Data Exploration and Multiple Linear Regression (MLR) using SAS

The “Votes" data set, records properties of 732 County level results for percent voting for Bill Clinton in 1992 Presidential Election and Demographic variables. For a description of the data, see Canvas (votes data and attribute information). Main task is to check if the vote percentage is dependent on demographics data.

1. Generate box-plots of the savings (Mean Savings in $) and poverty (% in poverty) attributes and identify/remove the cutoff values for outliers.

2. Try 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, (use LINCOME = log(INCOME)) and then use LINCOME as one of predictor. Keep the first 500 records as a training set (call it VOTETRAIN) which you will use to fit the model; the remaining 232 will be used as a test set (VOTETEST). Use only the following variables in your model:

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

(a) Report the coefficients obtained by your model. Would you drop any of the variables used in your model (based on the t-scores or p-values)?

(b) Report the MSE obtained on VOTETRAIN. How much does this increase when you score your model on VOTETEST?

(c) (Bonus 2 points). Do you think your MLR model is reasonable for this problem? You may look at the distribution of residuals to provide an informed answer.