1. Load “patients.csv” for patient self-evaluation dataset.
2. Use variables **Age, Gender, Height, Weight, Smoker, Location,**

**SelfAssessedHealthStatus** to build a linear regression model to predict the

systolic blood pressure.

load patients;

Patient\_data=table(Age, Gender, Height, Weight, Smoker, Location,SelfAssessedHealthStatus,Systolic,'VariableNames',{'Age', 'Gender', 'Height', 'Weight', 'Smoker', 'Location','SelfAssessedHealthStatus','Systolic'});

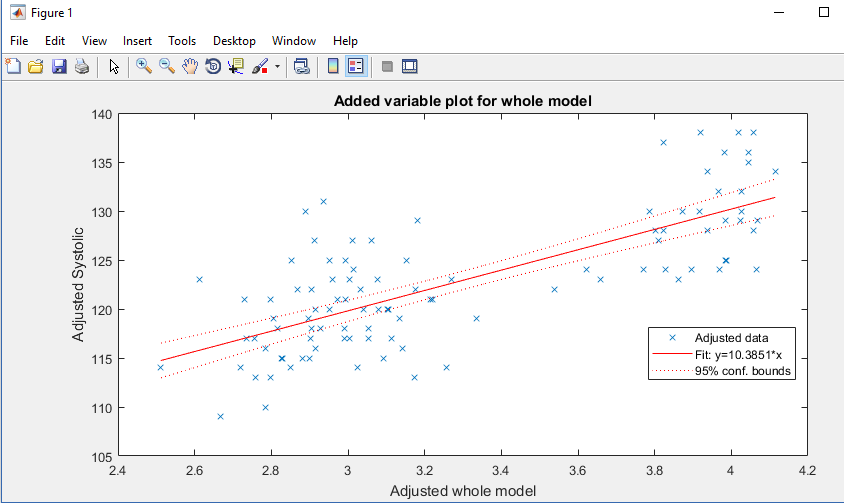
Patient\_data.Gender=categorical(Patient\_data.Gender);

Patient\_data.Smoker=categorical(Patient\_data.Smoker);

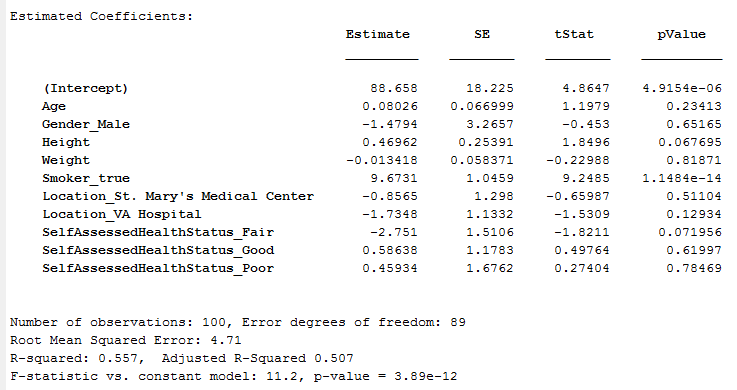
Patient\_data.Location=categorical(Patient\_data.Location);

Patient\_data.SelfAssessedHealthStatus=categorical(Patient\_data.SelfAssessedHealthStatus);

LRModel=fitlm(Patient\_data,'Systolic~Age+Gender+Height+Weight+Smoker+Location+SelfAssessedHealthStatus')



1. **What are the regression coefficients (thetas)?**



1. **How do you interpret those numbers?**

The equation is as follows:

Systolic = 88.66 + 0.08 \*Age - 1.48 \*Gender\_Male + 0.47\*Height -0.013\*Weight + 9.67\*Smoker\_true - 0.8565\*Location\_St. Mary’s Medical Center - 1.7348\*Location\_VA Hospital - 2.751\*SelfAssessedHealthStatus\_Fair + 0.59\*SelfAssessedHealthStatus\_Good + 0.46\*SelfAssessedHealthStatus\_Poor

Above linear equation says that, if all the values are zero, then systolic pressure will be 88.66.

1. As age increases by one unit, Systolic BP will increase by .08 units
2. If gender is female then Systolic BP will reduce by 1.5 units
3. If the height of the person increases by one unit then the Systolic BP will increase 0.47 units
4. If the weight of the person increases by one unit then the Systolic BP will increase 0.013 units
5. If person is a smoker then Systolic BP will increase by 9.67 units
6. If Location is St. Marys Medical Center then Systolic BP will decrease by 0.856 units
7. If Location is St. VA Hospital then Systolic BP will decrease by 1.734 units
8. If Self assessed Health Status is fair then Systolic BP will decrease by 2.751 units
9. If Self assessed Health Status is good then Systolic BP will increase by 0.59 units
10. If Self assessed Health Status is poor then Systolic BP will increase by 0.46 units
11. **If you need to identify one outlier record, which records are potential outliers?**

**How do you reach this conclusion?**

%Identify the Outliers

figure,subplot(2,2,1)

plot(LRModel)

subplot(2,2,2)

plotResiduals(LRModel,'fitted')

subplot(2,2,3)

plotResiduals(LRModel)

subplot(2,2,4)

normplot(LRModel.Residuals.Raw)

xlabel('residuals')

%Cook Distance

figure, subplot(3,1,1)

plot(LRModel)

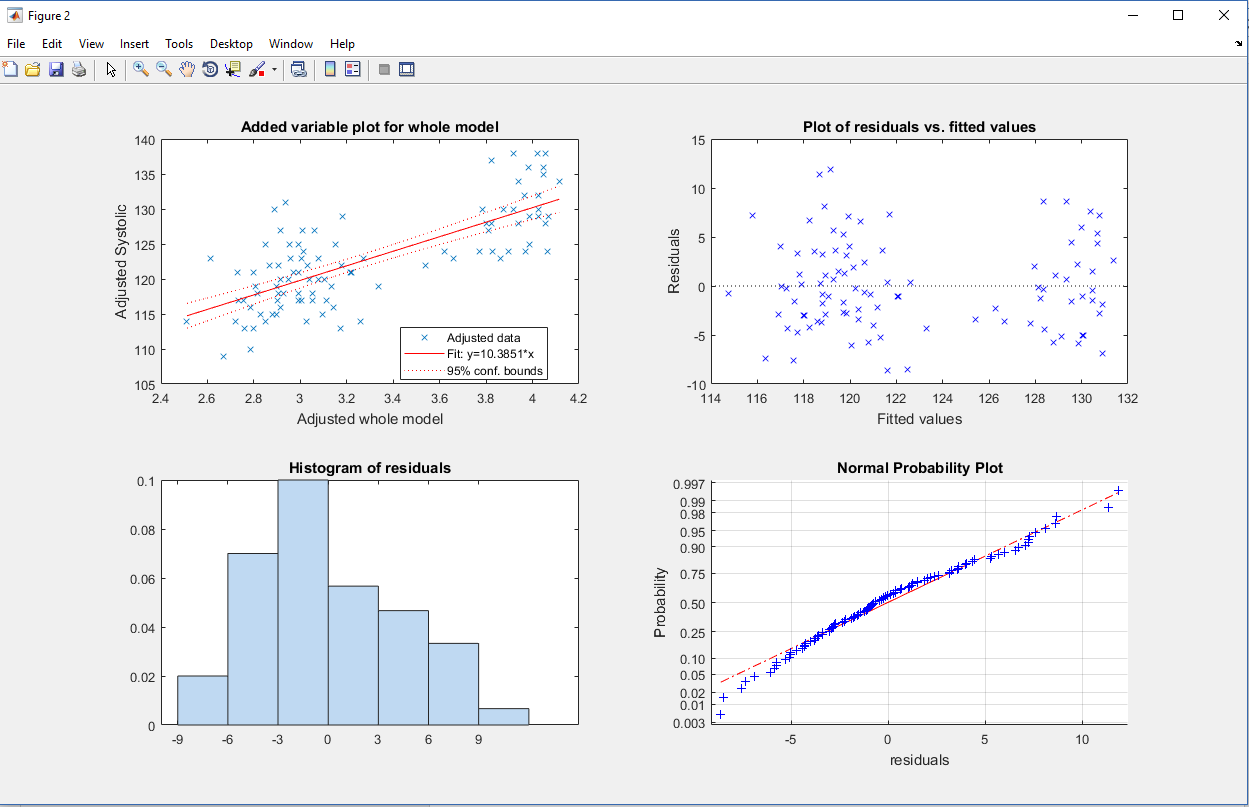
subplot(3,1,2)

plotDiagnostics(LRModel)

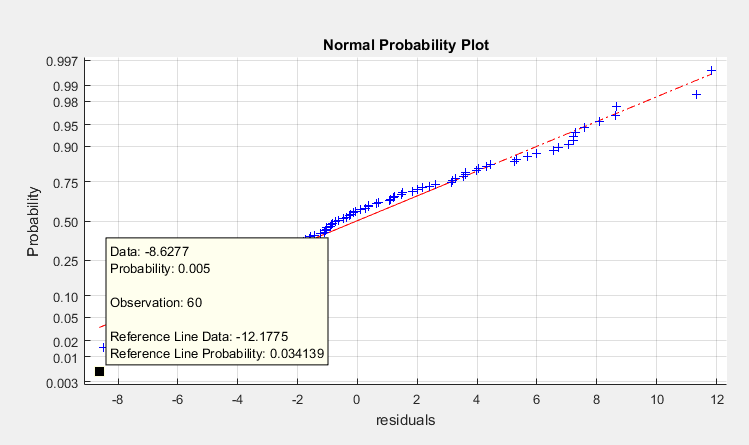
subplot(3,1,3)

plotDiagnostics(LRModel,'cookd')

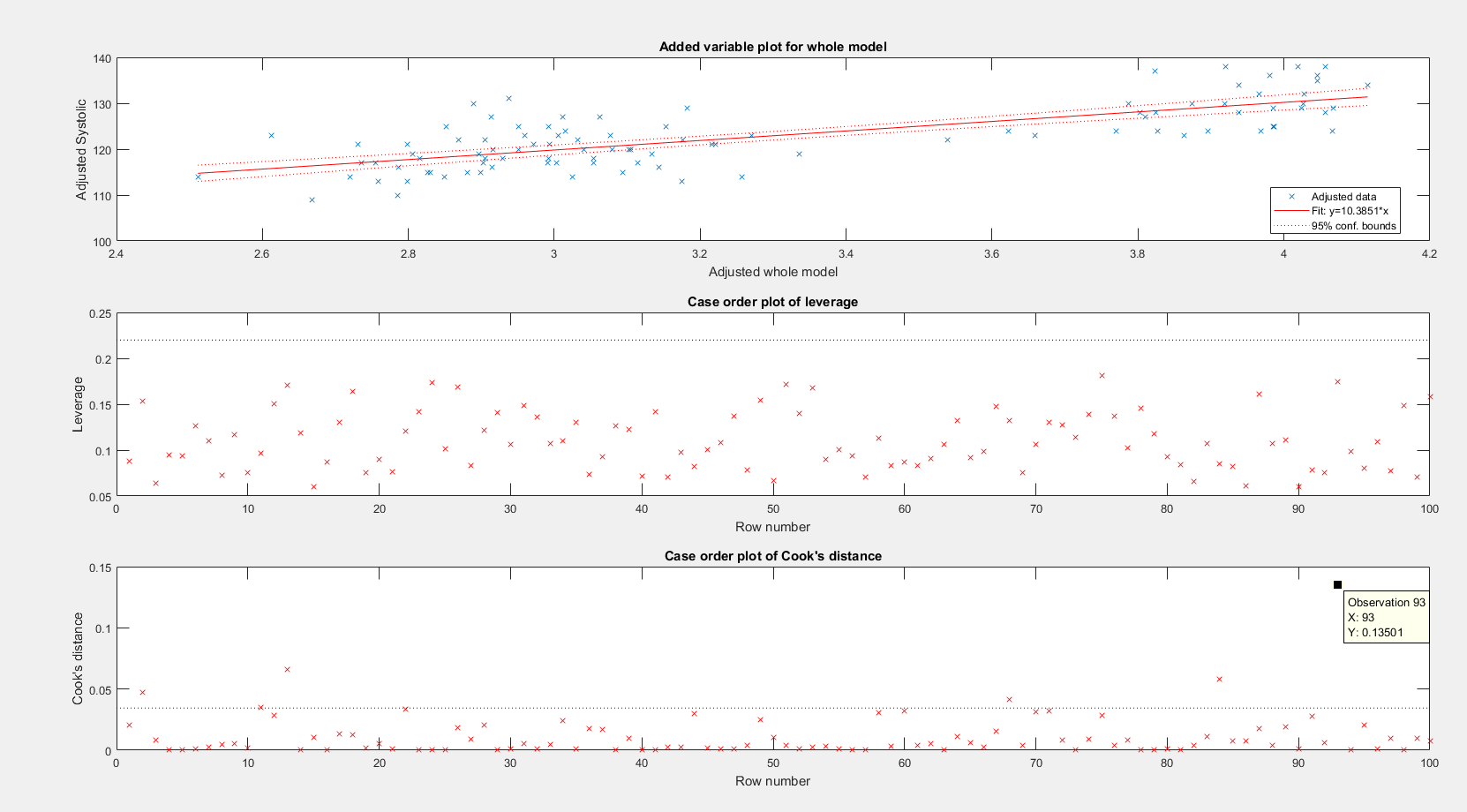
LRModel.Diagnostics;



Based on the normal probability plot, I do not see any clear outlier but Observation 60 looks like a potential outlier



Now, we look at the Cook distance and Observation 93 looks like an outlier as it’s almost 3 times the average cooks distance



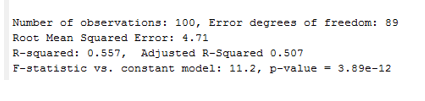
1. **If you need to identify one or few useless features (independent variables or**

**predictors), which one(s) will you choose? Why do you reach this conclusion?**

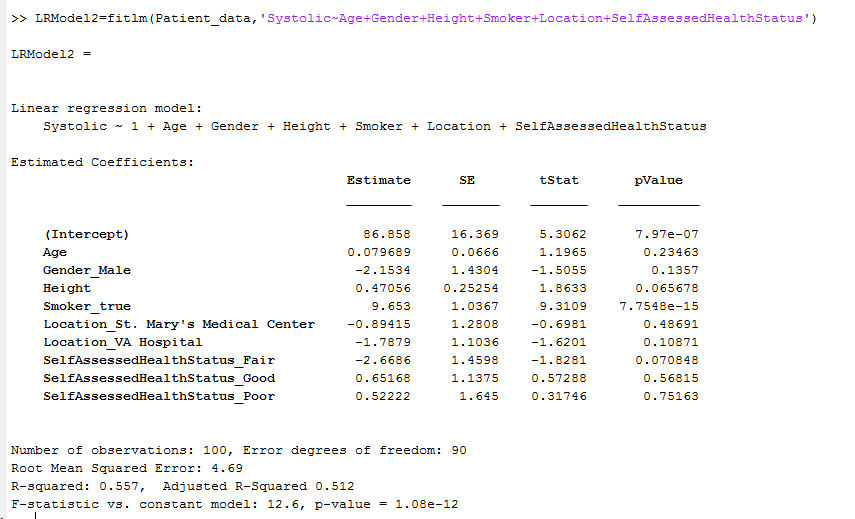
Weight has the highest p-value of 0.82, so we can try removing this predictor and build a new model

The Root mean square decreased from 4.71 to 4.69 and adjusted R2 increased from 0.507 to .512

**Original model values:**



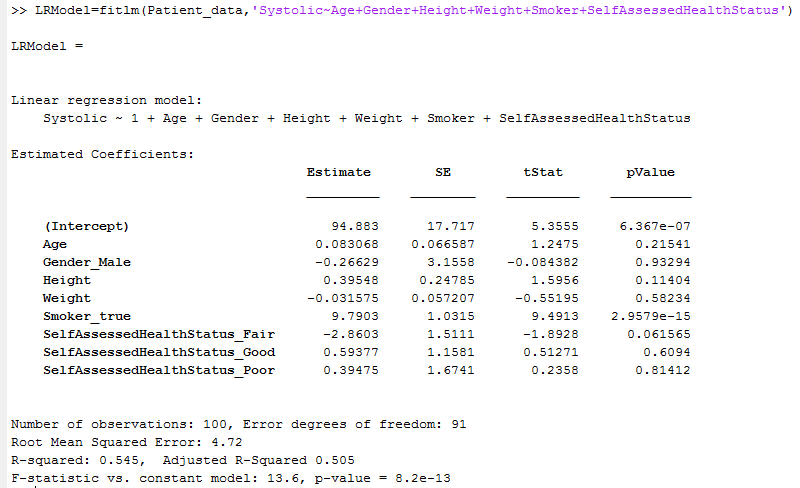
**New Model**



Now we will try removing the outlier, observation 93 and then build the model again.

Removing observation 93 did not help to improve the RMSE or Adjusted R2. So, the one not so useful feature is **WEIGHT** that could be removed.

Tried removing the Location but that increased the RMSE and also reduced the Adjusted R2. **But logically, it sounds more meaningful to remove the location to predict the Systolic BP.**



**I tried creating a lasso plot to see the important variables but could not produce the lasso plot.**

prednames = Patient\_data.Properties.VariableNames(1:end-1);

X = Patient\_data(:,1:end);

y = Patient\_data(:,end);

[beta,FitInfo] = lasso(X,y,'Standardize',true,'CV',10,'PredictorNames',prednames);

lassoplot(beta,FitInfo,'PlotType','Lambda','XScale','log');

hplot = get(gca,'Children');

