SEIS 763: Machine Learning
Garth Mortensen, mort0052@stthomas.edu

Graduate Program in Software
SEIS 763: ML
Assignment #3 (100 points)
Due Date: June 18th, 2018

STANDARDIZE

Write a MatLab (or a programming language of your choice) program with excellent comments to perform AND provide answers the following tasks:

1. Use Matlab command "load patients" to load patient self evaluation dataset.

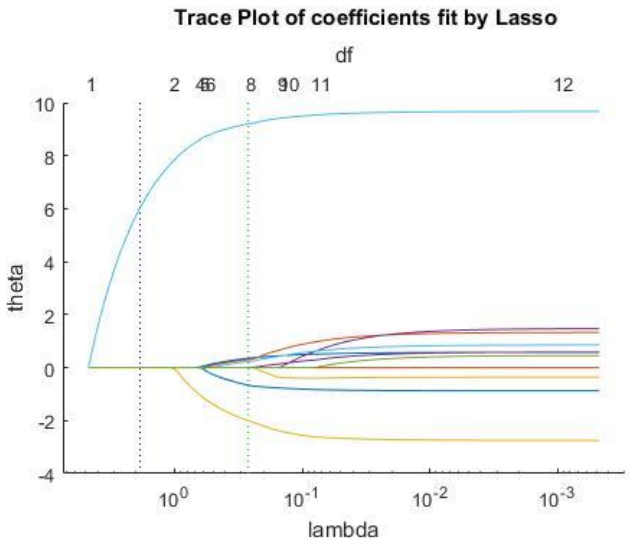
2. If you use other programming languages or tools, save the data to a file so your tool can read.

3. Use variables Age, Gender, Height, Weight, Smoker, Location, SelfAssessedHealthStatus to build a linear regression model to predict the systolic blood pressure.

```
%Clear previous variables, wipe screen, close windows
clear all
clc
% close all

% 3. Use variables  Age, Gender, Height, Weight, Smoker, Location,
% SelfAssessedHealthStatus to build a linear regression model to predict the
% systolic blood pressure.

% https://www.mathworks.com/help/matlab/matlab_prog/create-a-table.html
load patients;

%Target-------------------------------------------------------------
Y = Systolic;

%Standardize numericals---------------------------------------------
XNumeric = [Age Height Weight];
XNumeric_scaled = zscore(XNumeric);

%One-hot encode categoricals----------------------------------------
%Because these attributes are single-columns containing many values, we
%need to break them into binary attributes, one for each value.
Gender = nominal(Gender);
GenderCateg = dummyvar(Gender);

Location = nominal(Location);
LocationCateg = dummyvar(Location);

SelfAssessedHealthStatus = nominal(SelfAssessedHealthStatus);
SelfAssessedHealthStatusCateg = dummyvar(SelfAssessedHealthStatus);

% Bring Categorical together
%Now that we've broken each attribute value into a seperate binary vector,
%we need to bring them all back together into a single matrix.
XCateg = [GenderCateg LocationCateg SelfAssessedHealthStatusCateg Smoker];

% Merge numerical with categorical matrices----------------------
XAll = [XNumeric_scaled XCateg];
```

4. 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 coordinates 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.**

```
% 4. 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 coordinates
% 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.

%Lasso=============================================================
%We need to determine the number of k-folds and alpha value.
%With those values set, we can run our lasso linear regression.
%[B, FitInfo] = lasso(X, Y,Name, Value)
% Set cross validation k-fold, k
kfold = 10;
% 'Alpha', alpha value, where alpha = 1 is lasso, and = 0.00001 approaches
% ridge
alpha = 1;
% Don't set lambda. It's a vector, not a scalar.
% Default lambda count (steps) = 100
[B FitInfo] = lasso(XAll, Y, 'CV', kfold, 'Alpha', alpha);


% Lasso Plot of Coefficients=========================================
lassoPlot(B, FitInfo, 'PredictorNames', {'Age', 'Height', 'Weight',...
    'Female', 'Male',...
    'County General Hospital', 'St Marys Medical Center', 'VA Hospital',...
    'Excellent', 'Fair', 'Good', 'Poor',...
    'Smoker'},...
    'PlotType', 'lambda',...
    'XScale', 'log'),...
    ylabel('theta'),...
    xlabel('lambda')
```



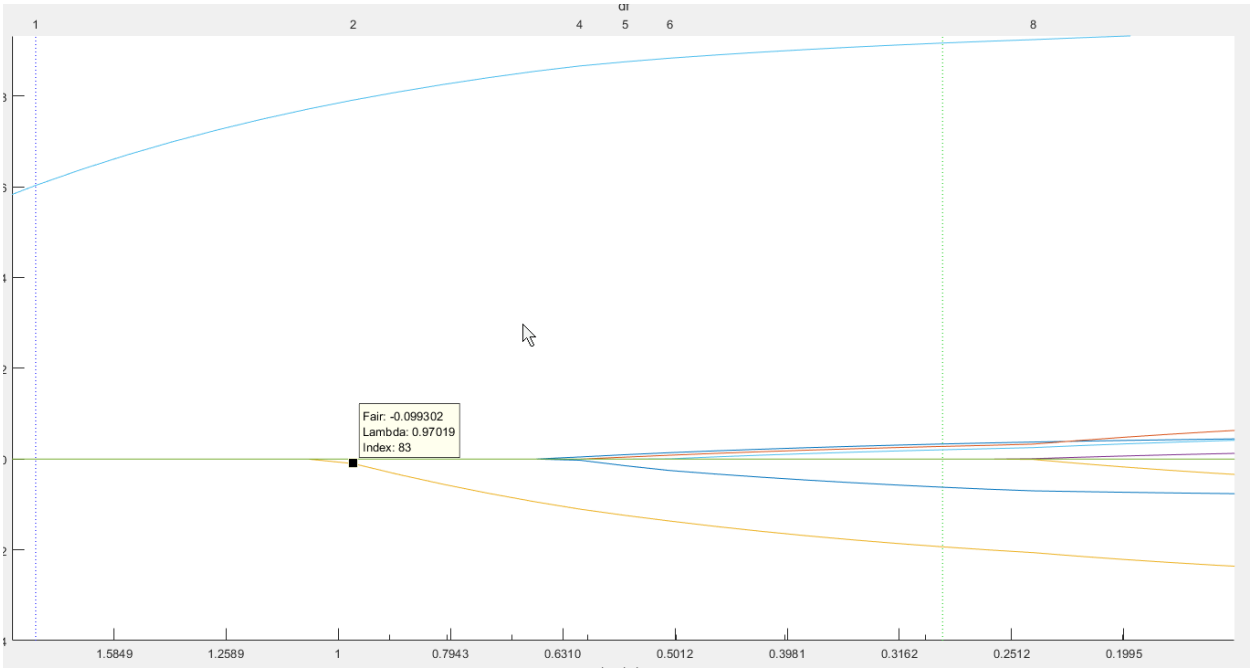
5. Which top **TWO predictors are you going to select after the lasso analysis?**

The coefficient for Age is 0, which means that for every 1 standard deviation change of Age, the Y-target response variable changes by 0. It remains 0 within the two vertical lines. For that reason, I do not choose Age as a top two predictor.

That leaves Smoker = 1 and SelfHealthAssessment = 'Fair' as the two leading predictors.

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

Lambda = 0.97, near to where $theta_{Fair}$ intercepts the x-axis and $theta_{Smoker}$ is also not zero, but all other theta values are.



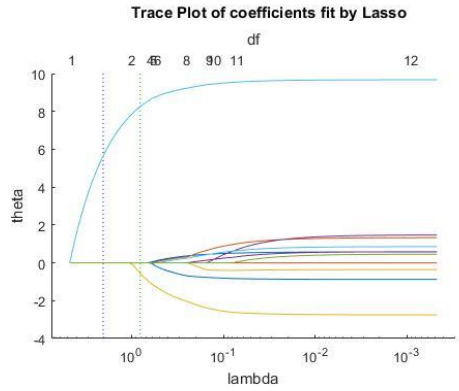
7. What are the q values for the two selected predictors at the lambda (l) value you identified in the last question?

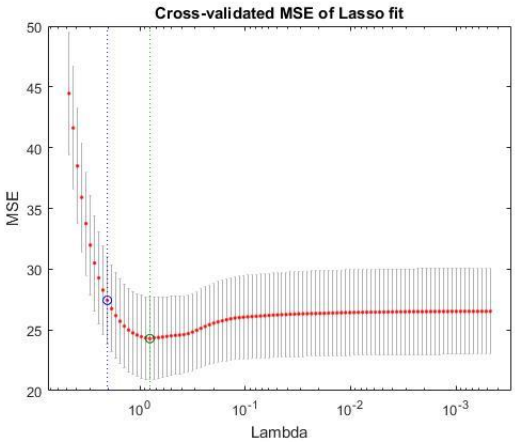
$Theta_{Fair}$ = -0.993

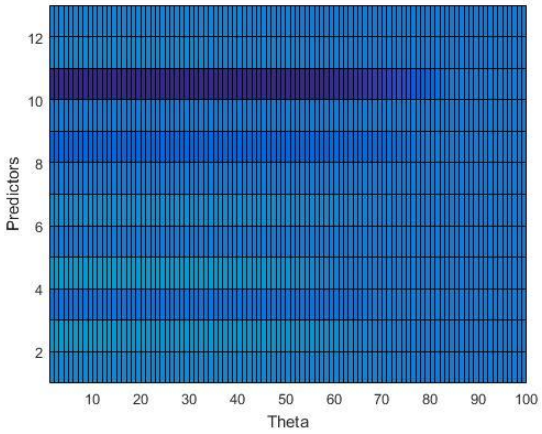
$Theta_{Smoker}$ = 7.9105

```
%SEIS763 Machine Learning
%Garth Mortensen, mort0052@stthomas.edu
%Assignment 3

%Clear previous variables, wipe screen, close windows
clear all
clc
% close all

% 3. Use variables  Age, Gender, Height, Weight, Smoker, Location,
% SelfAssessedHealthStatus to build a linear regression model to predict the
% systolic blood pressure.

% https://www.mathworks.com/help/matlab/matlab_prog/create-a-table.html
load patients;

%Target-------------------------------------------------------------
Y = Systolic;

%Standardize numericals---------------------------------------------
XNumeric = [Age Height Weight];
XNumeric_scaled = zscore(XNumeric);

%One-hot encode categoricals----------------------------------------
%Because these attributes are single-columns containing many values, we
%need to break them into binary attributes, one for each value.
Gender = nominal(Gender);
GenderCateg = dummyvar(Gender);

Location = nominal(Location);
LocationCateg = dummyvar(Location);

SelfAssessedHealthStatus = nominal(SelfAssessedHealthStatus);
SelfAssessedHealthStatusCateg = dummyvar(SelfAssessedHealthStatus);

% Bring Categorical together
%Now that we've broken each attribute value into a seperate binary vector,
%we need to bring them all back together into a single matrix.
XCateg = [GenderCateg LocationCateg SelfAssessedHealthStatusCateg Smoker];

% Merge numerical with categorical matrices----------------------
XAll = [XNumeric_scaled XCateg];


% 4. 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
coordinates
% 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.
```

```
%Lasso=============================================================
%We need to determine the number of k-folds and alpha value.
%With those values set, we can run our lasso linear regression.
%[B, FitInfo] = lasso(X, Y,Name, Value)
% Set cross validation k-fold, k
kfold = 10;
% 'Alpha', alpha value, where alpha = 1 is lasso, and = 0.00001 approaches
% ridge
alpha = 1;
% Don't set lambda. It's a vector, not a scalar.
% Default lambda count (steps) = 100
[B FitInfo] = lasso(XAll, Y, 'CV', kfold, 'Alpha', alpha);


% Lasso Plot of Coefficients=========================================
lassoPlot(B, FitInfo, 'PredictorNames', {'Age', 'Height', 'Weight',...
    'Female', 'Male',...
    'County General Hospital', 'St Marys Medical Center', 'VA Hospital',...
    'Excellent', 'Fair', 'Good', 'Poor',...
    'Smoker'},...
    'PlotType', 'lambda',...
    'XScale', 'log'),...
    ylabel('theta'),...
    xlabel('lambda')


% Cross-validated Deviance of Lasso Plot=============================
%Product an extra graph to display cross-validation
lassoPlot(B, FitInfo, 'PlotType', 'CV');

% Theta vs Predictors================================================
%Product an extra graph to display theta vs predictors
figure, pcolor(B), xlabel('Theta'), ylabel('Predictors')

% Interpretation-----------------------------------------------------
% Identify number of nonzero coefficients are minimum deviance plus one
% standard deviation.
indx = FitInfo.Index1SE;
B0 = B(:,indx);
nonzeros = sum(B0 ~= 0)
```


Trace Plot of coefficients fit by Lasso
df
1
2
4
6
8
9
10
11
12
theta
lambda
10^0
10^{-1}
10^{-2}
10^{-3}


Cross-validated MSE of Lasso fit
MSE
Lambda
10^0
10^{-1}
10^{-2}
10^{-3}


Predictors
Theta