clear; clc; close all;

load features.mat

Scatter plot matrices between the variables along with a univariate histogram for each variable.

features = [notes, pitch, census, provenance];

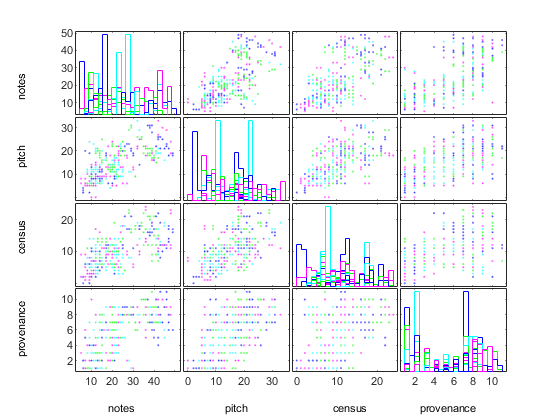
featuresNames = {'notes', 'pitch', 'census', 'provenance'};

figure

gplotmatrix(features,[],groundtruth,['c' 'b' 'm' 'g'],[],[],false);

text([.08 .35 .6 .8], repmat(-.1,1,4), featuresNames, 'FontSize',8);

text(repmat(-.12,1,4), [.8 .6 .35 .05], featuresNames, 'FontSize',8, 'Rotation',90);



Preparing Data with four Predictor variables (notes, pitch, census, and provenance) and a response variable (groundtruth). Also, creating the code as categorical data. Plotting the histogram to look for the outliers.

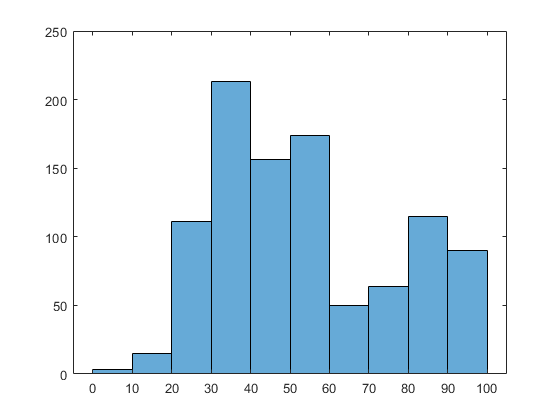
tb1 = table(notes, pitch, census, provenance, groundtruth, code, ...

'VariableNames', {'notes', 'pitch', 'census', 'provenance', 'groundtruth', 'code'});

tb1.code = categorical(tb1.code,1:4,["Totally different","Small", "Inspired", "Strongly Similar"]);

% wordcloud(similarity\_keyword); % require text analytics toolbox

histogram(tb1.groundtruth,10) % check the outlier



Partition the data set into a training set and testing set (30%).

rng('default')

c = cvpartition(height(tb1),"holdout",0.3);

trainData = tb1(training(c),:);

testData = tb1(test(c),:);

Fit a linear regression model. The model displays the model formula, estimated coefficients, and summary statistics. The R-squared value shows that the model explains approximately 97% of the variability in the response variable.

mdl = fitlm(trainData, "predictorVars",["notes","pitch","census","provenance"],"ResponseVar","groundtruth")

mdl =

Linear regression model:

groundtruth ~ 1 + notes + pitch + census + provenance

Estimated Coefficients:

**Estimate** **SE** **tStat** **pValue**

**\_\_\_\_\_\_\_\_** **\_\_\_\_\_\_\_\_** **\_\_\_\_\_\_** **\_\_\_\_\_\_\_\_\_\_\_**

**(Intercept)** 2.1717 0.36628 5.929 4.8201e-09

**notes**  1.1048 0.020196 54.704 6.2884e-253

**pitch**  1.0033 0.030476 32.92 1.5882e-143

**census**  0.37196 0.037046 10.04 3.1316e-22

**provenance**  1.1563 0.069131 16.726 5.8422e-53

Number of observations: 694, Error degrees of freedom: 689

Root Mean Squared Error: 3.79

R-squared: 0.971, Adjusted R-Squared: 0.971

F-statistic vs. constant model: 5.81e+03, p-value = 0

Coefficient of determination (R-squared) shows the proportionate amount of variation in the response variable explained by the independent variables in the linear regression model. The larger the R-squared is, the more variability is explained by the linear regression model. In general, the higher the R-squared, the better the model fits the data.

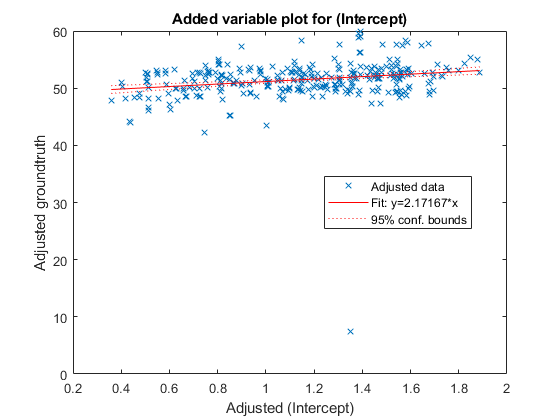
The model displays the p-value of each coefficient. The p-values shows which variables are significant to the model. The p value is least for notes variable which says that it is statistically significant. If p value larger than 0.05 (5% significance level), then those variables are not significant and, therefore, we consider removing them.

tStat is the t-test statistic. It should be high enough for a significant variable.

SE is the squared error. Notes lead to least error among all.

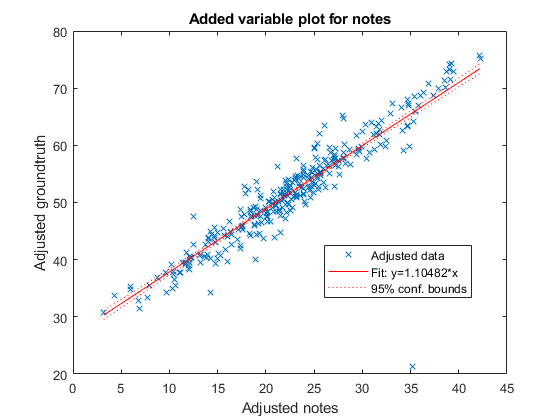
% Effect of intercept on response variable (groundtruth)

plotAdded(mdl,[1])



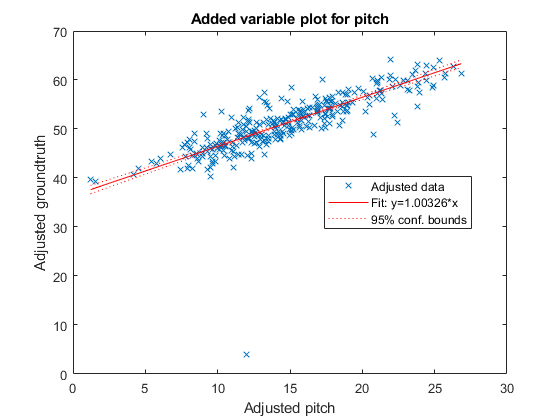
% Effect of notes on response variable (groundtruth)

plotAdded(mdl,[2])



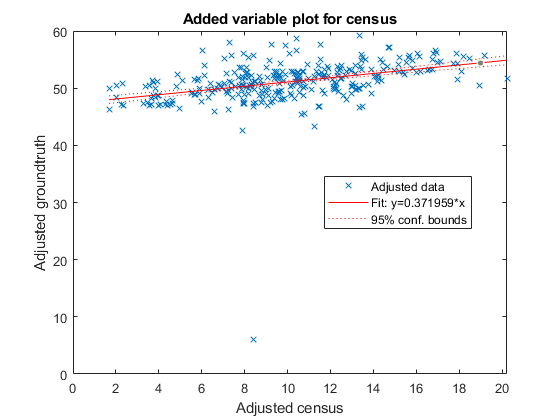
% Effect of pitch on response variable (groundtruth)

plotAdded(mdl,[3])



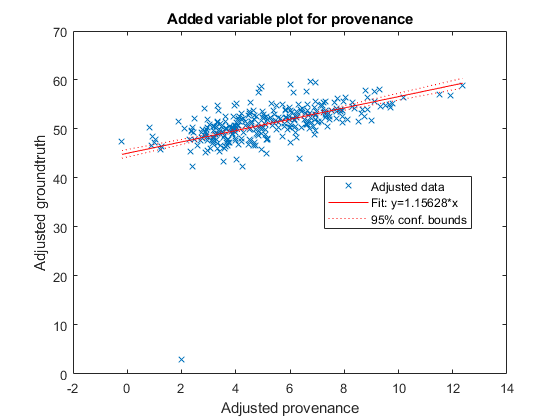
% Effect of census on response variable (groundtruth)

plotAdded(mdl,[4])



% Effect of provenance on response variable (groundtruth)

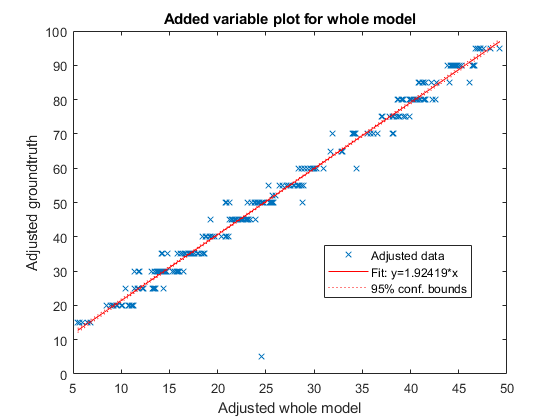
plotAdded(mdl,[5])



The dominant variable can be found by the slope of the fitted line. As the notes has high slope, it is more significant feature i.e., the groundtruth is highly dependent on the notes. It is also proved in terms of p values and squared error.

% Effect of all features on response variable (groundtruth)

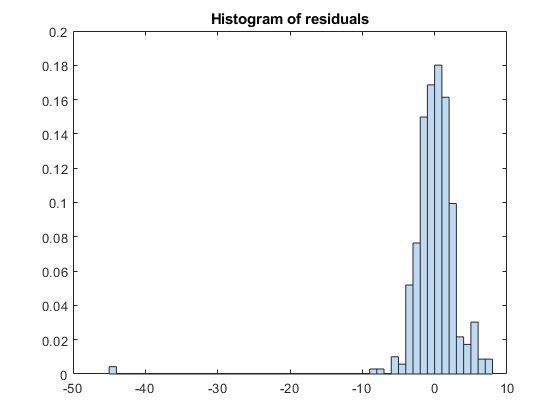
plot(mdl)



The fitted line is how the model, as a group of variables, can explain the response variable. The slope of the fitted line is not close to zero, and the confidence bound does not include a horizontal line, showing that the model fits better than a degenerate model consisting of only a constant term. The test statistic value shown in the model display (F-statistic vs. constant model) also shows that the model fits better than the degenerate model.

% Identify the outlier

plotResiduals(mdl)



A few residuals that are smaller than -10 are found. If the model has very high p values, then it might have showed more outlier. The histogram is free of outlier.

Predict responses to new data: Predicts responses to the test data set (testData).

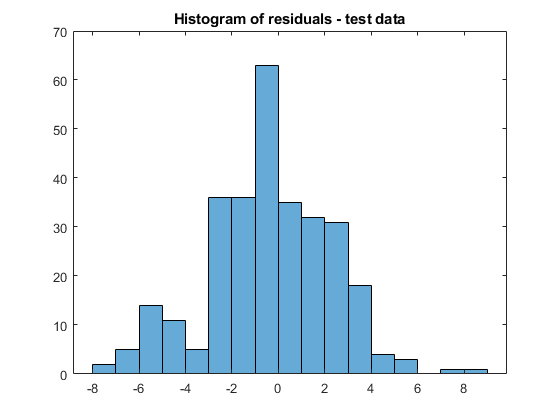
ypred = predict(mdl,testData);

% Plot the residual histogram of the test data set

errs = ypred - testData.groundtruth;

histogram(errs)

title("Histogram of residuals - test data")



errs(isoutlier(errs,'grubbs'))

ans =

0×1 empty double column vector

There is zero outlier in the test set too.

% Error plot

boxplot(ypred,testData.groundtruth, 'PlotStyle',"compact")

