**DD#1 – Multiple Regression**

Multiple regression is a very powerful tool that creates an equation that models a dependent/response variable based on multiple independent/explanatory variables. In this team-level analysis for last season MLB season (data from SABR#5 & 6), you will be regressing Wins (y-variable) again with several different x-values for all 30 teams. Then, you will be comparing your results to what you found in SABR#5 & 6. You will need to use the Data Analysis Toolpak in Excel to run the Multiple Regression. For instructions, use the ‘Multiple Regression Class Notes for Sabermetrics Seminar’ from the Schoology folder.

**Part I (Modeling Wins based on Runs Scored and Runs Allowed)**

1. Create a multiple regression equation for the season to estimate Wins based on Runs Scored and Runs Allowed. Do each of your variables have p-values below 0.05? If not, remove the variable with the highest p-value and rerun the regression. What is your model/equation (round to the thousandths place)?
2. What is the ‘Adjusted R-squared’? How does it compare to the r2 in your single-variable regressions in SABR#5 & 6 (Runs Scored, Runs Allowed, Payroll) that were also created to estimate Wins?
3. Give the real-world interpretation of the slope for each of the variables in your new model. Do the signs make sense? How many runs scored equals a win? How many fewer runs allowed equals a win? How do these compare to your answers in SABR#5 & 6?
4. Based on the regression model/equation, how many Wins would you expect a team that had 780 Runs Scored and 780 Runs Allowed? Does this seem reasonable for a 162-game season?

**Part II (Modeling Wins based on Runs Scored, Runs Allowed AND Payroll)**

How much does money matter? This time include Payroll and start the multiple regression process over with Runs Scored and Runs Allowed to create a new equation for the season. Do each of your variables have p-values below 0.05? If not, remove the variable with the highest p-value and rerun the regression. What is your new model/equation (round to the thousandths place)? How did it compare to your model in Part I? Why do you think that it is the same/different?

**Part III (Moneyball)**

It's all about winning. Using over 40 years of MLB team-level statistics that are provided, you will be modeling wins (W). The 14 candidate explanatory variables {don’t use Games (G)} are commonly used baseball statistics so you can look up their definitions, if needed.

* From Schoology, use the **training dataset** to create your model to predict wins: ‘train (use this to create your model).xlsx’
* Do not use G (Games) in your model – this is there if you’d like use any of the stats on a per game basis.
* Once you have finalized your model, apply it to the test dataset in Schoology to see how well the model works: ‘test (apply model to this to estimate W).xlsx’

1. Create a multiple regression equation using the training dataset. Do each of your variables have p-values below 0.05? If not, remove the variable with the highest p-value and rerun the regression. What is your model/equation (round to the thousandths place)?
2. What is the ‘Adjusted R-squared’?
3. Give the real-world interpretation of the slopes for each of the variables in the model. Do the signs make sense?
4. Next, email your estimated wins (W) (i.e., whole column of values) to me at [dbenesh@gsgis.k12.va.us](mailto:dbenesh@gsgis.k12.va.us).  Your results will be judged by [Mean Absolute Error](https://en.wikipedia.org/wiki/Mean_absolute_error).  You must get MAE<8. If you don’t, ask Yoda what to do.

**Part IV (Modeling Wins in another sport)**

Pick another sport (NBA, NFL, etc.), and do a similar analysis with both Points Scored and Points Allowed in the multiple regression model. You should already have the data.

A. What is your model/equation (round to the thousandths place)?

B. Based on the r2 which is the better model - this one or those you did previously? Why do you think that this is the case?