Statistical Analysis of MLB Pitchers

Sebastian DiPrampero

November 2021

1 Introduction

For my project, I am using data from Lahman's Database, which is an extensive collection of Major League Baseball statistics that ranges from 1871 to 2020 and covers pitching, hitting, and fielding statistics for every MLB player in history. It also includes statistics about coaches, managers, Hall of Famers, and even every MLB park. I have extracted a list of 500 pitchers throughout MLB history and I am using the total number of wins that they earned throughout their entire career as my response variable. I am using a group of pitching statistics recorded for each player as my design matrix. The goal of my project is to determine which pitching statistics have the most predictive power over the total number of wins a pitcher will achieve in their career, and I want to create a model that can predict the player's total wins using only a few of these forecasted statistics. The purpose of this model will be to assist managers in deciding which pitchers to keep on their team's roster and which players to cut. MLB pitchers are often some of the highest paid athletes in the world, so it is of crucial importance that when a manager decides to keep an expensive player, they can be confident that they will receive a return on investment, meaning that the pitcher will accumulate a sum of wins over a certain period of time. I first performed Simple Linear Regression on a select few statistics that I gathered, then I moved on to a more comprehensive method of Multiple Linear Regression to analyze my data.

2 Data Configuration

I took a subset of the data and used 500 arbitrary pitchers in MLB history. I provided a short sample of my filtered data in the figure below:

Player	Wins	ERA	HR	Games	Strikeouts	WP	Losses	Shutouts
1	87	4.496	154	263	888	53	108	6
2	86	3.496	138	732	1030	53	81	0
3	75	3.211	123	467	856	35	72	9
4	97	3.540	87	248	526	24	80	8
5	82	4.835	264	291	723	22	83	4
6	110	4.284	157	261	1350	76	79	5

My response variable is represented by the wins column which corresponds to the total wins for the pitcher in their career. My design matrix values include ERA which stands for earned run average, HR which is home runs allowed, total games played, total strikeouts, total wild pitches, total losses, and total shutouts. Each statistic is aggregated over the pitcher's full career, and I had to remove a few pitchers that had NA values reported for some of their statistics. In total, my data has 498 observations with 7 explanatory variables.

My general plan was to perform multiple linear regression using all of my pitching statistics, but I decided to first test some simple linear regression to see if a pitcher's wins could be predicted with only one variable. The first statistic I decided to test was ERA, because this is widely recognized by managers in the MLB as one of the best stats to determine a pitcher's efficiency. I also decided to test strikeouts since this is a good representation of a pitcher's skill level and overall ability without influence from his team. The final statistic I chose to test was games played since it is indicative of a pitcher's durability.

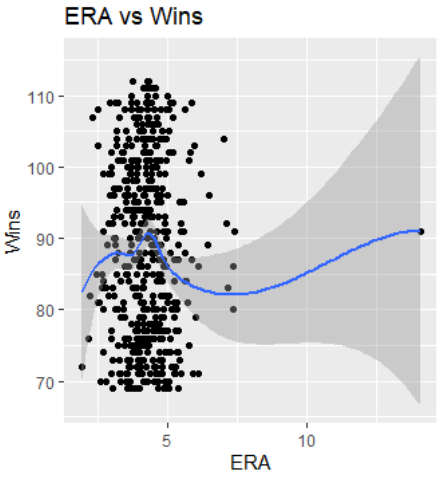
3 Simple Linear Regression

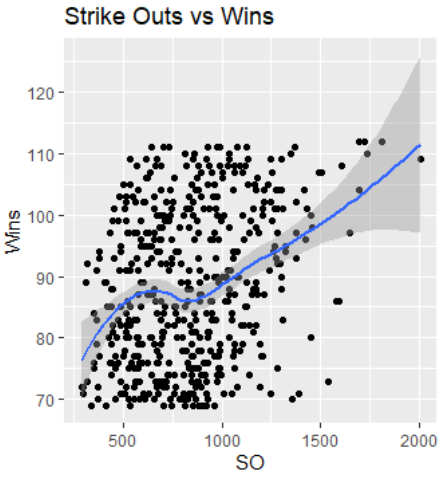
In my first test of ERA vs Wins, the graph produced clearly shows that my model was not very successful in being able to predict the amount of wins for each pitcher linearly.

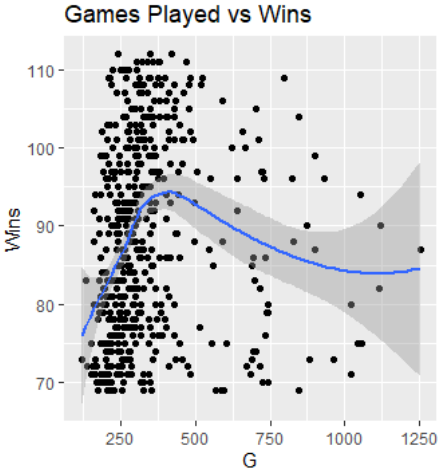
The differences in ERA among my sampled pitchers was too small to generate a very linear model. The points seem to be almost randomly distributed around an ERA value of 3.5, with the exception of a few outliers. Regardless, the shape of the fitted line compared to the observed values makes it obvious that simple linear regression is not going to be a very good predictor of MLB pitchers wins when the regressor chosen is ERA.

My second test of Strikeouts vs Wins worked far better than ERA vs Wins, but it still was not perfectly linear or particularly accurate in doing its job of predicting wins. It is also evident that the regression line is influenced greatly by only a few data points at high strikeout values, and if I happened to sample a different group of pitchers, these points may have been absent, causing a distinctly nonlinear relationship between Strikeouts and Wins

Finally, I tested Games vs. Wins, which clearly follows a nonlinear pattern as well. At first glance it seemed like simple linear regression was not going to be the best way of predicting wins for MLB pitchers over their careers.





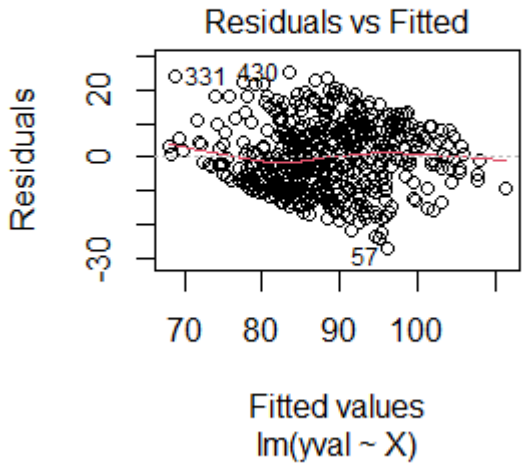


Looking at the values associated with my Simple Linear Regression models, we can see that all three regressors have p values that are not significant at the 0.05 level. They each also have very low adjusted r squared values. This confirms that simple linear regression will not work for variables I chose in my data set.

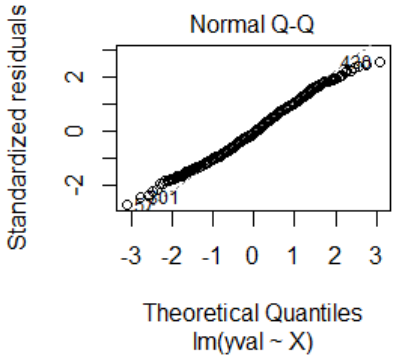
Simple Linear Regression			
Regressor Tested	**P Value**	**F Statistic**	**Adj R Squared**
ERA	0.0652	0.2676	-0.001482
Strikeouts	9.352e-11	43.83	0.07964
Games Played	0.217	1.528	0.001066

4 Multiple Linear Regression Assumptions

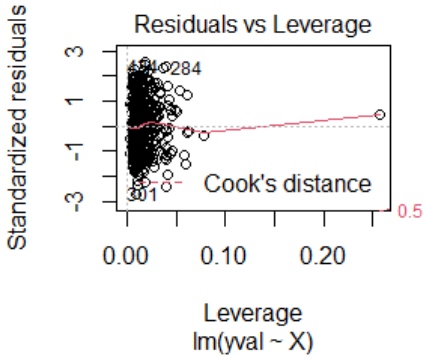
Before I began testing my data with Multiple Linear Regression, I wanted to check the assumptions of uncorrelated errors, homogeneity, normal errors, and linearity. I first generated a plot of my residuals vs fitted values, and the data points seem to be randomly distributed around zero.



I then created a Normal QQ plot with my data, and I found that the plotted values generally follow the normal line, with only some small deviations around the tails.



I also generated a plot of the residuals vs the leverage values, which demonstrates that there are no outliers with a large influence on my data, since none of my points lie outside of Cook's distance.



Based on these graphs, I decided to accept the assumptions for Multiple Linear Regression.

5 Full Model Description

$y = 36.20 + -4.45x_1 + -0.02x_2 + 0.05x_3 + 0.01x_4 + 0.09x_5 + 0.45x_6 + 1.43x_7$

ERA	x_1
HR	x_2
G	x_3
SO	x_4
WP	x_5
L	x_6
SHO	x_7

Using all seven of my regressors, I generated an initial model that I intended to compress into a simpler model later. One initial observation from my full model's equation was that ERA had the largest coefficient and therefore the largest impact on my model. It also had a negative coefficient meaning that a larger ERA results in less wins. This makes sense since a lower ERA corresponds to a pitcher allowing the opposing team to score less runs, resulting in more wins.

6 Checking for Multicollinearity

Covariance Table							
	ERA	**HR**	**G**	**SO**	**WP**	**L**	**SHO**
ERA	1.00	0.40	-0.09	0.09	0.09	0.19	-0.24
HR	0.40	1.00	0.16	0.66	0.04	0.29	-0.37
G	-0.09	0.16	1.00	0.21	0.01	0.16	-0.35
SO	0.09	0.66	0.21	1.00	0.35	0.19	-0.30
WP	0.09	0.04	-0.01	0.35	1.00	0.23	-0.14
L	0.19	0.29	0.16	0.19	0.23	1.00	0.17
SHO	-0.24	-0.37	-0.35	-0.30	-0.14	0.17	1.00

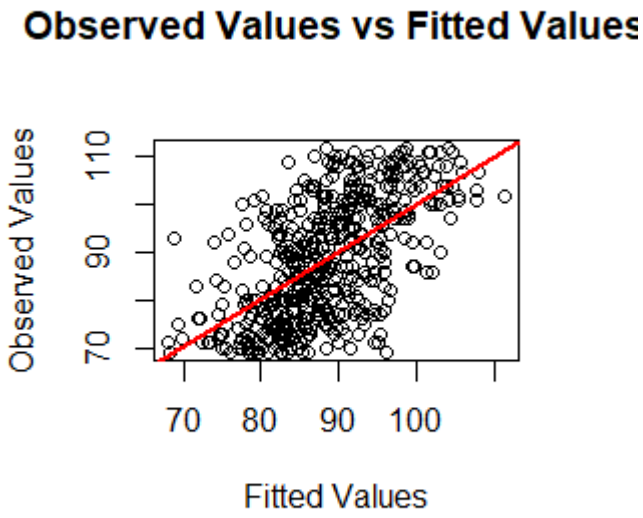
Before I tried to reduce my model, I wanted to check for multicollinearity in my data. The fact that no pair of variables generates a large value in the covariance table indicates that there is no issue with multicollinearity. The largest value in the table is 0.66 which occurs when looking at strikeouts with home runs, but the cutoff for issues with multicollinearity occurs at about 0.9 so this value is acceptable.

I also calculated the Variance Inflation factor for each of my regressors, and every value is below the cutoff of 5, which means that it is reasonable to assume little to no multicollinearity in my model.

Multicollinearity Analysis	
Regressor	**V.I.F.**
ERA	1.420148
Home Runs	2.870985
Games Played	1.349837
Strikeouts	2.433856
Wild Pitches	1.458530
Losses	1.512123
Shutouts	1.699806

7 Full Model Tests

To evaluate how well my full model predicts my data, I plotted the observed values vs the fitted values and found that my data has about an equal amount of points above and below the fitted line.



I found that my model has an R squared value of 0.3652 and an Adjusted R squared value of 0.3561. This means that about 35 percent of the variation in total wins is explained by my model.

I then wanted to do an F-Test on my data to make sure that at least one of my regressors is statistically significant, and I found that my model has an F value of 40.106 which results in a p value that is close to zero, meaning that we can reject the null hypothesis stating that all of my regressors are equal to zero.

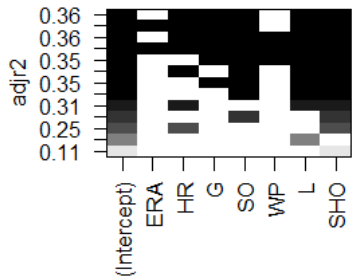
F-Test $H_0 : B_1 = B_2 = B_3 = B_4 = B_5 = B_6 = B_7 = 0$	
F Value	40.106
Degrees of Freedom	7, 488
P Value	2.2e-16

I also looked at each regressor's individual t-test to see which variables were most impactful in the full model. Based on these tests, we can see that ERA and Wild Pitches have the largest p values, indicating that removing them may result in a more efficient model.

Individual T-tests		
Regressor	**T Value**	**P Value**
ERA	-0.840	0.4013
Home Runs	2.133	0.0334
Games Played	1.966	0.0498
Strikeouts	4.905	1.27e-06
Wild Pitches	-0.563	0.5737
Losses	5.824	1.05e-08
Shutouts	9.252	2e-16

8 Model Reduction

I ran multiple f tests and was able to create a graph that shows each combination of regressors and what their corresponding Adjusted R squared values are.



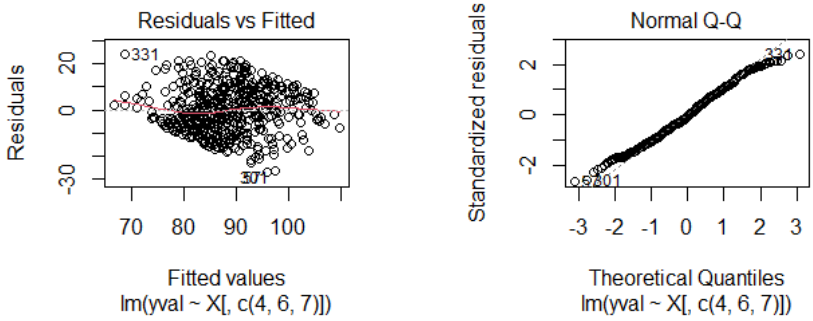
Based on this graph, I decided to further explore the model with the highest Adjusted R squared value, and the model with the least regressors at approximately 0.35 adjusted R squared

New Models				
	F Value	**P Value**	**Adj R Sq**	**Total Vars**
Full Model	40.106	2.23e-16	0.3561	7
Removed ERA, WP	56.01	2.23-16	0.3572	5
Used SO, L, SHO	88.41	2.23-16	0.3463	3

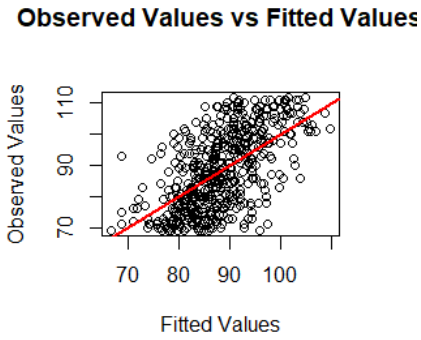
I ran some initial tests on my two new models and found that both have very high f values, very low p values, and negligible differences in adjusted r squared values. Because of this, I decided to choose the third model that removes ERA, home runs, wild pitches, and games played since it only uses three variables and its adjusted R squared value is only marginally smaller than the full model.

9 New Model

I wanted to check the assumptions for multiple linear regression once again before fully exploring my new model. I created a plot of the residuals vs the fitted values and found that it seems to be randomly distributed around zero, and I created another QQ plot to test the normality of my errors and found that there is only slight deviations at the tail ends of the plot. My new model passes the assumptions for multiple linear regression.



Finally, I wanted to visualize my new model and also test the significance of each of the regressors. I plotted the observed values vs the fitted values and it fits my data well. It looks similar to the plot of my full model, but it is using less than half of the number of regressors, which makes it far easier to implement in a real world scenario.



I then performed t tests on each of my regressors and found that they are all significant at the 0.01 level. I have also included the final equation for my reduced model.

$$y = 53.814 + 0.0146x_1 + 0.1760x_2 + 0.8149x_3$$

New Model: Individual T-tests		
Regressor	**T Value**	**P Value**
Strikeouts (x_1)	8.795	2.23e-16
Losses (x_2)	7.665	9.59e-14
Shutouts (x_3)	9.815	2.23e-16

10 Conclusions & Future Work

One interesting conclusion I found in my research was that ERA, Home runs, and Games played all ended up being removed from my final model, even though across the MLB, most managers would highlight these stats as important in predicting a pitcher's success. Additionally, some applicable future work includes possibly assessing a larger number of pitchers or considering more intricate statistics that my database did not cover.

11 Appendix

Link to Executable R Code in Github

12 References

I used the following websites for information regarding general MLB inquiries, as well as specific statistics about the salaries of MLB pitchers:

- Avg Pay for MLB Pitchers
- MLB salaries vs NFL, NHL, MLS salaries
- Postseason Bonus Money

I also used a pdf that details the content of Lahman's Database:

- Lahman's Baseball Database

13 Self-Reflection

When looking back at the process of creating my project, I can recognize a few instances where I ran into some avoidable issues. For example, when taking a sample of pitchers from my data set, I initially just extracted the first five hundred recorded pitchers and gathered their statistics. However, the data set that I was using had some pitchers repeated due to their coming out of retirement or, in some cases, switching teams. There were also some NA values scattered throughout my data due to a lack of information being recorded for pitchers that played in the MLB in the early 1900's. I tried to manually remove each NA value and skip repeated pitchers at first, but this led to a snowball effect of more issues with how my data was configured. If I was to do this project over again, I would spend more time cleaning my data at the very beginning of the process, so that I didn't have to waste time during the analysis process. Overall, I devoted more time to this project than any of my other previous projects, but I also feel like I gained a deeper understanding of the concept and I found myself very interested in the application of linear regression to a subject that captivates my interest in the real world.