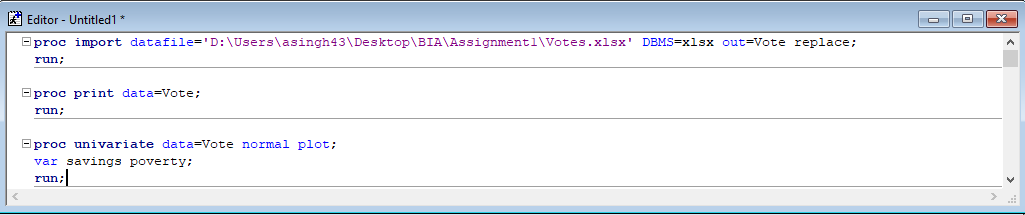
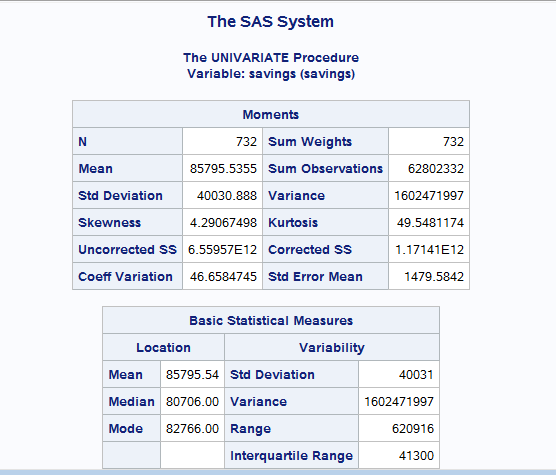
# Data Exploration and Multiple Linear Regression (MLR) using SAS

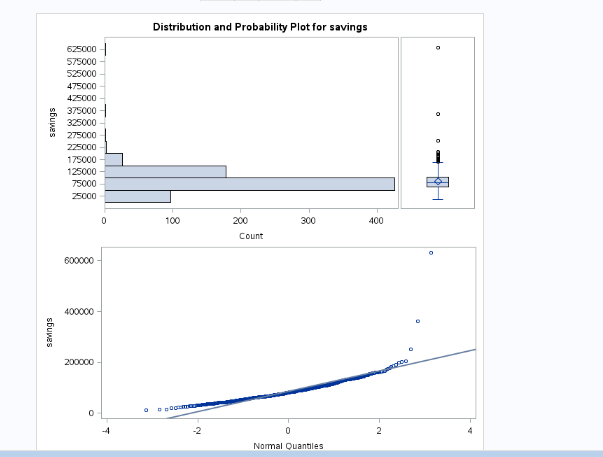
1. Generating box-plots of the savings (Mean Savings in $) and poverty (% in poverty) attributes and identifying the cutoff values for outliers.



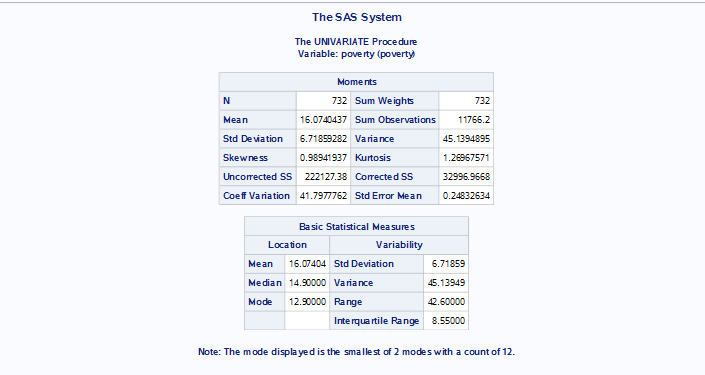
We will get our Box-Whisker Plot for savings and poverty which looks like this-

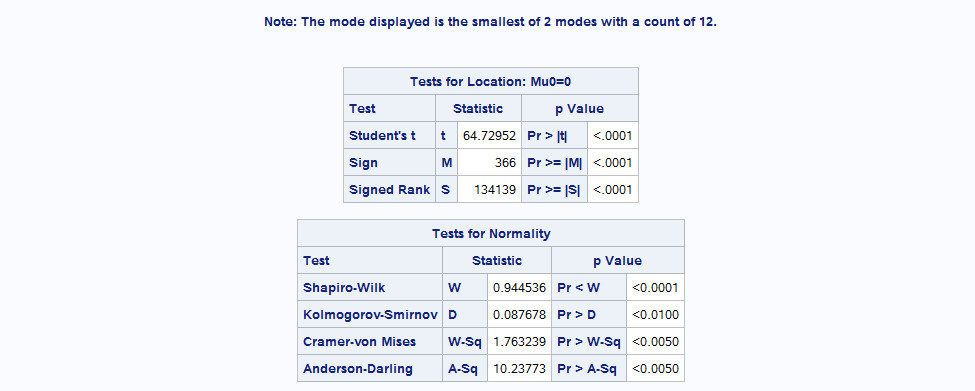


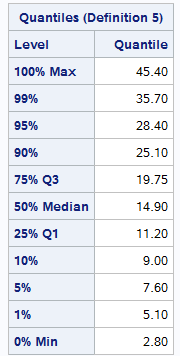
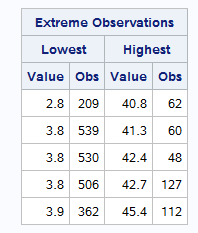


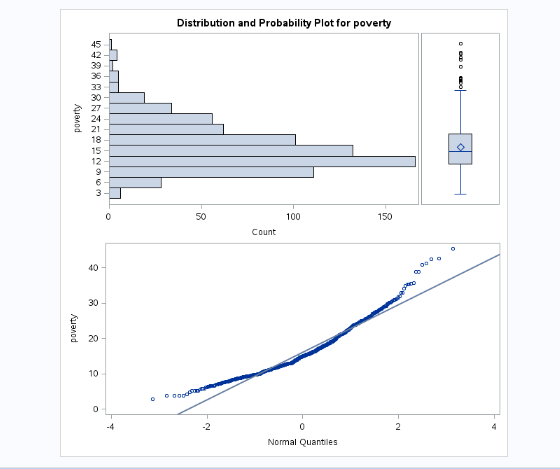


**For Poverty-**







This shows that they have outliers which we needs to remove. For that, we will calculate cutoff for the data.

**For Savings-**

**Cut off Q4 (as we can see that outliers are towards Q4)**=1.5\*Interquartile Range + Q3

=1.5 \* 41300 +103658.5

= **165,608.5**

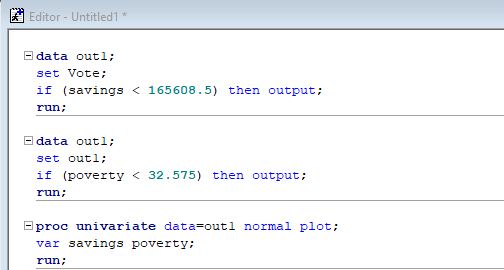
**For Poverty-**

**Cut off Q4** =1.5\*Interquartile Range + Q3

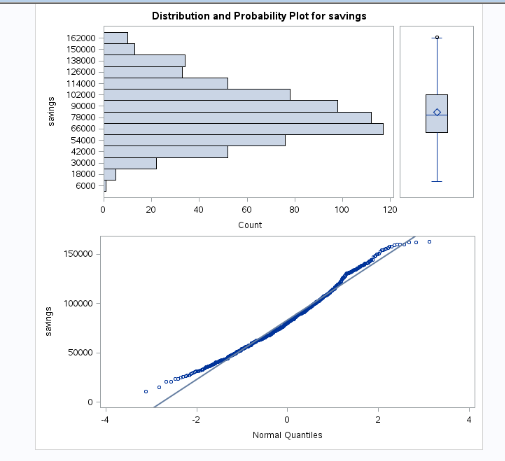
=1.5 \* 8.55 +19.75

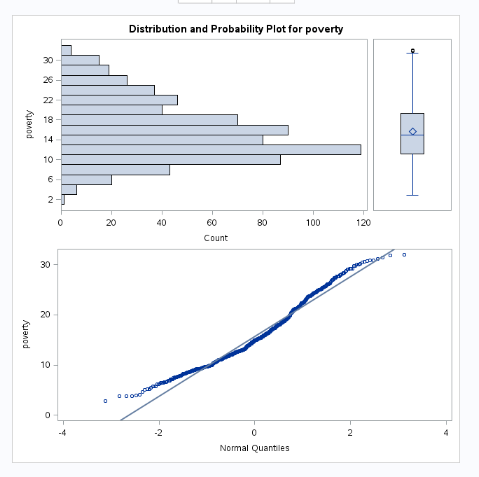
= **32.575**

Now, since we get our cut off value, we will drop the outliers.



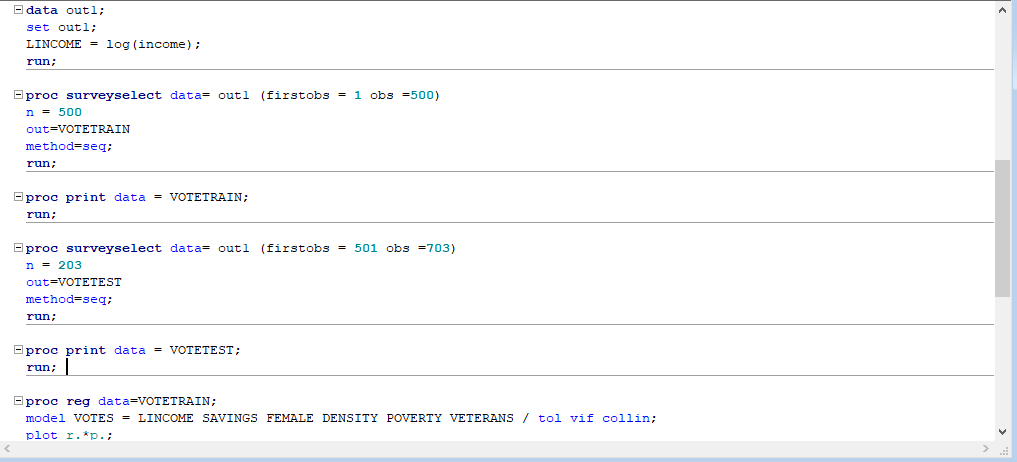
After this, we can generate the plot and see that now it will not have any outliers.

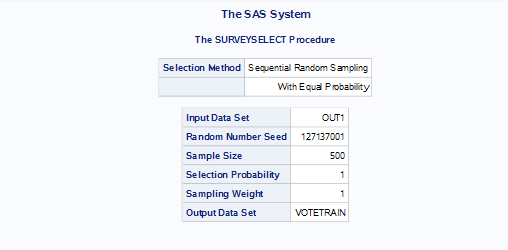


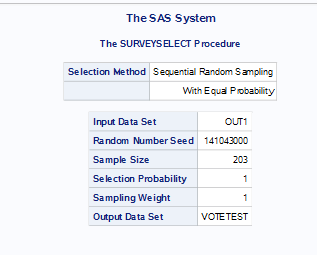


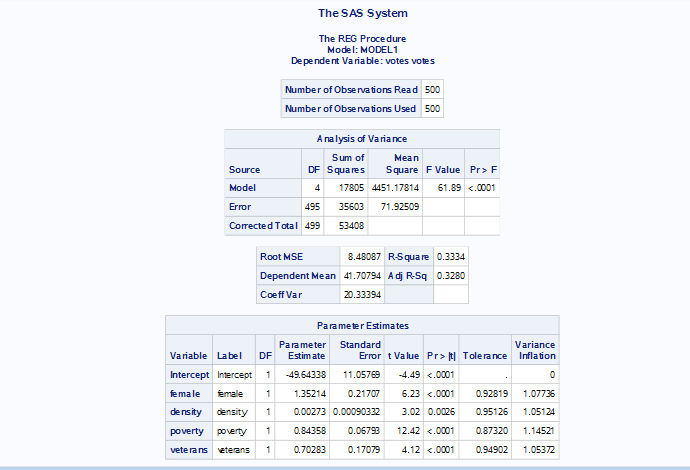
1. Trying to fit an MLR to this dataset, with VOTES as the dependent variable. INCOME has somewhat longish tail, so we will take a log transform. We are keeping the first 500 records as a training set,call it VOTETRAIN which will be used to fit the model and the remaining 232 will be used as a test set VOTETEST. Using equation

VOTES = LINCOME + SAVINGS + FEMALE +DENSITY +POVERTY + VETERANS









The coefficient obtained in my model is –

Beta0 = -37.70371

Beta1 = -0.74106

Beta2 = 0.00002667

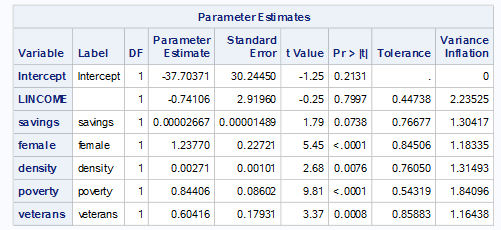
Beta3 = 1.23770

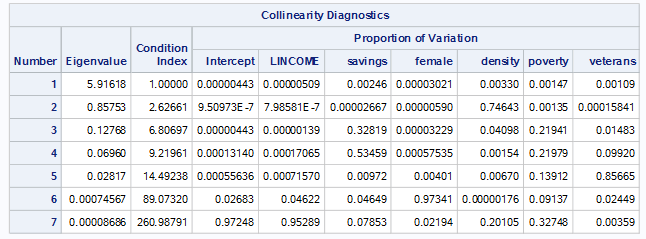
Beta4 = 0.00271

Beta5 = 0.84406

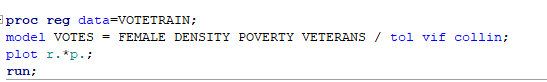
Beta6 = 0.60416

As we can observe, variables LINCOME and savings have p-value 0.7997 and 0.0738 respectively. Having tested these variables for p>0.5, it is better to drop them further.

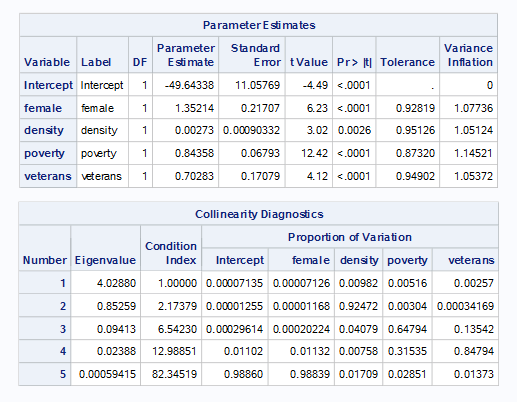


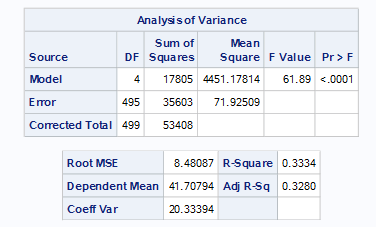


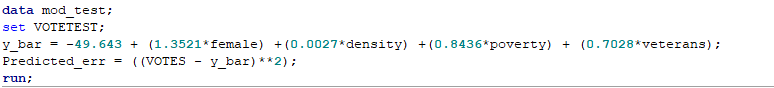
After dropping the variables,



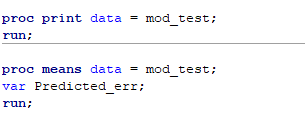
The p-value after dropping them-



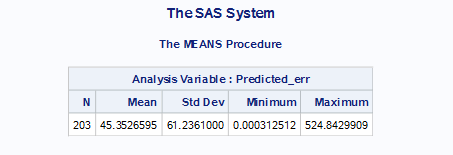




Calculating the MSE(Mean Square Error)-



Output for this-

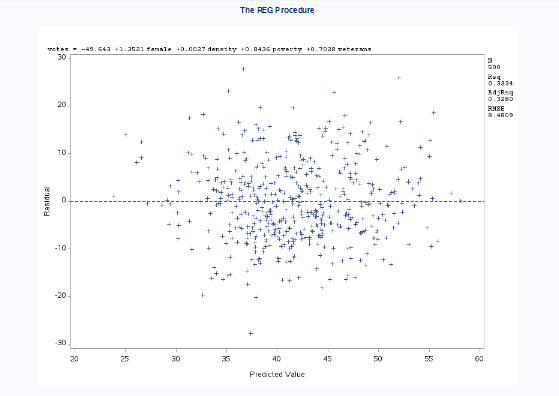


RMSE for Train Data Set = 8.48

Mean for test data MSE= 45.35

RMSE for Test Data Set = 6.734

Therefore, we see there is an decrease in the MSE by **1.746.**



As we look at the residual plot, it is confirmed that the model is not biased at all. The graph between Residual and Predicted value is reasonably distributed and therefore my MLR model is fit for this problem.