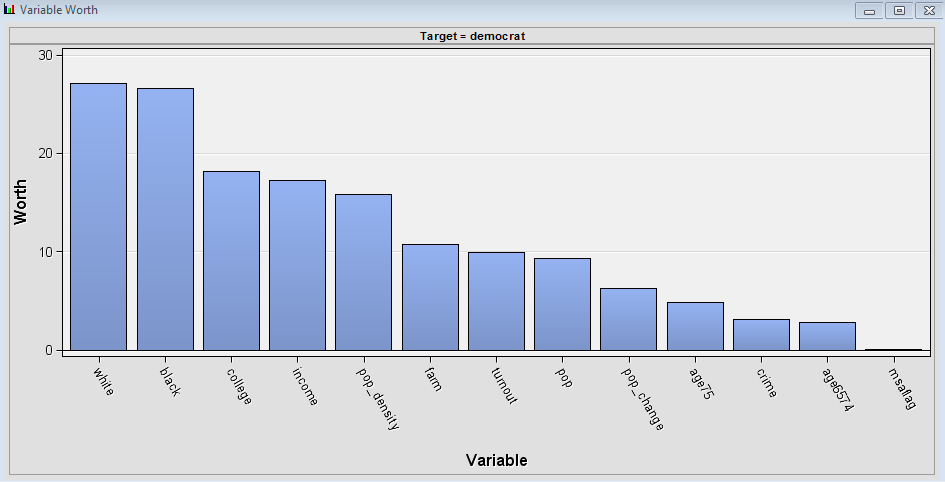
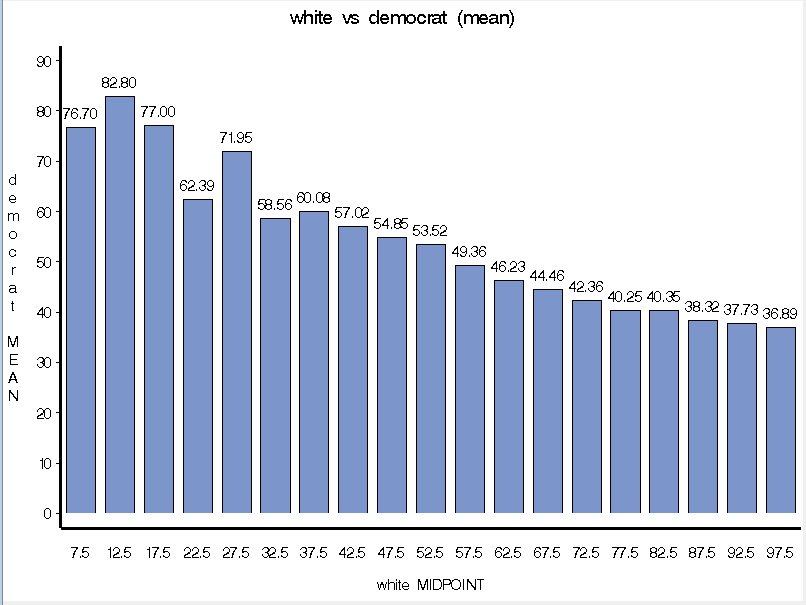
**ISE 365/465 – Data Mining Homework 2 (Due Mar. 10, 2016 in class)**

**Name: Bolun Xu**

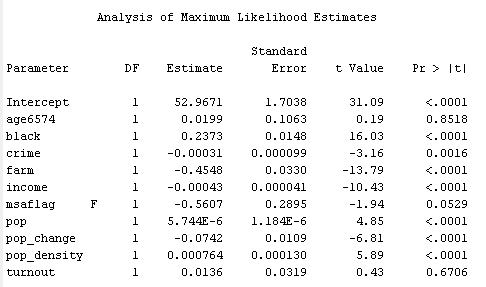
1. Provide the following (hand in the items after b on paper in class):
   1. Your project file electronically– (10 points for general correctness)
   2. A screen dump of the worth graph from your StatExplore Node (5 points)



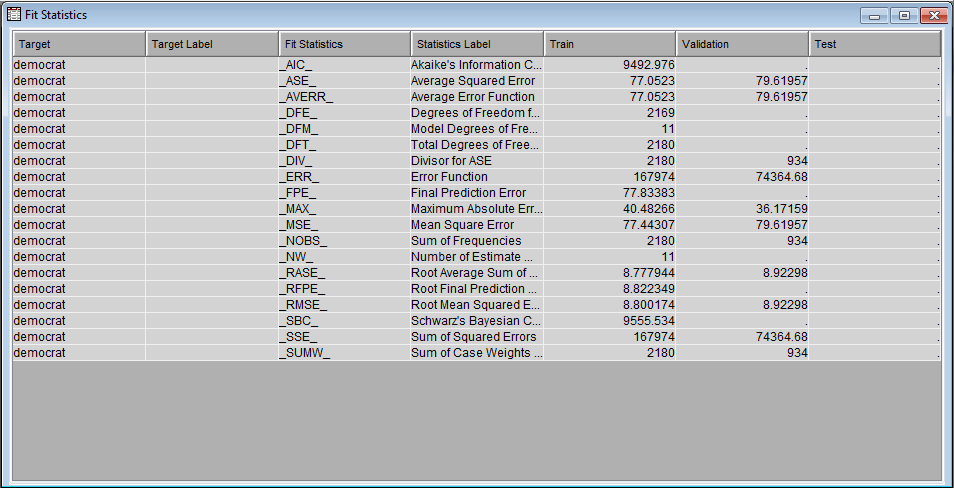
* 1. A screen dump of the multiplot graph for the variable with the highest worth versus democrat from 9b. (5 points)



* 1. The SAS output of the regression node showing the input variables in MODEL2 to predict democrat and their coefficients and statistical significance (5 points)



* 1. A screen dump of the Fit Statistics Window from the Regression Node for MODEL2 (5 points)



* 1. Answers to the following questions:
     1. Which input variables have missing values? In 1 or 2 sentences, how does this affect the regression in Enterprise Miner? (5 points)

A: msa, pmsa, republican, Perot and turnout have missing values. Loss of training data lowers the confidence level of regression.

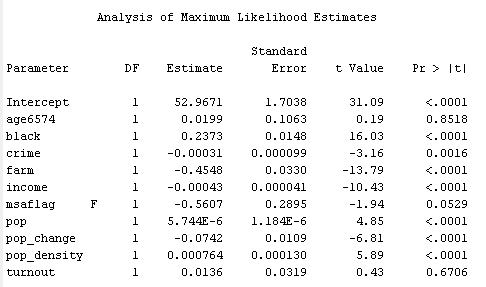
* + 1. Which IBM SPSS Modeler node could have been used to calculate all the correlations in step 8? (5 points)

A: Statistic node, Auto Numeric node and PCA/Factor node.

* + 1. Interpret the Multiplot graph in 9c. (5 points)

A: With more percentage of white, the mean democrat is more likely to be lower.

* + 1. Interpret the table in 9d. Make sure to explain the effect and significance of each input variable relative to the target in words. (15 points)

A: 

Low Signficance (p-value <0.05) indicates input has significant linear relationship with dependent y target variable, so Intercept, black, crime, farm, income, pop, pop\_change and pop\_density have a significant linear relationship with dependent y target variable. Others don’t have significant linear relationship.

If we increase age6574 by one unit, the democrat will increase by 0.0199.

If we increase black by one unit, the democrat will increase by 0.2373.

If we increase crime by one unit, the democrat will decrease by 0.00031.

If we increase farm by one unit, the democrat will decrease by 0.4548.

If we increase income by one unit, the democrat will decrease by 0.00043.

If we turn msaflag from F to T, the democrat will decrease by 0.5607.

If we increase pop by one unit, the democrat will increase by 5.744E-6.

If we increase pop\_change by one unit, the democrat will decrease by 0.0742.

If we increase pop\_density by one unit, the democrat will increase by 0.000764.

If we increase turnout by one unit, the democrat will increase by 0.0136.

* + 1. What is the Validation Root Mean Square Error and Adjusted R-square for MODEL1 and MODEL2? Is a model with lower Root Mean Square Error and higher adjusted R-square always the better model? Explain why or why not. (10 points)

A:Model 1

Validation Root Mean Square Error = 8.86

Adjusted R-square = 0.3456

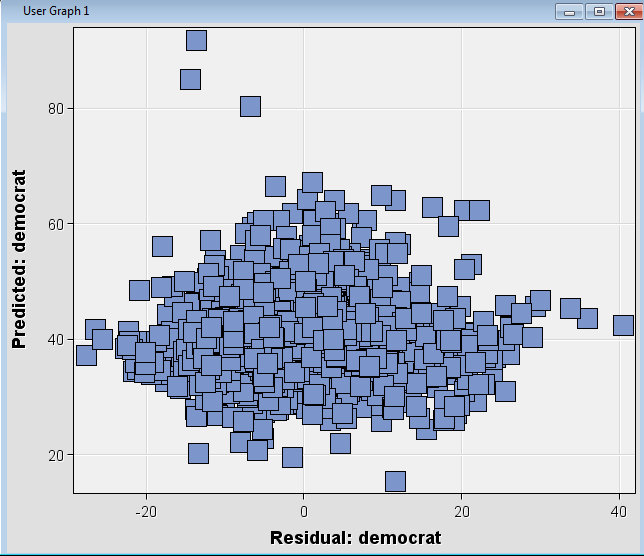
Model 2

Validation Root Mean Square Error = 8.92

Adjusted R-square = 0.3375

A model with lower Root Mean Square Error is always the better model. For adjusted R-square, it depends. Adjusted R-square will natural raise with increased number of variables. So if we fix the number of variables, then higher adjusted R-square makes better model. Otherwise, with higher adjusted R-square, it is not “always” a better model.

* + 1. Are the homoscedasticity of errors versus prediction and the normality of residuals assumptions for linear regression met in MODEL2? Explain how you determined this. **You must show evidence for your answers with output from your Enterprise Miner project.** (10 points)

A:

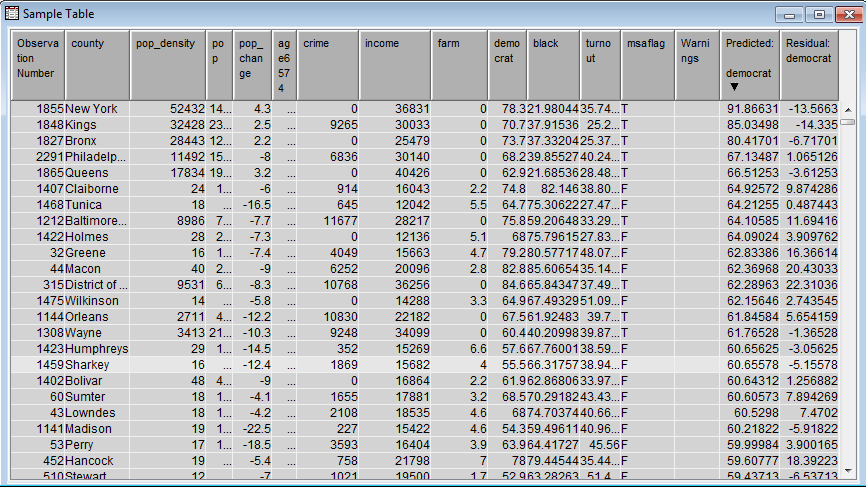
Residuals vary randomly around zero and the spread of residuals is about the same throughout the plot. Thus, we can say that residuals satisfy homoscedasticity assumption across predictions.

* + 1. If you chose to hold out all the counties in Kansas as your validation data set and train your model on all other counties, is this a good idea? Explain your answer. (10 points)

A: No, I don’t think so. Votes in the same state have some kind of trend. That is different in another state. It will lead to a deviation in prediction.

* + 1. Which county has the highest predicted Democrat % vote in your validation data set? Do you believe the prediction for this county is less or more accurate than the average predictions for other counties in your validation data set? Explain your answers referencing output from your model. **You must show evidence for your answer with output from your Enterprise Miner model.** (10 points)

A:



New York has the highest predicted Democrat % vote. The prediction for this county is less accurate than the average predictions for other counties in that the root mean square error is 8.92 less than 13.5663. This shows the conclusion above.