**DATA 621 Homework 1**

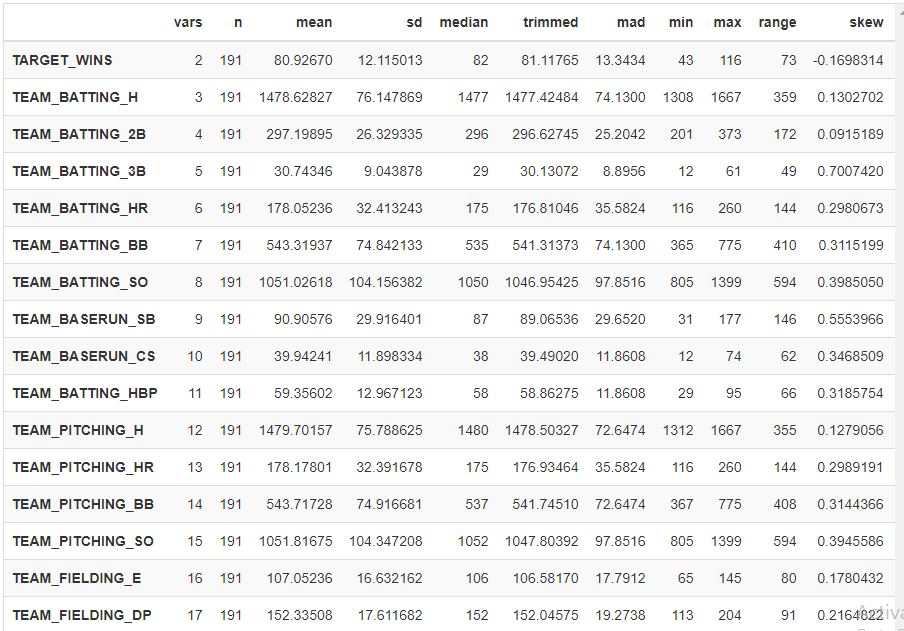
**Linear Regression**

**Trishita Nath**

# DATA EXPLORATION

The training data set contains 17 columns and 2276 rows, covering baseball team performance statistics from the years 1871 to 2006 inclusive. The data has been adjusted to match the performance of a typical 162 game season. The data-set was entirely numerical and contained no categorical variables.

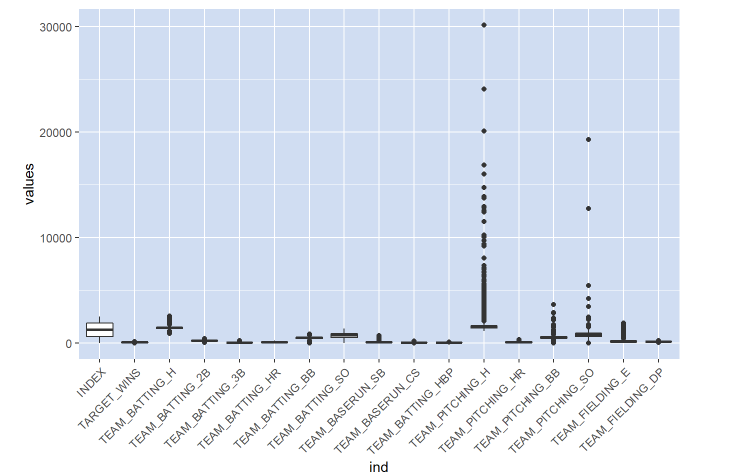
## Summary stats



From the data above, there are multiple variables with missing values, with TEAM-BATTING\_HBP being the highest.

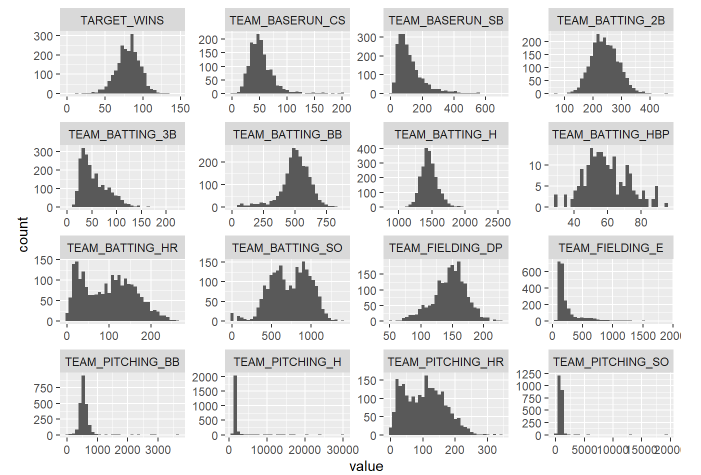
## Boxplots

The boxplots below help show the spread of data within the dataset, and show various outliers. TEAM\_PITCHING\_H seems to have the highest spread with the most outliers.

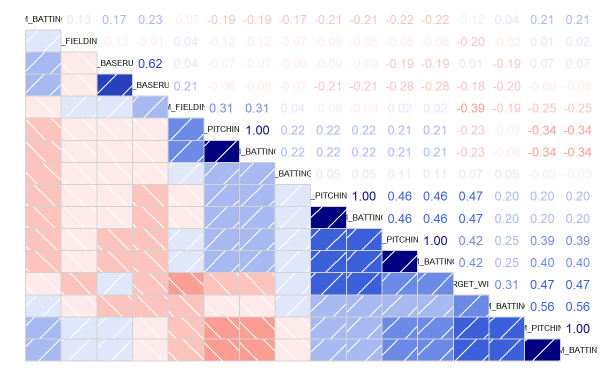


## Histograms

Histograms below show distributions of the variables:



## Multicollinearity



When considering features for my models, I will take into account the correlations among features so as to avoid including pairs with strong correlations.

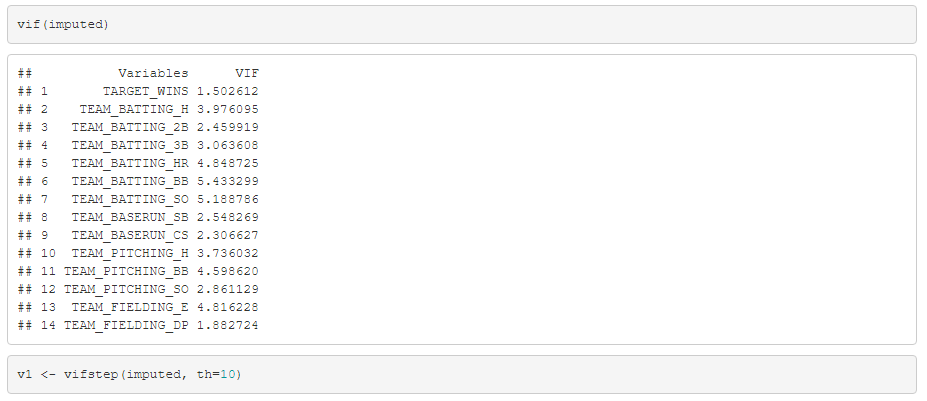
# DATA PREPARATION

## Removed Fields

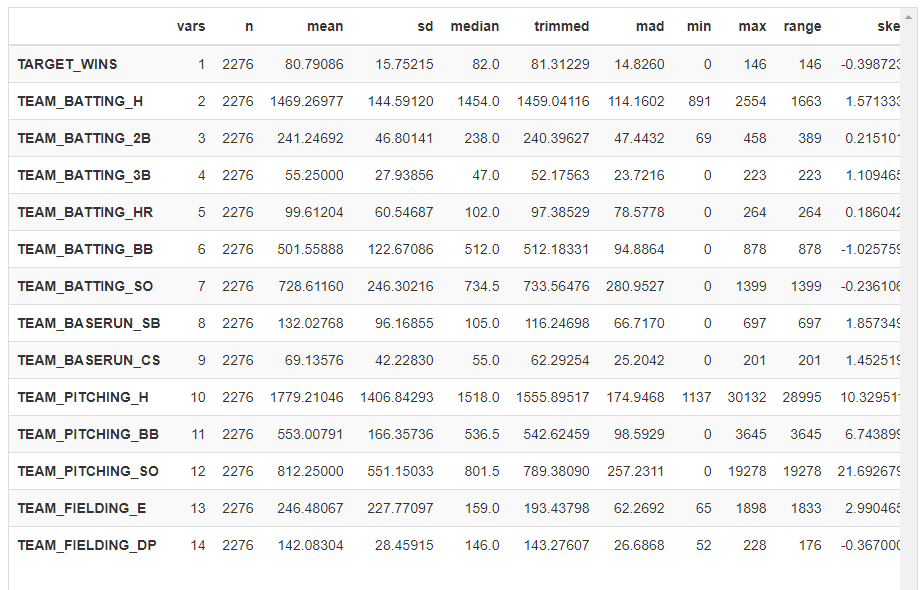
TEAM\_BATTING\_HBP is missing over 90% of its values. I will remove this variable. Variables TEAM\_PITCHING\_HR and TEAM\_BATTING\_HR are also very closely correlated with each other. I will drop the TEAM\_PITCHING\_HR from the dataset.

## Imputation

I performed imputation via prediction using the MICE (Multivariate Imputation) library using a random forest prediction method. Variables that exceed the established threshold will be discarded to avoid collinearity issues.



Final output is as follows:



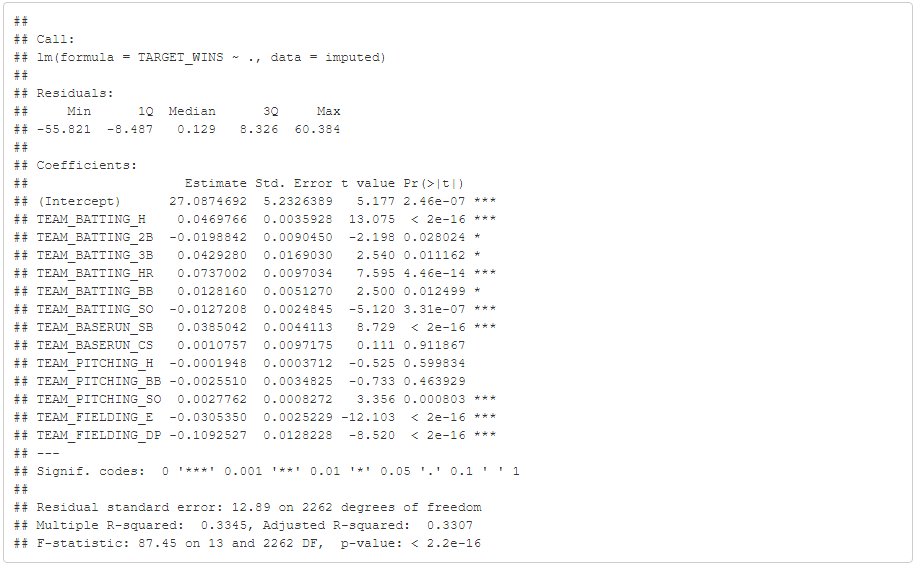
# Build Models

I will build 3 different linear regression models, to determine which one provides the best prediction for the number of wins. These are:

* All variables
* Only significant variables
* Backwards elimination of each variable

## Model 1: All Variables

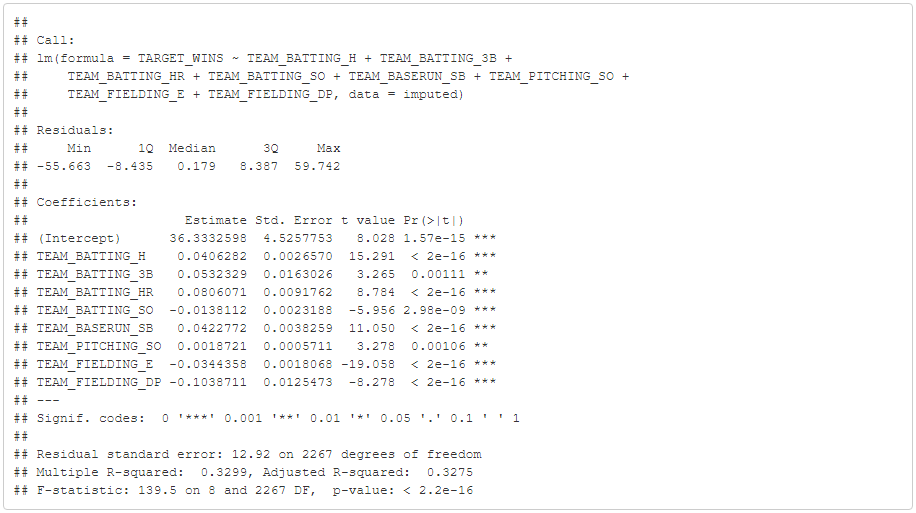
After the data has been imputed, all of the variables will be tested to determine the base model they provided:



F-statistic is 87.45, R-squared is 0.3307.

## Model 2: Highly Significant Variables Only

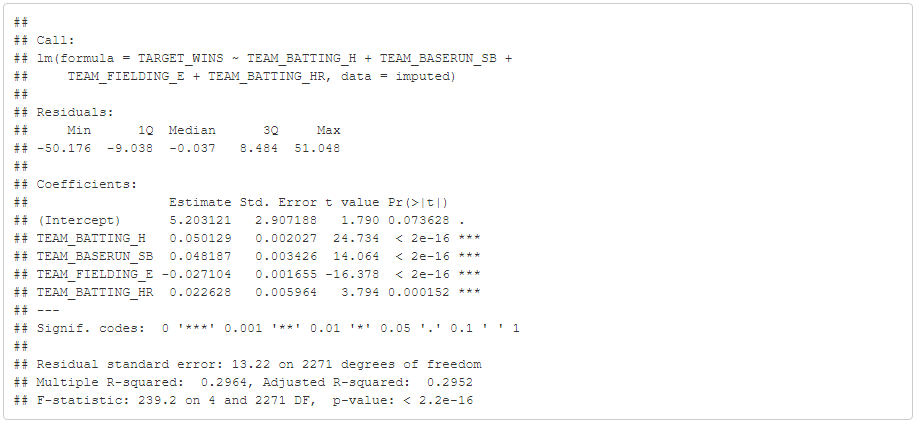
This model will focus only on the variables that are statistically significant. Variables will be chosen based on their significance level from the R output:



F-statistic is 139.5, R-squared is 0.3275. The F-statistic is better than the first model but the R-squared drops slightly.

# Model 3: Backwards Elimination

Variables are removed one by one to determine best fit model. After each variable is removed, the model is re-run until the most optimal output values are produced. Only the final output is shown.



F-statistic is 239.2, R-squared is 0.2964 The F-statistic is larger than both of the other two models, however the R-squared is slightly lower than the other two

# SELECT MODELS

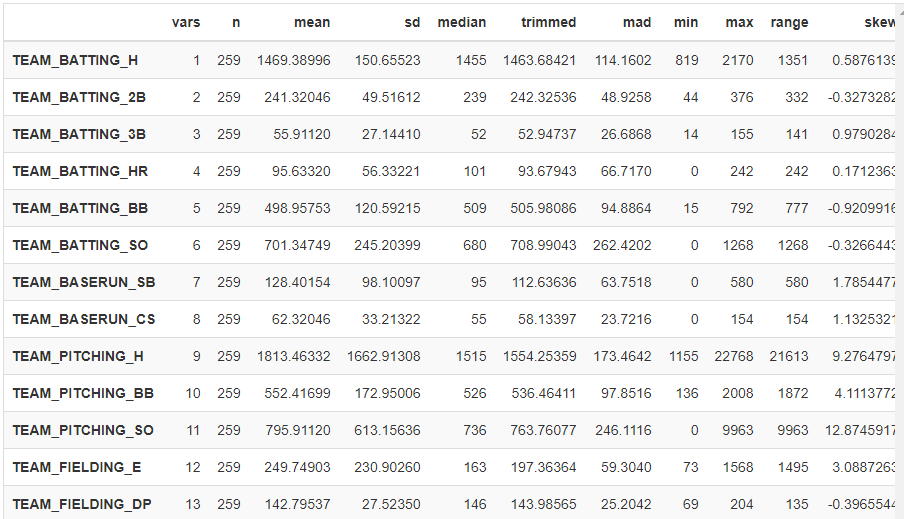
My model of choice would be Model 3 due to the following reasons:

* There is greater statistical significance with the third model relative to the others and uses less unnecessary variables to compute our prediction without compromising the adjusted R^2 value.
* Model 3 seems to have lower VIF scores than Models 1 and 2.
* It has better adjustment for multicollinearity

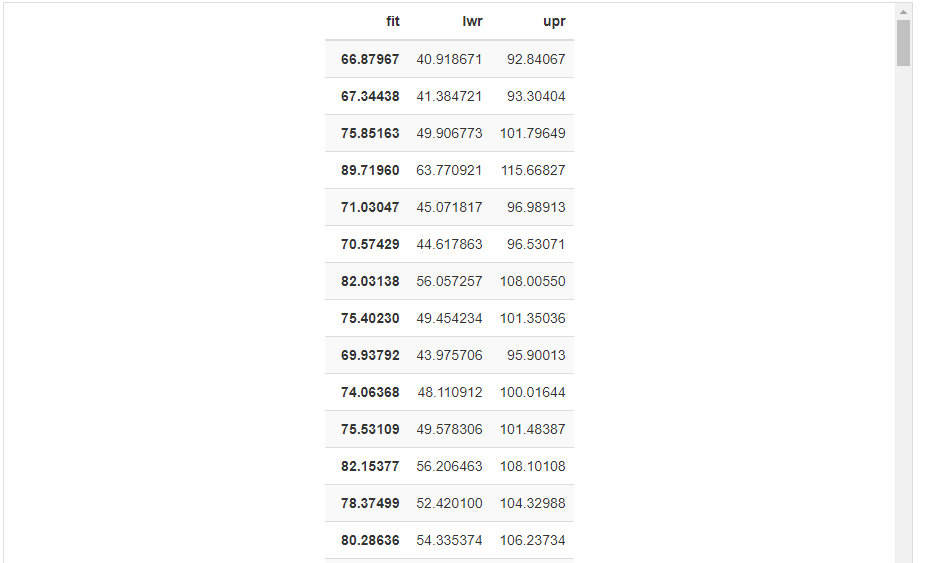
Predictions

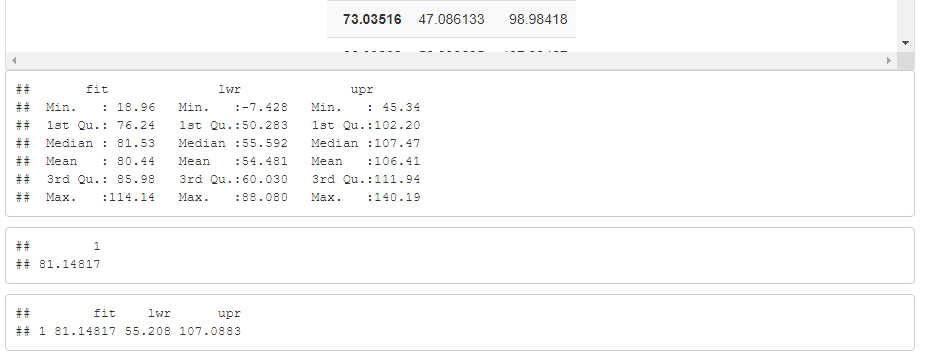
The evaluation dataset would be subjected to the same processing as the training data set.

Stats



After imputing and cleaning the data, using the predict function and Model 3, the following are the predicted values including prediction intervals:





# References

* A Modern Approach to Regression with R: Simon Sheather
* Linear Models with R: Julian Faraway.
* R package vignette, [mixtools: An R Package for Analyzing Finite Mixture Models](https://cran.r-project.org/web/packages/mixtools/vignettes/mixtools.pdf)
* [Detecting Multicolinearity with VIF](https://online.stat.psu.edu/stat462/node/180/)

# Appendix

## Moneyball Dataset Columns

* INDEX: Identification Variable(Do not use)
* TARGET\_WINS: Number of wins
* TEAM\_BATTING\_H : Base Hits by batters (1B,2B,3B,HR)
* TEAM\_BATTING\_2B: Doubles by batters (2B)
* TEAM\_BATTING\_3B: Triples by batters (3B)
* TEAM\_BATTING\_HR: Home runs by batters (4B)
* TEAM\_BATTING\_BB: Walks by batters
* TEAM\_BATTING\_HBP: Batters hit by pitch (get a free base)
* TEAM\_BATTING\_SO: Strikeouts by batters
* TEAM\_BASERUN\_SB: Stolen bases
* TEAM\_BASERUN\_CS: Caught stealing
* TEAM\_FIELDING\_E: Errors
* TEAM\_FIELDING\_DP: Double Plays
* TEAM\_PITCHING\_BB: Walks allowed
* TEAM\_PITCHING\_H: Hits allowed
* TEAM\_PITCHING\_HR: Homeruns allowed
* TEAM\_PITCHING\_SO: Strikeouts by pitchers

## R Code

GitHub: https://github.com/nathtrish334/Data-621/blob/main/Homework01.rmd

RPubs: https://rpubs.com/trishitanath/814750