Alexander Booth

Predict 411, Section 58

Professor Ott

October 8, 2017

**Unit 01: "Moneyball Baseball Problem"**

**Introduction:**

In this assignment, I will use OLS (“Linear Regression”) and the given statistics to predict the number of wins for a given baseball team. I will only use the variables given to me, or variables that I derive from the data provided. The data for this assignment contains approximately 2,200 records. Each record represents a professional baseball team from the years 1871 to 2006, inclusive. Each record contains the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season. I plan to use Ordinary Least Squares (OLS) regression to predict the number of wins for a team. I will construct multiple models and select a single model based on both model comparison criteria, such as AIC, BIC, and Adjusted R^2, as well as interpretability. The data manipulation steps, as well as the parameters from the model, will comprise a strategy for predicting wins on a test set.

I am provided with two data sets, one for training the model, the other to feed into the predicting strategy. The training data seems to have occurred during a long span of time, yet the data itself does not provide a variable to indicate time. The game of baseball has evolved over the period of 1871 to 2006, from the deadball era, through pitcher friendly and batter friendly periods. It seems unlikely that this dataset is comprised of homogeneous records from relatively similar eras. As such, I expect what is being measured will have experienced some changes over the history of baseball and these changes will likely manifest in the data. As I move further into examining the data, I will keep an eye out for outliers since they may be indication of previous ways in which baseball was being played.

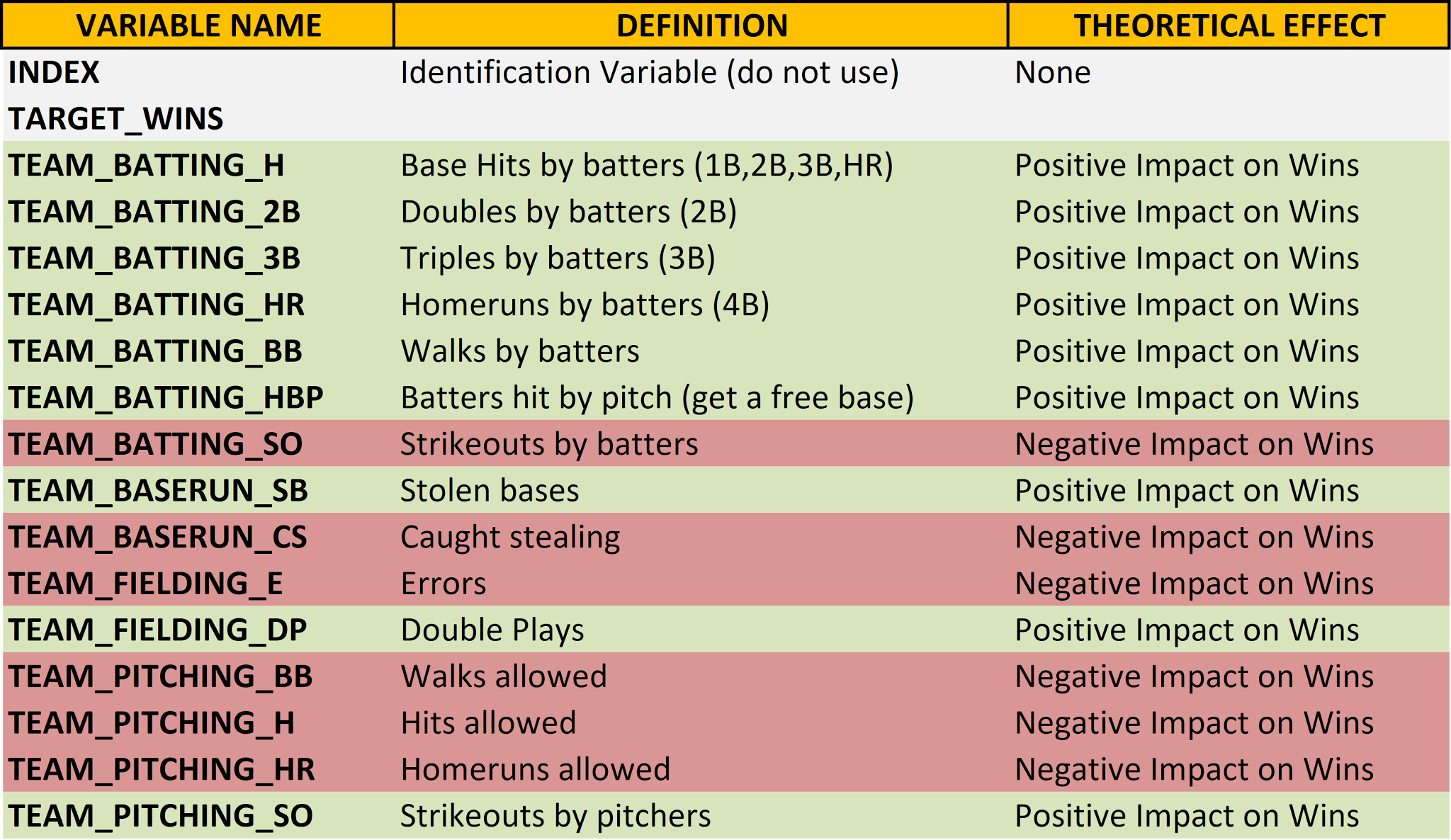
I will use techniques learned in the, 410, such as dummy coding, automatic variable selection, and dimensionality reduction as well as those learned in the first few weeks of this class, such as variable transformation, fixing missing data and outliers, as well as linear regression. The hope is that through exploratory data analysis I can become intimate with the data, and use some of the above techniques to create a set of variables that will perform well when moving into model selection.

**Exploratory Data Analysis:**

My plan to learn and explore the data set will consist of the following steps:

* Examine the data, as well as the data dictionary. Fix any errors (e.g. label-switching)
* Consider the arithmetic relationship between variables (e.g. combining or decomposing), as well as contextual relationships between variables (e.g. variables that represent conceptually opposite measurements)
* Understand what data is missing
* Understand the initial relationship variables have with our dependent variable
* Impute variables that have missing data
* Examine variable distributions and consider further imputation or indicator coding
* Re-examine relationship between re-expressed/imputed independent variables and dependent variable
* Begin model construction

There are two data sets provided, one is the training data set comprised of 2276 observations. The other is a testing data set comprised of 259 observations. I will begin my exploratory data analysis (EDA) by examining the variables provided to us in the data dictionary.

**Table 1: Data Dictionary with Proposed Theoretical Effect**

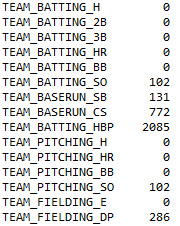
First, it is apparent that this data has been prepared for analysis since all of the variable labels are consistent. Additionally, each variable is continuous, and appears to be a count of a specific metric. Some of these metrics could likely be arithmetically combined, such as computing the number of “1B” hits by taking the difference between overall hits and “2B”, “3B”, “HR”. Another metric called “Total Bases” can be calculated by 1\*1B + 2\*2B + 3\*3B + 4\*HR. Finally, a metric called “On-Base Count” can be calculated by adding together “Hits” and “Walks” for batting and pitching respectively. I will create all four of those variables after missing data has been imputed and outliers have been fixed.

It is imperative to remain diligent to potential relationships between variables that can be explored, where one variable can help to provide context for examining another variable. Some examples of these presumed relationships are:

* TEAM\_BATTING\_H & TEAM\_PITCHING\_H
* TEAM\_BATTING\_BB & TEAM\_PITCHING\_BB
* TEAM\_BATTING\_HR & TEAM\_PITCHING\_HR

I believe that by looking at the descriptive statistics of one of these variables, I could compare those to the descriptive statistics of the other variable and have contextually informed observations. However, first I will examine the training data for missing values.

**Table 2: Missing Values**

****

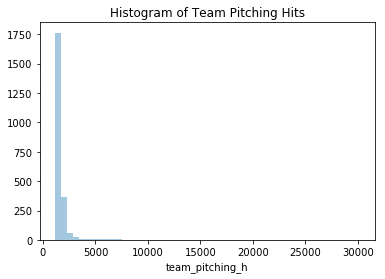
A few of the variables have missing data. We will create two new variables in this process; one with the IMP\_\* prefix for the imputed variable, leaving the original variable

untouched, and one with the m\_\* prefix as an indicator variable for the imputed variable. Sometimes, the fact that a variable was missing can actually be predictive. This means the indicator variables might be entered into the predictive model. For instance HBP and CS have only recently been kept track of. If they are missing, the data could be from an earlier period in baseball’s history, which could have predictive capability. I will impute each of these variables with missing data with their median, since as I will soon show, most of these distributions are highly skewed and irregular. The median, unlike the mean, is less susceptible to extreme values.

Next, after imputing these values, I will examine the distribution of each variable via a histogram and boxplot for extreme values, outliers, or wrongly inserted data. As I said earlier, the visual examination of the variables as histograms and boxplots shows wildly differing from expected normal shapes for several variables, some with straight-up impossible values. I will attach all of these graphics in the Appendix.

Some of these variables have an alarming shape to their distribution, with indication that there are some extreme values. I am looking for values that appear to be so extreme that I should simply impute them. With this line of reasoning I found several unreasonable seasonal averages for a couple variables. Here are some of my favorite outliers. One team in one season had their pitchers apparently log 19,278 strikeouts. Averaging that out per game gives 19,278/162 = 119 strikeouts per game. Unfortunately, there are only 9\*3 = 27 outs that are possible in a game making this point impossible. As such, this is not an outlier, but wrong data. Another one is 30,132 pitching hits given up over one season. This equates to 186 hits given up a game which is just absurd. These are a just a couple of impossible examples. However, many other points, while not impossible, break records listed on baseball-reference and baseball-almanac. For instance, one team apparently had 0 wins in a season, when the record for single season losses is 20.

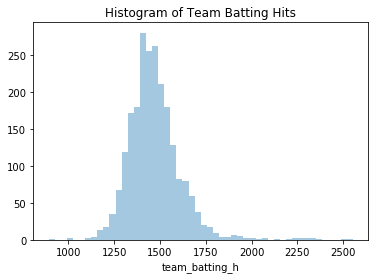
I am now going to consider historical reference data from Michael Bein’s ‘Century at a Glance’ Graphical History of Baseball to further validate this point for a subset of variables. While it is scaled slightly differently, the distributions should still be similar.

**Figure 1 Distribution of Pitcher Hits Allowed**

**Figure 2: Historical Hits Per Game**

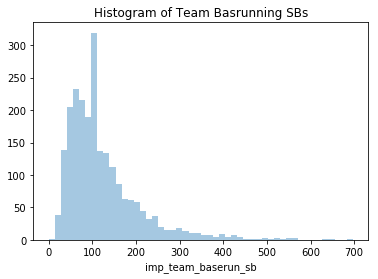
The maximum hits allowed is 11. Over 162 games, this provides a maximum reference of 1,782 hits. This means that any points over that number is unreasonable to say the least. The possible relationship TEAM\_BATTING\_H & TEAM\_PITCHING\_H should provide context for one another.

**Figure 3: Batting Hits**

****

This distribution looks more normal, but also contains extreme values, well over the 1,782 number determined above. Action needs to be taken to fix the outliers in both of those variables. Next is stolen bases per team per season.

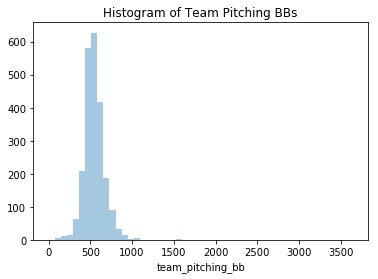
**Figure 4: Stolen Bases**

****

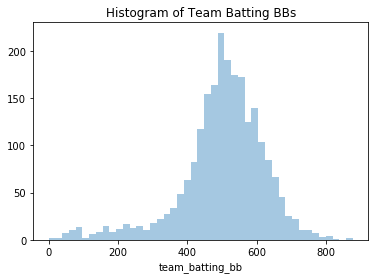
**Figure 5: Historical Stolen Bases per Game**

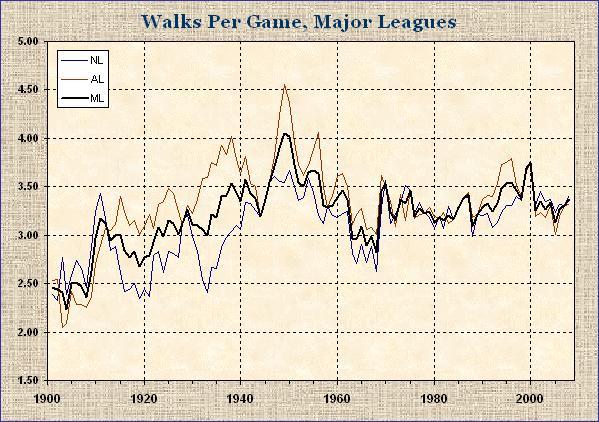
Again, the maximum rate per game is 1.5, which extends to 243 steals over a 162 game season. Our training data distribution of stolen bases has a long right tail extending up to 700! All of those points between 250 and 700 are outliers, and impossible. Next is pitching walks allowed, which should again be similar to batter walks.

**Figure 6: Pitching Walks Allowed**

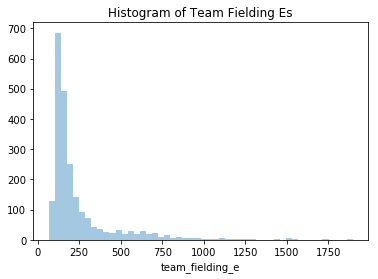
****

**Figure 7: Batters Walked**

****

**Figure 8: Historical Walks per Game**

There are a couple of things wrong here. First, the maximum rate is around 4.5 walks per game which extends to 730 for a full season. Both walks allowed and batter walks have seasons greater than this number. The lowest historical rate is around 2 per game which extends to 324 for a season. Both walks allowed and batter walks have seasons less than this number. Finally, I will end with fielder’s errors.

**Figure 9: Distribution of Fielder’s Errors **

**Figure 10: Historical Fielder’s Errors**

Again, the same kind of bad data can be seen. The maximum rate of errors was 2.5. Over a full season, that equates to 405 errors a season. However, our histogram shows a large right tail with many values greater than 500 errors. After analyzing the data in this way, almost all of the variables had issues. Here is the full list. Wins, Batting Hits, Batting 2B, Batting 3B, Batting Walks, Batting SOs, Baserunning SB, Baserunning CS, Fielding Errors, Pitching Walks, Pitching Hits, and Pitching Strikeouts.

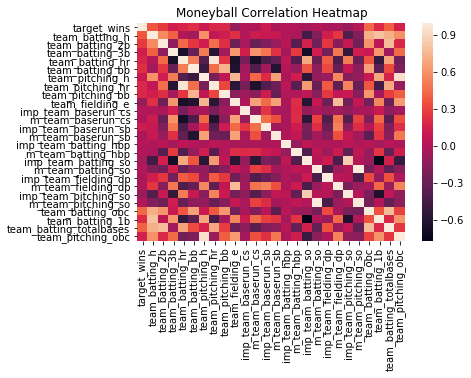
From our studies, I know that we have several options to fix this data:

* Truncating at a specific value (or set of values, e.g. ranges)
* log transformations
* Standardization (e.g. Z-Score)
* Binning (e.g. Buckets, Quantiles)
* Combining above techniques (e.g. log followed by binning)

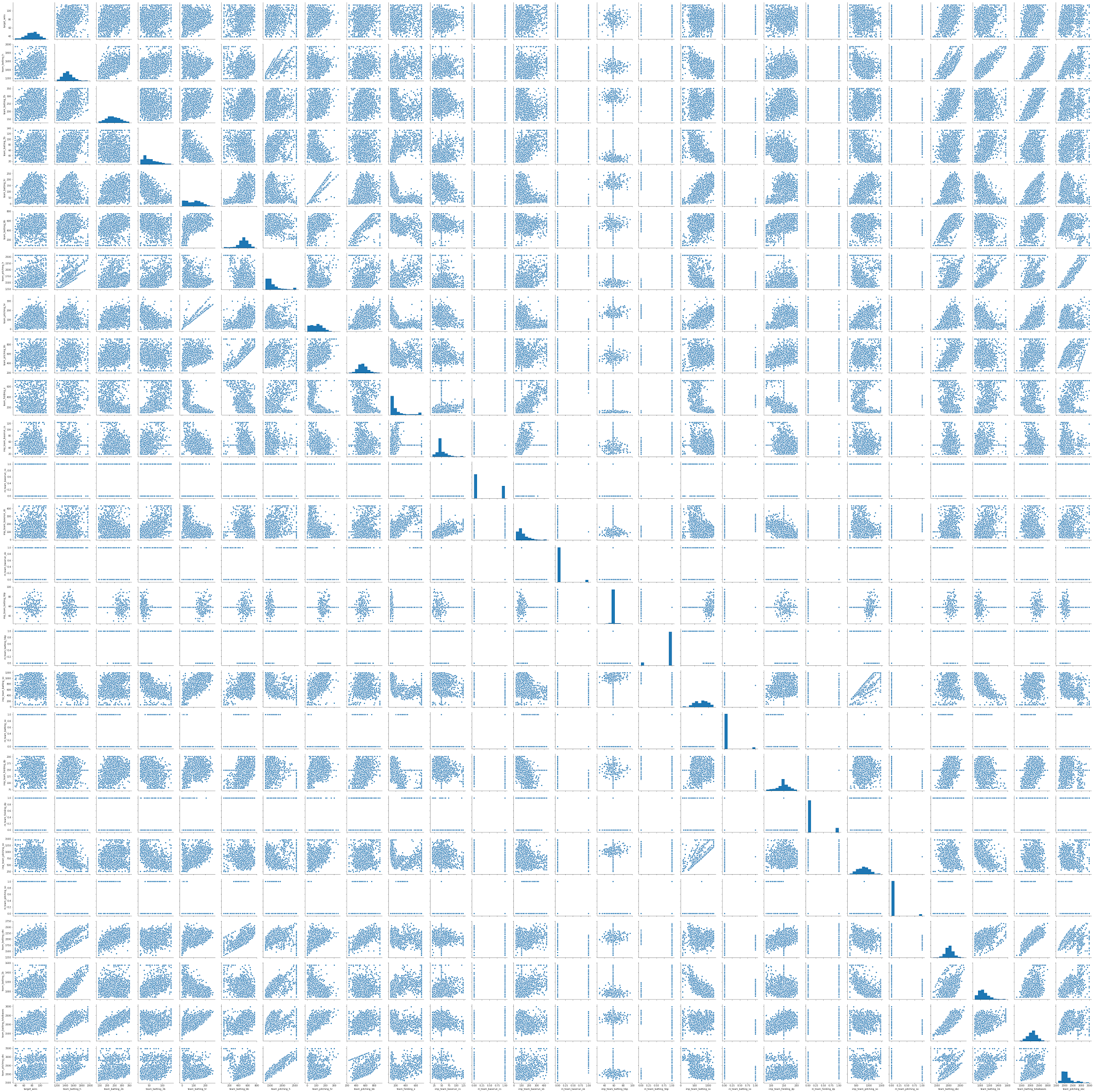
Since I am committed to using the OLS regression as our model, I cannot simply choose to alter the modeling technique in the face of the extreme values. Even though this is an assignment, there are many industries that have regulatory limitations that would prevent one from pursuing a different model out of convenience. I chose to use the truncating strategy at this time based off of quantiles. For the variables listed above, if any values exceeded the 99th percentile, then they were replaced with the value of the 99th percentile. Likewise for values less than the 1th percentile. Finally, with missing values imputed and outliers mostly fixed, I created the four variables of 1Bs, Total Bases, and Batting and Pitching On-Base Count, as defined earlier. It was important to remember to do these same actions with the test data. As such, I imputed missing data with the medians from the training data and truncating the variables using the original 99th and 1th percentiles of the same variables from the training set.

**Modeling:**

Before moving onto to creating models, I first looked at the correlations of all of our variables with a heatmap and a pair plot.

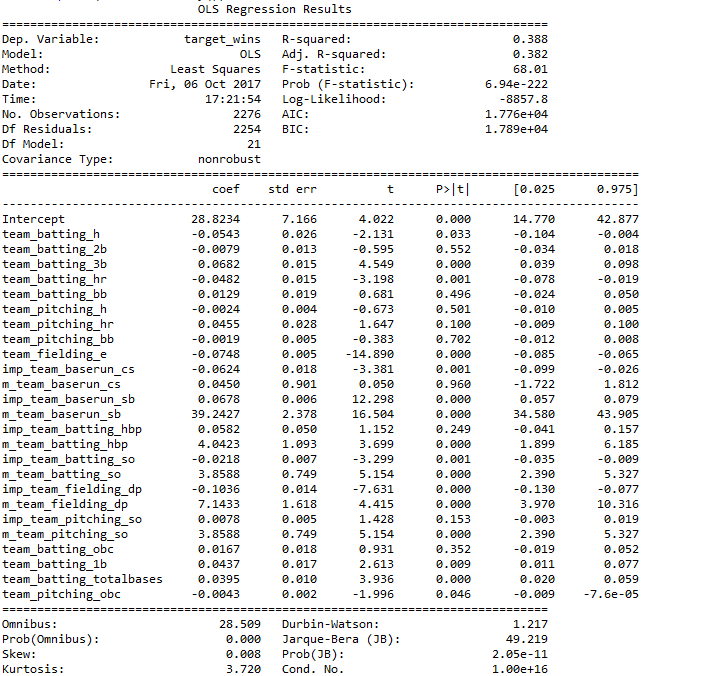
**Figure 11: Correlation Heatmap**

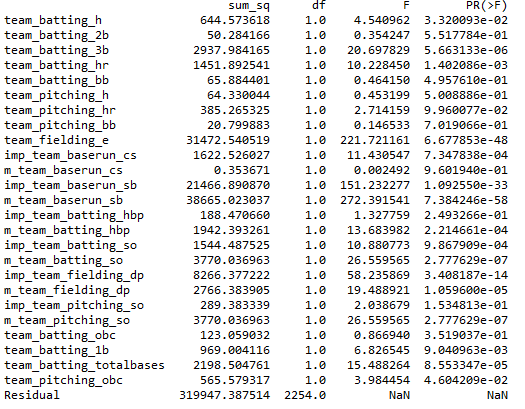
**Figure 12: Correlation PairPlot**

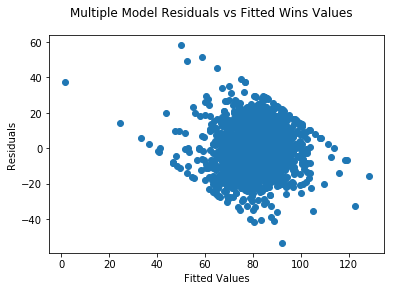
****

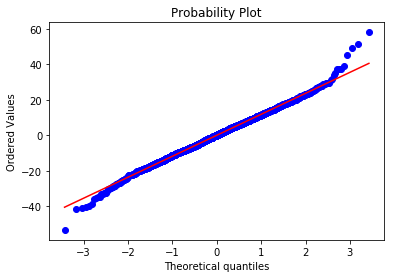
Perhaps unsurprisingly, total bases and the batting on base count had the highest correlation with target wins. These statistics are similar to On-Base Percentage and Slugging Percentage which sabermetricians have proved are correlated to runs scored, a precursor of wins and the numbers behind the OPS baseball statistic. Unfortunately, the correlation maps hint at potentially strong multi correlation issues in models developed, as many variables seemed to have a relationship with each other.

The first predictive model I created was a simple OLS regression model comprising of all 26 independent variables currently in the data set, including all imputed, fixed, and indicator variables. Here are the results from that model.

**Table 3: Model 1 Summary**

**Table 4: Model 1 Anova**

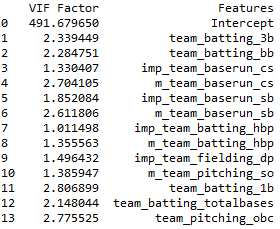
**Figure 13: Model 1 Residual vs Predicted Plot**

**Figure 14: Model 1 Residual QQ-Plot**

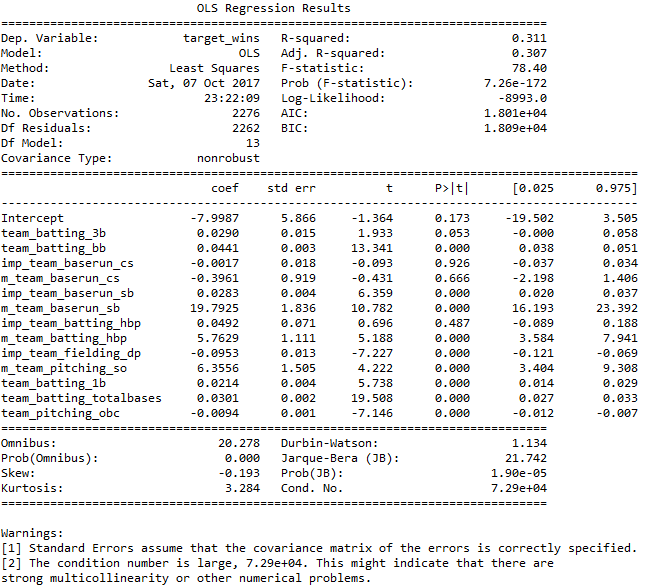
While this model seems to perform adequately, based off its relatively high, at least to the rest of the models, Adjusted-R^2 value, there a few things to comment. First, the many variables used present a good case for overfitting. Essentially, this generally takes the form of making an overly complex model to explain idiosyncrasies in the data under study. This model is quite complex. Additionally, not all of the independent variables are significant at the .05 alpha level. Furthermore, not all of variables are actually independent and multicollinearity does manifest itself in this model. The residual vs predicted graph shows two very clear linear limits, and is not randomized at the more extreme game wins. The QQ-Plot supports this as it shows that the residuals detour away from normality at the extremes. Finally, the mean absolute error of this model was 9.36. This model suffers from overfitting, multi-collinearity, and poor estimations for extreme values of the dependent variable.

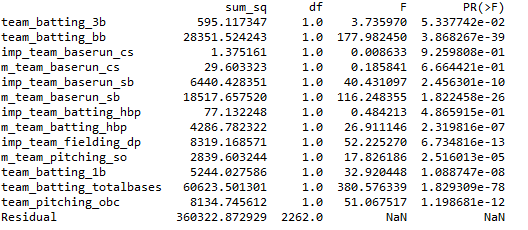
For the second model I created, I decided to eliminate multicollinearity as defined by the Variance Inflation Factor or VIF. After creating a feature auto-selection function, I set the VIF limit to be three. The following features were then used in the second model:

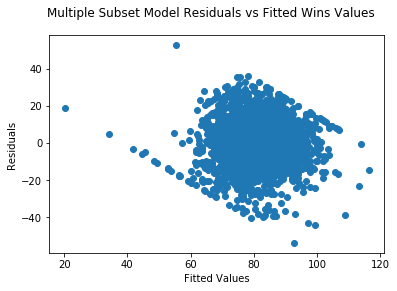
**Table 5: Features with VIF less than 3**

****

The output from a model created from these features are described below.

**Table 6: Model 2 Summary**

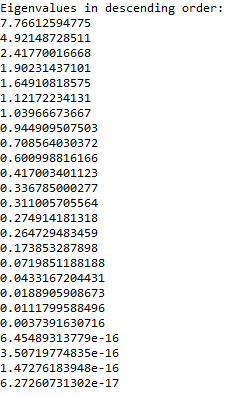
**Table 7: Model 2 ANOVA Table**

**Figure 15: Model 2 Residual vs Predicted Plot**

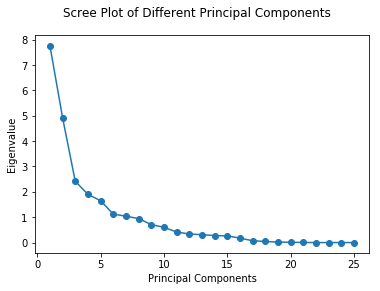
While this model reduced the effects of multicollinearity, it did so in expense of the ability to predict target wins. The Adjusted-R^2 decreased, the AIC and BIC both increased and the mean absolute error increased to 10.02. This model also seems to suffer most from predicting wins at the extremes, as smaller values are overpredicted and larger values are underpredicted. Additionally, a couple of variables are used which are deemed to be statistically insignificant. These are baserunning caught stealing and batting hit by pitch.

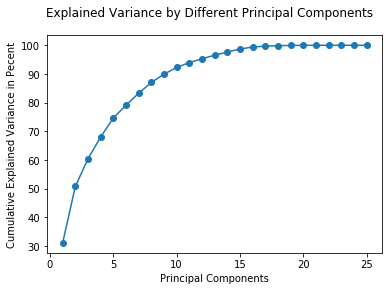
For my third model, I decided to reduce the dimensionality of the data by performing Principal Component Analysis, or PCA. The results of this analysis indicated that 7 vectors were significant, or had an eigenvalue of over 1.0. Using these 7 vectors, I performed a third regression analysis. The results are enumerated below.

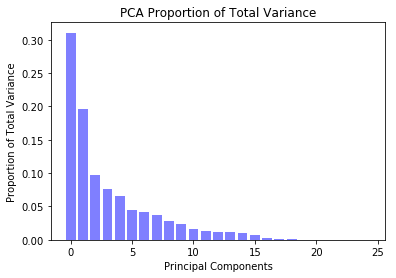
**Table 8: PCA Eigenvalues**

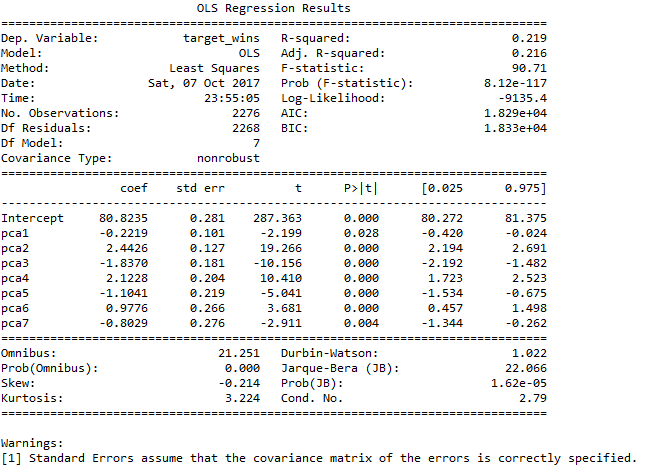


**Figure 16: Scree Plot**

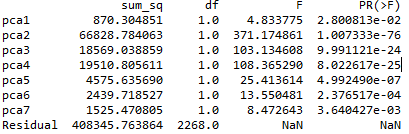
****

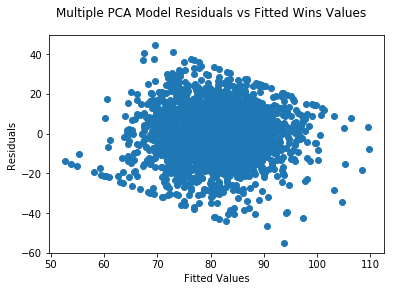
**Figure 17: PCA Cumulative Explained Variance**

**Figure 18: PCA Proportion of Explained Variance**

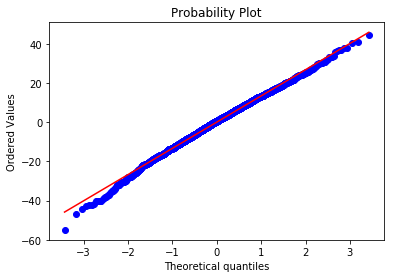
**Table 9: Model 3 Summary**

**Table 10: Model 3 ANOVA**

****

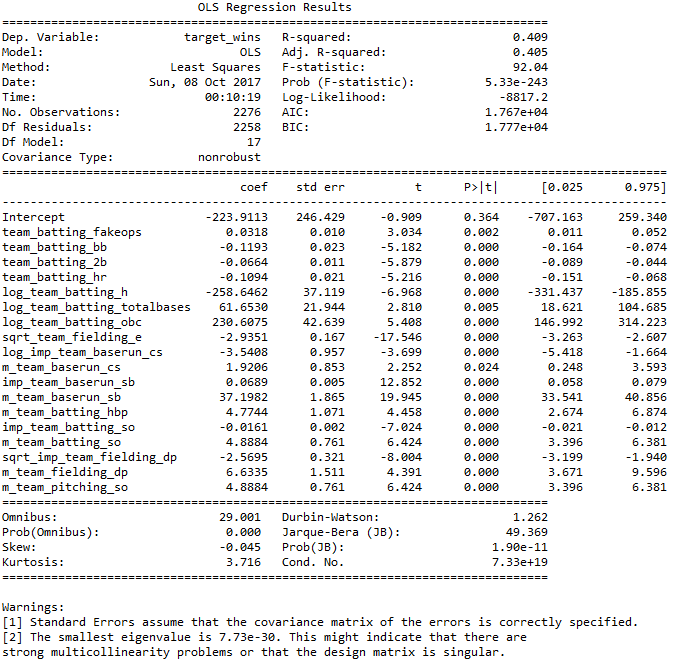
**Figure 19: Model 3 Residual vs Predicted **

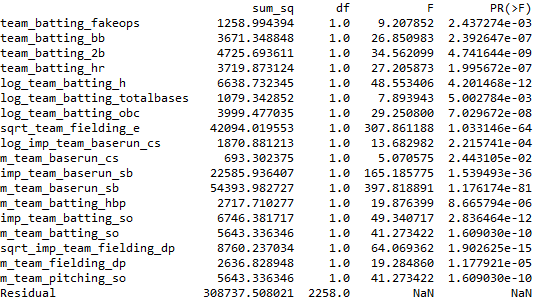
**Figure 20: Model 3 QQ-Plot**

****

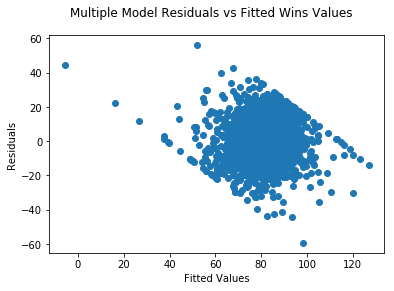
This model completely eradicated multicollinearity. However, only using the top 7 PCA vectors reduces the Adjusted-R^2 tremendously as well as increases the AIC and BIC. However, the residual plot is a lot more normal and the qq-plot shows less of an issue at the extremes. In actuality, this model performs poorly in its predictive ability.

For my final model, I used a combination of variable transformations and automatic feature selection of statistically significant variables. I also created a variable called Batting\_Fake\_OPS which was just the total bases added to the on base count. The variables I transformed are the following: Hits, Total Bases, On Base Count, and Caught Stealing were all transformed by the log function, and Fielding Errors and Fielding Double Plays were transformed by the square root function. Here are the results from that final model:

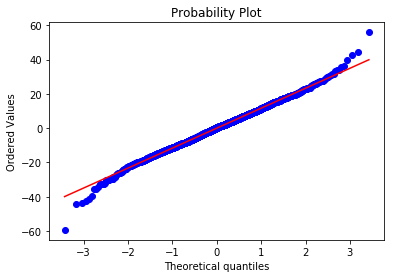
**Table 11: Model 4 Summary**

**Table 12: Model 4 ANOVA Table **

**Figure 21: Model 4 Residual Plot**

****

**Figure 22: Model 4 QQ\_Plot**

****

This model performs the best out of the bunch. The Adjusted\_R^squared has jumped above .40, the AIC and BIC have decreased, and every predictor variable is statistically significant at the alpha = .05 level. Unfortunately, this model does suffer from multicollinearity and predicts the edge cases of wins wrong quite drastically. The QQ-Plot shows a large deviation in normality at the extreme values, reinforcing these idea. However, this model seems to predict target wins the best. In fact, when used to predict wins on the test data set, it produced the best score on Kaggle as of early Sunday morning, 10/08/2017.

Here is a quick caveat. I did try and bin the fake OPS variable, however the resulting variable was not statistically significant and reduced the Adjusted-R^2 value and increased the AIC and BIC. As such, I did not include it in the final model. Additionally, of this final model, a few coefficients did not make much sense. For instance, the log of hits is quite negative. I would expect this value to be positive. However, due to collinearity with OBC and fake OPS, it makes some sense that this value is backwards. Other wrong coefficients are walks, doubles, and homeruns, which can be attributed to the strong multicollinearity issue, as well as fielding double plays, which is off by itself. Double plays have no collinearity with other variables and should increase the amount of wins a team has, however this model says the opposite is true.

**Model Selection:**

The best model in effect can account for 40% of the variability that is apparent

in the training data set. I am more accustomed to working with data sets where I can build models that have much higher goodness- of-fit diagnostics without nearly the amount of effort put into this model. Between all of the models constructed, a few of which are not documented in this report, I have to consider that the expected predictive ability on this data set is likely going to be quite low. It is highly likely that Simpson’s Paradox in in effect, which states, “what is true for the whole population may not be true for any of the subpopulations”. I am pretty confident that there are multiple sub-populations within this data, most likely due to how long of a period the data was being collected over. My attempts at trying to isolate this phenomena have only served to convolute the interpretation of the best model.

In dealing with wins in baseball, most analysts have determined that they are quite random and hard to predict. This is why many statistics used in sabermetrics deal with runs scored and runs allowed. Runs are the precursor to wins, and while wins are hard to predict, runs are much easier to correlate with and predict by simple statistics. In fact, if this study was done with runs scored, on-base percentage and slugging percentage percentage would perform the best in tandem to predict runs scored.

However, I am going to select the best model based off of Adjusted-R^2, AIC, BIC, and statistically significant predictor variables. In this case, I am forced to recommend my last model In this model, target wins is equal to 0.318 \* fake\_OPS - 0.1193 \* batting\_bb - .0664 \* batting\_2b - 0.1094 \* batting\_hr - 258.6462 \* log(batting\_hits) + 61.653 \* log(batting\_totalBases) + 230.6075 \* log(batting\_OBC) - 2.9351 \* sqrt(fielding\_errors) - 3.5408 \* log(baserunning\_cs) + 1.9206 \* m\_team\_baserun\_cs + .0689 \* baserun\_sb + 37.1982 \* m\_team\_baserun\_sb + 4.7744 \* m\_team\_batting\_hbp - 0.0161 \* batting\_so + 4.8884 \* m\_team\_batting\_so - 2.5695 \* sqrt(fielding\_dp) + 6.6335 \* m\_team\_fielding\_dp + 4.8884 \* team\_pitching\_so - 223.9113

**Conclusion:**

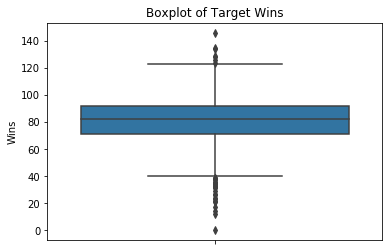
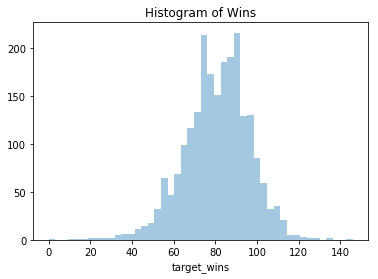
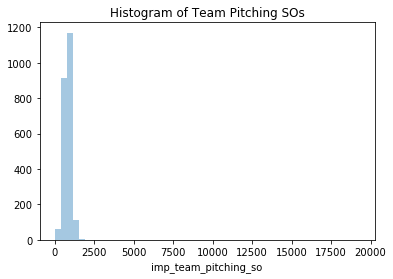
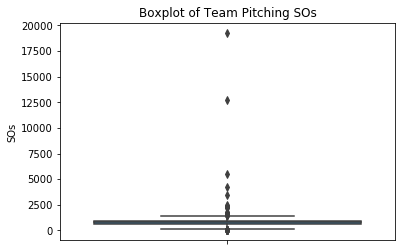
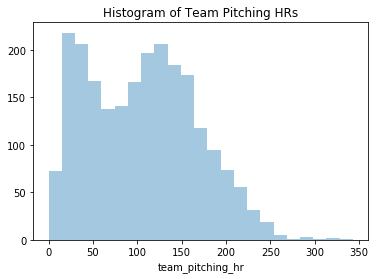
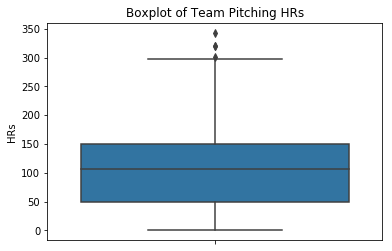
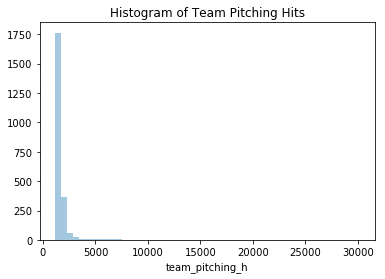
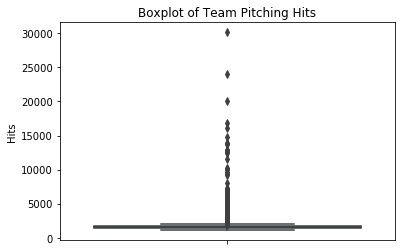
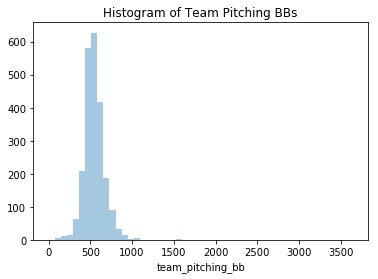
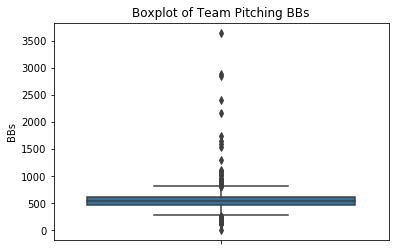
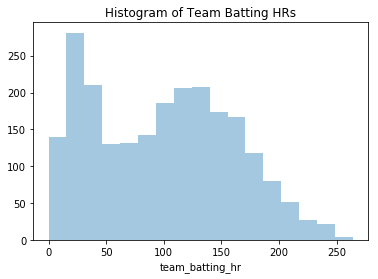
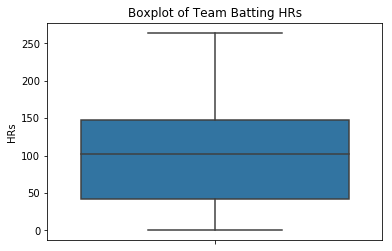
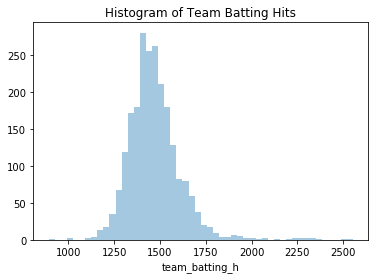
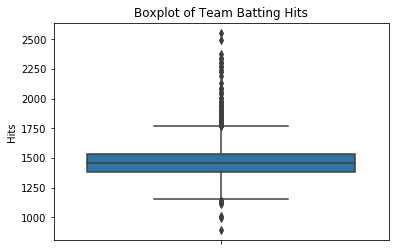
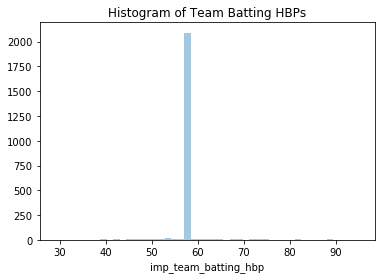
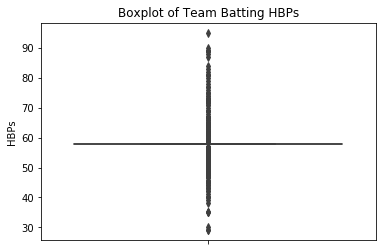
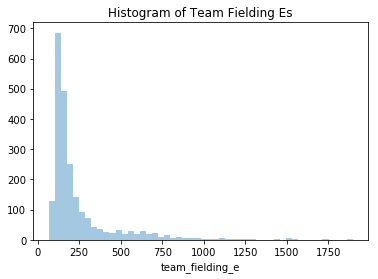
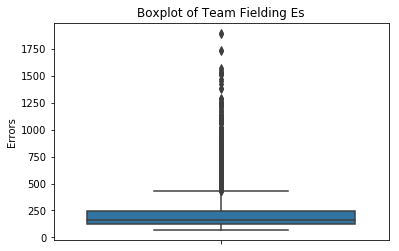
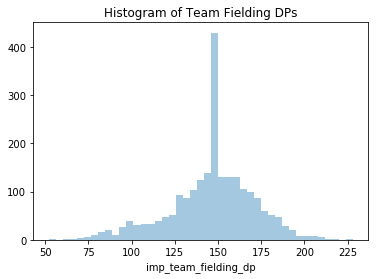
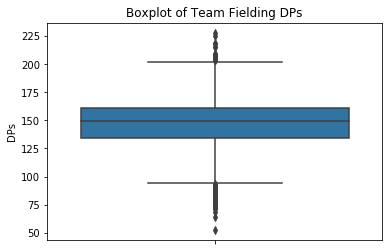
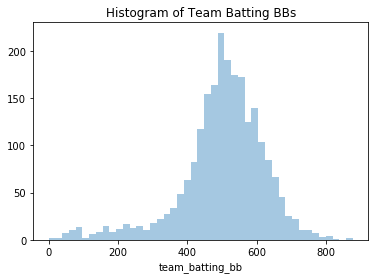
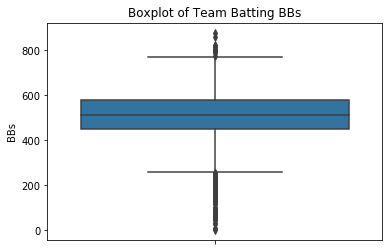
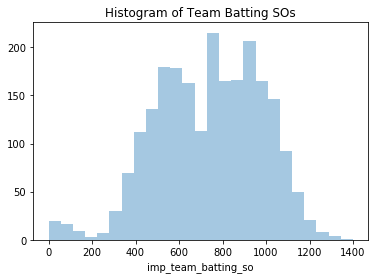
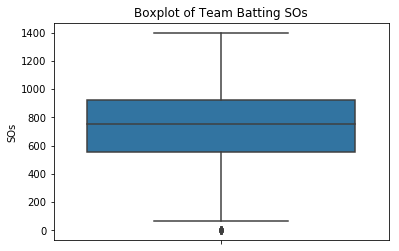
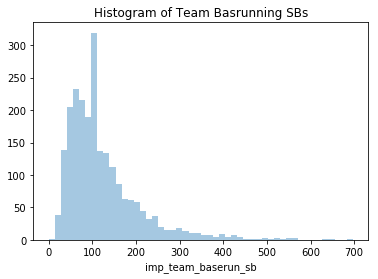
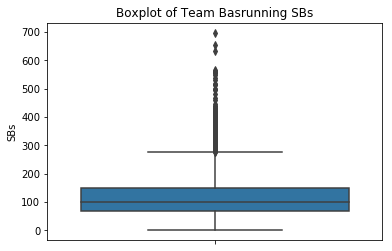
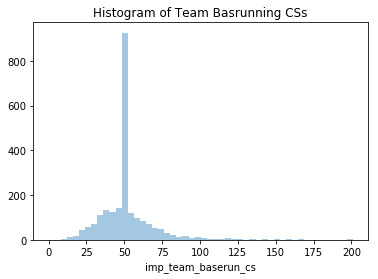
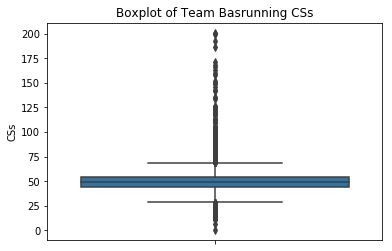
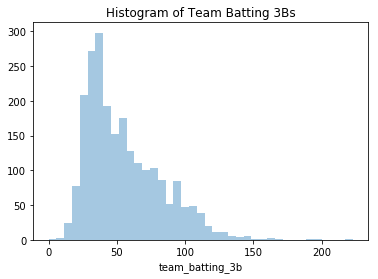
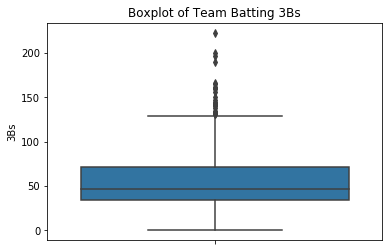
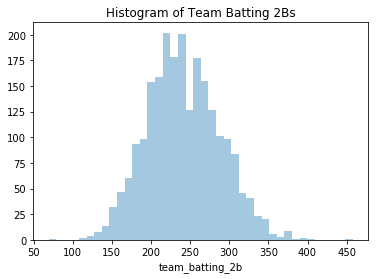
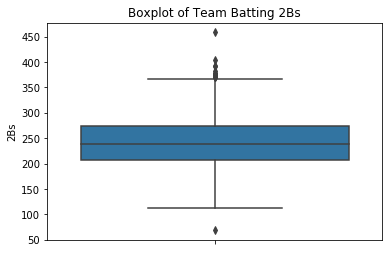
Obviously, this model is quite complex. However, wins in baseball are a complex entity. Interestingly, none of the pitching statistics made it into the final model. This is confusing, as pitching should make up about 50% percent of the games and have an effect on wins. This model does enforce that players with a high OPS and low error rate should be sought after on offense. This makes intuitive sense as more people on base, or getting more bases per plate appearance, creates more runs scored which should increase the number of wins for a team.

Unfortunately, I did not even comprehensively exhaust all of the modeling options at my disposal. The confidence in deploying this model for predictive value is relatively low. Given a data set that provided context, such as year or league, would have allowed me to build multiple models for what is likely very different subpopulations. Further, if the data set was not so error ridden, a more comprehensive and accurate model could have been created.

It may be best for teams in the future to predict runs scored or allowed through various statistics. OPS for runs scored and opponent OPS and pitcher WHIP are known to be good and correlated variables. Predicting runs provides a robust estimate for predicted wins, through formulas such as the Pythagorean formula. Additionally, predicting runs eliminates the noise and random error associated with predicting wins. I would recommend further analysis into the study of predicting runs, and then using that to predict wins, instead of this analysis to solely predict wins from basic baseball statistics.

**Appendix**

**EDA Graphics**

****