MLB Team Wins Analysis

Tom Seifert

2024-05-09

Introduction

In Major League Baseball, a wide array of traditional and advanced statistics can be used to evaluate players, teams, or the league as a whole. With such a diverse set of metrics to evaluate performance, there is no consensus among players, executives, or fans about which of these metrics measure performance most successfully. While advanced statistics like wRC+ (weighted runs created plus) attempt to quantify a player or team's performance with a single value, the foundations for these metrics always begin with traditional statistics such as hits, strikeouts, and walks. Similarly, this article will consider purely traditional team metrics while applying statistical regression methods to identify what metrics correlate to wins and to what degree these correlative metrics affect a team's win count.

The data set I will analyze contains season data for all MLB teams for 2019, 2021, and 2022 (2020 is omitted because of the shortened season due to COVID-19), a total 90 observations. My outcome variable will be the number of games won by a team in a particular season, and I will attempt to create a multiple linear regression model to assess the significance of the linear relationship between a team's number of wins and several of a team's offensive, defensive, and miscellaneous team statistics from a given season.

The offensive predictors include a team's number of hits, number of home runs, number of walks, and number of strikeouts by batters in a season. The defensive predictors are number of hits allowed, number of home runs allowed, number of walks allowed, number of strikeouts by pitchers, and fielding percentage (% of plays where an error is not made) in a season. The last predictor will be the number of fans who attended a team's games in the season. I will utilize statistical testing, assumption verification, and model selection to identify which combination of these ten predictors most effectively predicts an MLB team's number of wins in a season.

Table 1 shows a description of each variable used in the regression analysis:

Table 1: Variable Descriptions

W	Number of wins
H	Number of hits
HR	Number of home runs
BB	Number of walks
SO	Number of strikeouts (by batters)
HA	Number of hits allowed
HRA	Number of home runs allowed
BBA	Number of walks allowed
SOA	Number of strikeouts (by pitchers)
FP	Fielding percentage
attendance	Cumulative attendance for all home games)

Table 2: Summary Statistics for All Variables

	Mean	**SD**	**Variance**	**Min**	**Max**
W	81	15	221	47	111
H	1347	83	6941	1147	1554
HR	199	42	1722	110	307
BB	517	65	4180	378	645
SO	1398	104	10775	1122	1596
HA	1347	103	10624	1107	1576
HRA	199	34	1138	132	305
BBA	517	58	3333	384	617
SOA	1398	120	14335	1177	1671
FP	1	0	0	1	1
attendance	1981736	773149	597759103441	642617	3974309

Descriptive Analytics

Summary Statistics

Looking at the summary statistics for all of the variables in table 2, it can be seen that there is a variety of distributions among the variables, all of which are numerical. Numbers like attendance can go as high as almost 4 million with a twelve digit variance, while fielding percentage can only reach 1 at maximum and has a near 0 variance.

I will now construct a multiple linear regression model and perform a t-test for slopes to test whether each individual predictor is significant in predicting a team's number of wins in a season. The results of the t-tests can be seen in table 3.

Model Summary and T-tests

Using a significance level of 0.05, the t-test results show that six of our predictors have p-values below 0.05 and are thus significant, while attendance, FP, SO, and SOA have p-values above 0.05, rendering them insignificant. This indicates that the estimated slopes of of the significant variables (H, HR, BB, HA, HRA, and BBA) differ significantly from 0, and they thus have a significant effect on wins, the outcome variable in our model. Conversely, FP, SO, and SOA do not have a significant effect on the outcome variable because their estimated slopes do not significantly differ from 0.

Table 3: Model Summary

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-423.761	261.341	-1.621	0.10890
H	0.056	0.011	4.877	0.00001
HR	0.096	0.022	4.449	0.00003
BB	0.044	0.013	3.481	0.00082
SO	0.004	0.007	0.481	0.63209
HA	-0.038	0.010	-3.651	0.00047
HRA	-0.134	0.023	-5.819	0.00000
BBA	-0.047	0.012	-4.042	0.00012
SOA	0.004	0.008	0.545	0.58758
FP	487.148	258.789	1.882	0.06346
attendance	0.000	0.000	0.139	0.89017

Model Equation:

Wins = -423.761 + (0.056 x H) + (0.096 x HR) + (0.044 x BB) + (0.004 x SO) + (-0.038 x HA)
+ (-0.134 x HRA) + (-0.047 x BBA) + (0.004 x SOA) + (487.148 x FP) + (0 x attendance)

The model resulted in an R-squared value of 0.897 and a 68.8 F-statistic with a p-value of 2.2x10^{-16} for the F-test. These results communicate that the model is highly likely to have at least one significant predictor, and the predictor variables explain 89.7% of the variation in wins.

Verifying Assumptions

In order to further assess the model's goodness of fit, the assumptions for multiple linear regression must be verified. The assumptions that must be verified include linearity between the response variable and the predictors, normality of the error terms, constant variance of the error terms, and no multicollinearity between predictors. It is also worthwhile to check for any outliers or influential points using standardized residuals, leverage, and Cook's Distance.

Figure 1 shows four diagnostic plots that will help check the model assumptions:

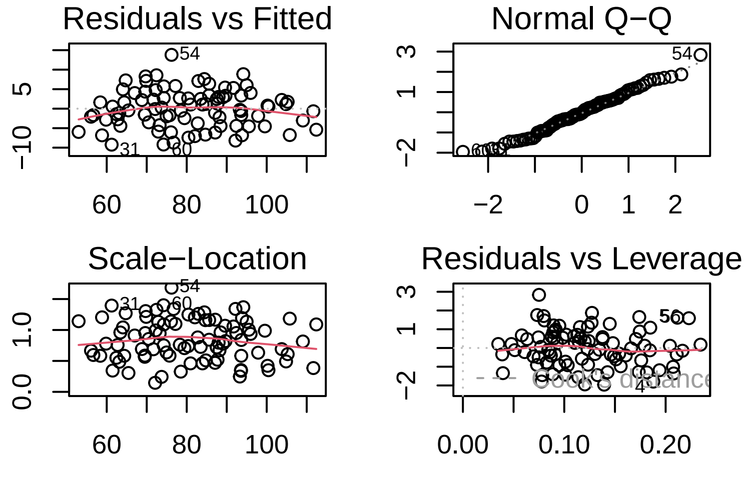

Figure 1: Diagnostic plots for final model

Table 4: Model VIF Values

	VIF
H	3.134918
HR	2.785325
BB	2.319427
SO	1.986496
HA	4.042930
HRA	2.104482
BBA	1.563744
SOA	2.826971
FP	1.365195
attendance	1.583266

Linearity

To satisfy the assumption of linearity, the residuals must be randomly scattered about the line y = 0 in the residuals vs. fitted values plot. This holds true, so the linearity assumption is met by the model.

Normality

In order to verify the assumption of normality of the error terms, the normal Q-Q plot must form an approximately straight line. This holds true for the normal Q-Q plot shown, verifying the normality of the error terms.

Constant Variance

The assumption of constant variance of the error terms can be checked in the residuals vs. fitted plot. In this plot, the spread of the residuals around the line y = 0 is fairly consistent and does not significantly change with the fitted values, so the constant variance assumption is met.

Non-Multicollinearity

To check for multicollinearity, the Variance Inflation Factor (VIF) for each variable in the model can be calculated. If none of the VIF values are greater than 5, then the correlation between predictors does not disrupt the coefficient estimates. The VIF values for each predictor are shown in table 4.

```
##         4         5         9        10        11        13        14        17
## 0.1878275 0.2116056 0.1706957 0.1814346 0.1940016 0.1658705 0.1759051 0.2344713
##        22        24        27        41        45        56        57        59
## 0.1848640 0.2109423 0.1606416 0.1556790 0.1734510 0.2229680 0.1744550 0.2073362
##        62        68        74        82        85        86
## 0.1792617 0.1739832 0.2166127 0.2042697 0.2079407 0.1845100
```

```
##          4          5         18         31         54         56         60
## 0.06783290 0.06358441 0.04645991 0.04700313 0.05929910 0.06589670 0.05646367
##         68
## 0.05221901
```

None of these values exceed 5, so the assumption of no multicollinearity between predictors is satisfied.

Outliers and Influential Points

Using standardized residuals, leverage, and Cook's Distance, outliers and influential points can be identified. By looking at the standardized residuals vs. leverage plot, there is only one with standardized residuals above

2 or below -2, making it an outlier. There are twenty-two high leverage points, which are categorized by having leverage values higher than 2(p+1)/n (p = number of predictors, n = number of observations), or higher than two times the average leverage value in the model, (p+1)/n. These leverage points have a significant effect on the model, but are not necessarily outliers, as a good leverage point will have high influence on the model while still following the regression pattern. There are eight points with Cook's Distances greater than 4/(n - 2), making them outliers.

Despite the outliers and influential points present in the model, none of the influential points are outliers, so they do not disrupt model estimates very much. In addition, these points do not cause any of the model assumptions to be violated. Therefore, it will be not necessary to transform this model to account for outliers.

After verifying the assumptions, the model can be improved using model selection to ensure the best combination of predictor variables are included in the model. This method may also help reduce the number of outliers and influential points in the model.

Variable Selection

The model is shown to meet the assumptions of multiple linear regression. Now, model selection can be used to find the best combination of predictors in predicting the outcome variable, wins. The forward regression model selection method searches for the best fitting model by starting with no predictors and continuously adding the most significant predictor not yet in the model, eventually resulting in the original full model. The best of the ten models can then be identified by looking for the highest adjusted R-squared, lowest CP, and lowest BIC between the models. The best fitting models will have the best predicting power without introducing too many variables, as this complexity could cause over fitting, worsening the model's ability to generalize the model to new observations accurately.

Figure 2 shows graphical comparisons between the ten models based on their adjusted R-squared, CP, and BIC value.

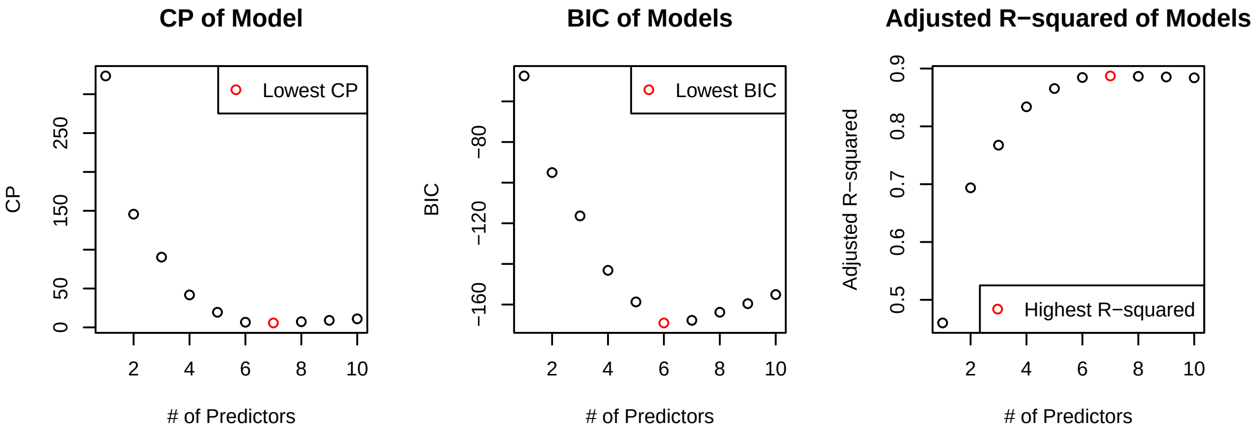

Figure 2: Comparison of R-squared, CP, and BIC across stepwise selected models

Using forward selection, the BIC and CP of the models stop significantly decreasing after the model with six variables is considered. R-squared stops significantly increasing after the model with six variables. The model with seven variables has a slightly lower value of CP and a slightly higher value of R-squared in comparison to the six variable model.

In order to choose between the six and seven variable models, I will conduct a partial F-test to analyze the significance of the predictor that is present the seven variable model and absent in the six variable model. The null and alternative hypotheses for the test are as follows:

H_0: The reduced model is more significant in predicting wins than the full model

H_1: The full model is a more significant predictor of wins than the reduced model

Table 5: Partial F-test Results

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
83	2117.220	NA	NA	NA	NA
82	2042.827	1	74.394	2.986	0.0877412

```
## Analysis of Variance Table
##
## Response: W
##           Df Sum Sq Mean Sq F value                Pr(>F)
## H          1 3171.9  3171.9 124.345 < 0.00000000000000022 ***
## HR         1 4381.9  4381.9 171.779 < 0.00000000000000022 ***
## BB         1 2555.8  2555.8 100.192 0.0000000000000006296 ***
## HA         1 5647.9  5647.9 221.411 < 0.00000000000000022 ***
## HRA        1 1399.9  1399.9  54.880 0.0000000000968616951 ***
## BBA        1  379.4   379.4  14.874             0.0002261 ***
## Residuals 83 2117.2    25.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = W ~ H + HR + BB + HA + HRA + BBA, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6613  -3.2454  -0.3722   3.3394  14.6192
##
## Coefficients:
##              Estimate Std. Error t value        Pr(>|t|)
## (Intercept) 68.546014  14.113082   4.857 0.0000055426540 ***
## H            0.056413   0.007565   7.457 0.0000000000776 ***
## HR           0.096566   0.019317   4.999 0.0000031594901 ***
## BB           0.049088   0.011687   4.200 0.0000666767172 ***
## HA          -0.042993   0.007765  -5.537 0.0000003526799 ***
## HRA         -0.138908   0.022313  -6.225 0.0000000187684 ***
## BBA         -0.043680   0.011326  -3.857        0.000226 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.051 on 83 degrees of freedom
## Multiple R-squared:  0.8923, Adjusted R-squared:  0.8845
## F-statistic: 114.6 on 6 and 83 DF,  p-value: < 0.00000000000000022
```

As shown in table 5, the p-value for this test is greater than the significance level of 0.05, so we fail to reject the null hypothesis that the reduced model outperforms the model using FP as a predictor, meaning that FP it is an insignificant predictor in our model and the reduced six variable model will be chosen as the final model. It is also worth noting that the six variable model has only two high leverage points, a large decrease from the twenty-two high leverage points present in the original model. The final model summary is shown in table 6.

Model Equation:

$$\text{W} = 68.546 + (0.056 \text{ x H}) + (0.096 \text{ x HR}) + (0.049 \text{ x BB}) + (-0.043) \text{ x HA} + (-0.139 \text{ x HRA}) +$$

Table 6: Reduced Model Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	68.546	14.113	4.857	0.00001
H	0.056	0.008	7.457	0.00000
HR	0.097	0.019	4.999	0.00000
BB	0.049	0.012	4.200	0.00007
HA	-0.043	0.008	-5.537	0.00000
HRA	-0.139	0.022	-6.225	0.00000
BBA	-0.044	0.011	-3.857	0.00023

(-0.044 x BBA)

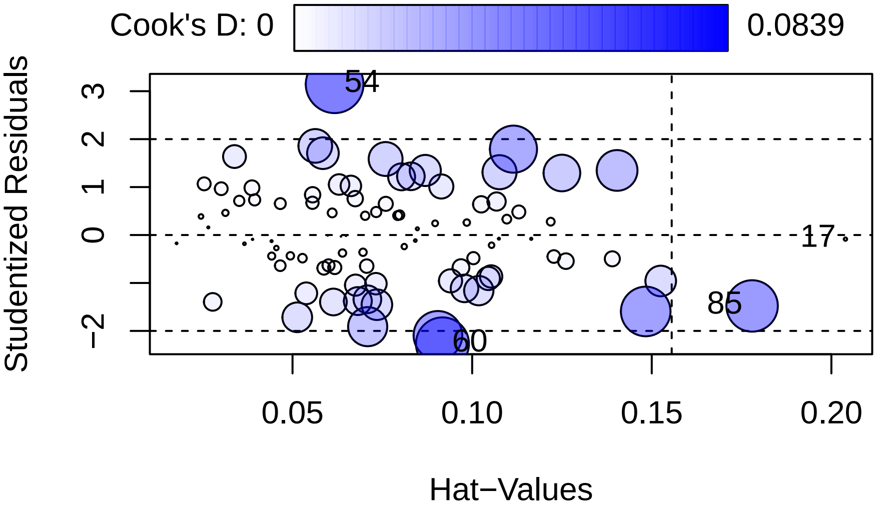

Figure 3: Cook's Distance indicated on graph of standard residuals vs. leverage

Conclusion:

After investigating the relationship between the amount of games an MLB team wins in a season and ten predictor variables with a goal of creating a multiple linear regression model to fit the relationship, six predictors were chosen for the final wins model: **hits**, **home runs**, **walks**, **hits allowed**, **home runs allowed**, and **walks allowed**.

These results suggest that the number of hits, home runs, and walks are important metrics of baseball that teams should try to maximize on offense and limit for opposing teams on defense in order to increase win potential.

However, the model does not fully explain the variation in team wins; only 89.23%. The outliers and influential points in the data set could be accounted for more effectively by utilizing the weighted least squares method, which could improve the model's goodness of fit. If given more time, further investigation of other predictor variables and more observations of data could yield an improved model.

Further application of the model could involve introducing new data to test the model on in order to evaluate how well it predicts a team's number of wins. Assuming that the model predicts team wins fairly accurately, further investigation of how to maximize and minimize metrics positively and negatively correlated with wins respectively could help MLB teams train and prepare their players and rosters to give them the best chance to succeed.