iteration we model "*target* ~ 1+�� ", where we loop through each of the predictors and store the resulting Mean

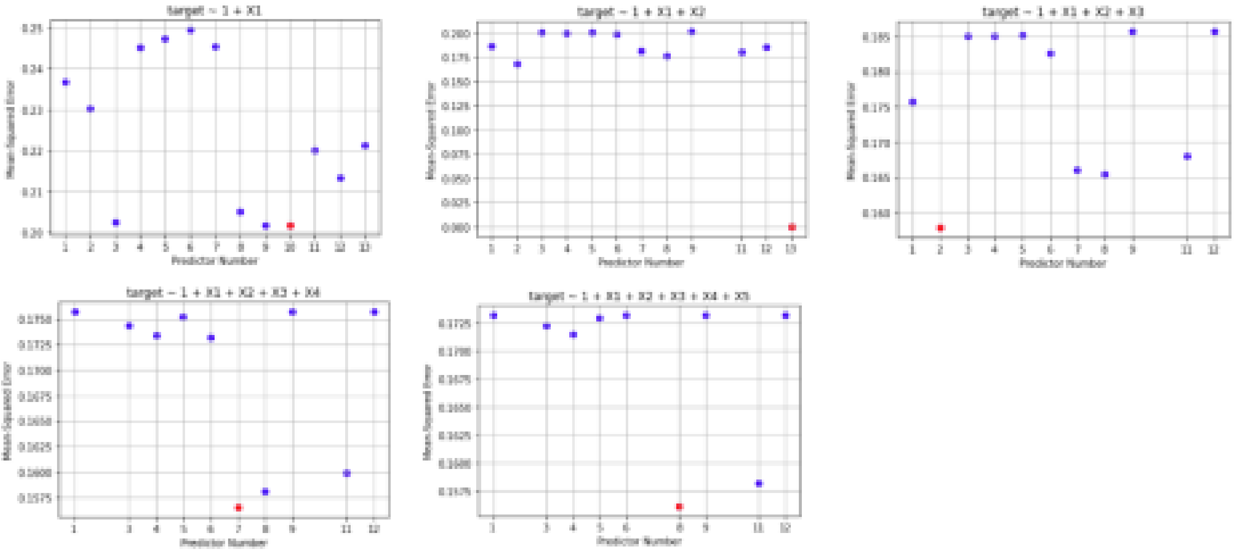
Squared 1

Error (MSE) when comparing the predictions made from that model to the true *target* values. We choose the predictor that results in the lowest Mean Squared Error to be our desired X1. Then in the second iteration we model "*target* ~ 1+�� + ", where we loop through each of the remaining predictors after

$_1$��$_2$

step 1 and use their resulting MSE's as our desired X2. This process is then repeated until a desired formula limit is reached.

The result of *Forward Subset Selection* for this project is shown below:



In the graphs above, the x-axis represents the predictor numbers, the y-axis shows the MSE for the model, and the lowest MSE model is highlighted in red. It is important to note that we are not performing K-Fold Cross Validation yet. Our models for the data above were trained using the entire dataset and then tested on the same dataset again; thus overfitting may be present. In fact, overfitting can be seen in the graph for the 2nd iteration as the MSE for "target ~ 1 + predictor 10 + predictor 13" gave us an MSE of 0.000. FSS is performed till a desired MSE is obtained and/or a desired number of predictors. Ignoring the overfitted model in the 2nd iteration, we decided to perform FSS until the change in MSE became negligible. As can seen in MSE between iterations 4 and 5, the MSE remains just below 0.1575 in both iterations and therefore, it was our decision to stop after 5 predictors were chosen from FSS. The resulting model formula was:

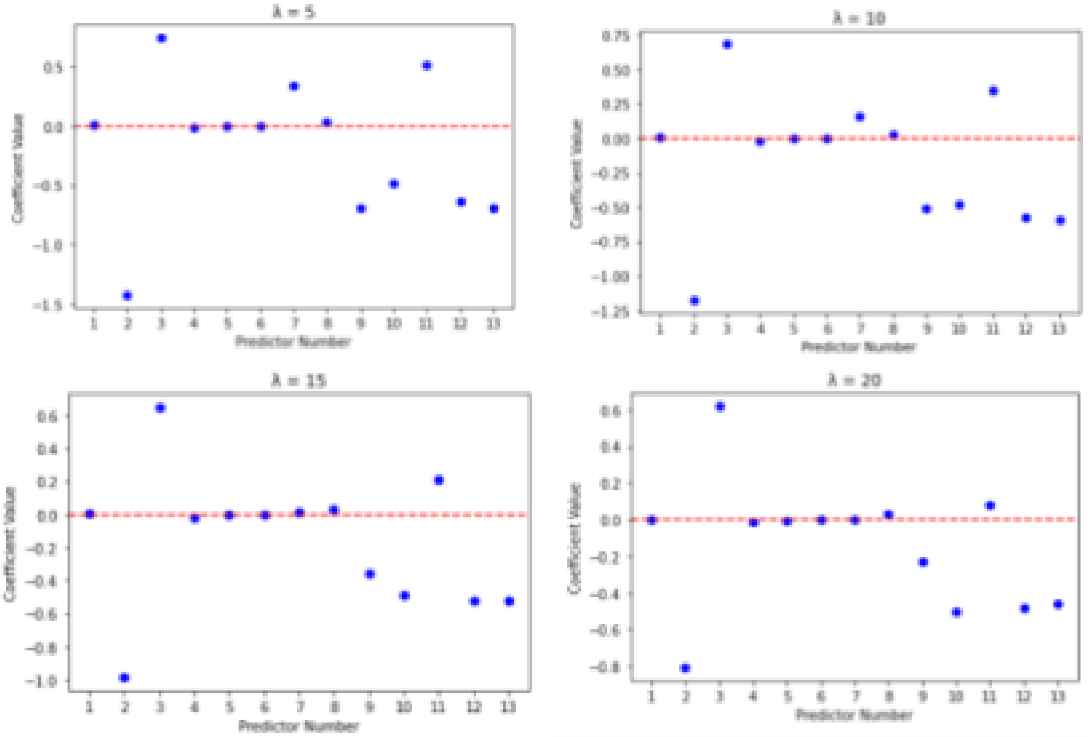
$$target \sim 1 + oldpeak + thal + sex + restecg + thalach$$

As seen, this model is sparse because it removes 8 predictors and this model is much more interpretable than the multiple linear regression model with 13 predictors.

Model Selection - LASSO:

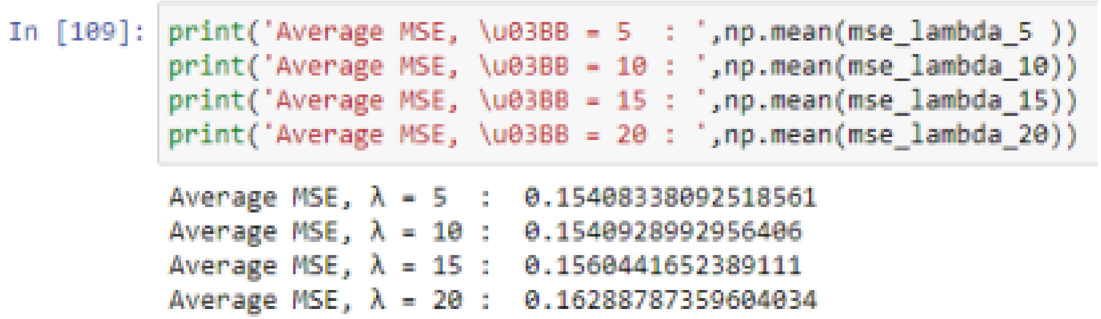
An alternative method of performing regularization to the parameters of a model is by using *LASSO* regression. *LASSO* is a sparse regression model in which all of the predictors are considered, and some are brought to zero. This increases the interpretability of the model by reducing the amount of variables that predict the output. *LASSO regression* differs from other forms of regularization by the *Penalty* term within its *Cost* function. *LASSO*'s *Cost* function includes the *L1 Penalty*, which is equal to the sum of the absolute value of the model's coefficients. The *L1 penalty* is weighted by a measure of regularization strength, *λ (lambda).* In practice, *LASSO* seeks to minimize the *L1 penalty*, and thus minimize the *Cost* function. This results in the shrinkage of the coefficients of the predictors, with some of the predictors reaching zero. The magnitude of the shrinkage, and how many coefficients are brought to zero, increases as λ increases. One method of performing this regression in Python is to employ *scikit-learn's LogisticRegression* library. This library supports the use of *LASSO* by setting the penalty parameter to *'L1'*, and inputting the inverse of the regularization strength (*λ^{-1}*). To determine the appropriate value for λ, we compared four different *LASSO* regression models with λ = 5, 10, 15, and 20.

The values of the coefficients for each parameter computed by *LASSO* regression for λ = 5, 10, 15, and 20 are shown below:



Inspecting the coefficients of the parameters for each regularization strength reveals that the coefficients are drawn closer to zero as λ increases. For each model, there are five coefficients that lie very close to zero. These coefficients correspond to the following predictors: age, resting blood pressure, cholesterol, fasting blood sugar level, and maximum heart rate. From λ = 5 onwards, the magnitude of the predictor ‘fasting blood sugar’ is valued at zero. Once λ increases to 20, the weight of ‘resting electrocardiographic results’ reaches zero. This can be interpreted as *LASSO* regression minimizing the least significant predictors in the model as the strength of the shrinkage increases.

The performance of the regularization can be evaluated by comparing each model’s *k-Fold* cross-validated *MSE* across 10 folds. Using cross-validation in this setting ensures that complexity is not the sole indicator of model performance. If the models are compared by using the *resubstitution error*, then the error will strictly increase as the regularization strength increases.



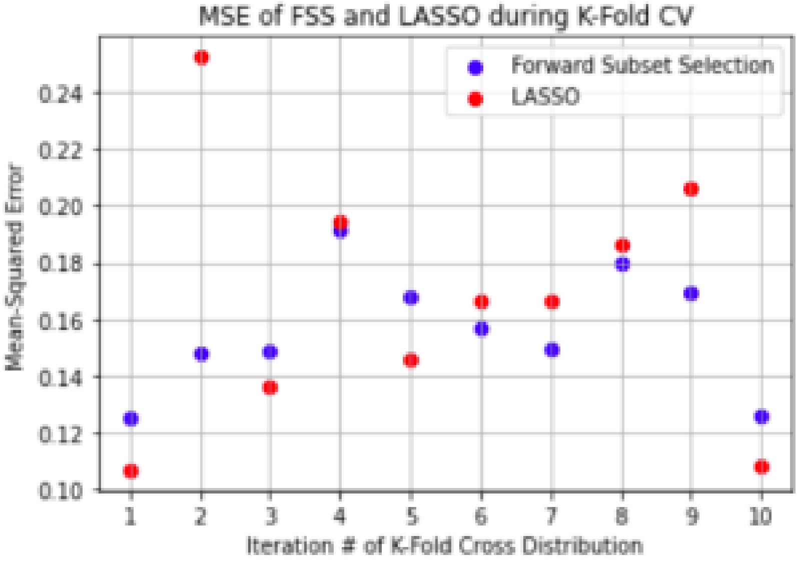
```
In [109]: print('Average MSE, \u03BB = 5  : ',np.mean(mse_lambda_5 ))
          print('Average MSE, \u03BB = 10 : ',np.mean(mse_lambda_10))
          print('Average MSE, \u03BB = 15 : ',np.mean(mse_lambda_15))
          print('Average MSE, \u03BB = 20 : ',np.mean(mse_lambda_20))

          Average MSE, λ = 5  :  0.15408338092518561
          Average MSE, λ = 10 :  0.1540928992956406
          Average MSE, λ = 15 :  0.1560441652389111
          Average MSE, λ = 20 :  0.16288787359604034
```

The average *MSE* derived from *k-Fold* cross-validation, k = 10 are shown below:

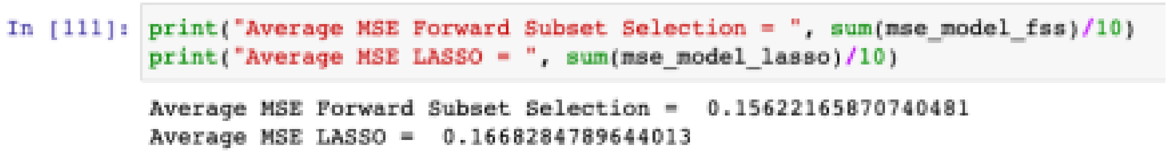
The results of the cross-validation indicate, unsurprisingly, that MSE increases with regularization strength. In addition to minimizing the cross-validated *MSE*, interpretability is an important factor when choosing the appropriate model, as it is a primary benefit of using *LASSO* regression. In this comparison, the maximum interpretability is attained when two coefficients are set at zero, which occurs when λ = 20. Since the objective of this analysis is to create an interpretable model, and the *MSE* of the model does not increase dramatically, we selected the *LASSO* regression with λ = 20 as the optimal choice of regularization strength. This regression represents a model in which ‘fasting blood sugar’, and resting electrocardiographic results’ are not significant features in predicting heart disease. Now, the optimal models from both *Forward Stepwise Subset Selection* and *LASSO* are compared to make the final model selection.

Model Selection - K-Fold Cross Validation between FSS and LASSO:

After two models were chosen using *Forward Subset Selection* and *LASSO*, their effectiveness in accurately predicting cardiovascular disease was evaluated by performing K-Fold Cross Validation. In each iteration of k-fold CV, the model's were trained using training data (approximately 90% of the total data in each iteration) and tested using the test data (approximately 10% of the total data in each iteration). The test involved the models predicting the binary value of the outcome variable, *target*, and then comparing this to the true value within the dataset to compute a mean square error. The mean squared error for *Forward Subset Selection* and *LASSO* in each iteration is shown in the model below:



As seen from the data, with the exception of iteration 2, the MSE's for FSS and LASSO are quite close to each other, as they are usually not more than 0.02 away from each other. The Mean MSE's for both models are shown below:



```
In [111]: print("Average MSE Forward Subset Selection = ", sum(mse_model_fss)/10)
          print("Average MSE LASSO = ", sum(mse_model_lasso)/10)

          Average MSE Forward Subset Selection =  0.15622165870740481
          Average MSE LASSO =  0.1668284789644013
```

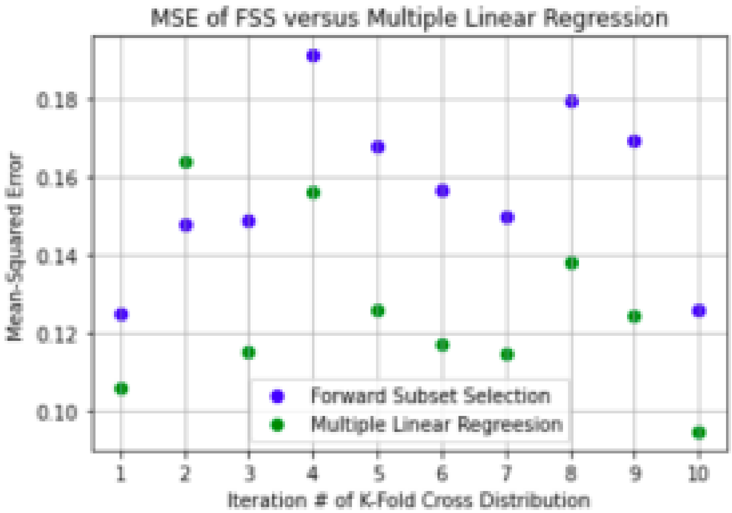
Choosing between the two models, the model with the lower MSE will ensure more accurate predictions. Although both model's MSEs are quite close, Forward Subset Selection has a mean MSE lower than LASSO by approximately = 0.1668 - 0.1562 = 0.0106. The objective of this project is to choose a sparse model to accurately predict the outcome variable. In this case, choosing FSS over LASSO is a sensible choice because the LASSO model was only able to remove 2 predictors, thus resulting in a model with 11 predictors which is still quite complex. By comparison, the FSS model was able to remove 8 predictors, making it less complex and more flexible, and still produce an MSE lower than LASSO. Therefore, for this project our sparse model is the result of our Forward Subset Selection, which was a model with the formula:

$$target \sim 1 + oldpeak + thal + sex + restecg + thalach$$

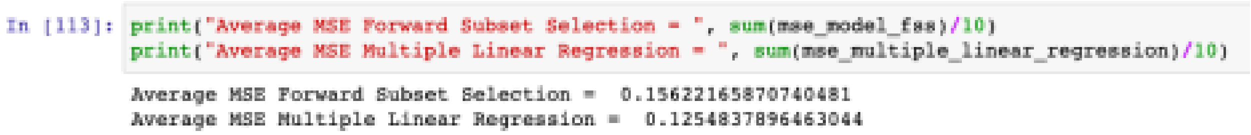
5. Results - Model Estimation

Our hypothesis for this project was: we can develop a sparse model, using *Forward Subset Selection* or *LASSO*, that is more interpretable and less complex than the *Multiple Linear Regression* model and the increase in average *Mean Squared Error* (MSE) will not be greater than 0.05.

Now that we have developed a sparse mode that is more interpretable and less complex than the multiple linear regression model, we must determine if the increase in mean MSE is greater than 0.05. In the same k-fold cross validation we performed in section 4, the MSE for the multiple linear regression model was also calculated. The plot below shows the difference between the MSE of our chosen sparse model (the Forward Subset Model) and the base multiple linear regression model:



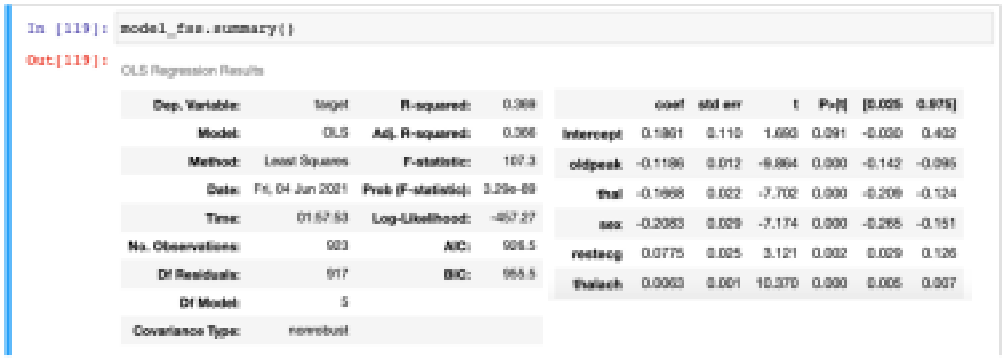
The Mean MSE’s for both models are shown below:



```
In [113]: print("Average MSE Forward Subset Selection = ", sum(mse_model_fss)/10)
          print("Average MSE Multiple Linear Regression = ", sum(mse_multiple_linear_regression)/10)

          Average MSE Forward Subset Selection =  0.15622165870740481
          Average MSE Multiple Linear Regression =  0.12548378964630444
```

As can be seen above, the MSE for our chosen sparse model results in an increase in the mean MSE of approximately: 0.1562 - 0.1255 = 0.0307. This is less than the 0.05 limit we set for ourselves. Thus, our sparse model is not only more interpretable and less complex, the resulting accuracy is not very different from the baseline Multiple Linear Regression model.

To obtain the final parameter estimates of the model, the model was again trained using the entire dataset. The resulting parameters are:

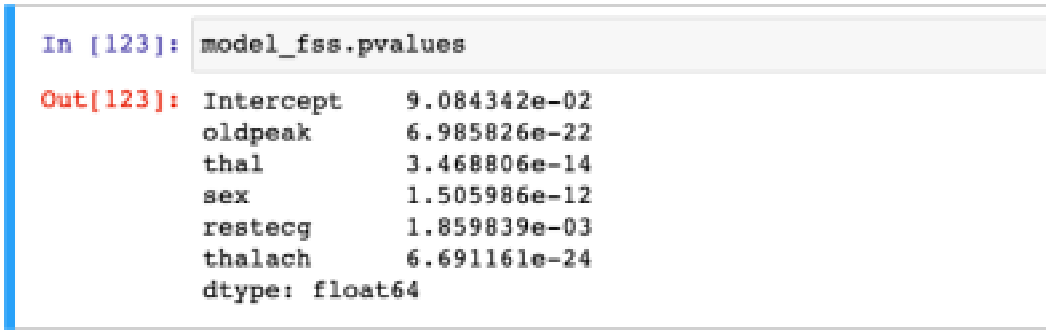


```
In [119]: model_fss.summary()
Out[119]:
```

OLS Regression Results

Dep. Variable:	target	R-squared:	0.369
Model:	OLS	Adj. R-squared:	0.366
Method:	Least Squares	F-statistic:	107.3
Date:	Fri, 04 Jun 2021	Prob (F-statistic):	3.29e-89
Time:	01:57:53	Log-Likelihood:	-457.27
No. Observations:	923	AIC:	926.5
Df Residuals:	917	BIC:	955.5
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1861	0.110	1.693	0.091	-0.030	0.402
oldpeak	-0.1186	0.012	-9.864	0.000	-0.142	-0.095
thal	-0.1668	0.022	-7.702	0.000	-0.209	-0.124
sex	-0.2083	0.029	-7.174	0.000	-0.265	-0.151
restecg	0.0775	0.025	3.121	0.002	0.029	0.126
thalach	0.0063	0.001	10.370	0.000	0.005	0.007



```
In [123]: model_fss.pvalues

Out[123]: Intercept    9.084342e-02
          oldpeak      6.985826e-22
          thal         3.468806e-14
          sex          1.505986e-12
          restecg      1.859839e-03
          thalach      6.691161e-24
          dtype: float64
```

The p-values for each predictor are significantly smaller than 0.05 and by some orders of magnitude. Thus, we can conclude that each predictor chosen in the final Forward Subset Selection model has a statistically significant effect on the outcome variable, *target*. This is expected because FSS isolated the predictors with some of the most weight comparative to the other predictors.

6. Discussion

After completing the process of model selection and estimation, we reached a conclusion that is in agreement with our hypothesis. Our final model fits our criteria for an appropriate model by retaining a low mean squared error, and greatly increasing interpretability through sparse regularization. Furthermore, our model indicated a statistically significant correlation between heart disease, and five health features. These predictors are exercise-induced ST depression, thallasemia, sex, resting electrocardiographic results, and peak heart rate. Since the model satisfied our statistical goals in this analysis, it will also be helpful in fulfilling the greater purpose of this report.

Applied to the real world, the goal of this model is to provide insight into which health factors are the most significant precursors to cardiovascular disease, and to spread understanding of the importance of certain health measurements. As the leading cause of death in the world, any information about the causes of cardiovascular disease can be greatly valuable. This information plays a critical role in diagnosis, which is the first step to treating any disease. Misunderstanding the precursors of heart disease, and which are most important, can lead to the misdiagnosis of an onsetting ailment. Research has shown that, especially among women, misdiagnosis of cardiovascular disease is a widespread health threat. In a

study done on malpractice claims regarding patients with cardiovascular disease, 70 percent of patients with misdiagnosed heart disease ended up dying from their sickness *(American College of Cardiology)*. This problem raises serious health concerns for at-risk individuals, even if they are taking all the right steps in monitoring for cardiovascular disease, and are being treated by a professional. These misdiagnoses are often due to symptoms being hidden, or being attributed to a more benign condition, such as indigestion.

Thus, a model that does not rely on subjective symptoms such as pain, and rather on health measurements, could have great use in the medical field. If a model such as the one derived in this report were to be expanded, or made more robust, it could play an important role in the diagnosis of disease. One course for improving this model would be to simply have more observations. A larger sample size would increase the size of the training data, and represent more of the total population. This could be done by recording more cases alongside these measurements, and appending them to a large, nationwide database. Another method is to expand the variety of predictors used in the model. This would allow researchers to conclude which measurements are the most important, and could maybe disprove some of the statistically significant correlations found in this report. Such research could also expose issues such as multicollinearity among the predictors. It is likely that multicollinearity is one of the most problematic factors among health measurements and symptoms, as these features are often caused by one another. In addition to expanding the scope of this model, it would also be important to define the difference between benign conditions and cardiovascular disease, as they often exhibit similar symptoms. Together, these contributions could lead to a major decrease in misdiagnosis in the clinic, and a greater understanding of health risks in the domestic space.

Works Cited

“CDC - Cardiovascular Disease.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 13 Jan. 2021, www.cdc.gov/heartdisease/about.htm.

Lapp, David. “Heart Disease Dataset.” *Kaggle*, Kaggle, 6 June 2019, www.kaggle.com/johnsmith88/heart-disease-dataset.

Marcot, B.G., Hanea, A.M. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?. Comput Stat (2020).

Taking the Risks to Heart: Misdiagnosis of Heart Disease. American College of Cardiology. 2017, February 20.
https://www.acc.org/membership/join-us/benefits/additional-member-only-benefits/acc-and-the-d

octors-company/the-doctors-company-updates/2017/02/20/12/55/taking-the-risks-to-heart-misdiagnosis-of-heart-disease