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.

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.

load patients;

%Create dummy variables

D\_Smoker = dummyvar(nominal(Smoker));

D\_Sex = dummyvar(nominal(Gender));

D\_Loc = dummyvar(nominal(Location));

D\_SelfEval = dummyvar(ordinal(SelfAssessedHealthStatus));

%Standardize variables

ZX = [zscore([Age Height Weight]) D\_Sex(:, 2) D\_Smoker(:,2) D\_Loc(:, 2:3) D\_SelfEval(:, 2:4)];

Y = Systolic;

PN={'Age';'Height';'Weight';'Sex';'Smoker';'Loc1';'Loc2';'SE1';'SE2';'SE3'};

[b fitinfo] = lasso(ZX, Y, 'CV',10, 'Alpha', 1,'PredictorNames',PN)

lassoPlot(b,fitinfo,'PlotType', 'Lambda','PredictorNames',PN, 'XScale','log');

lassoPlot(b,fitinfo,'PlotType','CV');

lassoPlot(b,fitinfo,'PlotType','L1');

BestValue = find(fitinfo.Lambda == fitinfo.LambdaMinMSE)

%Build the model

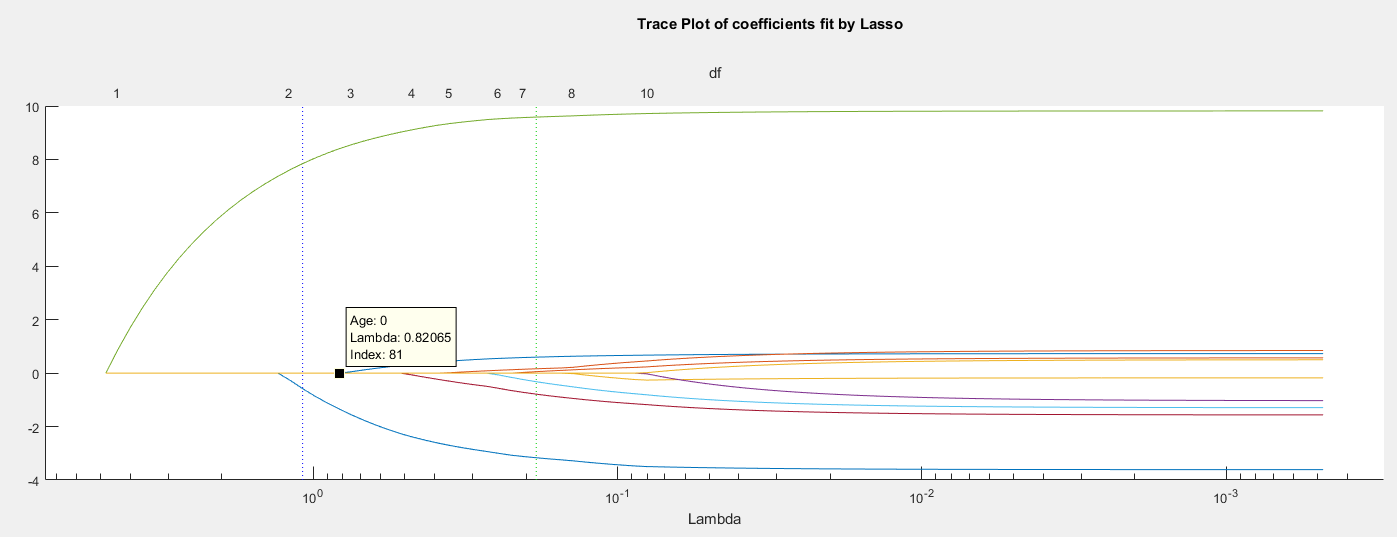
model = fitlm(ZX,Y)

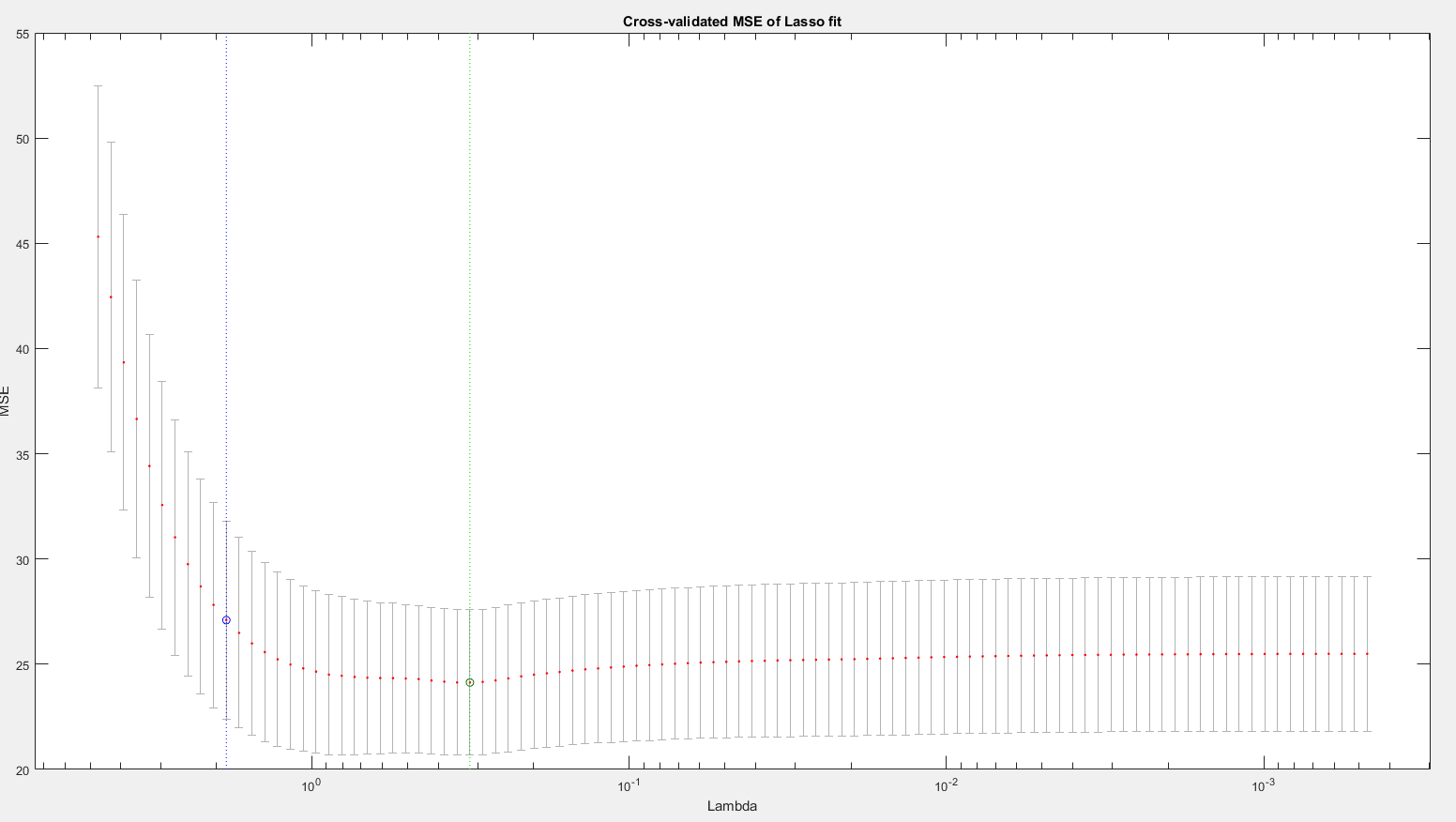
%Find the outliers and remove them

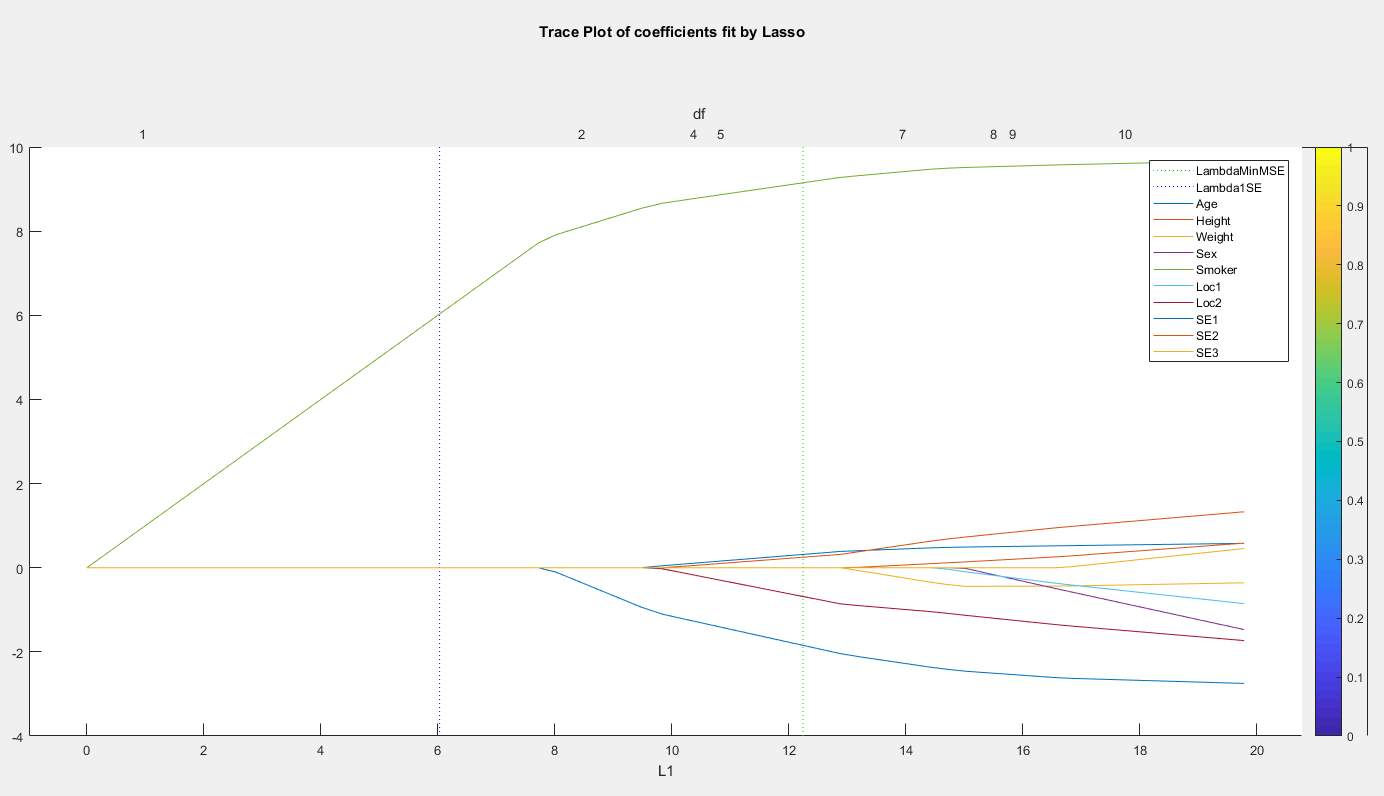
rown = find((model.Diagnostics.CooksDistance) > 3\* mean(model.Diagnostics.CooksDistance))

%Build the model again

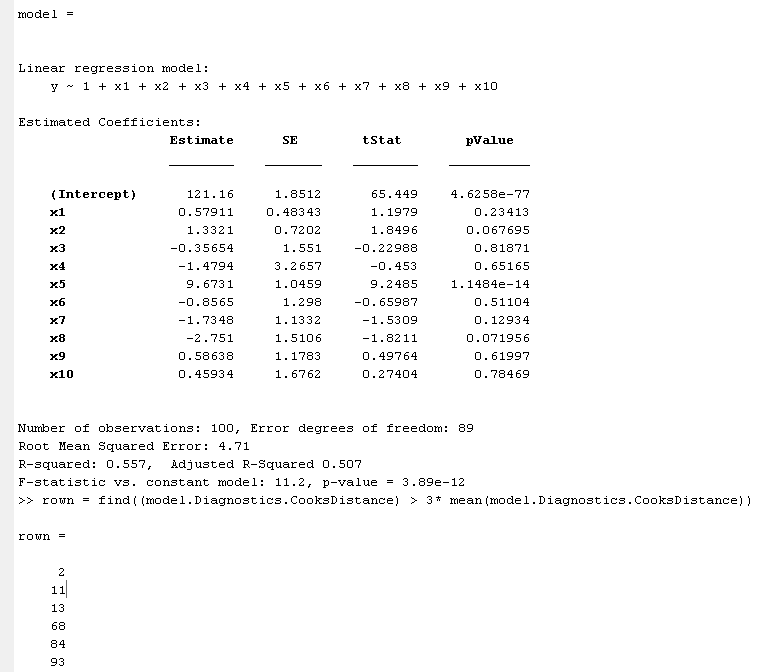
Model2 = fitlm(ZX,Y)



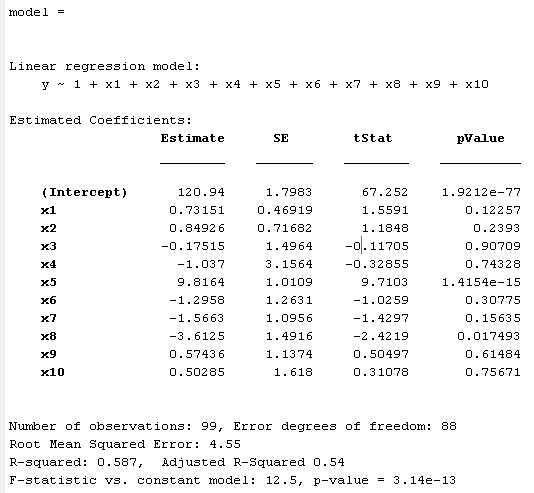




**Model:**



**Now build the new model:**

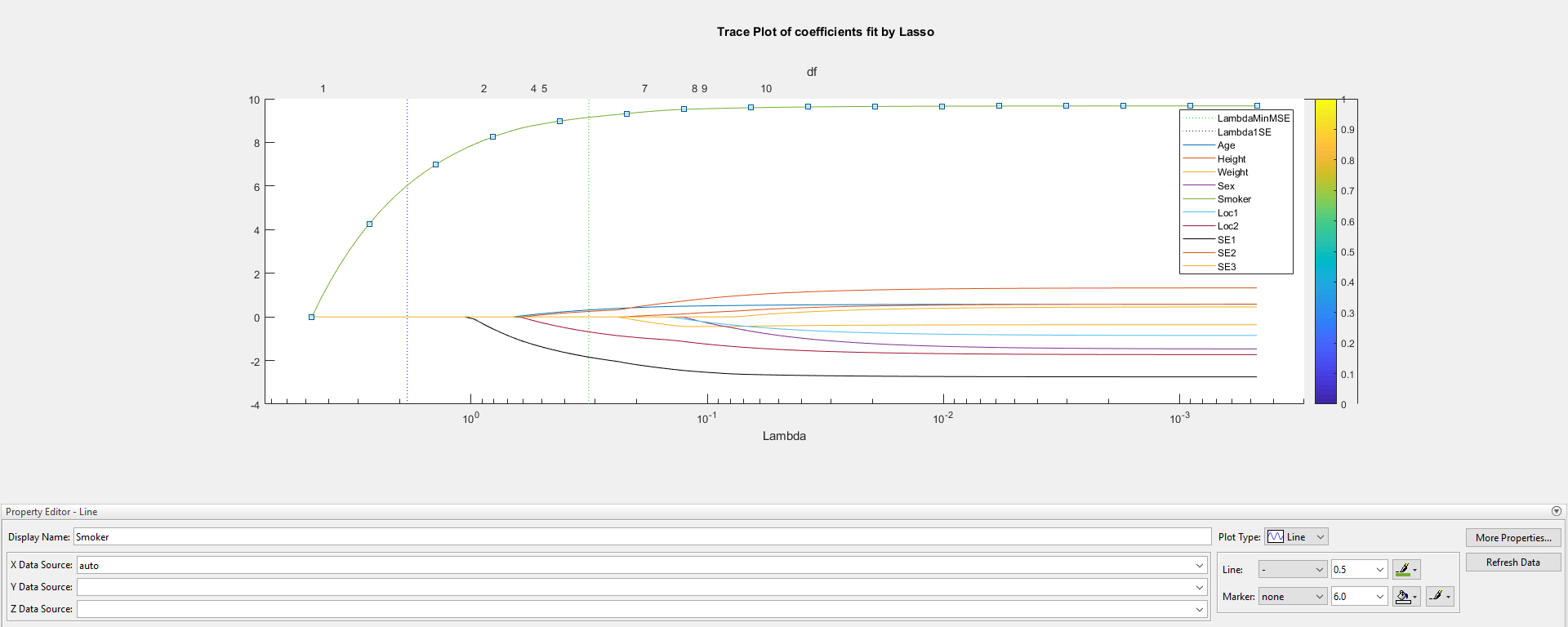


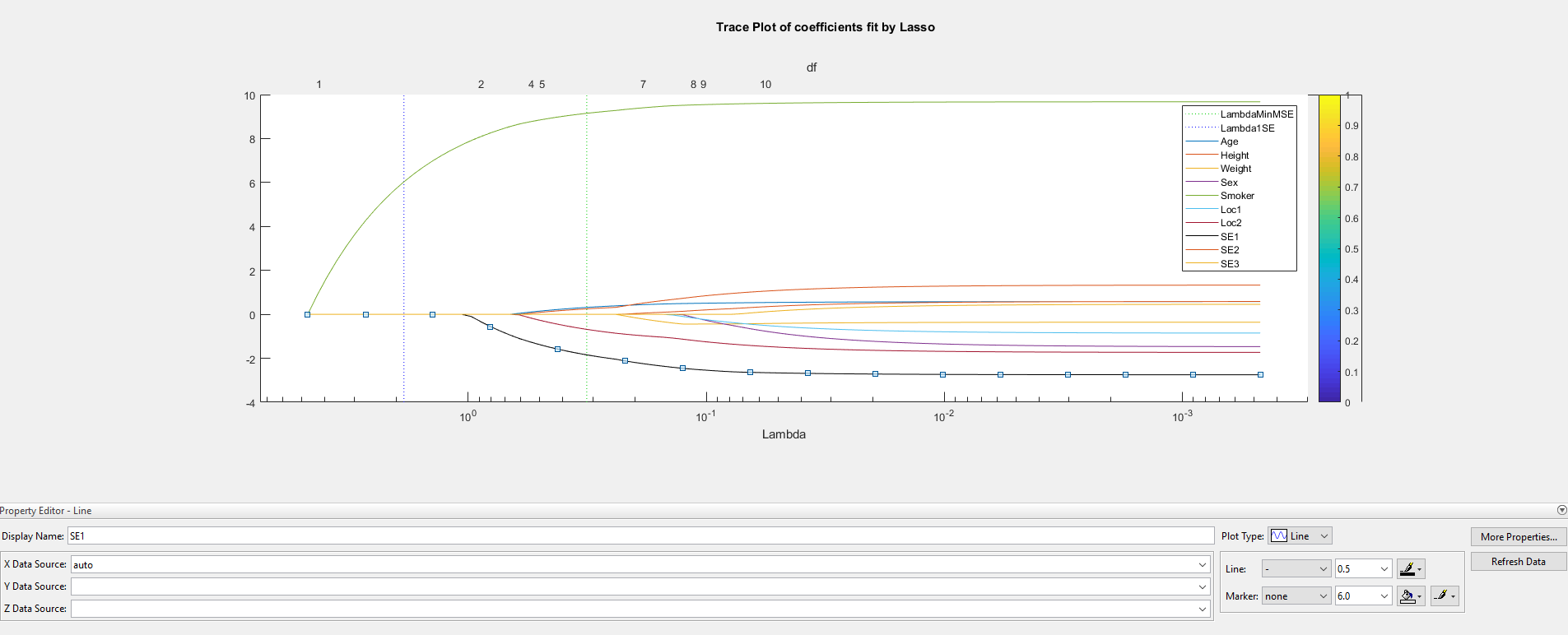
We see an improvement in the Adjusted R2 and the error reduced too.

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

**Solution:**

**SMOKER and SE1(SelfAssessedHealthStatus=fair) are the top 2 predictors**



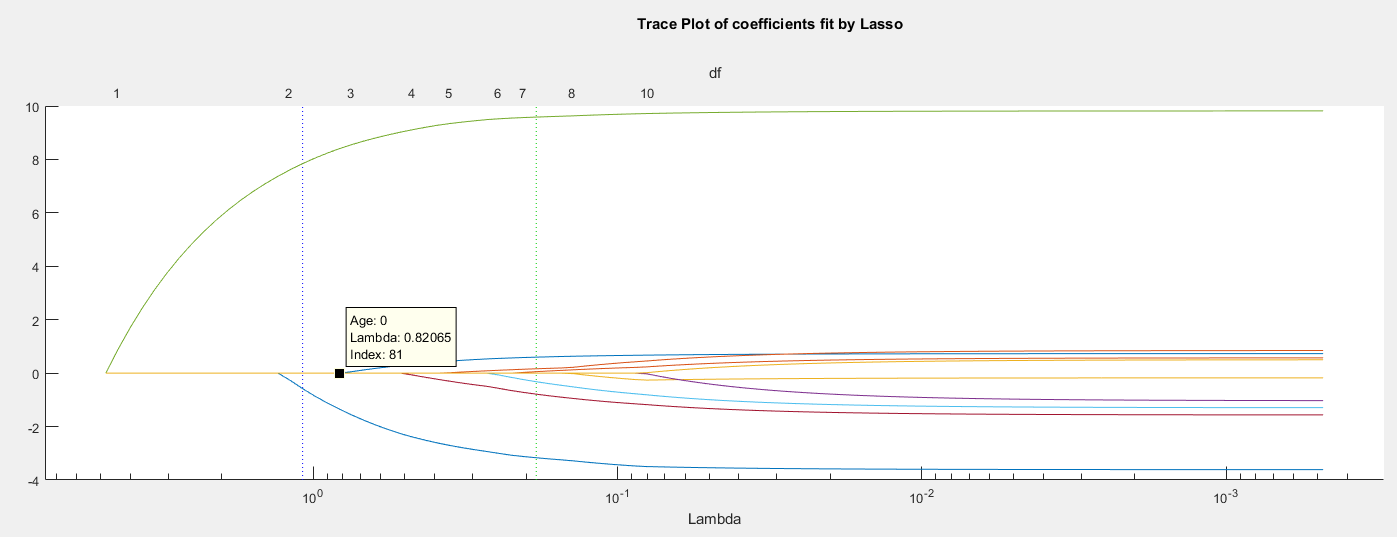


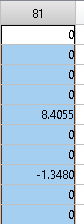
**6. What is the lambda (l) value you choose in order to select the top two predictors**

**you identified in the last question?**

**Solution:**

Lambda = 0.82065





**7. What are the q values for the two selected predictors at the lambda (l) value you**

**identified in the last question?**

**Solution:**

SMOKER: 8.4055

SE1(SelfAssessedHealthStatus=fair): -1.3480