Name: Byunghun Lee

1. Load the patient data from “ML\_HW\_Data\_Patients.csv” file.

Patients\_df = pd.read\_csv('Patients.csv')

1. Use variables **Age, Gender, Height, Weight, Smoker, Location, SelfAssessedHealthStatus** to build a linear regression model to predict the systolic blood pressure. Graphical user interface

   Description automatically generated
2. Use **\*\*lasso regression\*\*** with **\*\*10-fold cross-validation\*\*** to identify useful predictors. Plot a lasso plot with readable tick labels on the X and Y axes in your plot for easy visualization and verification. Missing clear and readable tick labels in your plot will cost you significant points for this assignment. Graphical user interface

   Description automatically generated
3. Which top **\*\*TWO\*\*** predictors are you going to select after the lasso analysis?

Since it is not true that most of our dataset is useful for predicting Systolic blood pressure, lasso regression can exclude useless variables from equations, it is a litter better than Ridge regression at reducing the Variance in models that contain a lot of useless variables.

So my two predictors are Gender & Location

1. What is the lambda (l) value you choose in order to select the top two predictors you identified in the last question?

R\_squared 34.376749

lambda 0.200000

Name: 19, dtype: float64

1. What are the q values for the two selected predictors at the lambda (l) value you identified in the last question?

# Best Model

reg\_best = Lasso(alpha = 0.144737)

reg\_best.fit(X\_train, y\_train)

Lasso(alpha=0.144737, copy\_X=True, fit\_intercept=True, max\_iter=1000,

normalize=False, positive=False, precompute=False, random\_state=None,

selection='cyclic', tol=0.0001, warm\_start=False)

from sklearn.metrics import mean\_squared\_error

mean\_squared\_error(y\_test, reg\_best.predict(X\_test))

32.54105584096965

reg\_best.coef\_

array([ 1.11565225, -0. , 0.87507912, -0.06084871, 0.41829045,

8.571702 , -0.78587678])

My interpretation from HW 4

Since lasso regression is very similar to Ridge Regression but it has some very important differences. For each predictor, I split the data set training (using least squares by minimizing sum of square residuals) which is law bias and high variance and testing data. Using lasso regression, I tried to minimize the sum of the squared residuals that I already did for HW2 (linear regression) + lambda\*lthe slopel (lasso regression penalty). Even though lasso regression had more Bias than least squares, but in return for that small amount of Bias, the lasso regression line has a significant drop in Variance for testing set which means that lasso regression seems starting with a slightly worse fit but provided better long-term predictions at the end.

Just like with Ridge Regression, lambda can be any value from 0 to positive infinity and is determined using Cross validation.

Just like the Ridge Regression Penalty, the Lasso Regression Penalty contains all of the estimated parameters(predictors) except for the y-intercept.

As I increase the value for lambda, lasso regression may shrink parameter (predictors) a lot more than they shrink the slope which means that lasso regression can shrink the slope all the way to 0.