Winning Statistics: Analysis of Variables Indicative of Team Success in the MLB

By: Collin Fabian, Kyle Ventura & Arjav Shah

MA252 - Regression Analysis

Prof. Reagan Mozer

12/11/20

Contents

Introduction and Motivation:

Baseball is America's pastime and the sport is one of the most intricate, strategy based games in society today. There are several factors that lead a team to a winning baseball season and our team wants to investigate what those factors may be.The purpose of this statistical analysis was to try to understand which baseball statistics associate with wins in the Major League of Baseball. To begin our analysis, the team used the Baseball Databank from Kaggle. After identifying the dependent variable, win percentage, and independent variables, the team statistic columns from the databank, the team constructed a custom database with the necessary rows and columns. The custom dataset includes MLB offensive team data from 2010 to 2015.

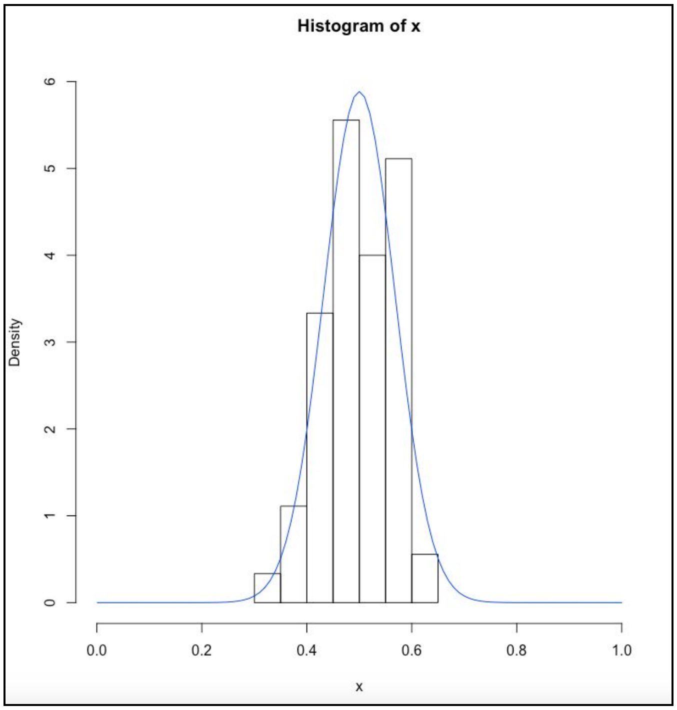
Our motivation for undertaking this analysis is to create a model that can identify which specific statistics are most indicative of a team's future success. One reason for developing this model is to be able to make smarter predictions about a team's success throughout the season. This would be very useful for individuals gambling on a team, the model would help identify which teams are most likely to do well as the season continues. Another reason for developing the model was so that other baseball teams within the MLB can use this as a guide for a winning season. A coach could look at the model and determine which aspects of the game they should focus on in order to have the greatest success as a team. Lastly, this model could help teams determine if the results of their season were reflected on the results of their statistics. Using this model, they can judge if they over or underperformed based on their statistical output.

EDA Methods (Exploratory Data Analysis Methods):

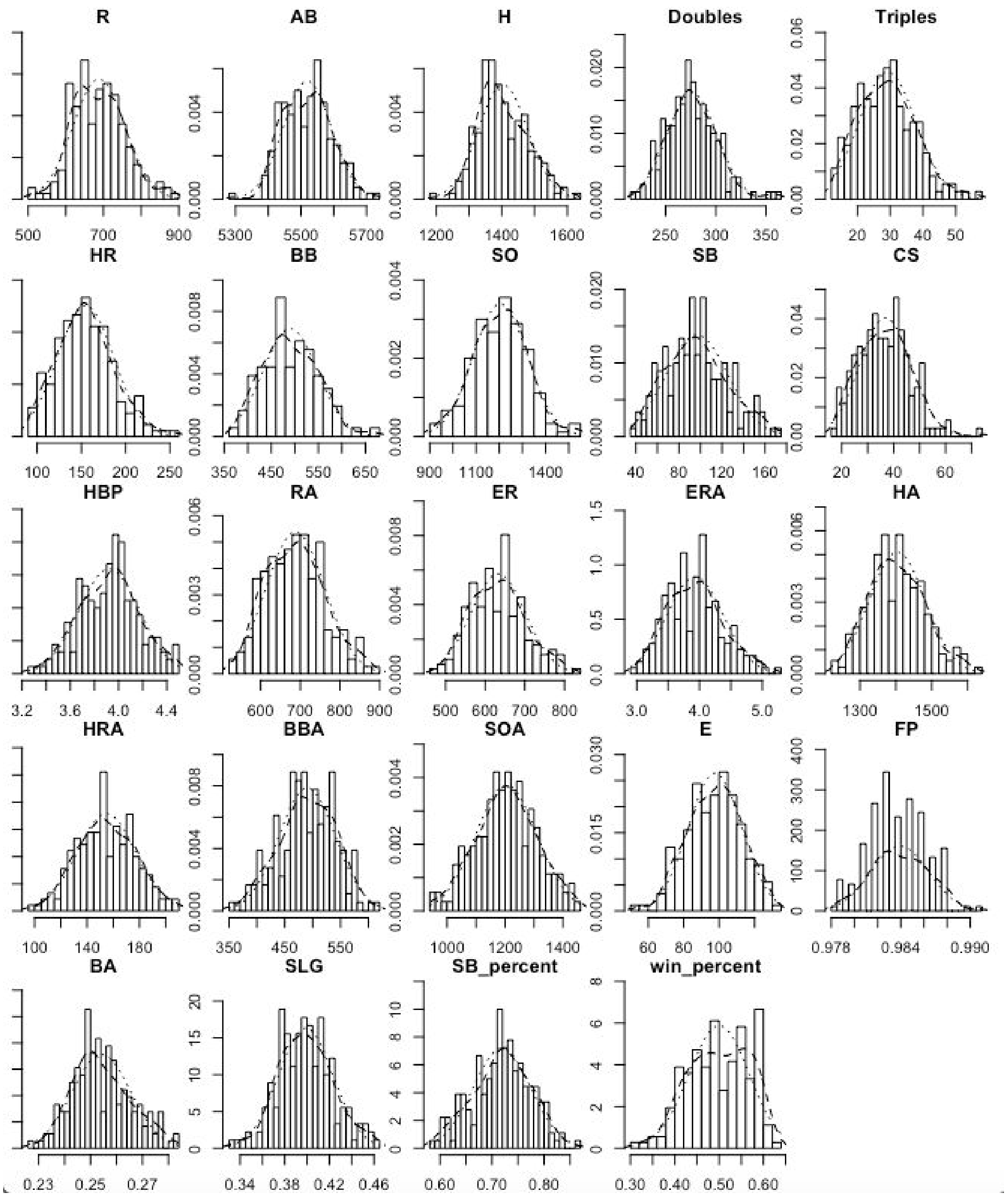
The dataset constructed from Kaggle took the winning percentage column as our dependent, y, variable. Our independent variables include R (runs scored), AB (at bats), H (hits by batters), 2B (doubles), 3B (triples), HR (home runs by batters), BB (walks by batters), SO (strikeouts by batters), SB (stolen bases), CS (caught stealing), HBP (batters hit by pitch), RA (opponents runs scored), ER (earned runs allowed), ERA (earned run average), HA (hits allowed), HRA (home runs allowed), BBA (walks allowed), SOA (strikeouts by pitchers), E (errors), FP (fielding percentage), BA (batting average), SLG (slugging percentage), and SB% (stolen base percentage). Our team cleaned this dataset by removing columns that were not points of interest according to the study and removing the years from 2015 onward. We also had to calculate averages for columns such as earned run average, batting average, and stolen percentage. Some factors provide insight into offensive team attributes, while the others provide insight into defensive team attributes. Each of the independent variables are classified as numerical data and the study does not include the analysis of any categorical factors.

In order to run regression analysis with R and further analyze the MLB dataset, the following libraries were used: glmulti, rJava, moments, normtest, ggplot2, plyr, psych, and readr. Each of these libraries makes running regression models and certain testing easier and quicker to interpret, with a great level of accuracy. To better understand the data at hand, our team used tools in R such as histograms, tests of skewness, and tests of kurtosis.

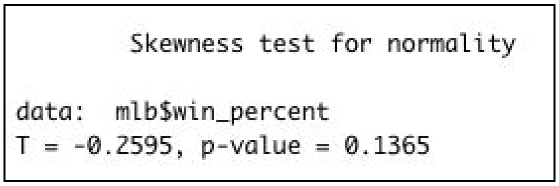
Beginning with our data, we set out to do a preliminary analysis of the response variable. We created a histogram of winning percentage to determine some of the assumptions for regression. We used the data winning percentage instead of wins due to the case that there might be a change in the number of games played over years.



After running a preliminary histogram model on our response variable we did the same to our explanatory variables to better view the data and visually see if there was any skew in the variables. The output of this can be seen below.

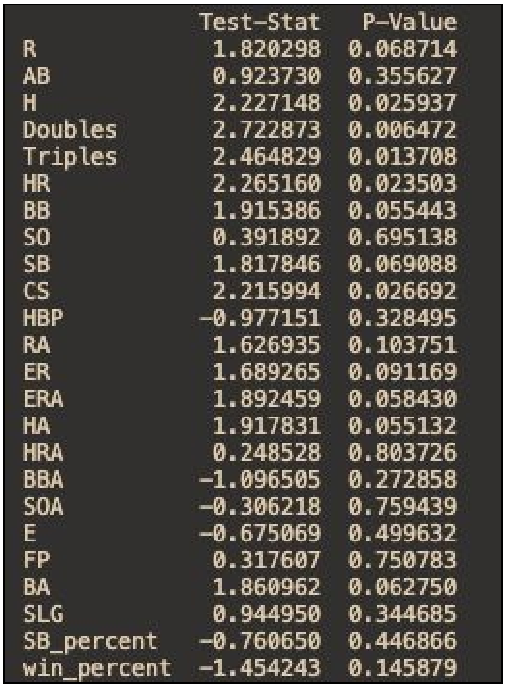


With this dataset, we ran a test on the variables to see if they were skewed. On win percentage, the p-value was greater than 0.05, meaning that there was no significant skewness to the data. For the rest of the variables, we found that hits(H), doubles, triples, home runs(HR), and caught stealing(CS) had a p-value less than 0.05.



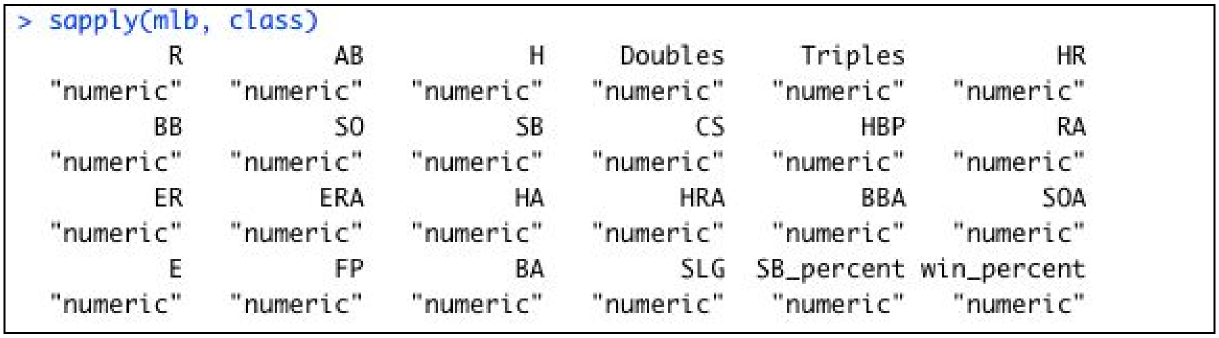
```
        Skewness test for normality

data:  mlb$win_percent
T = -0.2595, p-value = 0.1365
```



```
             Test-Stat   P-Value
R             1.820298  0.068714
AB            0.923730  0.355627
H             2.227148  0.025937
Doubles       2.722873  0.006472
Triples       2.464829  0.013708
HR            2.265160  0.023503
BB            1.915386  0.055443
SO            0.391892  0.695138
SB            1.817846  0.069088
CS            2.215994  0.026692
HBP          -0.977151  0.328495
RA            1.626935  0.103751
ER            1.689265  0.091169
ERA           1.892459  0.058430
HA            1.917831  0.055132
HRA           0.248528  0.803726
BBA          -1.096505  0.272858
SOA          -0.306218  0.759439
E            -0.675069  0.499632
FP            0.317607  0.750783
BA            1.860962  0.062750
SLG           0.944950  0.344685
SB_percent   -0.760650  0.446866
win_percent  -1.454243  0.145879
```

We can also see below that all of our variables are numeric, which you can see below. Initially, we had kept certain categorical variables like team franchise ID. However, when we tried to incorporate it into our model, it created too many variables for the model to be feasibly used in the future. For that reason, we decided to leave franchise ID out of the equation and focus solely on the numerical data. To account for potential differences among franchises (which we can assume will be the case), we will analyze the difference between franchises once we are analyzing our final model. Only then can we determine what variables are needed to be analyzed on a team-by-team basis.



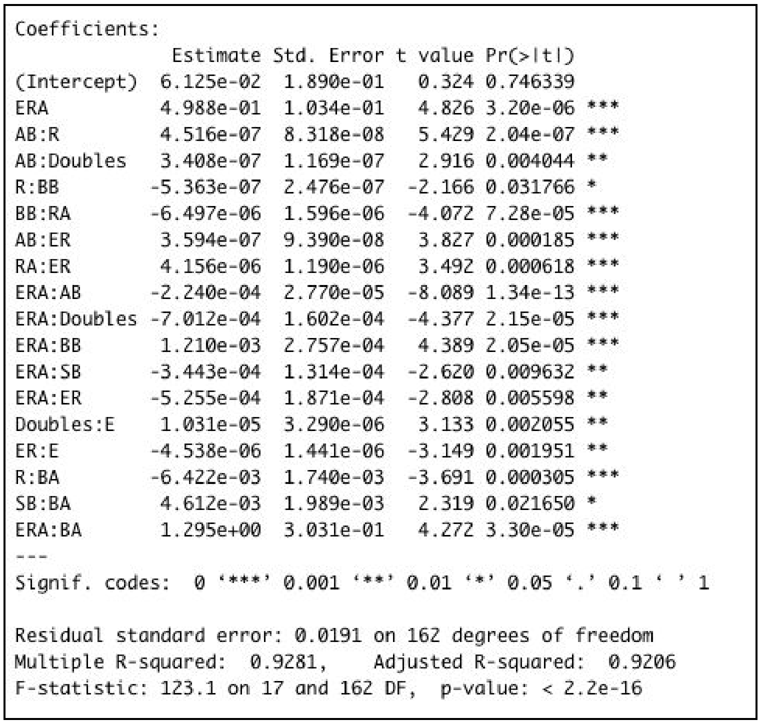
```
> sapply(mlb, class)
          R          AB           H     Doubles     Triples          HR 
  "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric" 
         BB          SO          SB          CS         HBP          RA 
  "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric" 
         ER         ERA          HA         HRA         BBA         SOA 
  "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric" 
          E          FP          BA         SLG  SB_percent win_percent 
  "numeric"   "numeric"   "numeric"   "numeric"   "numeric"   "numeric" 
```

For our model selection, we decided to use the package ‘glmulti’ to make our model selection. Glmulti is an automated model selector. Taking in a regression equation as input, it provides a wrapper for glm and other functions. It automatically generates all possible models using all variables in the given regression equation. From there the package uses Information Criterion (AIC, AICc or BIC) to choose the best model. Since we wanted to incorporate interaction terms into our model, it would not be feasible for the package to run an exhaustive search, meaning it would run through every single possible model. Therefore, we decided to use a genetic algorithm. A genetic algorithm is used when we want to find the best model, but cannot afford to search through all possible model combinations. It uses certain rules, generally based on some information criterion, to determine an appropriate termination point. For our case, we decided to use AIC as our information criterion as it is the only one we have used before in class. We decided that we would set our genetic algorithm to halt the model search automatically if it did not produce a new best model after 100 consecutive models.

Results:

Final Model

At the end we developed the following model:



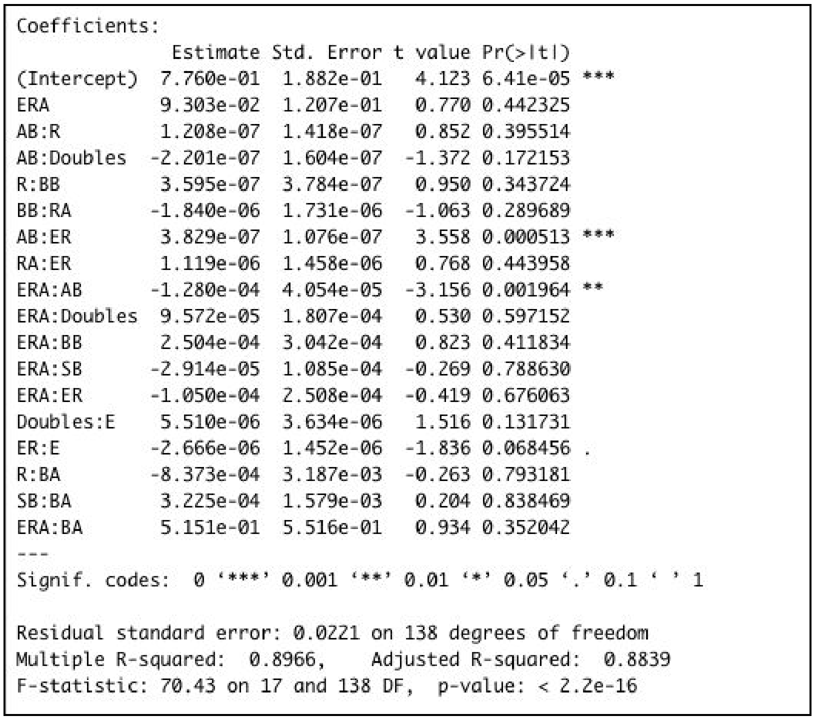
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.125e-02  1.890e-01   0.324 0.746339    
ERA          4.988e-01  1.034e-01   4.826 3.20e-06 ***
AB:R         4.516e-07  8.318e-08   5.429 2.04e-07 ***
AB:Doubles   3.408e-07  1.169e-07   2.916 0.004044 ** 
R:BB        -5.363e-07  2.476e-07  -2.166 0.031766 *  
BB:RA       -6.497e-06  1.596e-06  -4.072 7.28e-05 ***
AB:ER        3.594e-07  9.390e-08   3.827 0.000185 ***
RA:ER        4.156e-06  1.190e-06   3.492 0.000618 ***
ERA:AB      -2.240e-04  2.770e-05  -8.089 1.34e-13 ***
ERA:Doubles -7.012e-04  1.602e-04  -4.377 2.15e-05 ***
ERA:BB       1.210e-03  2.757e-04   4.389 2.05e-05 ***
ERA:SB      -3.443e-04  1.314e-04  -2.620 0.009632 ** 
ERA:ER      -5.255e-04  1.871e-04  -2.808 0.005598 ** 
Doubles:E    1.031e-05  3.290e-06   3.133 0.002055 ** 
ER:E        -4.538e-06  1.441e-06  -3.149 0.001951 ** 
R:BA        -6.422e-03  1.740e-03  -3.691 0.000305 ***
SB:BA        4.612e-03  1.989e-03   2.319 0.021650 *  
ERA:BA       1.295e+00  3.031e-01   4.272 3.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0191 on 162 degrees of freedom
Multiple R-squared:  0.9281,    Adjusted R-squared:  0.9206 
F-statistic: 123.1 on 17 and 162 DF,  p-value: < 2.2e-16
```

We can see that we got an adjusted R-squared value of 0.9206, which indicates that about 92.6% of the variation in our data can be represented by the model. Our model used a large amount of interaction terms, particularly involving the variable ERA and ER. Since these two variables revolve around the same statistic, earned runs, it tells us that earned runs play a very significant role in determining the amount of success a team will have. One would think that earned-run-average and win percentage would be negatively correlated; because after all, the more runs you give up, the more you have to score in order to win. However, our model actually has ERA associated with a positive coefficient. This means that, according to our model, the higher earned-run-average a team has, the more successful they will be on average. What this tells us is that teams that give up more earned-runs per game, must also be the teams that are scoring a lot. However, we must also consider the fact that there are many terms in our model where ER or ERA are interacting with another variable. All of the interaction terms regarding ERA or ER and some offensive statistic, are all positive. Therefore, we can conclude that these offensive statistics like: doubles, batters walked, at-bats and batting average play a significant role in the model as well. Thus, we can then interpret that the teams with high earned-run-averages but also have successful offensive statistics (especially regarding doubles, walks, at-bats and batting average) will be the teams that are the most successful on average.

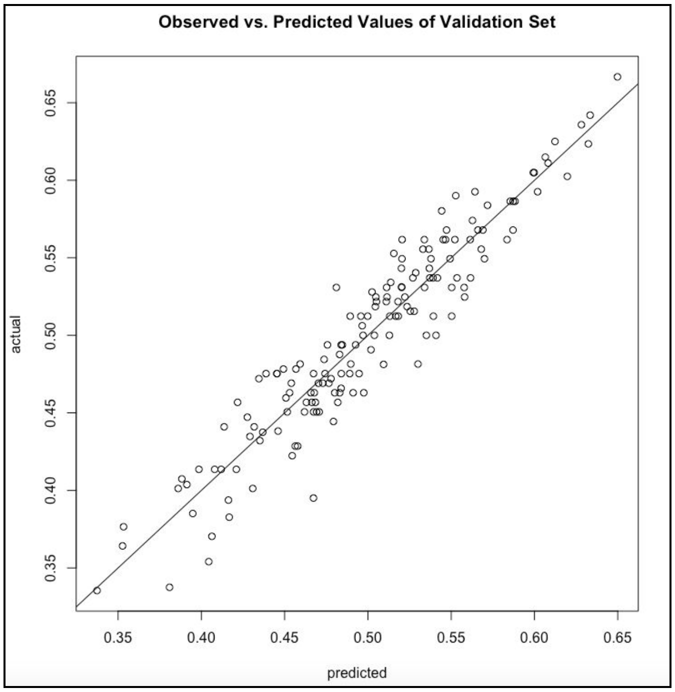
Check for Overfitting

Since our R-squared value is so high, it is essential that we suspect overfitting in our model. In order to check for this, we collected the same data from the years 1985-1990 as a validation set for our model. When we plotted the observed vs. predicted values of the validation data using our final model, we got the following:



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.760e-01  1.882e-01   4.123 6.41e-05 ***
ERA          9.303e-02  1.207e-01   0.770 0.442325
AB:R         1.208e-07  1.418e-07   0.852 0.395514
AB:Doubles  -2.201e-07  1.604e-07  -1.372 0.172153
R:BB         3.595e-07  3.784e-07   0.950 0.343724
BB:RA       -1.840e-06  1.731e-06  -1.063 0.289689
AB:ER        3.829e-07  1.076e-07   3.558 0.000513 ***
RA:ER        1.119e-06  1.458e-06   0.768 0.443958
ERA:AB      -1.280e-04  4.054e-05  -3.156 0.001964 **
ERA:Doubles  9.572e-05  1.807e-04   0.530 0.597152
ERA:BB       2.504e-04  3.042e-04   0.823 0.411834
ERA:SB      -2.914e-05  1.085e-04  -0.269 0.788630
ERA:ER      -1.050e-04  2.508e-04  -0.419 0.676063
Doubles:E    5.510e-06  3.634e-06   1.516 0.131731
ER:E        -2.666e-06  1.452e-06  -1.836 0.068456 .
R:BA        -8.373e-04  3.187e-03  -0.263 0.793181
SB:BA        3.225e-04  1.579e-03   0.204 0.838469
ERA:BA       5.151e-01  5.516e-01   0.934 0.352042
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0221 on 138 degrees of freedom
Multiple R-squared:  0.8966,	Adjusted R-squared:  0.8839
F-statistic: 70.43 on 17 and 138 DF,  p-value: < 2.2e-16
```

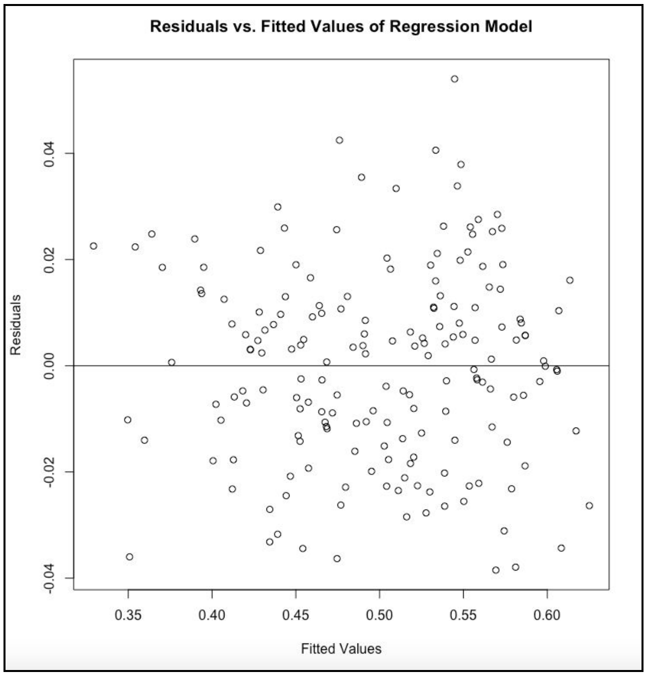


As we can see, our model produces an adjusted R-squared of 0.8839, even when we use our validation data. This tells us that our model isn’t just good for predicting amongst its own input data, but it can also be used as a good predictor when it comes to other data. However, what is interesting is that many of the variables in this instance have p-values that are above 0.05. But when looking at the graph of our observed vs. predicted values, the model still seems to do a good job at predicting win percentage. Therefore, we can be sure that our model will be effective going forward with predictions about current and future seasons. Next we will also make sure that none of the model assumptions for linear regression are being violated.

Model Assumptions

1. *Constant Variance*

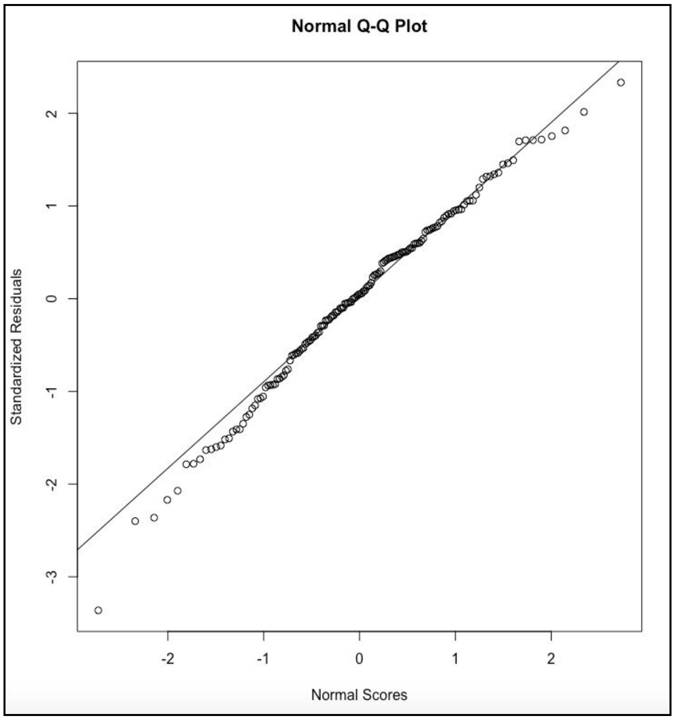
Our first assumption is that we must have constant variance amongst our data. To check this assumption we must look at a scatter plot of our residuals vs. our fitted values.



As we can see, our points are randomly scattered about the line y=0, and show no signs of patterns. Therefore, we can assume that our data has a constant variance.

2. *Normality of Residuals*

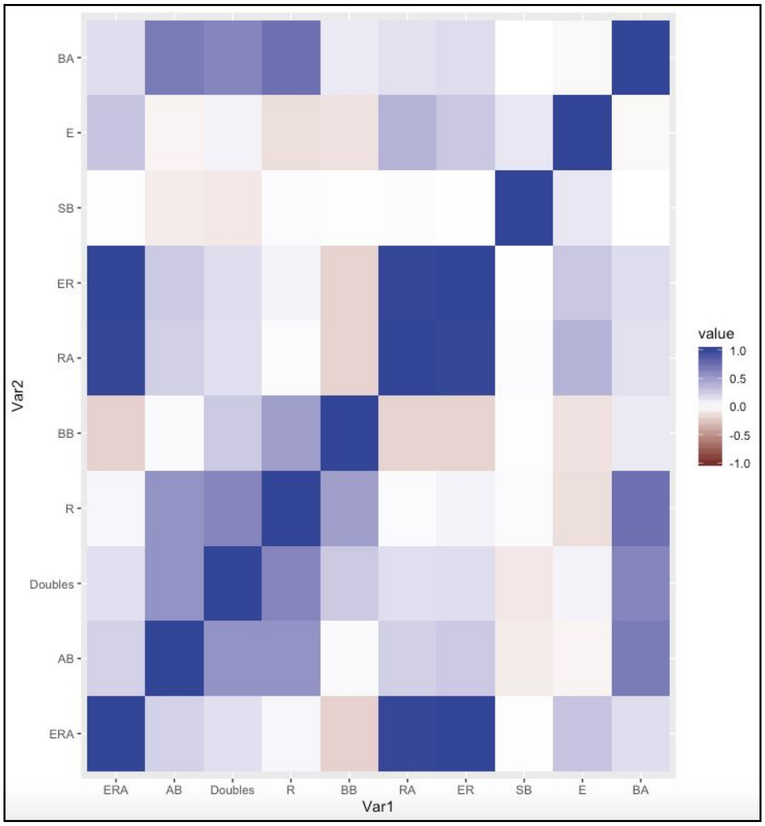
The next assumption we had to validate with our model is that our residuals are normally distributed. To check this, we created a QQplot of the residuals' standard quantiles vs their corresponding quantiles if it were perfectly normal.



As we can see, our residuals’ tails do not flair off nor do they develop any definite patterns. This means that we can assume that our residuals are normally distributed.

3. *No Multicollinearity*

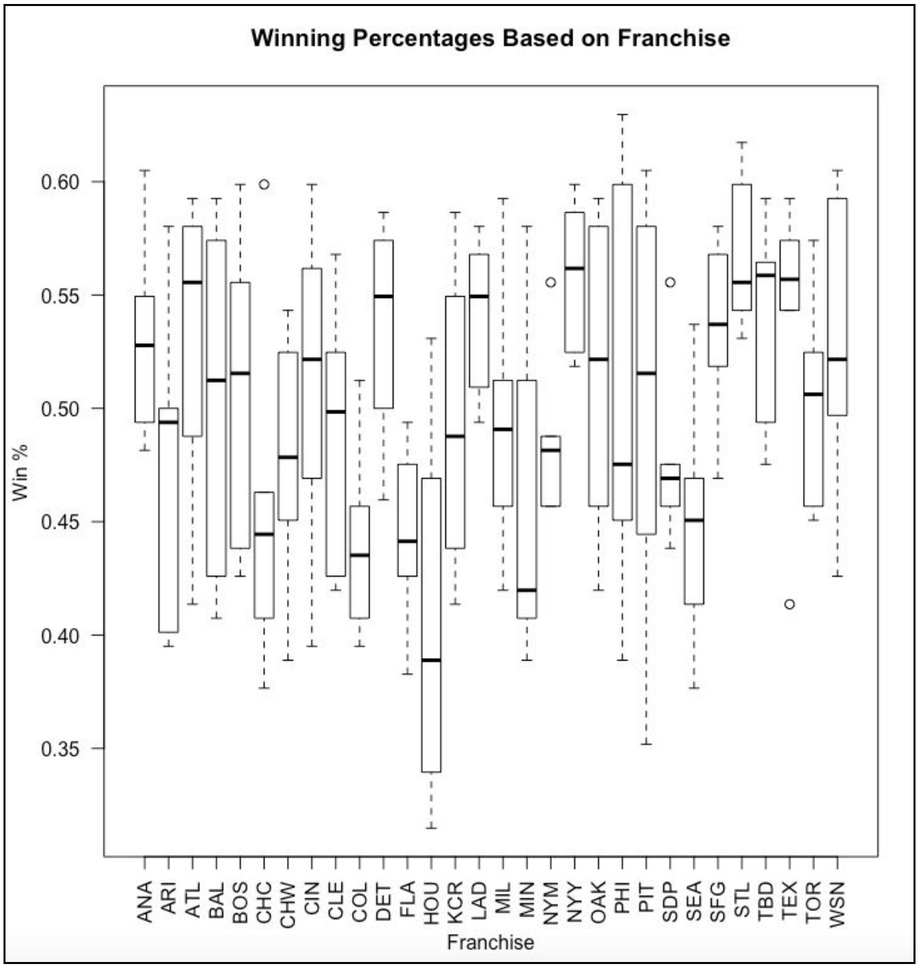
The next assumption is that there is little-to-no multicollinearity. Multicollinearity occurs when the explanatory variables are too highly correlated with each other. In order to test this, we must produce a correlation heatmap of all the variables in the model.



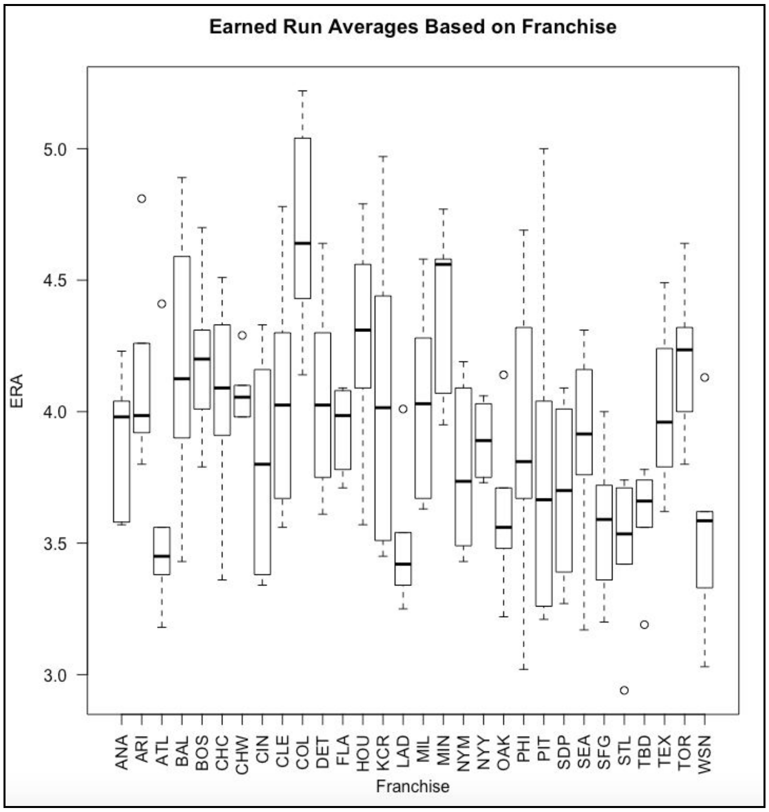
We can see that there are a few variables that seem to be somewhat positively correlated with each other. However, there are no instances that raise any alarms. Therefore, we can assume that our explanatory variables have little-to-no multicollinearity.

Team-by-Team Analysis

Another part of our analysis was to figure out if the significance of certain variables differ from one team to the next. To do this, we grouped our data based on franchise ID. We first looked at the variance in distributions of winning percentages over the years.



When we look at our plot, it is clear that some teams have seen more consistent success than others. For instance, it is clear that the St. Louis Cardinals, New York Yankees and Texas Rangers have been the three most consistently dominant teams while the Chicago Cubs, Colorado Rockies and Florida Marlins have been the consistently worst teams. What we now wonder is how some of the significant predictors in our model differ between these teams as well. First, let’s take a look at the same boxplot, but with ERA used for our y-values.



We can see that the distribution of ERA's across all of the teams in our data vary tremendously. There are certain teams like Colorado and Toronto who have small distributions with ranges less than 0.75 ERA's. However, there are other teams like Kansas City, Baltimore and Philadelphia whose distributions' ranges are as much as 1.75 ERA's. What this means to us is that the effects of ERA on a particular team will probably differ depending on what franchise we are making predictions about. The only way to test this theory for certain is to try and incorporate franchise ID into our regression model. However, because there are about 30 different teams, it would probably be too extensive to feasibly do. So, if one is to use the model for future purposes, it would be best to pay attention to what franchise you are making predictions for.

Conclusion:

Overall, using our model showed us that the variables most indicative of a team's success are ERA, at-bats and batting average. These findings, along with our overfitting test and models for assumption, can be used to help understand future statistics in baseball and can help teams with decisions involving the future of the team. Individual teams can monitor disiencies in their team and where an increase in certain areas can improve them the most. This along with roster construction and management can lead to a better team on the field. While not perfect, baseball can definitely use statistics and regression to their advantage in trying to predict the future based on the past.