Pre Live Session Unit 2 Assignment

In golf there is an expression, “drive for show putt for dough.” This data set contains average winnings for a given year of profession golfers along with additional performance statistics such as driving distance and accuracy, average number of putts per hole, as well as others including some with variable names that are only labeled arbitrarily. The idea here is that although it is quite fun and impressive to watch some one hit a golf ball very hard and far, it actually matters more when they are closer to the whole and chipping and putting. Can we build a model to predict player’s average winnings given their various performance metrics?

For this assignment, we are going to play around a little bit with glmselect to get a better feel for issues when using metrics such as R-squared in terms of reporting predictive ability of a model and the idea of overfitting. Using the data set and SAS code run the models and generate the output. There are 5 total models that are run in the code.

# Question 1

Examine the glmselect output from the first two proc glmselect (labeled M1, M2 in the code) calls and compare them in the following way.

1. What is different between the two OLS models in terms of the predictors? (note we have tricked glmselct in doing OLS by specifying Forward feature selection with no stopping criterion)
   1. **M1 uses 6 variables**
   2. **M2 uses 26 variables**
2. What are the two models R-square values and adjusted R-squared values?
   1. **M1**
      1. **R-Square: 0.6350**
      2. **Adj R-Square: 0.6098**
   2. **M2**
      1. **R-Square: 0.6977**
      2. **Adj R-Square: 0.5804**
3. Examine the Fit criteria and ASE plots. In terms of prediction, do you think there is much harm in using all of the predictors versus using a feature selection approach to reduce the model down?
   1. **Both M1 and M2 follow the same pattern on the fit criteria and ASE plots. In combination with the adjusted R-square values, model M1 seems like a more appropriate model since it accomplishes a slightly better fit with significantly fewer explanatory variables.**

# Question 2

Compare the second and third proc glmselct calls (M2, M3). These both have the same predictors, but one is OLS and the other is using LASSO feature selection using cross validation.

1. Note the R-squared and Adjusted R-squared and compare them.
   1. **M2**
      1. **R-Square: 0.6977**
      2. **Adj R-Square: 0.5804**
   2. **M3**
      1. **R-Square: 0.5830**
      2. **Adj R-Square: 0.5564**
   3. **The M2 model has a slightly better adjusted R-square value.**
2. What variables are included using the LASSO as a feature selection technique?
   1. **DriveAcc**
   2. **Greens**
   3. **Avg Putts**
   4. **Save**
   5. **V23**
   6. **V25**
3. Suppose now that I told you that all of the predictors with generic names are just a bunch of random numbers, how does that piece of information potentially change your feeling on whether it matters or not to do feature selection.
   1. **For this example, if the generically-named predictors are random numbers, I don’t think they’re worth including in the model. However, the question only asks “whether it matters or not to do feature selection.” We don’t have enough information to answer that beyond the scope of this single question.**

# Question 3

Compare the fourth and fifth glmselect calls. These models include interaction terms, so the model is even more complex and the potential for overfitting becomes even greater.

1. In model 5, examine the CVpress fit criterion panel and compare it to the ASE plot for the test set. Does the CV fit panel mimic the ASE test performance pretty well?
   1. **The CV fit panel does appear to mimic the ASE test performance fairly well, although the CL Press panel shows a steeper, earlier drop.**
2. In model 5 that uses the CV approach for feature selection, if we have used Adj-Rsquared rather than CV press, how good would you feel about the predictions you made with that particular model?
   1. **I may be reading this incorrectly, but the best criterion value for the adjusted r-square approach appears to be #34 which is the removal of AvgPutts\*AvgPutts. This doesn’t seem to be the best strategy.**

# Bonus/Critical Thinking

When comparing ASE plots of OLS and LASSO from our given code, you may have noticed that OLS seems to yield smaller test error values than LASSO. That may seem contradictory. Why do you think this is happening and why the actual values of the ASE for the OLS and LASSO models we ran are not directly comparable?

**Is this because the LASSO model penalizes the variables, therefore increasing the test error values?**