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

**The purpose of this homework is to allow you to use Enterprise Miner to solve a prediction problem using linear regression. Variable Description is posted on the Course Site. The data is from the 1992 US Presidential Election. Your job is to determine the characteristics of counties who are likely to vote for a Democrat.**

**Follow these steps:**

1. Create a project and read the “counties” SAS File into Enterprise Miner after creating a data source for the file. Set the “democrat” variable to “Target” role
2. Create a new variable called “msaflag” using the SAS Code Node assigning a value “T” if msa > 0 and a value of “F” otherwise. Use the following code:

data &EM\_EXPORT\_TRAIN;

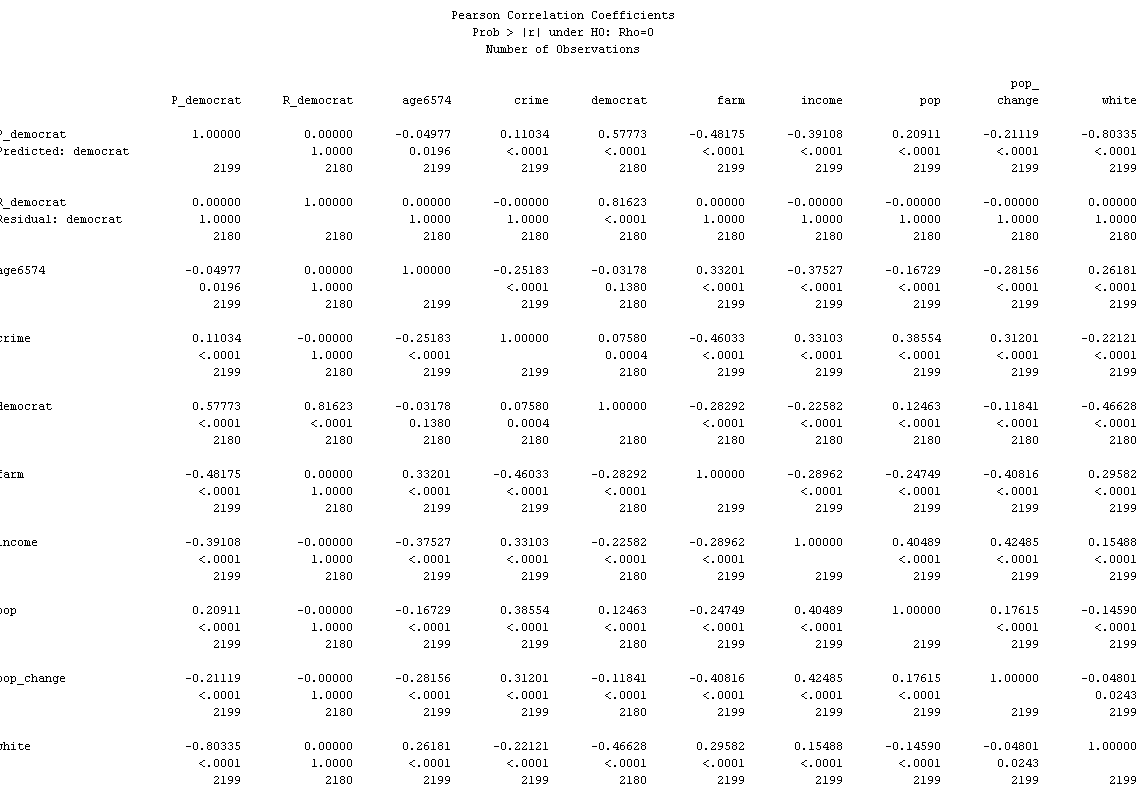
set &EM\_IMPORT\_DATA;

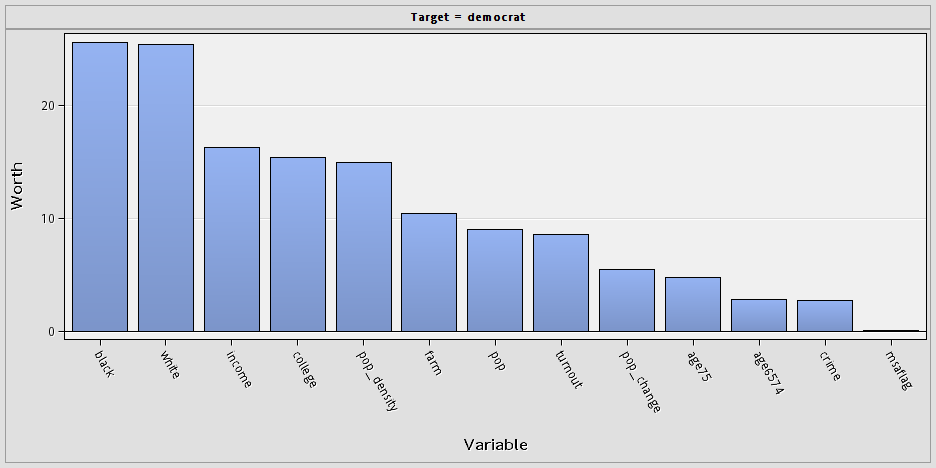
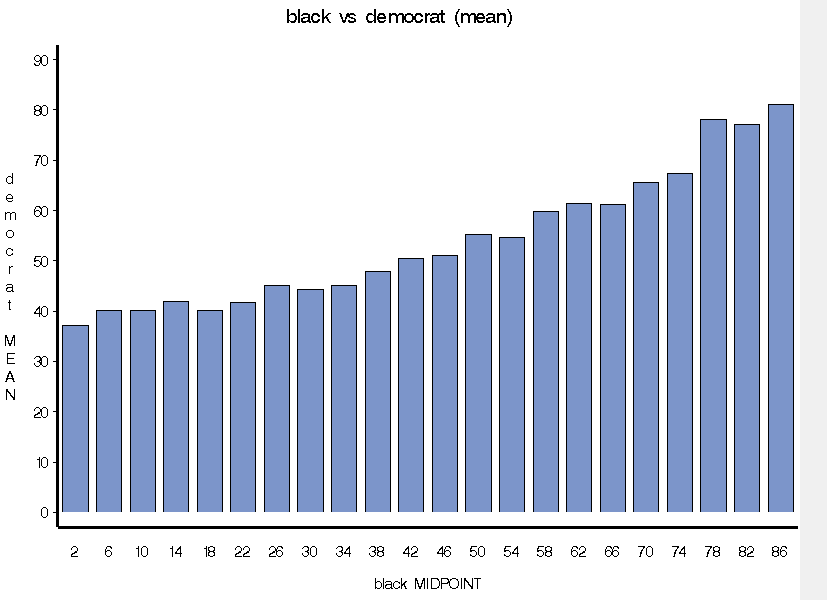
IF msa > 0 THEN msaflag = "T";

ELSE msaflag = "F";

run;

1. Use the Drop Node to drop state, msa and pmsa variables – Make sure to drop these from the table as well as metadata using the appropriate node setting to do this. These dropped variables should not show up in the later node output or models (or the screenshots you provide for the homework described below). YOU MUST ALSO DROP OTHER VARIABLES THAT ARE NOT VALID PREDICTORS FOR DEMOCRAT % VOTE so you don’t use variables that are proxy variables for the target in your final models.
2. Run a Stat Explore Node.
3. Run a Multiplot Node
4. Partition the data into a training set with 70% of the data and a validation set with 30% of the data using all other default settings in the Partition Node.
5. Run the Regression Node using Linear Regression with Stepwise model selection to predict democrat % vote. Use all default settings in the Regression Node other than the settings needed to execute the previous sentence requirements. This is MODEL1.
6. Now create another regression model predicting democrat % vote **making sure there are no interval variables with correlations with an absolute magnitude above 0.5 with each other in the final model as predictors. Hint: look at Lab 1 for one way to determine how to evaluate correlations. This is MODEL2.**



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) 
   3. A screen dump of the multiplot graph for the variable with the highest worth versus democrat from 9b. (5 points) 
   4. The SAS output of the regression node showing the input variables in MODEL2 to predict democrat and their coefficients and statistical significance (5 points) The selected model is the model trained in the last step (Step 8). It consists of the following effects:

Intercept age6574 crime farm income msaflag pop pop\_change white

Analysis of Variance

Sum of

Source DF Squares Mean Square F Value Pr > F

Model 8 85016 10627 135.96 <.0001

Error 2171 169697 78.165343

Corrected Total 2179 254713

Model Fit Statistics

R-Square 0.3338 Adj R-Sq 0.3313

AIC 9511.2229 BIC 9513.3053

SBC 9562.4066 C(p) 8.0527

Type 3 Analysis of Effects

Sum of

Effect DF Squares F Value Pr > F

age6574 1 536.7816 6.87 0.0088

crime 1 1499.3320 19.18 <.0001

farm 1 16678.1826 213.37 <.0001

income 1 6017.8935 76.99 <.0001

msaflag 1 533.4700 6.82 0.0091

pop 1 3114.7505 39.85 <.0001

pop\_change 1 5453.0864 69.76 <.0001

white 1 22724.1065 290.72 <.0001

Analysis of Maximum Likelihood Estimates

Standard

Parameter DF Estimate Error t Value Pr > |t|

Intercept 1 73.5137 1.6079 45.72 <.0001

age6574 1 0.2638 0.1007 2.62 0.0088

crime 1 -0.00044 0.000100 -4.38 <.0001

farm 1 -0.4680 0.0320 -14.61 <.0001

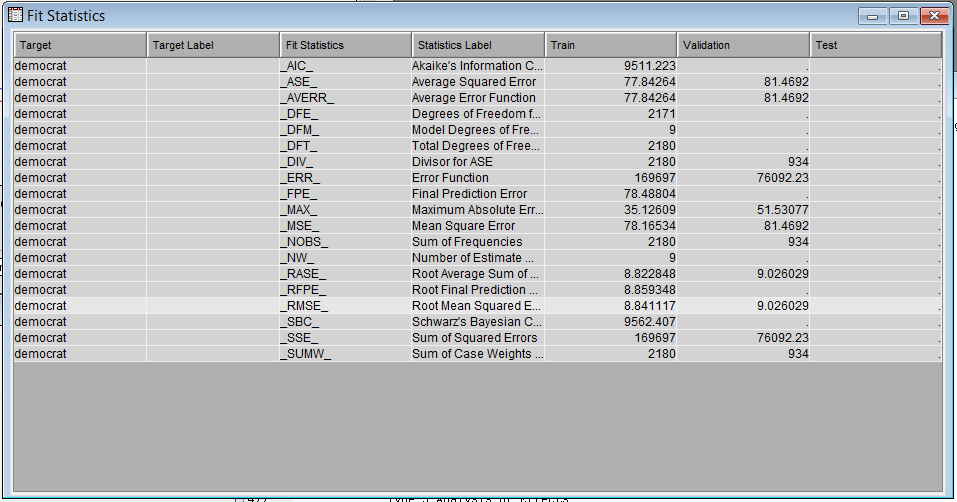
income 1 -0.00035 0.000039 -8.77 <.0001

msaflag F 1 -0.7571 0.2898 -2.61 0.0091

pop 1 6.878E-6 1.09E-6 6.31 <.0001

pop\_change 1 -0.0908 0.0109 -8.35 <.0001

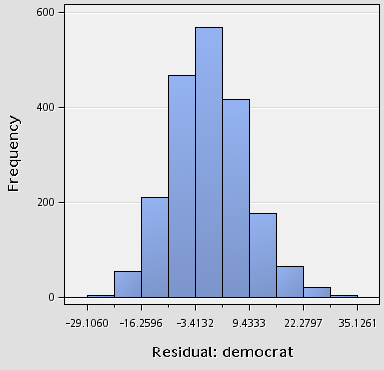
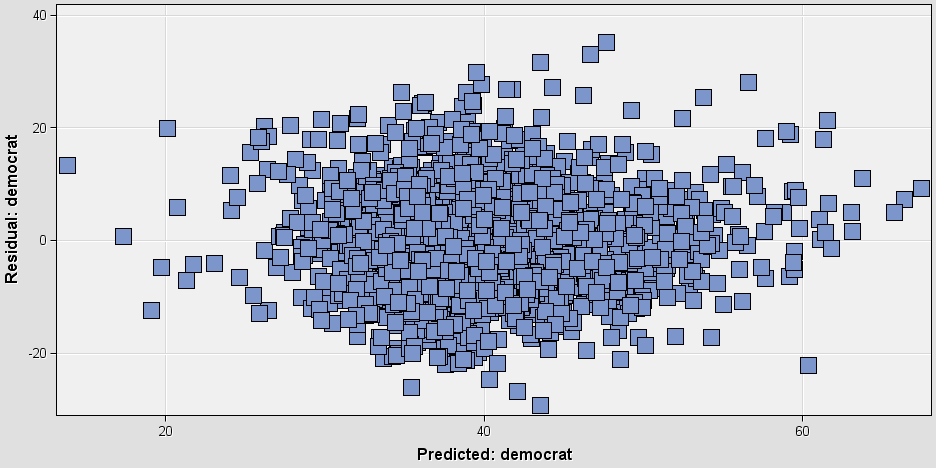
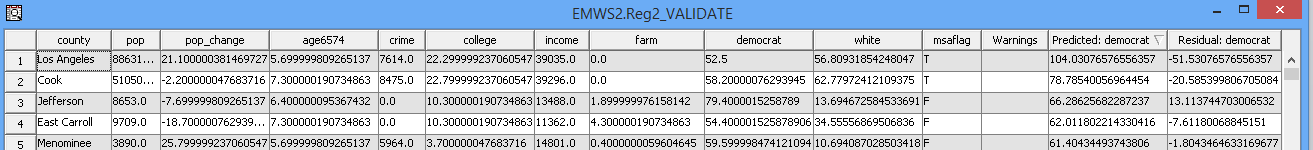
white 1 -0.2459 0.0144 -17.05 <.0001

* 1. A screen dump of the Fit Statistics Window from the Regression Node for MODEL2 (5 points) 
  2. 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)**MSA, Republican, Democrat, and Perot. We dealt with the MSA missing values with the new flag we created so we could drop the MSA variable. The other 3 variables are missing in the same rows, so we will not be able predict democrat in these 27 rows unless we impute the missing value. Since Republican and Perot are proxies for democrat, we dropped them as well. However, if they were valid predictors the rows where they were missing would have been dropped unless we imputed the missing values.**
     2. Which IBM SPSS Modeler Node could have been used to create the msaflag variable? (5 points) **Statistics Node**
     3. Interpret the Multiplot graph in 9c. (5 points) **It appears that rising % black population has an increasing likelihood to vote for democrat. It looks like a fairly linear relationship.**
     4. 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) **All variables shown are statistically significant (p <=0.05). The intercept baseline for % democrat vote by county is 73.5 percent, an increase of 1% in population between ages 65 to 74 increases democrat % by 0.26, an increase of 1 crime per 100,000 residents increases county democrat % by 0.00044, every percent increase in the farm population in a county decreases democrat % by 0.468, each dollar increase in income decreases democrat by .00035%, having msaflag false decreases democrat % by 0.75 compared to msa\_flag = true, an increase of one person in population increases democrat % by 6.878E-6**  **(YOU MAY WANT TO CREATE THIS VARIABLE TO REPRESENT EVERY 1000 population increase TO MAKE IT MORE INTERPRETABLE), every percent increase in population change decreases democrat % by 0.09, and each increase of 1 percent of white population decrease democrat % by 0.245.**
     5. 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)

**Model1: Adj R-square = .3456, RMSE = 8.86**

**Model2: Adj R-square = .3313, RMSE = 9.03**

**While these statistics are used to guide model development and evaluate model quality, lower RMSE and higher adjusted R-square are not the only considerations in choosing the best model. Issues such as people issues, model complexity, practical business issues must also be taken into account.**

* + 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)  **Looking at the histogram of residuals, the normality assumption is met from visual inspection. We will also say the homoscedasticity assumption is met although there is a bit of concern at the upper end of the predicted scale as the variance seems to reduce relative to the residual.**
    2. 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) **This is not a good idea as Kansas may not be representative of all the data and would show a large inaccuracy in the model when the model may be good for the entire US. If you held out Kansas because you were interested in predicting Kansas with data from the rest of the US, this would also be a problem as the training data is not representative of Kansas and may not predict it very well. If you could find variables in the training data that were consistent in value for the other 49 states and Kansas, the model may be acceptable.**
    3. 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) **Los Angeles county is the highest with 104% Democrat, but it has a high residual value of -51 which means this prediction is less accurate than the average prediction error of 9.03 for Model 2. Also, the percent is 104% which is a bit worrying since 100% is the theoretical maximum. This model may need to be re-worked to ensure the prediction is 100% or less.**