Kristina White
Seung Yi
ACMS 30600-01
12/12/19

Linear Regression Project: 2019 MLB Pitching Statistics

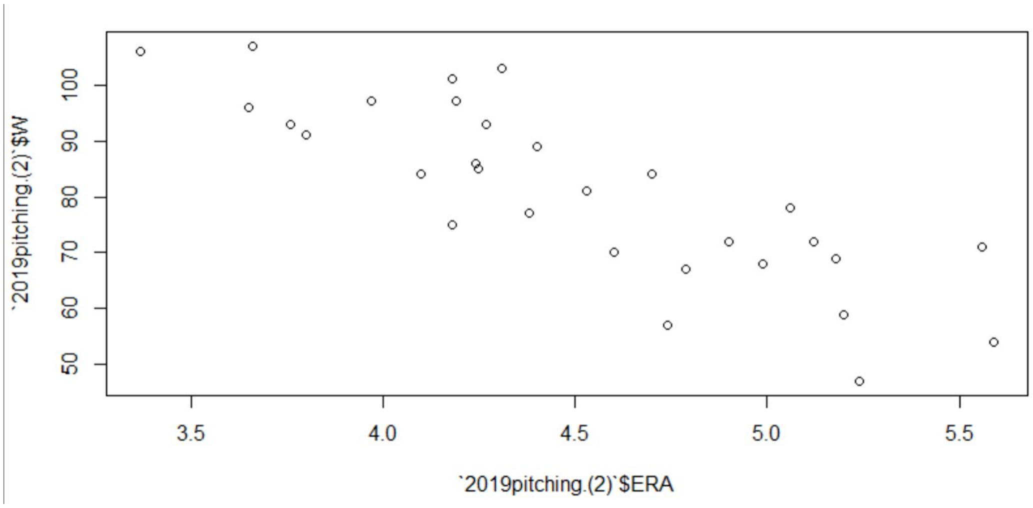
1. Introduction

   We are analyzing the MLB Team Pitching Statistics for 2019. The variables we will include in our model are ERA, R, SO, AVG, and WHIP. The sample size is 30 teams. From this analysis, we hope to know how these variables influence the number of regular-season wins for a team. In particular, we want to learn which best predict the number of games won.

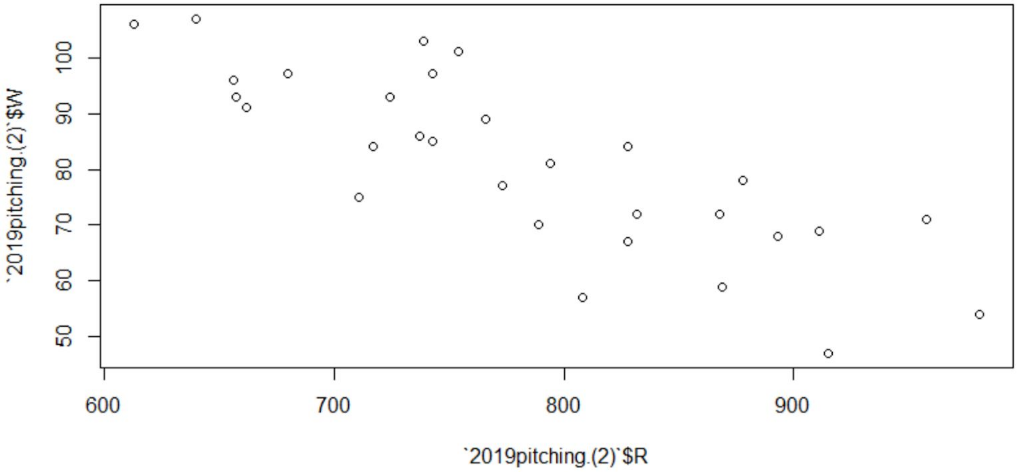
2. Data

   Our data was taken from the MLB website that featured the pitching statistics for the 2019 teams in the MLB. Below is the link.
   http://mlb.mlb.com/stats/sortable.jsp?c_id=mlb#elem=%5Bobject+Object%5D&tab_level=child&click_text=Sortable+Team+pitching&game_type='R'&season=2019&season_type=ANY&league_code='MLB'&sectionType=st&statType=pitching&page=1&ts=1574806369112
   Here is a copy of the text file containing the data.
   https://drive.google.com/open?id=1koEZqr_xMBfP2rrrtMuGCqcDGx1otwXu

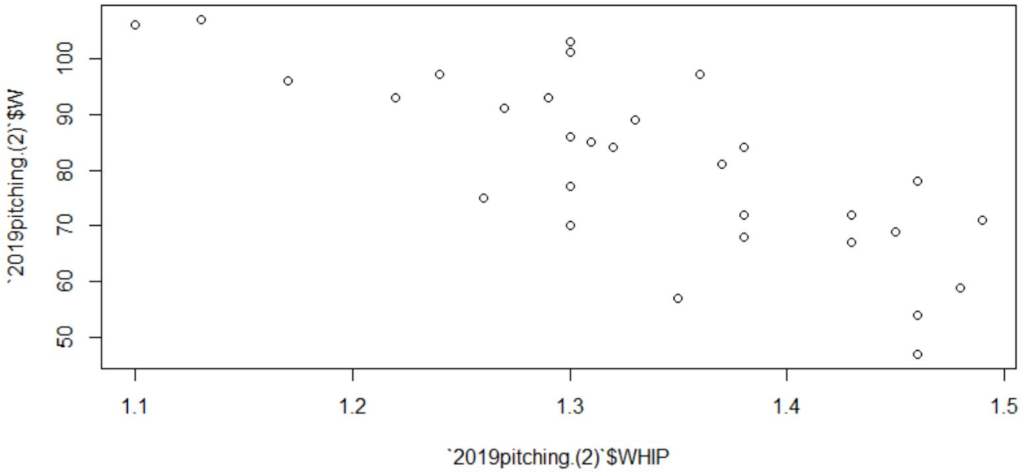
3. Regression Analysis
   a. Exploratory Data Analysis
      A. Scatterplot of Wins Predicted by ERA



B. Scatterplot of Wins Predicted by R



C. Scatterplot of Wins Predicted by WHIP



As a result of performing a test for multicollinearity using the cbind() function, the X variables that are most highly correlated with each other are ERA and R, AVG and WHIP, ERA and WHIP, R and WHIP, and R and AVG. We defined "highly correlated" as having a correlation of at least 0.90.

b. Linear Regression Analysis

The full model with all of the X predictors has a corresponding R^2 of 0.7327 and adjusted R^2 of 0.677.

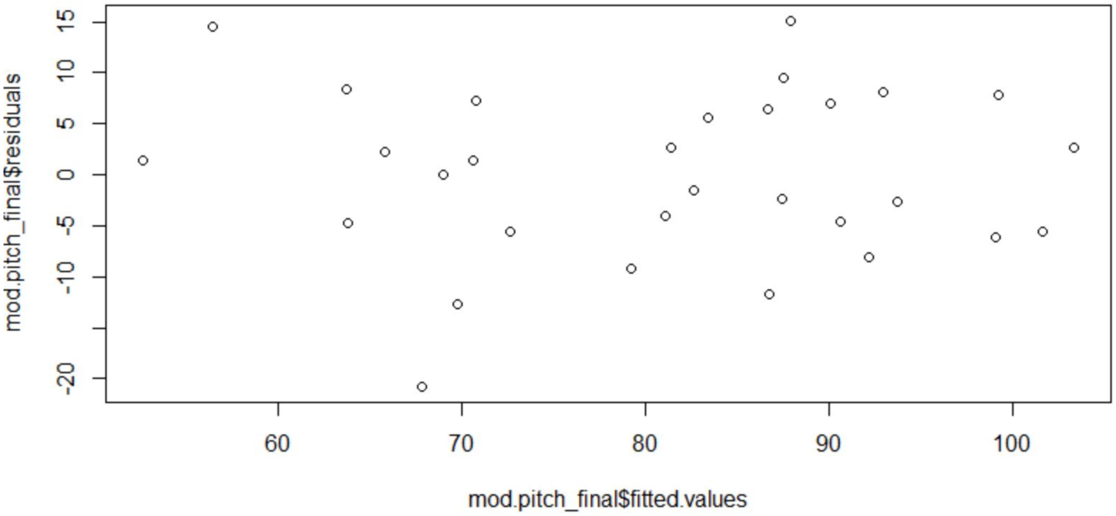
We performed an F-Test for nested models to remove a subset of variables from the full model. Based on the scatterplots and their correlations, we expected to remove three predictors: R, SO, and AVG. The F statistic had a value of 1.1042, and the p-value was 0.3667. Using α=0.05, we failed to reject the null hypothesis and concluded these predictors were significant and should remain in the model.

However, after using the stepAIC() function to find the optimal subset of predictors, we found that the optimal model predictors are ERA, SO, and AVG with a minimized AIC of 133.73.

Using the rstandard() and cooks.distance() functions, we found that there are no outliers or influential observations in the full model.

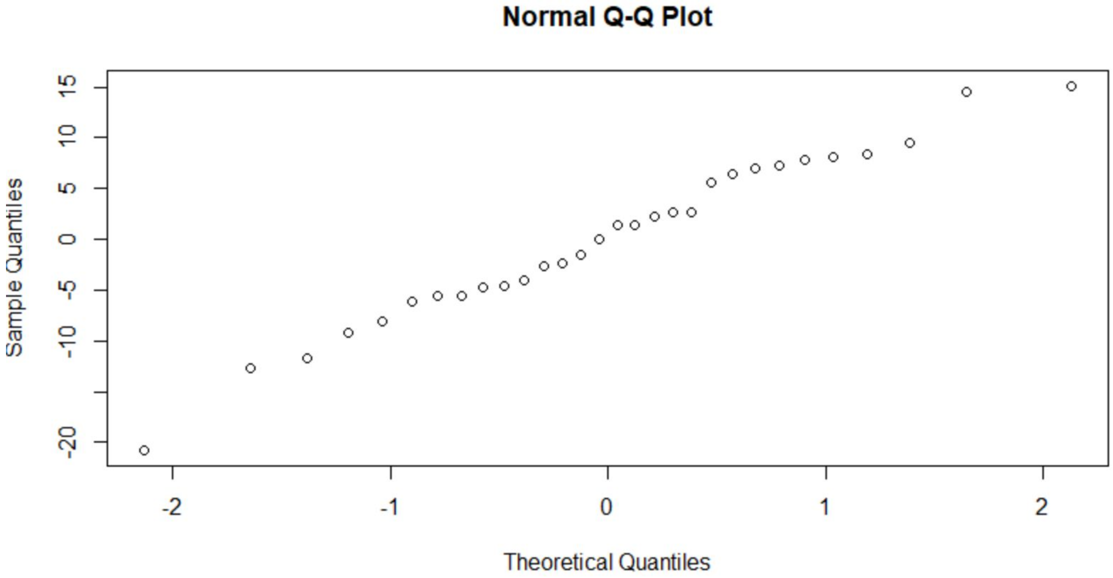
Based on the results from the stepAIC() function, and considering that the F-Test and the scatterplots showed ERA as a promising predictor, we chose ERA, SO, and AVG as predictors for our final model. We confirmed this by fitting the final model and seeing R^2 = 0.7295 with an Adjusted R^2 of 0.6983.

D. Residual Plot



Seeing that there are no significant patterns in the residual plot, the model assumptions are upheld. These assumptions are that the residuals have a mean of zero, follow the normal distribution, have constant variance, and are independent.

E. Normal QQ Plot



Seeing that the QQ plot shows a significantly straight pattern, the model assumption that the residuals follow the normal distribution is upheld.

The 95% confidence intervals, with a critical value of 2.0555, for each of the predictors are as follows:
ERA: upper bound=-15.0073 lower bound=-40.6199
SO: upper bound=0.0726 lower bound=-0.0095
AVG: upper bound=972.4051 lower bound=-151.9528

4. Results
Taken from the final model summary, the p-values for the three predictors are as follows: ERA=0.000137, SO=0.126358, AVG=0.145677. The only significant p-value is that of ERA because it is less than α=0.05. Therefore, we reject the null hypothesis that β_{ERA}= 0 and conclude that the predictor is significant. For SO and AVG, we fail to reject the null hypothesis that β_{SO} and β_{AVG} are each equal to zero.

The 95% confidence interval for ERA was constructed in the section above and is (-40.6199, -15.0073). The confidence intervals for SO and AVG are (-0.0095, 0.0726) and (-151.9528, 972.4501), respectively. These are plausible ranges for β_{ERA}, β_{SO}, and β_{AVG}, respectively.

5. Conclusion
Our final model predicts Wins fairly well. The R^2 is 0.7295 and the Adjusted R^2 is 0.6983, which implies pretty decent predicting power. The model could still have been

improved with additional predictor variables we chose not to include due to assumed redundancy. Perhaps including those predictors might have yielded a better final model despite the sacrifice to simplicity. It would be interesting to initially model the data with all of the given possible predictors and to see if it is possible to fit a better final model. Lastly, we were also curious about the AVG predictor yielding such a large slope and standard error, though these are proportional to each other and yield a reasonable t-value. Since β_{AVG} represents the expected increase in Wins on average for every unit increase in AVG, it could be beneficial to multiply the small decimal values for AVG by 100 before fitting a model to produce a more reasonable slope and standard error.

6. Appendix

After downloading the 2019pitching.txt file from Section (2) Data, the code below should run, assuming that the downloaded file is still named 2019pitching.
https://drive.google.com/open?id=1KALM6UNGo6mq7IxhXDEMuxDBaIystoB1