

MLB Pitching Stats For Predicting Team Wins

Andrew Floyd, Charlie Duvall, David Gillcrest, Raylynn Swift

STAT 5346/CS 5204

MLB Pitching Stats For Predicting Team Wins

Background

As data becomes a bigger part of the world, it naturally has expanded to almost every industry. One of the more popular and profitable of these industries is sports, in which Major League Baseball is at the forefront. Statistical analysis and the use of data to improve teams' performances is the new status quo in baseball as teams have gotten farther away from the traditional scouting techniques. In the past, in-person scouting and using the eye test was the way individual players and teams were evaluated. But this has come to change as all professional organizations now use data analytics and analysis (often called "Sabermetrics" when referring to baseball), with less qualitative scouting. Statistics are used especially at the MLB level as the large sample size of a full 162 game season provides for a rich arsenal of numbers to be analyzed. There are thousands of statistics out there that are now be used to better teams' performance. This project's goal was to look at teams total pitching stats and see whether they impact a team's overall win record for a season. While pitching is only half of the game, we hoped to find a high correlation between good pitching and winning.

Data

The data, retrieved from baseball-reference.com, included both the 2016 and 2017 MLB seasons and thirty team, leads to a total of $n=60$ observations. There was some initial concern that the data from the two seasons might be correlated, but upon quick observation this fear was quelled as the data points appeared to be independent enough for our purposes. Also, our knowledge of baseball leads us to know that teams change enough for the two years to be combined. The response variable was the number of team wins per season (W), which ranged from 59 to 104 in the 2016/2017 seasons. Five pitching stats were investigated as possible quantitative predictor variables: earned run average (ERA), strikeouts (SO), saves (SV), home runs (HR), and WHIP (Walks and Hits per Innings Pitched). The data used in this analysis is observational in nature.

Statistical Methods

Analysis first started with a standardized MLR using all five predictor variables. Because the variables vary across different ranges (for example, ERA and WHIP are closer to 1-4 while HRs are around 200), we decided from the start to standardize the model to make interpretation of the regression coefficients more meaningful. This full regression model had a good R^2 (0.7102) and adjusted- R^2 (0.6833) values, but only SV was a marginally significant predictor of team wins. Also, the VIF test indicated that ERA (VIF=12.734) and WHIP (VIF=9.987) had issues with multicollinearity. Collinearity diagnostics indicated WHIP, SO, and HR had also had issues with multicollinearity. Multicollinearity was an expected problem when we began, due to the nature of ERA and WHIP. These two variables tend to be closely related as the more hits and walks a teams pitching staff gives up, more runs are also given up. Despite this fact, we decided to use both in our initial MLR as they are two of the most important and recognized pitching statistics in modern baseball.


The “best” models selection method and stepwise selection method were used to pick an appropriate model. We expected the final model to contain either ERA or WHIP, but probably not both due to problems with multicollinearity. The “best” models selection method found the top three models of each size. We picked the models base first on the highest R^2 within their size group, and then looked to the models with the highest adjusted- R^2 . We then eliminated models that had high C_p values while looking for low SBC and AIC values as well. We also noticed that models that contained both ERA and WHIP tended to be slightly worse than those that only had one of the two. With these criteria, we eliminated all models, except a three-variable model containing SV, SO, and WHIP. This model had the lowest AIC and SBC values, the best C_p value and was very close to having the highest adjusted- R^2 . The stepwise model selection method also chose this same model, with WHIP being first followed by SV and SO. Multicollinearity and outlier checks were then performed on this model.

Results

After switching to the three-variable model, we ran the standardized MLR again. We found that while the R^2 (0.7041) and adjusted R^2 (0.6883) were very similar to those of the

original model, there was a drastic decrease in p-values for each variable individually. As seen below, WHIP has become very significant along with SV. SO lowered as well, but not quite to the level of significance. VIF values also decreased, presumably due to the elimination of ERA in this model, which was highly correlated with WHIP. This decrease is shown in Figure 1.

Variable	Pr > t	Standardized Estimate	Variance Inflation
Intercept	0.0001	0	0
ERA	0.3570	-0.24288	12.73380
SV	0.0526	0.17413	1.43799
HR	0.9246	0.01046	2.25421
SO	0.0953	0.17904	2.07239
WHIP	0.1028	-0.38421	9.98718



Variable	Pr > t	Standardized Estimate	Variance Inflation
Intercept	<.0001	0	0
SV	0.0138	0.20547	1.23743
SO	0.0854	0.18295	2.06651
WHIP	<.0001	-0.58866	2.34050

Figure 1: Selected values from both the full MLR ANOVA table and the three-variable model (with outliers)

When we ran multicollinearity tests, we saw that collinearity also decreased drastically with this model. WHIP had a condition index of 155.42 in the full model and reduced to a condition index of 91.73 in our selected model. This is still borderline severe, but it is a significant improvement on the full MLR model and acceptable given the observational nature of the data.

As can be seen in Figure 2, the residuals vs predictor variable graphs show that assumptions of constant variance and linearity are upheld.

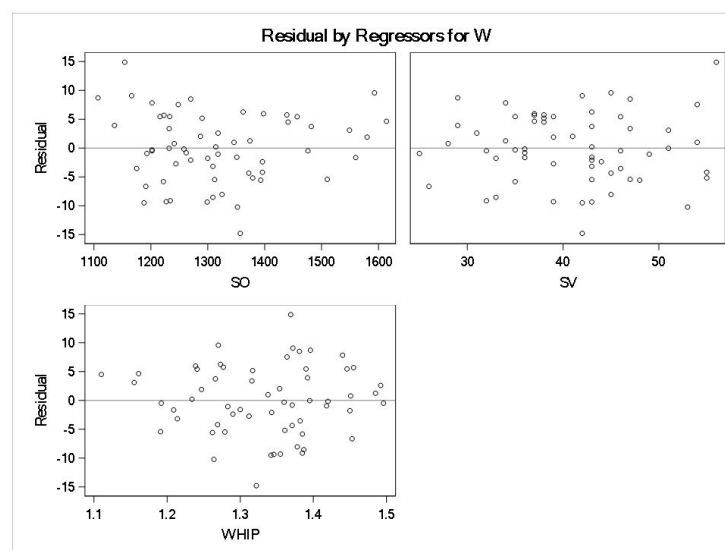


Figure 2: Residual by regressors for Wins in the three-variable model (with outliers)

Figure 3, below, shows the fit diagnostics for W in the three-variable model. The residual vs quantile graph shows the assumption of linearity is mostly upheld. The normal probability plot of residuals seems to uphold the assumption of normality.

It is clear that there may be some outliers by looking at the leverage vs residuals graph and particularly the Cook's D vs observation graph, so we decided to do some tests for outliers. The first test was ran in terms of outlier analysis was looking at the output statistics for all data points in our model. This included the Cook's D, RStudent, DFFITS and DFBETAS which helped us to visualize which values had the very high Cook's D that we saw in the graph. Three values were identified as outliers using the 50th percentile of Cook's D method, but only one of these variables had a DFFITS value over 1. A DFFITS value cutoff of 1 was used because the dataset we used is small. We then decided to create a new data set without the three values and tested if their removal influenced the model. All in all, however, this 3-variable model looked to have a very strong linear correlation based on the fit diagnostics.

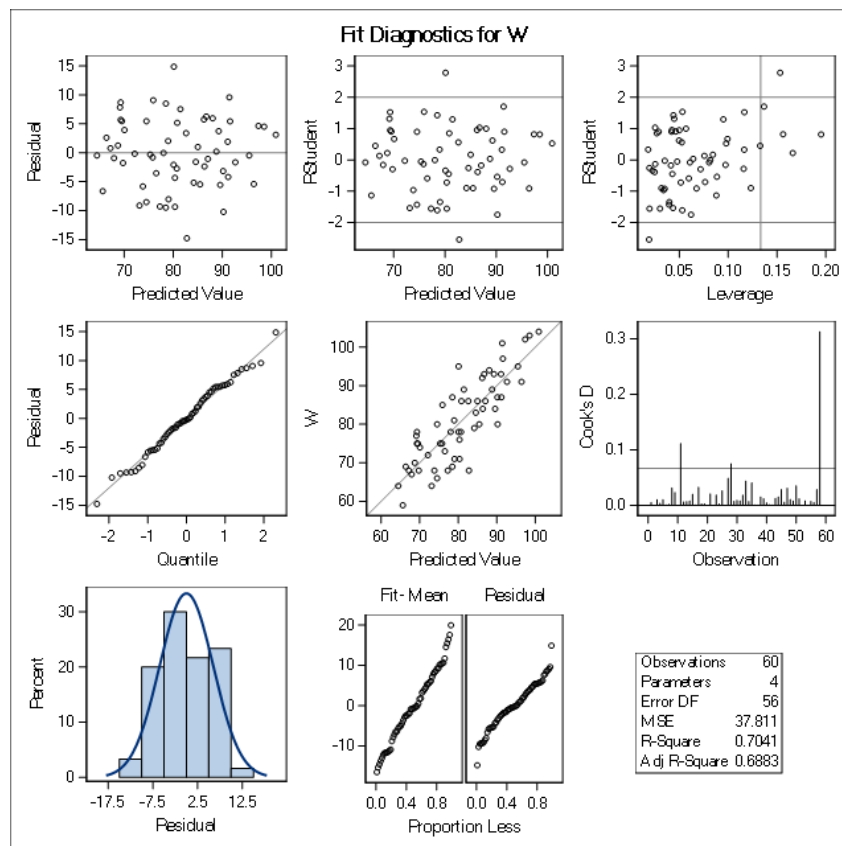


Figure 3: Fit diagnostics for Wins in the three-variable model (with outliers)

After removing the three outliers, we re-ran the MLR test and saw an improvement in our three-variable model's R^2 (0.7412) and adjusted R^2 (0.7265) values. The individual p-value for SO became significant ($p=0.0440$), but the p-value for SV increased to borderline significance ($p=0.0590$). The parameter estimates changed slightly, but these slight changes were determined to be significant given their relative size to the parameters themselves. Based on these results, we believe that each of the three data points removed were anomalies.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	Intercept	1	144.21805	30.11209	4.79	<.0001	0
SO	SO	1	0.01673	0.00956	1.75	0.0854	0.18295
SV	SV	1	0.29870	0.11754	2.54	0.0138	0.20547
WHIP	WHIP	1	-73.14444	13.81702	-5.29	<.0001	-0.58866

Figure 4: Parameter Estimates Table for the three-variable model (with outliers)

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	Intercept	1	142.60876	29.67941	4.80	<.0001	0
SO	SO	1	0.02036	0.00986	2.06	0.0440	0.21198
SV	SV	1	0.22370	0.11594	1.93	0.0590	0.15159
WHIP	WHIP	1	-73.66673	13.14755	-5.60	<.0001	-0.61439

Figure 5: Parameter Estimates Table for the three-variable model (outliers removed)

After we found our best model by removing the three outliers, we decided to test its predictive abilities. We randomly selected two teams' data from two different years, and the results of our model were very close to that of the actual on-field performances. The first prediction ran was for the St. Louis Cardinals in 2014. Their actual number of wins was 90, and our model predicted 88.92 ± 12.9094 using a 95% confidence interval for predictions. The second prediction was for the Kansas City Royals in 2012, with their actual number of wins being 72. Our model predicted the number of wins was 72.58 ± 12.3880 for a 95% confidence interval. While the range of the predictions was quite high, the predicted values were extremely close to the actual values, leading us to believe that this is a reasonable model.

Conclusions

We found that some pitching statistics are good predictors of a MLB team's number of wins for a season. WHIP and the number of saves and strikeouts over the season are good predictors for the number of team wins that season. We found that WHIP and ERA (earned run average) were highly related to each other, which we expected. Surprisingly, the number of homeruns in the season was not a significant predictor of team wins. Overall, this project showed that pitching statistics, while only representing half of the game, can be used to predict the number of games a MLB team wins in a season.

References

- Sports Reference LLC. *2016 MLB Team Statistics*. Retrieved from
<https://www.baseball-reference.com/leagues/MLB/2016.shtml>
- Sports Reference LLC. *2017 MLB Team Statistics*. Retrieved from
<https://www.baseball-reference.com/leagues/MLB/2017.shtml>