**Week 3 Discussion Board:**

Consider the automated variable selection procedures of forward, backward, and stepwise variable selection.  Should any method be generally preferred over the others?  Are there scenarios where one method should be preferred over the other?

As Dan and Ryan mentioned, RABE states that backwards selection is the optimal method for stepwise selection due to it running through all possible predictor variables and best addressing collinearity. Another observation is that each procedure has different default confidence settings using the SLENTRY command (that can be manually changed).  For example, the default significance for forward selection is 0.50 for a variable to be included in the model, whereas with backwards selection the default significance setting is 0.10.      
  
If you want to quickly and easily identify “optimal” variables for a model, these procedures are a great tool, but by no means perfect and should be used with caution. Automated variable selection, regardless of which procedure is selected, can potentially lead to omitting key variables and/or including unnecessary ones. In my opinion AVS is most valuable when we have a large number of predictor variables. In this case these automated selection techniques can help point the analyst in the right direction by narrowing the field of potential variables to include in the model.

As Ratner stated in the natural seven-step cycle of statistical modeling and analysis, having competing techniques increases the odds that a thorough analysis has been conducted. As Ratner states often, variable selection in regression could be argued to be the hardest part in model building. Comparing variables in models across the different variable selection procedures ensures that the analyst can be comfortable with the selected model to perform adequately within a given purpose. The objectives of the variable selection method for regression models is to find an optimal model that has accuracy, stability, parsimony, interpretability, and avoidance of bias in drawing inferences. As Ratner has noted, none of the automated variable selection methods satisfy most of these goals. This would give reasoning that there would be no specific scenario where one method would be preferred over the other and the best chance of selecting an optimal model would be to use all three methods for comparison.

I would also note that we should begin the variable selection process by reviewing relevant literature, applying expert knowledge, etc to the problem we are considering and thus identifying a subset of variables for inclusion in a model prior to leveraging an automated technique such as Forward, Backward, Stepwise, etc.

Metrics such as Mallows C*p*, Residual Mean Square, Akaike Information Criterion (AIC), and Bayes Information Criterion (BIC), all help analysts detect which equation is best suited for a situation.

What are the adjusted R-squared, AIC, and BIC metrics?  How can we use these metrics in the modeling process and what advantage do they have over the R-squared metric?

    The adjusted R-squared, AIC, and BIC metrics are all helpful tools when evaluating equations and looking to find the best equation for a specific problem. The modeling process for regression analysis has eight main steps (Chatterjee & Hadi 2012 p13). Step seven consists of model validation and criticism, which is where the above stated metrics are utilized to discern the best model. The advantages of each metric is stated below:

* Adjusted R-squared: Similar to the R-squared test, adjusted R-Squared is used to discern the goodness of fit. This metric compares different models that have a different number of predictor variables. It tries to adjust the score for unequal variables when comparing two different sized models. R-squared is calculated by subtracting one from the Sum Standard Errors divided by Total Sum Squared. This model does not take into consideration different sized variables, where the adjusted R-squared equation subtracts out the larger number of variables, thus leveling the playing field for R.
* Akaike Information Criterion (AIC): This model balances accuracy and simplicity (2012). Through its equation, it penalizes models with more variables and a higher AIC is not desired when comparing two models. It has an edge over the R-squared metric because it adjusts for different sized equations.
* Bayes Information Criterion (BIC): AIC and BIC are very similar, but BIC has a much harsher penalty for models that have over 8 predictor models. This helps to control overfitting that AIC does not adjust for in evaluating models (2012).

All three metrics listed above have an edge on R-squared given that they adjust the score or validation value when comparing equations that do not have an equal number of predictor variables.

R2 indicates the amount of variation in our dependent or response variable Y can be explained by variation in our independent variable (predictor) X and therefore is a good measure of fit when we are dealing with simple linear regression problems.  However, as discussed and as seen in fitting a regression line to our building prices dataset in the assignment, as we add more predictor variables to our equation, R2 will always increase, regardless of whether or not the additional variables actually improve the fit of our model.  The adjusted R2 adjusts for the number of predictors included and only increases when additional predictors improve model fit.   
  
AIC, The Akaike Information Criterion, is a useful metric for comparing models, assuming that models are non-nested and there are no missing values.  As pointed out by Chaterjee and Hadi, "in selecting a model (AIC) tries to balance the conflicting demands of accuracy (fit) and simplicity (small number of variables)." (Chaterjee and Hadi, 2012) Models can be ranked by their AIC to facilitate comparison and assist in selecting the best model in terms of predictive accuracy and efficiency.   
  
BIC, Bayes Information Criterion, is also a criterion for model selection but includes a larger penalty for the number of parameters in candidate models.  The purpose is to prevent overfitting.