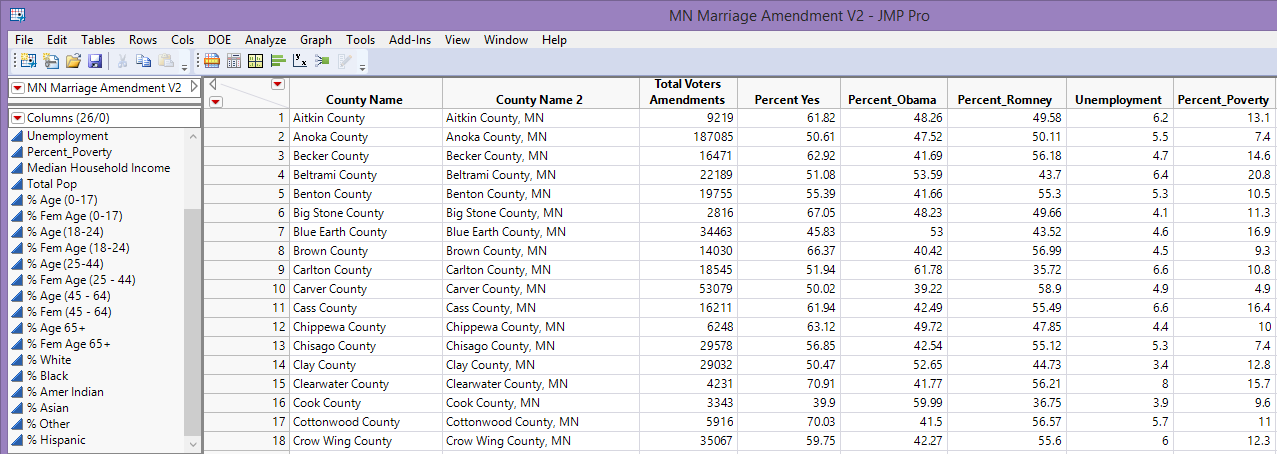
**16.1 – Introduction to Model Selection Procedures**

Model selection procedures in multiple regression automatically, though with guidance, find reasonable sub-models starting with a full model consisting of all candidate terms/predictors. The general principle in model development is to fit the smallest (least complicated) model that adequately explains the response, i.e. the most parsimonious model. Furthermore what constitutes a “good” model depends on the goal of the regression analysis. If the goal is interpretation of the model effects, i.e. interpreting the estimated coefficients, then simpler models will certainly be easier to discuss than an excessively complicated ones. If prediction accuracy is the goal, we also will generally find that simpler models will predict future values of the response more accurately due to the fact that a large complicated model may fit the available data very accurately but will not generalize well to future observations.

We have previously used the Backward Elimination procedure where we start with the full model and eliminate terms one-at-a-time until no further terms can be removed by using a p-value based threshold criterion. Backward Elimination is an example of a Stepwise model selection procedure where terms are deleted (or added) sequentially one-at-a-time. Stepwise selection is generally necessary because the number of possible models when we have a total of nonconstant/non-intercept terms is . As we will see in our first example with 21 predictors/terms this is 2,097,152 possible models!!!

**Example 16.1:** Consider the MN Marriage Amendment data from our course website. This dataset contains several potential variables. The primary goal of this investigation is to identify which (predictor) variables are most associated with the percentage of people that voted for Amendment #1.



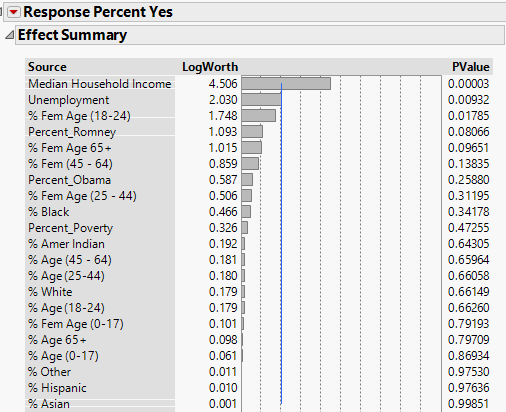
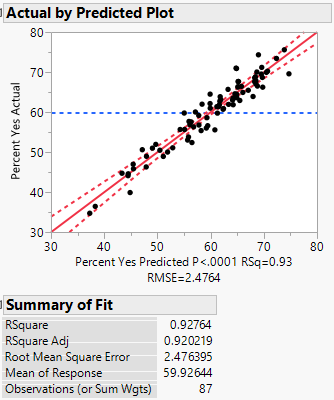
**The variables in the data file are:**

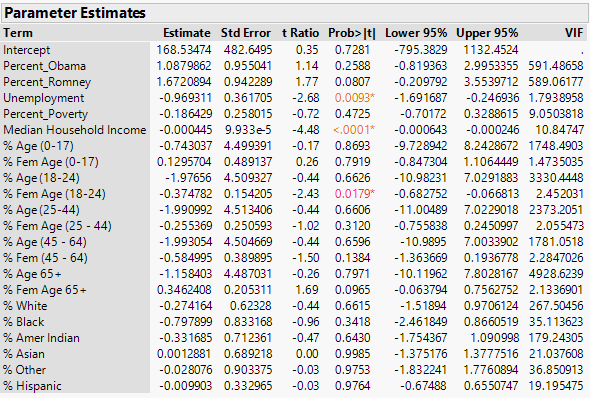
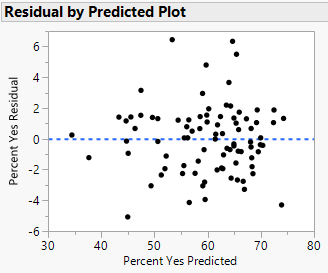
* - percent voting yes for the county
* – percent of votes cast for Obama.
* – percent of votes cast for Romney
* – percent of people unemployed in county
* - % of people in county living below poverty level.
* – median household income in county.
* – percent of pop. between 0 – 17 years of age.
* – 17) – % of pop. between 0 – 17 yrs. of age that are female
* – percent of pop. between 18 – 24 years of age.
* − 24) – % of pop. between 18 – 24 yrs. of age that are female
* – percent of pop. between 25 – 44 years of age
* – 44) - % of pop between 25 – 44 yrs. of age that are female
* – percent of pop. between 45 – 64 years of age
* – 64) - % of pop. between 45 – 64 yrs. of age that are female
* - percent of pop. 65 or over
* - % of pop. 65 or over that are female
* – percent of pop that is Caucasian
* – percent of pop that is Black
* – percent of pop that is American Indian
* – percent of pop that Asian
* – percent of pop that is Other
* – percent of pop that is Hispanic

The mean and variance functions for our preliminary “full” model is given below:

A summary of the Full Model (using all predictors) is shown below.

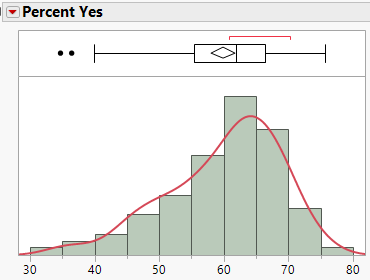
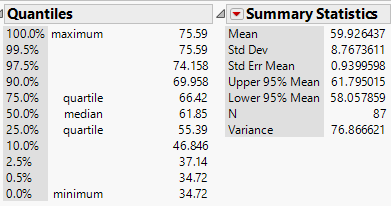
**Comments:**

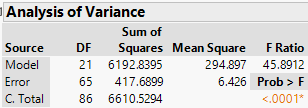
Question

1. Consider the variation in Percent Vote Yes in the following plot.

Average: 59.9, Std Dev = 8.77, Variance = 76.87, Count=87

What proportion of the variation in Percent Vote Yes can be explained by all of these   
 predictors? Discuss.



Comments Regarding R2:

* R2 measures the proportion of variation in the response that can be explained by the predictor variables in the model.
* Concern: Each time we add a predictor variable to the model, the sum of squared error can only decrease; however, the sums of squared total remains constant. Thus, R2 will always increase for each term(s) added to the model. This is true regardless of the worthiness of the added term(s).

**16.2 – Adjusted R-square**

As mentioned above whenever we add a term to a model the R-square increases even if that term is essentially useless. Thus using as model selection criterion is a poor choice. To fix the problem with the we have several other criterion at our disposal that penalize us for including unimportant terms to our model.

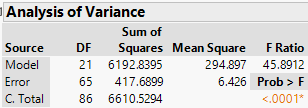
* The **Adjusted-R2** which is given by the quantity below. While it may not be clear from the formula, this measure penalizes us for incorporating unimportant terms to our model.

A second formula -- which is easier to use when computing *Adjusted-R2* is given by

* Adjusted-R2 can \*not\* be interpreted directly as the “proportion of variation in the response being explained by the predictors.” However, this quantity can be used to better quantify our desire to reduce the unexplained variation in the response using a *minimum* number of predictor variables. This concept is commonly referred to as having a **parsimonious** model.

Verify the calculation of the adjusted R2 quantity provided by JMP

ANOVA Table from Full Model



Consider once again the entire list of predictor variables used in our initial model above. Some of these predictors appear to be important in helping to explain the variation in Percent Vote Yes and the others appear not so important.

|  |  |
| --- | --- |
| Entire list of parameter estimates | A list, sorted by importance, is provided by JMP |

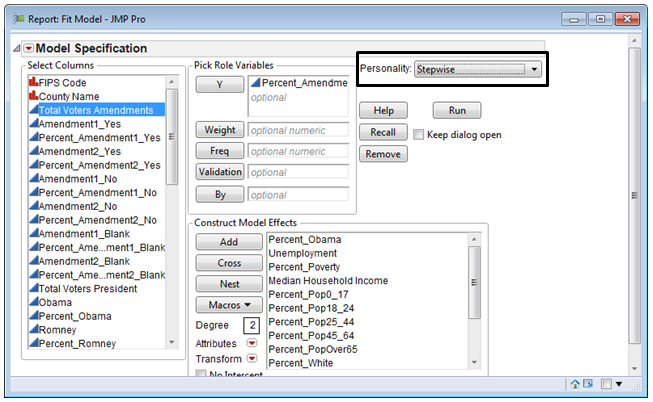
Questions:

1. Which predictor variables appear to be most important in explaining Percent Vote Yes?
2. Why can’t the absolute size of the estimated effect be used to rank the importance of a predictor variable?
3. What is a reasonable way to rank the importance of these predictors? Discuss.

**16.3 – Model Selection Procedures and Criterion**

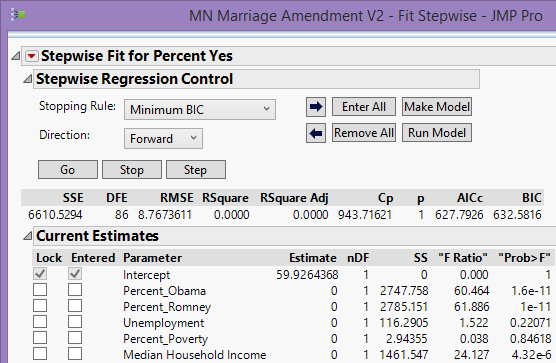
Starting with the mean and variance functions for the full model we can use model selection methods to reduce the size of the model.

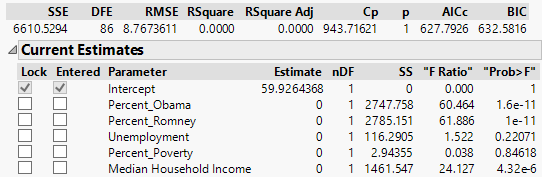
The**Stepwise**model building feature in JMP can be invoked by selecting **Stepwise** from the **Personality** box in the **Fit Model** window as shown below.

.

* Y box (Response): Percentage of People that Voted Yes for Amendment #1
* Model Effect box: The entire list of candidate predictor variables are used as terms. In general, any term type can be considered in a model selection process, e.g. categorical predictors, transformed predictor variables, interaction terms, etc.

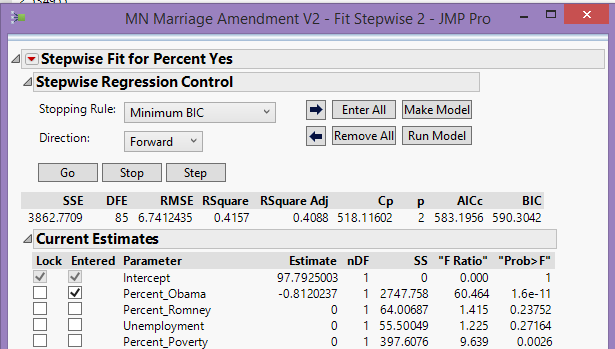
The following Stepwise Fit window is provided.



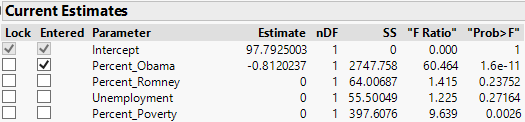
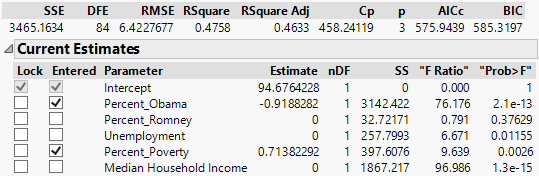
Let’s consider only the bottom portion of the output provided and consider building the model in a forward selection fashion, i.e. start with no terms and add them in stepwise.  


The effect of adding predictor variables to the model can be easily seen by selecting any number of predictor variables in the Entered Column.

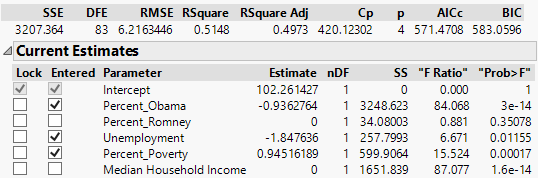
Model #1: Try Percent Obama as a predictor first.



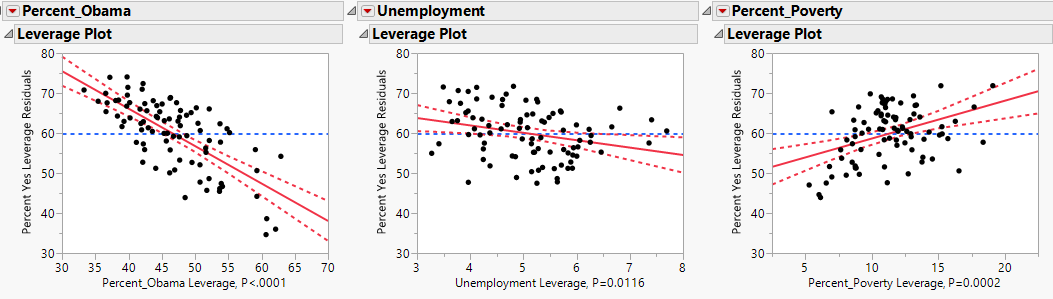
Model #2: Let’s add another predictor variable -- skipping over Unemployment and Percent Romney as both have p-values greater than .05.

  
After adding % Poverty  


Notice that after Percent Poverty was added to the model, the effect of Unemployment now appears to be important. Such anomalies are not uncommon. The and Adjusted- values suggest Unemployment does indeed produce a reduction in the unexplained variation.

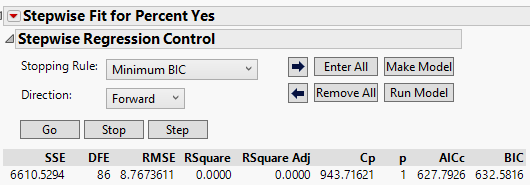


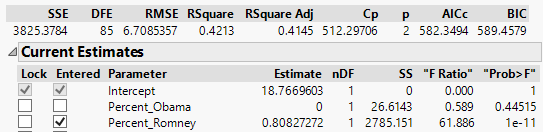
A quick check of the added variable plots suggest that Unemployment is indeed adding something to our model.

\

The Model Selection – Step by Step

JMP has automated the process of adding terms to a model. JMP has incorporated a variety of commonly used methodologies into its procedures. JMP is one of the best software packages that I have used for model selection.



1st Predictor to be put into the model is Percent Romney…

**Criteria Used in the Section Process**

* **Adjusted R2** – larger is better
* **Mallows’ Ck** -- closest to number of terms in model is best, i.e.

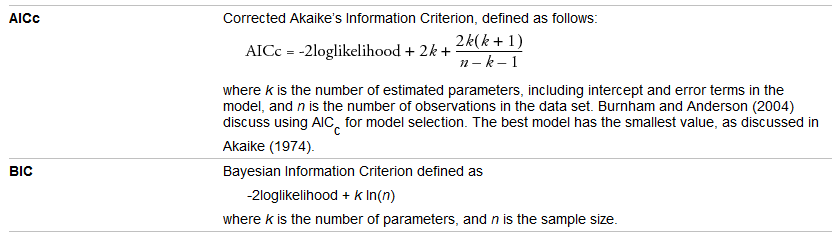
where,

Calculating Mallow’s for the model above with only Percent Romney added to the model, thus .

* **Akaike’s Information Criteria (AIC)** – smaller is better  
  AIC =

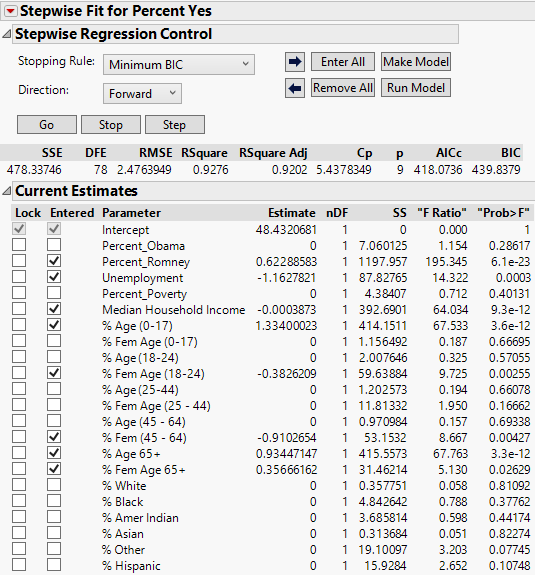
These criterion will generally choose different optimal models.

* **Bayesian Information Criterion (BIC)** – smaller is better  
  BIC =

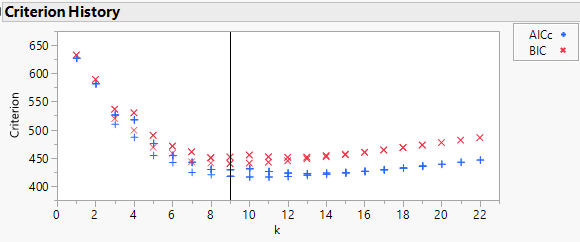


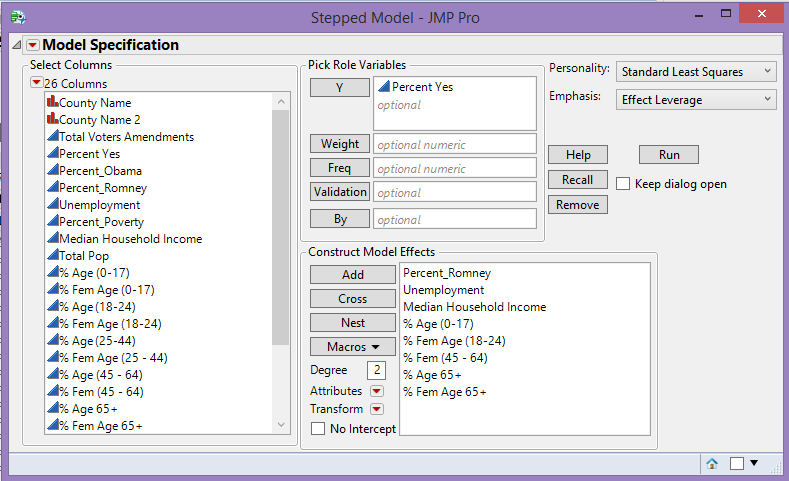
**Automating the Entire Selection Process**

Clicking the Go button instead of the Step button will proceed by **adding one variable at a time** until all predictor variables have been considered. This approach is called **Forward Selection**. After finishing this process, JMP will identify the terms deemed important via this process.

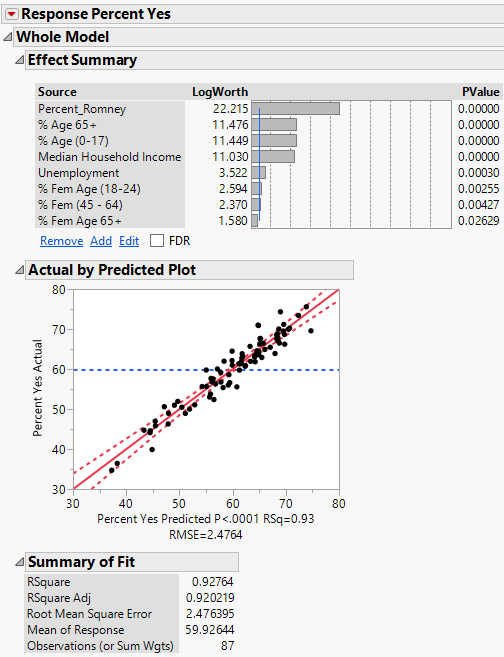
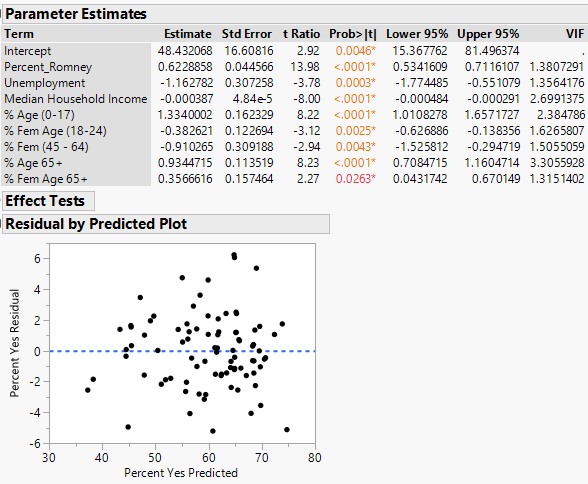


From the red drop down arrow by the heading Stepwise Fit for Percent Yes, you can select **Plot Criterion History**.

Clicking the **Make Model** button sets up a **Fit Model** dialog box with the selected terms in the Model Effects box. Clicking **Run Model** will fit the model and the standard model output.

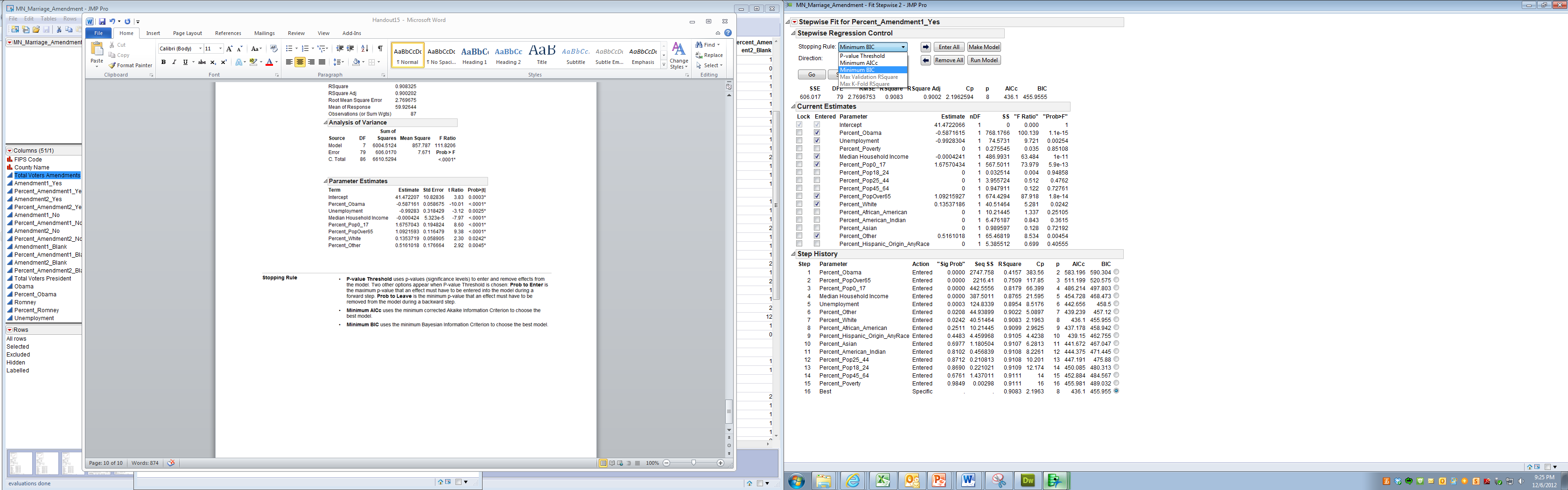


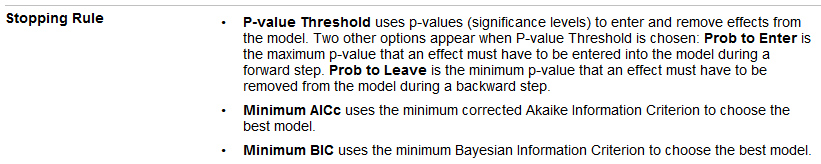
Using the BIC criterion the minimum obtained is for the k = 9 model with the 8 nonconstant/non-intercept terms: Percent Romney, Unemployment Rate, Median HH Income, % Age (0-17),   
% Fem Age (18-24), % Fem Age (45-64), % Age 65+, and % Fem Age 65+.

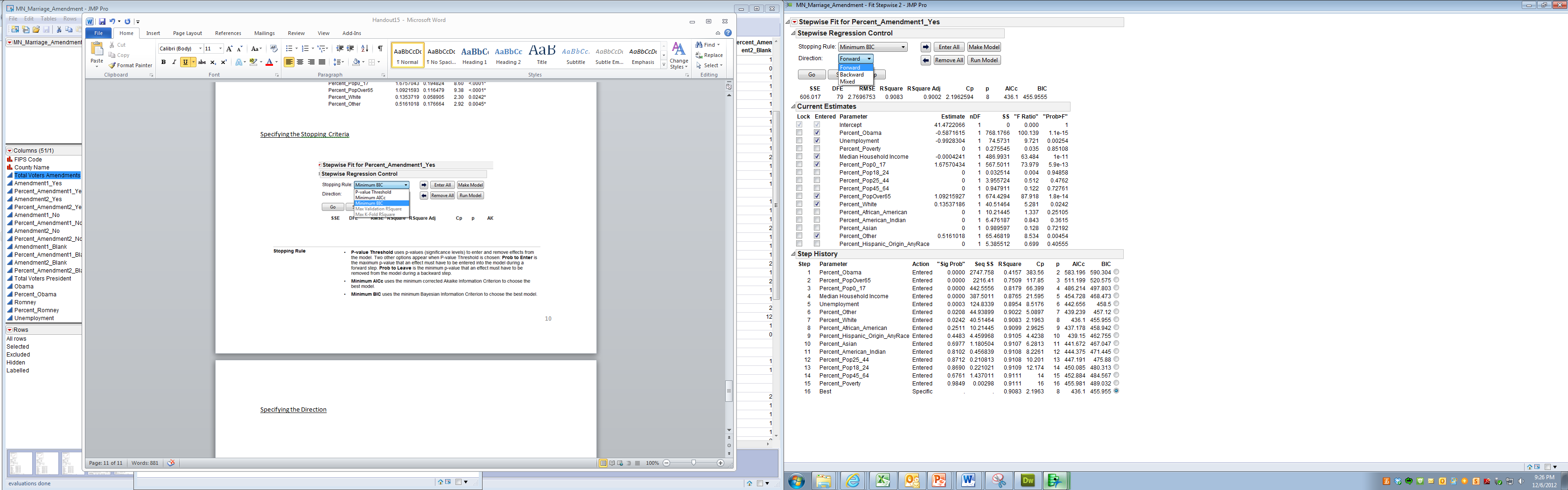
In performing a stepwise model selection in regression we need to determine what criterion will be used judge which model is “best”. The selection process will stop when the addition or deletion of additional terms does not improve the stopping criterion. In JMP we can use BIC, AIC/AICc, or p-value Thresholds to stop the search process. We can also use cross-validation which will discuss later in this section.

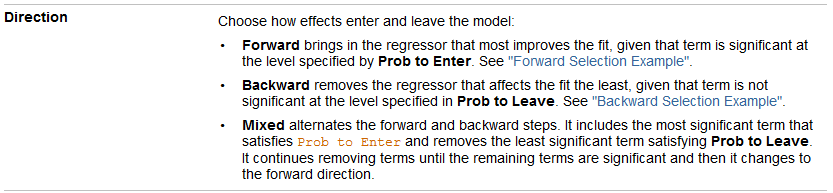
**Specifying the Stopping Criteria**





**Specifying the Direction**





Mixed selection is a hybrid of forward and backward selection methods and is discussed below.

**Procedure for Forward Selection:**   
The procedure is very similar to stepwise regression; however, once a term enters a model, it is never removed.

**Procedure for Backward Elimination:**   
The full model (with all potential terms) is fit first, and the coefficient with the largest p-value is identified. If this p-value is greater than αstay, then the term is dropped from the model. The new model (minus the one term) is then fit, and this process continues until no more terms can be dropped.

**Procedure for Mixed Direction**

1. Choose ≤ (be somewhat generous setting these, say 0.05 up to 0.25).
2. Fit a simple linear regression model for each of the potential predictor/terms. Then, obtain the p-value to test whether or not the regression is useful. Let denote the term with the smallest p-value.

* If p-value < then is retained.
* If p-value > then the procedure ends.

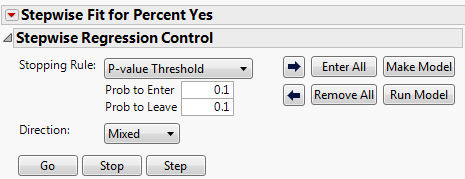
1. Fit all 2-term models, where is one of the pair. Obtain the p-value to test whether the coefficient of the second variable is zero, and let be the variable with the smallest p-value.

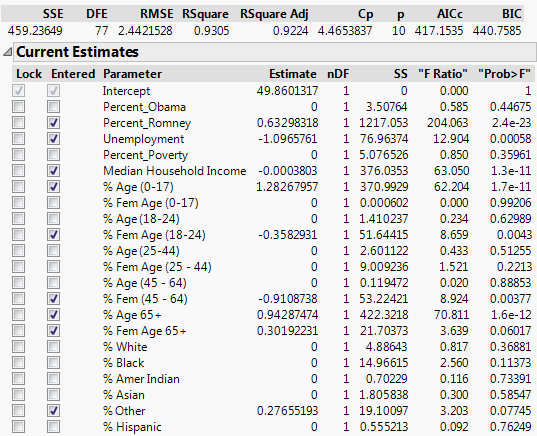
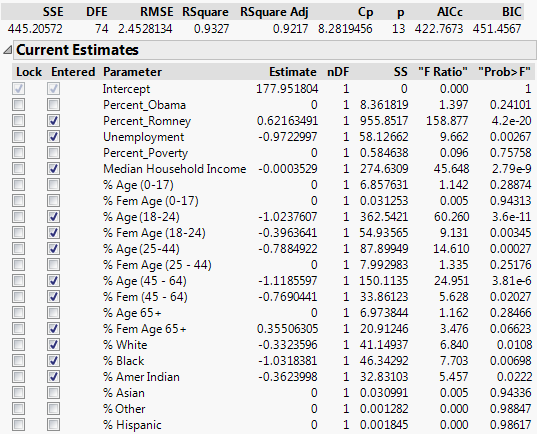
* If p-value > then the procedure ends.
* If p-value < then is retained. Then, we check to see if is still needed in the model. If p-value >, then is removed and we go to the beginning of Step 3. If the p-value is less than , then and are both retained.

1. Examine which term is the next candidate for addition, and then examine whether any other terms already in the model should be dropped. This procedure continues until no terms can be either added or deleted.

**Example 16.1 – MN Marriage Amendment (cont’d)**

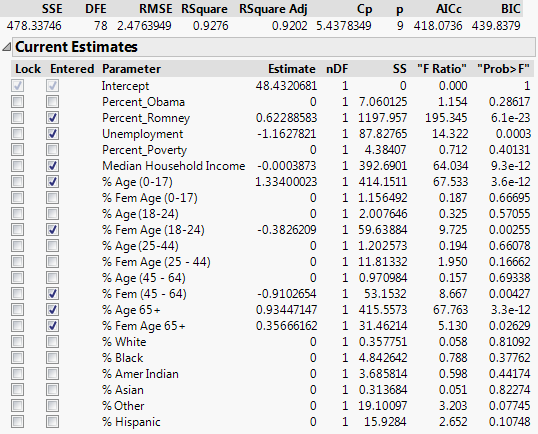
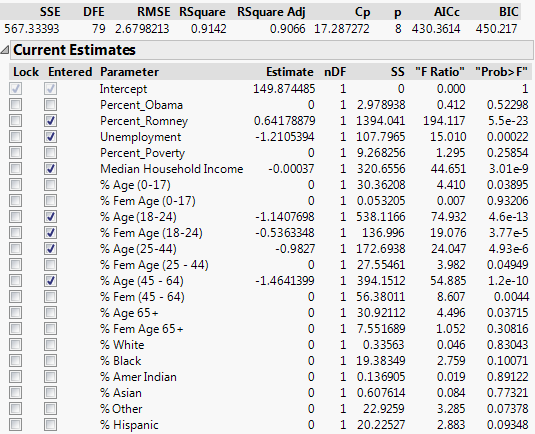
We will use the **Mixed Direction** method for these data using p-value thresholds and .

  
Starting the Mixed Direction selection process with nothing in the base model and starting with the full model containing potential predictors produces results below.

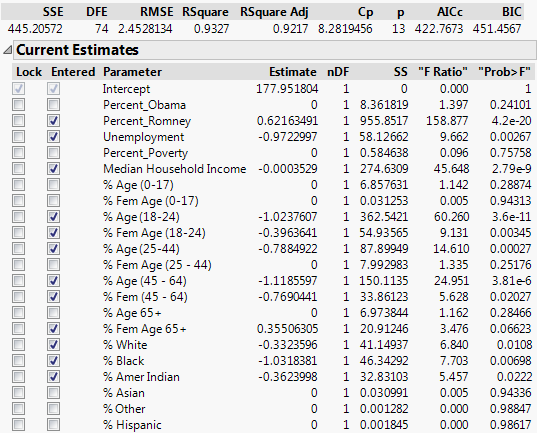
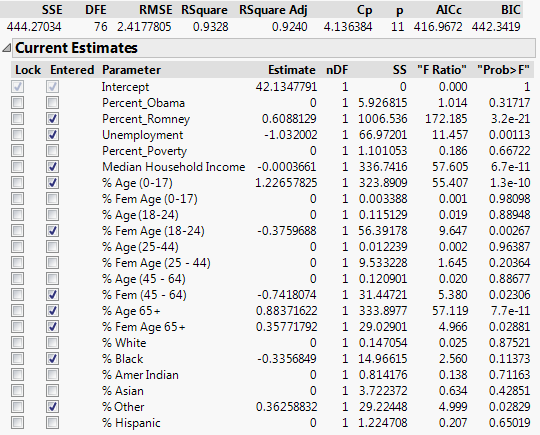
**Starting with Intercept Only (Remove All)** **Starting with Full Model (Enter All)** 

We can see that the two model do not agree! Thus direction matters in when using the Mixed Direction method. The AIC and BIC can also suffer from this problem as demonstrate below for the models chosen using the BIC criterion.

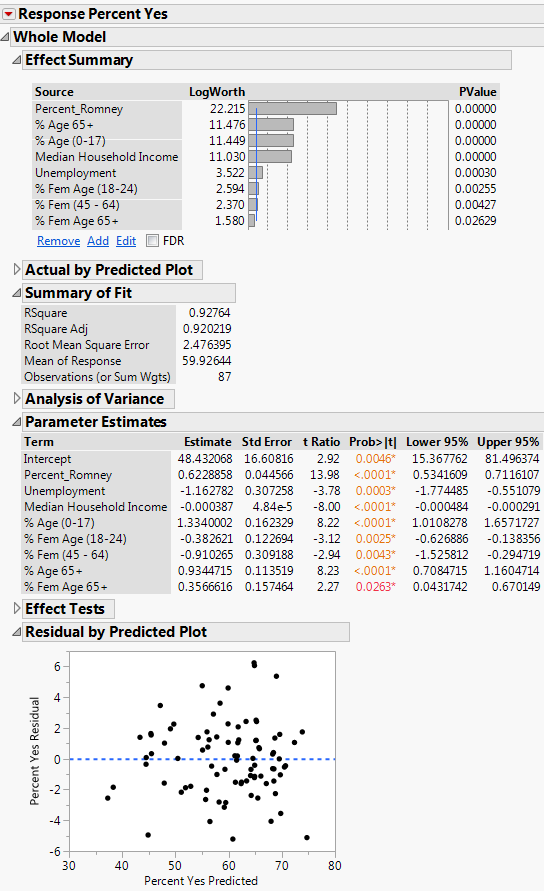
**BIC Forward Selection** **BIC Backward Elimination**

**AIC Forward Selection AIC Backward Elimination**



As the “best” models chosen by the three stopping criterion differ and the direction of the selection (backward or forward) matters for these data, we will have to make some subjective choices regarding which model to choose as “best” or possibly consider some of the metrics (i.e. AIC, BIC, ) simultaneously. The model below is the one I chose, but certainly others would be reasonable.



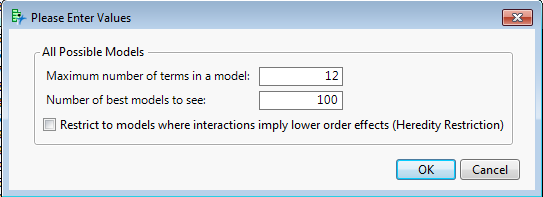
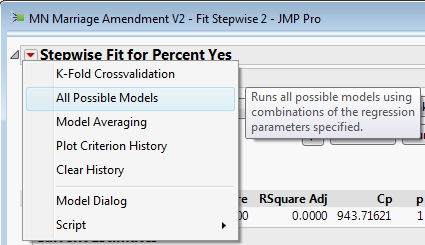
**Final Thoughts on Stepwise Selection Methods**

1. The model deemed “best” may be different depending upon the approach.
2. There is no clear rule as to how to use these procedures to choose a single “best” model. However, the outcomes from these search procedures can be used to identify a number of *possible* regression models that warrant further consideration.
3. Use experience, judgment, and the background science when model building! If you believe a predictor variable is fundamental, then include it regardless of the results of the search procedure.

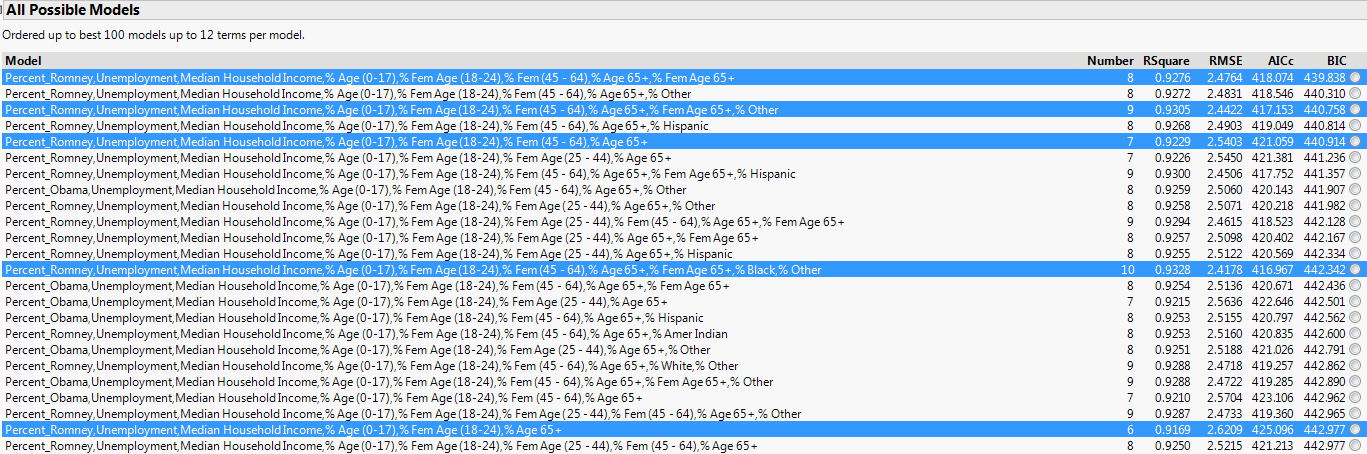
**16.3 – All Possible Regression Models**

As mention in 16.1 when we have a total of non-constant terms we wish to consider for using in a regression model there are a total of possible models to consider. This because each term is either in or out of the model (2 possibilities for each). When is large this will require fitting and examining a VERY LARGE number of models. JMP does have the option however to examine “all” possible sub-models up to a certain size presenting the top model candidates for that size. You can then sort the models examined using AIC or BIC to arrive at some good candidate sub-models.

**All Possible Models in Stepwise Platform** Here I have chosen models with up to 12 non-constant terms and have   
 specified to display the top 100 of each size, which is overkill!

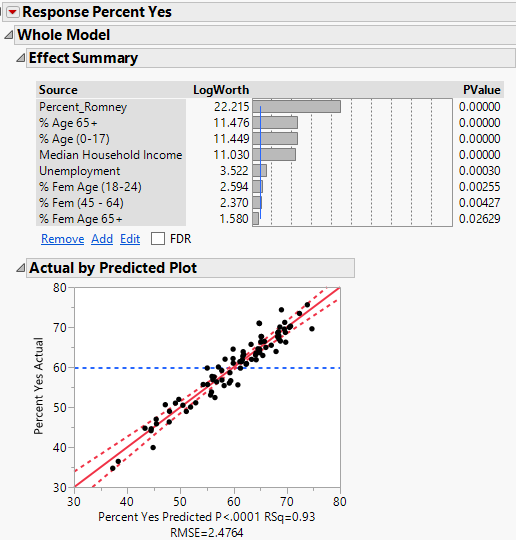
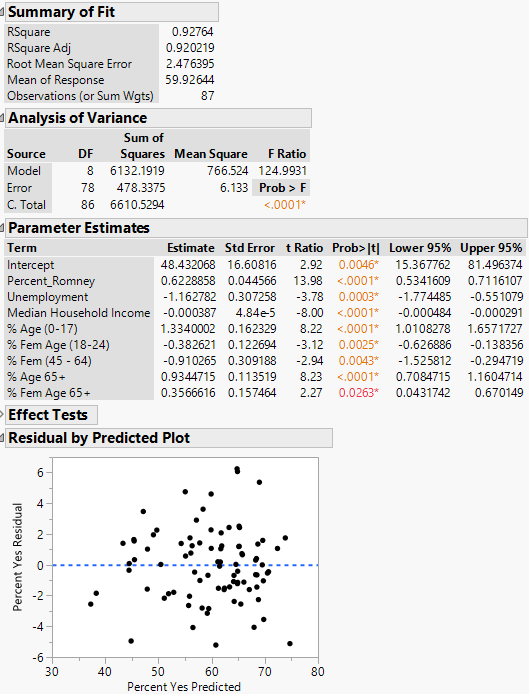


The resulting table is too large to display in this handout so only a portion of the results are shown below. The best models for each size are highlighted in the list. You can also sort by AIC/BIC as I have done below – right-click and **Sort by Columns…** checking the **Ascending** option.

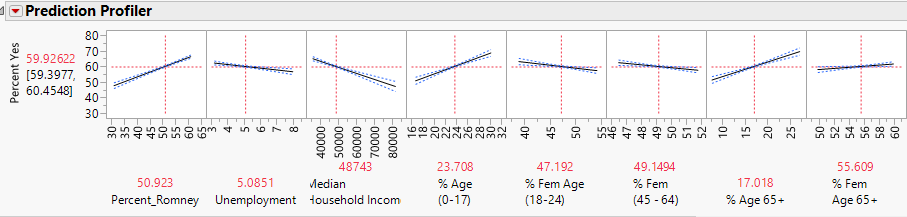


To select and fit a candidate model click on the button to the far right and Make/Run the Model.

The best 8 predictor model from this list is summarized below.

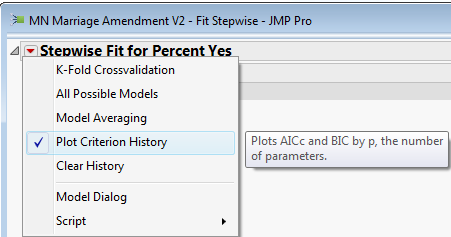
Using the Prediction Profile we can easily see all of the model effects graphically.



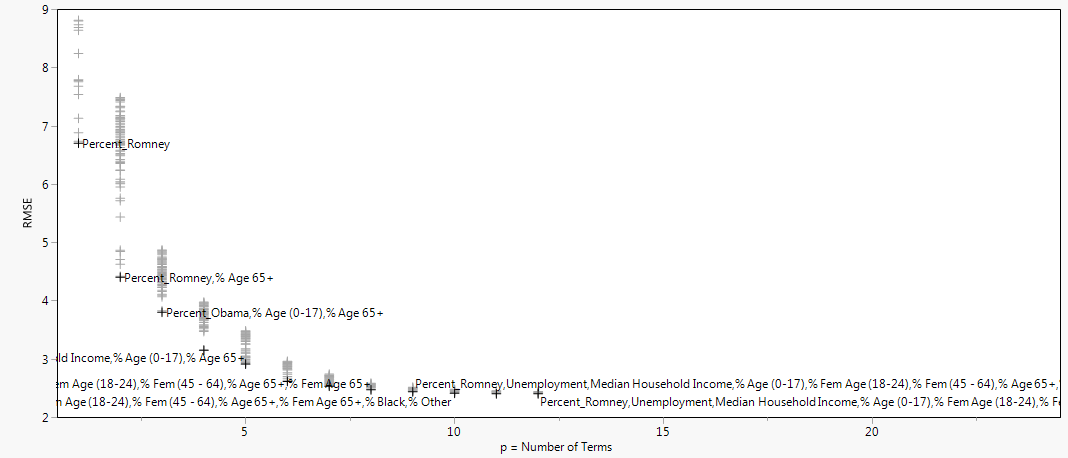
What characteristics do counties with a high percentage of “Yes” votes have?

What characteristics do counties with a low percentage of “No” votes have?

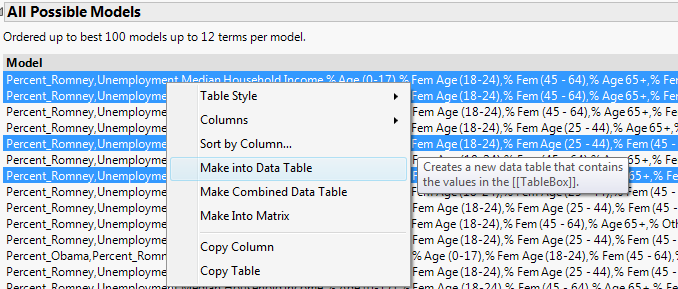
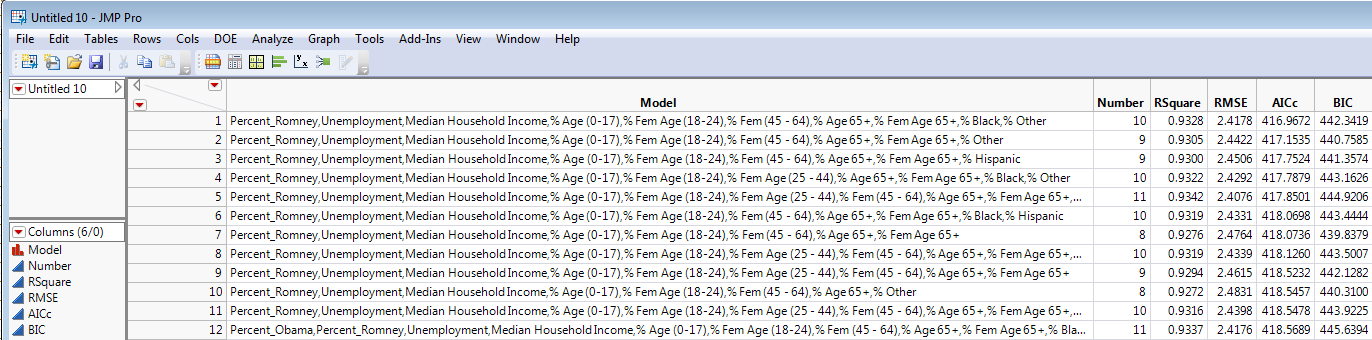
We can plot the results from exploring **All Possible Subsets** by selecting the **Plot Criterion History** option as shown below.



The resulting plot is shown below:



The “best model” for each model size (i.e. number of terms) is displayed although it is hard to read all of the terms included for the larger models in this plot.

Also you can save the results from All Possible Subsets into a JMP Data Table by right-clicking on the All Possible Models table and selecting **Make into Data Table**.  
 

In general, you can right-click on any table in JMP output and turn it into a Data Table.

**16.4 – Prediction Accuracy and Cross-Validation**

Sometimes the goal in developing a regression model is predict the response accurately, possible at the expense of interpretability. The stepwise methods above to some extent help identify models that may predict the response well but the criteria using in model selection are NOT considering how accurately our model will predict future values of the response.

In order to measure prediction accuracy we need to assess the ability of the model to predict the response using observations that were not used in the model development process. The reason why this is important is essentially the same reason why we cannot use the unadjusted in the model development process. Every time we add a term to the model the RSS goes down and the R-square goes up. A more complex model will always explain more variation in the response, however that does not mean it is going to predict the response more accurately.

To measure prediction accuracy we use **Cross-Validation**.In cross-validation we essentially divide our available data into disjoint sets of observations. One set of observations, called the ***training*** set, will be used to develop and fit the model. The model developed using the training set will then be used to predict the known response values in the other set, called the ***validation*** set. We use the validation set to choose the model that best predicts the response values in validation set. In order to judge the accuracy of future response predictions using the model selected by using the training and validation sets we may to choose have a third set of observations called the ***test*** set. The test cases are NOT used in the model development process at all, thus the accuracy of the response predictions for the test set cases should give us a reasonable measure of the prediction accuracy of our final model. If we do not have a large dataset we may not have enough observations to create these three sets, in which case we may want only create the training and validation sets. The validation and test sets are also called ***holdback*** sets as they contain observations held out in estimating the model. Most modern regression methods have some form of internal validation built into the algorithm that is used to “select” the model. There are different approaches one can take in forming training/validation/test sets for the purposes of conducting the cross-validation of a model. We will examine several schemes that are commonly employed below.

**16.5 - Split-Sample Cross-Validation Approaches**

Split-sample approaches simply split our original sample into the disjoint sets defined above. Splitting is usually done randomly, however we may choose to use a stratified sampling to take other factors into account when creating our sets. For example, if one of the variables in our data is the subject’s sex then we may want to make sure our sets are balanced in terms of the distribution of sex. We can also stratify on a numeric variable to make sure the distribution of this variable is roughly same in each set. For example if we are modeling home prices, we may want to make sure each set has a similar mixture of low and high price homes.

**Training/Validation Sets Only**

There is no definitive rule for the percentage of observations assigned to each set. Some common choices would 80/20, 75/25, 70/30, 66.6/33.4, or 60/40 (though if you are willing to use 40% of your data for validation purposes it would be better to use training/validation/test sets.) For multiple regression a rule of thumb that can be used is to assign p% to the validation set, where and   
 the largest number of parameters your model may contain.

Training Set

(100-p)%

Original Dataset

**Split randomly or stratified**

Validation Set  
  
(p%)

**Training/Validation/Test Sets**

Training Set

Original Dataset

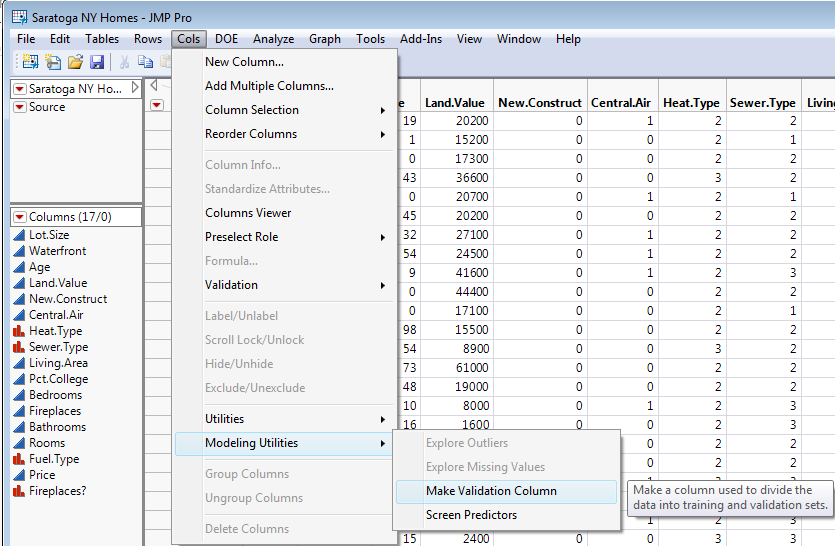
Again there is no definitive rule for the percentage of observations assigned to each set, however the most common are 60/20/20 or 70/20/10 with the former being the most common. Note the Test Set is in RED because it is not used in the model development process.

Validation Set

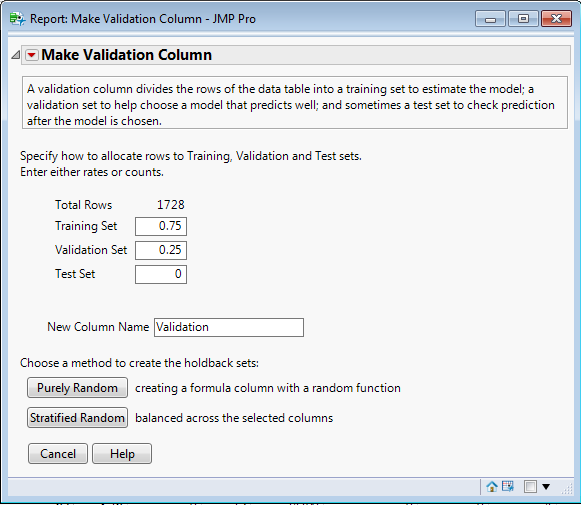
**Split randomly or stratified**

Test Set

JMP makes the process of generating these sets (either training/validation or training/validation/test) very easy by using **Cols > Modeling Utilities > Make Validation Column** option as shown below.



As you can see we have the option to make only training/validation sets or training/validation/test sets with percentages (proportions) we choose. The default is training/validation sets only with 75% training and 25% validation. We can also make the set assignments randomly or by incorporating stratification variable(s).



The model chosen gives the smallest prediction error when predicting the validation set response values.

Sum of Squared Prediction Errors

where the predicted value for *ith* response value in the validation set using the model under consideration. We choose the model that gives us the smallest value for this sum.

We can also consider the average squared prediction error by dividing the sum above by the number of observations in the validation set. We can also take the square root of this average to obtain the Root Mean Squared Error for Prediction (RMSEP).

where of observations in the validation set. We can also equivalently choose the model that maximizes the R-square for prediction which is the correlation between the predictions ( and the actual response values for the validation set squared, i.e.

If we have a test set, we can then apply any of these measures to the predictions made for the test set using the final model chosen using the training & validation sets to obtain a measure of the prediction accuracy for future values.

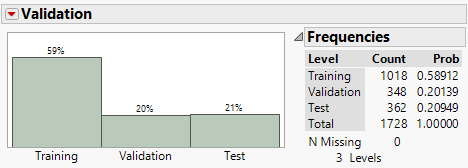
**Example 16.2 – Saratoga, NY Home Prices (Datafile: Saratoga NY Homes.JMP)**

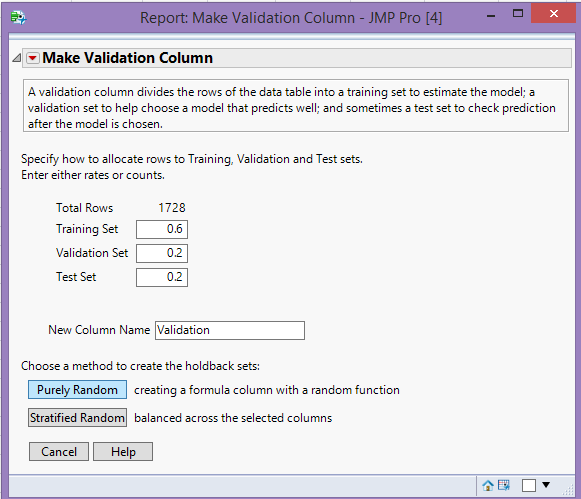
Consider again the Saratoga NY Homes dataset from Section 11.

**Variables:**

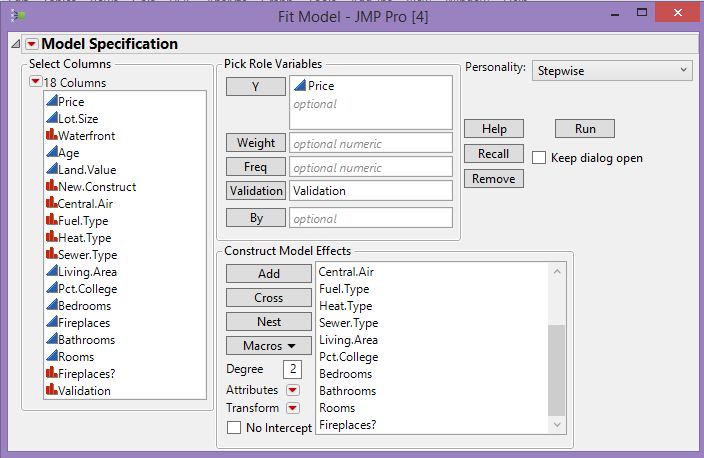
* Price – price of the home in dollars
* Lot.Size – acres
* Waterfront – is the home located on a waterfront (0 = no, 1 = yes)
* Age – age of the home (yrs.)
* Land.Value – assessed value of the land the home is on
* New.Construct – is the home a new construction (0 = no, 1 = yes)
* Central.Air – does the home have central air conditioning (0 = no, 1 = yes)
* Fuel.Type – type of fuel used to heat the home (2 = Gas, 3 = Electric, 4 = Oil)
* Heat.Type – type of heating system (2 = hot air, 3 = hot water, 4 = electric)
* Sewer.Type – type of sewer system (2 = none, 3 = private, 4 = public)
* Living.Area – living area (ft2)
* Pct.College – percentage of college housing in the home’s neighborhood
* Fireplaces – number of fireplaces in home (0,1,2,3, or 4) (WON’T USE, USE ONE BELOW)
* Bedrooms – number of bedrooms
* Rooms – number of rooms
* Fireplace? – does the home have at least one fireplace (0 = no, 1 = yes)

We will use split-sample cross-validation to select a model using all of the predictors above (with the exception of . We first form the training, validation, and test sets using a 60/20/20% split.

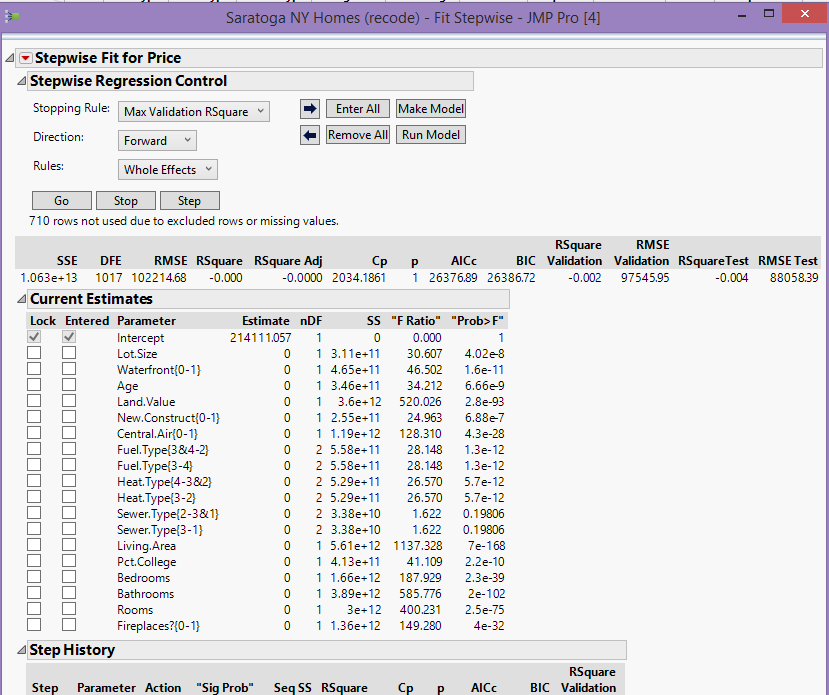
We have a total of *n = 1728* observations at our disposal, thus with a 60/20/20% split we have 1,010 homes in our training set, 341 in our validation set, and 362 in our test set.  




We use Fit Model to set up a dialog box with all of the predictors for Stepwise Selection.

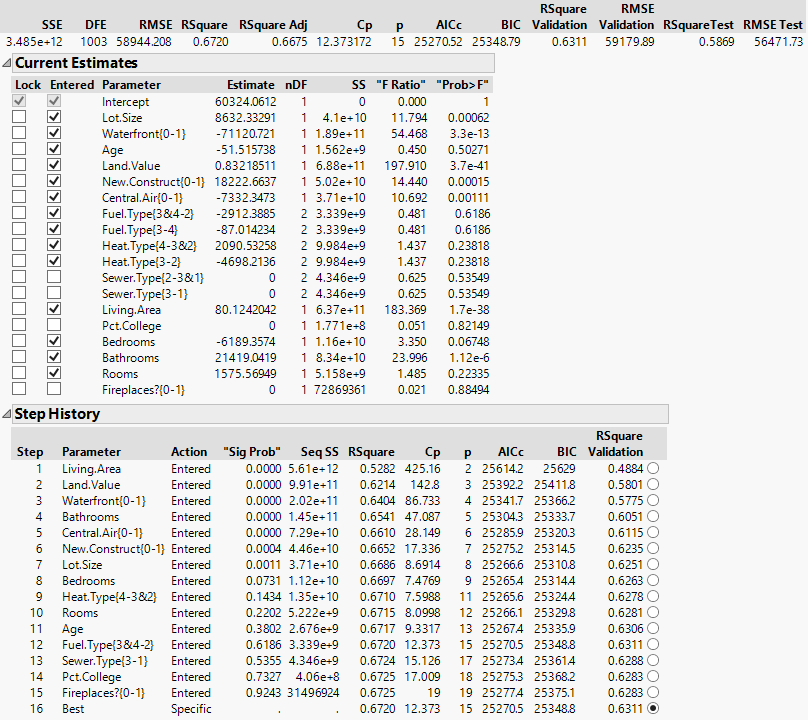


The Stepwise Platform is shown below – notice that we have changed Rules: Whole Effect which means that the factor predictors (heat type, fuel type, and sewer type) will not have their levels combined. If we have Rules: Combine the algorithm may combine fuel type levels, e.g it may combine gas and oil levels to create one dummy variable that is 1 if fuel type is gas or oil. This is not necessarily a bad thing but for this example it will not be allowed.



The resulting model chosen is shown in the table below created by using forward selection with **Max Validaton R-square** as the criterion to be optimized.

Results from Forward Stepwise selection.

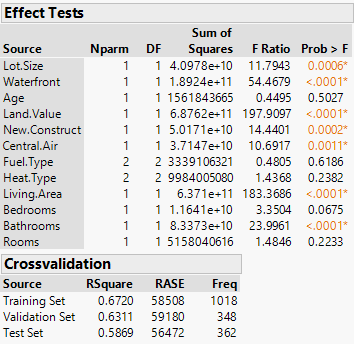


**Summary Statistics for the Fits to the Three Sets**

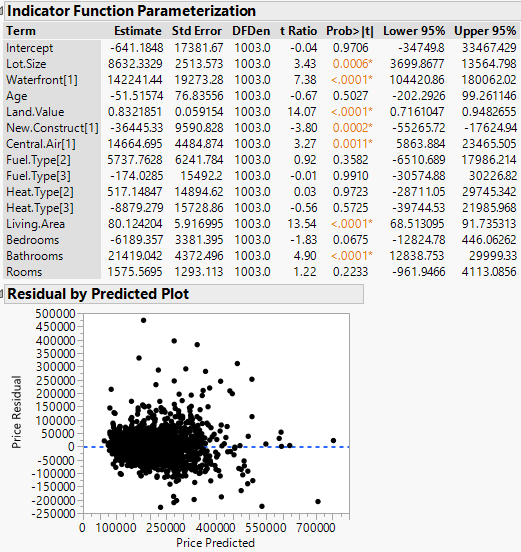
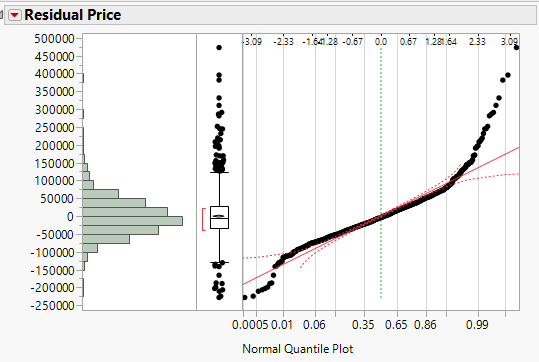
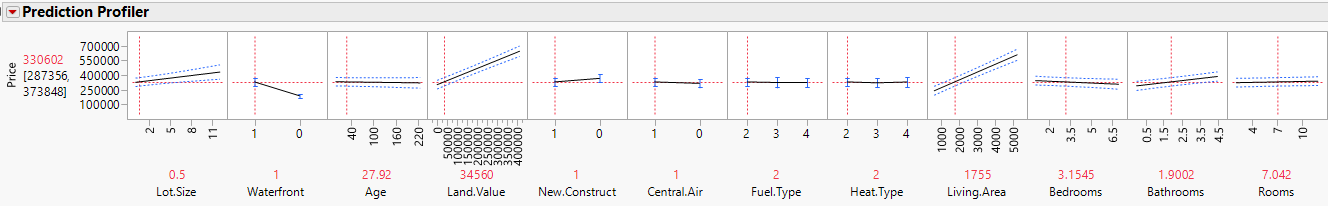
|  |  |  |
| --- | --- | --- |
| Set |  | – root mean squared error for prediction |
| Training | 67.20% | $58,944.21 |
| Validation | 63.11% | $59,179.89 |
| Test | 58.69% | $56,471.73 |

Forward Model Selected (note: Backward Selection yielded the same model)

Model Summary

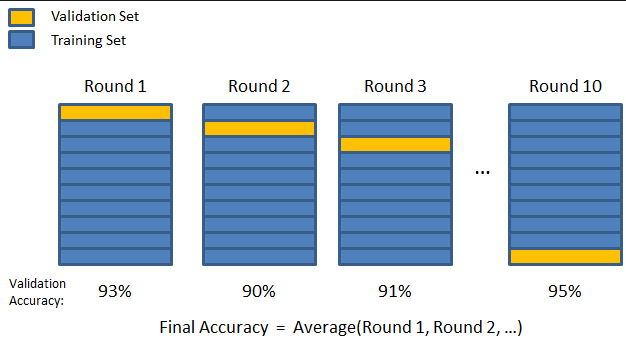
 

**Comments:**

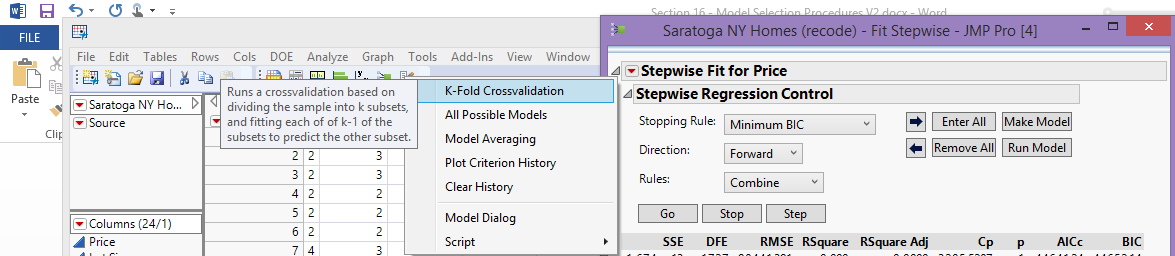
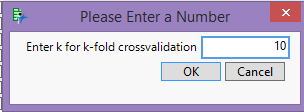
**16.6 – k-Fold Cross-Validation**

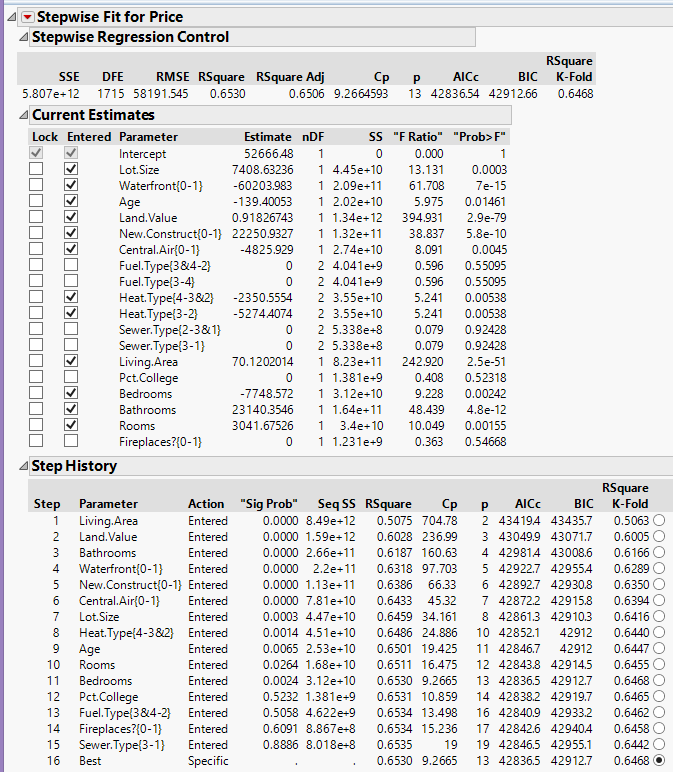
Another common cross-validation method used in model development is **k-fold Cross-validation**. In k-fold cross-validation the entire dataset is broken into roughly equal size disjoint sets (k = 5 or 10 typically). Then rounds of model fitting is done where the model is fit using (k-1) of the sets to predict the set left out with of the *k*-sets serving as the validation set. The diagram below illustrates a 10-fold cross-validation ().

10-fold Cross-Validation  


Using this method the model chosen is the one that has the best average prediction error over the subsets.

To perform k-fold cross-validation in JMP use Fit Model to set up the set of terms to consider and then choose Stepwise as usual. Then click on the main drop-down menu in the Stepwise platform as shown below and choose **k-Fold Cross-validation**.

Then choose the number of folds *k*.  
  
The results for these data are shown on the following page.

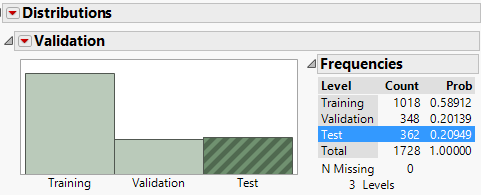
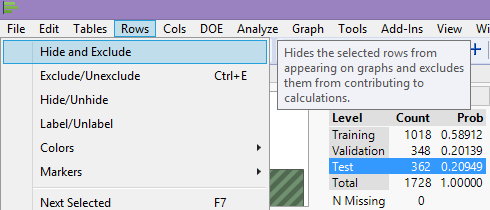
Results from k-Fold Cross-Validation  


The model chosen differs from the model chosen using the split-sample CV, it does not include the factor for fuel type.

Forward Model Selected (note: Backward Selection yielded the same model)

|  |  |  |
| --- | --- | --- |
| Set |  | – root mean squared error for prediction |
| Model | 65.30% | $58,191.55 |
| Validation | 64.68% | Not given |

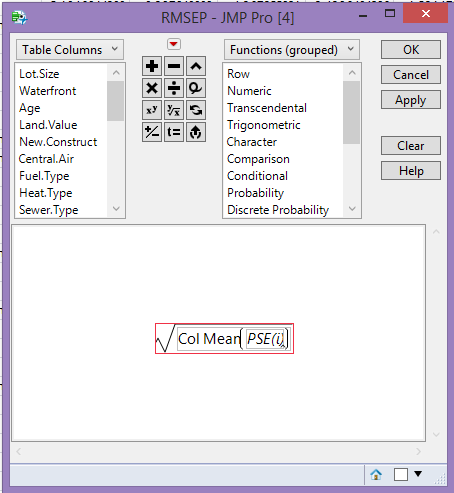
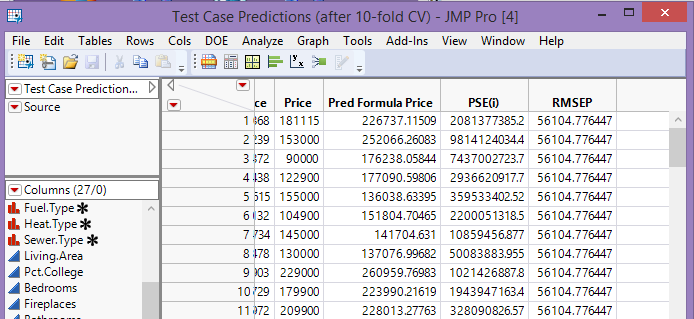
Rather than look at the summary of this model, we consider using our split-sample sets in conjunction with the 10-fold cross-validation. To do this we can use Rows > Hide and Exclude to temporarily remove the Test set cases from our data and then predict them back using the model selected via *k*-fold cross-validation. To remove the Test cases create use **Analyze > Distribution** to create a bar graph of the Validation column and click on the bar for the Test cases then hide and exclude them as shown below.

We use Stepwise selection with 10-fold cross-validation to choose a final model and then save the Prediction Formula. Even though the observations in the Test set were not used to fit the model the predicted selling prices of these homes is computed, which we can then the compare to the actual prices in the Test set using any of the measures discussed above.

The predictions for the Test cases are shown in the table below. I computed

for each case in the test set, which then be used to compute RMSEP.



The RMSEP = $56105.78 for predicting the test cases using the model chosen via 10-fold cross-validation using the observations contained in the training & validation sets.

**Leave-Out-One Cross-Validation (LOOCV)**

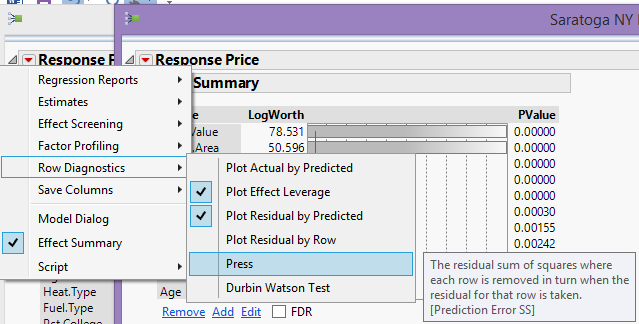
Leave-out-one cross-validation is a quick an easy way to assess prediction accuracy, though DEFINITELY not the best! LOOCV is equivalent to -fold cross-validation where # of observations in the data set.

Using the fact the predicted value for when the case is deleted from the model is equal to

Here and .

This is also called the jackknife residual and the sum of these squared jackknife residuals is called the PRESS statistic (Predicted REsidual Sum of Squares), one of the first measures of prediction error.

You can obtain the PRESS statistic for any fitted model in JMP by selecting PRESS option as shown below.



The PRESS statistic for the model chosen using 10-fold cross-validation fit to all cases is shown below.



**16.7 – Summary of Model Selection Procedures**

In this section we examined several methods for model selection in regression. Which method to use depends on the goals of the regression analysis. If building the simplest model that adequate explains the conditional expectation and our interest is in discussing the model effects (terms and predictors) then the forward, backward, and mixed model selection methods are fine. Using any of these generally results in several candidate models that we need to evaluate using more subjective criterion and the “best” model depends on the direction of our search (i.e. forward/backward) and the criterion used to compare rival models (i.e. AIC/BIC/).

If prediction accuracy is the goal then it is essential to perform some form of cross-validation. If you have enough observations available, then the use of training, validation, and test sets is recommended – possible in conjunction with other methods such -fold cross-validation. Also, even if prediction is not the goal, the model chosen by considering predictive performance may produce a reasonable model to use for interpretive purposes.