MLB Win Percentage Predictor for 2013 Season

Jesse Hendren-Mills

2022-07-10

Introduction

In 2013, MLB baseball saw the Boston Red Sox win their third World Series in 10 seasons over the St. Louis Cardinals with David Ortiz taking home World Series MVP honors. No one was really shocked by these two teams making it to the World Series. They were head and shoulders better than the majority of their respective leagues and truly dominated their way to the World Series. These two teams we great on the field and but had a seemingly significant difference in funding, Boston at 150+ Million and St.Louis at 92 Million. Additionally, they were both considered to be two of the better teams when it came to pitching and hitting. With future Hall of Famers such as Manny Rameriez, David Ortiz, and Pedro Martinez for Boston, to name a few. However, St. Louis had a few names of their own to include Albert Pujlos, Jim Edmonds, and Mark Mulder. But we are not here to discuss who made it to the World Series, but to discover what 'ingredients' it takes to "March into October".

The premise of this analysis is to understand if we can predict the number of expected wins a team recorded based on a teams runs scored, runs allowed, payroll, weighted on-base-average (wOBA), and fielding independent pitching (FIP). In theory, payroll enables teams to purchase the best players and the best players should be able to score the most runs while giving up the fewest. Those expected win totals will give us an understanding of the teams that should have made the playoffs in 2013. This would be based on the number of runs they scored versus the number of runs given up. Additionally, we look to observe if advanced metrics such as wOBA and FIP can explain any of the variation we observe with a team's win total in 2013.

Note: For everyone's knowledge, in 2013, 10 teams made it to the Major League Postseason. The Boston Red Sox, Oakland Atheletics, Detroit Tigers, Cleveland Indian, and Tampa Bay Rays represented the American League. In the National League the St. Louis Cardinals, Atlanta Braves, Los Angeles Dodgers, Pittsburgh Pirates, and Cincinnati Reds made the playoffs. In Major League Baseball, three teams in each league gain a playoff spot by winning their division, the other two spots are filled with the next highest records of a given team. So, if a team had 82 wins and led their division they would be in the playoffs over a team that may have won 89 games and finished 4th in another division.

Data Preparation

Reading & Loading Packages

- rmarkdown
- ggplot2
- dplyr
- Lahman
- formattable
- knitr
- car
- MVN
- broom
- lmtest
- stargazer

Data Table Variable Definitions

Batting

- playerID: A players unique I.D.
- yearID: Year of Record
- teamID: Team of Record
- AB: Plate appearances by a batter (minus walks or HBP)
- R: Total number of Runs recorded by a batter
- H: Total number of singles recorded by a batter
- X2B: Total number of doubles recorded by a batter
- X3B: Total number of triple recorded by a batter
- HR: Total number of Home Runs recorded by a batter
- BB: Total number of UNINTENTIONAL Walks recorded by a batter
- IBB: Total number of INTENTIONAL Walks recorded by a batter
- HBP: Total number of occurances a batter was HIT BY A PITCH
- SF: Total number of Sacrifice Flies by a batter
- wOBA: A batter's Weighted On-Base-Average
- TEAMwOBA: A given team's average Weighted On-Base-Average in 2013

Pitching

- playerID: A players unique I.D.
- yearID: Year of Record
- teamID: Team of Record
- HR: Home Runs allowed by a pitcher in a given year
- BB: Number of walks a pitcer gave up in a given year
- HBP: Number of hitters a pitcher hit with a pitch in a given year
- SO: Number of strikeouts a pitcher recorded in a given years
- IPouts: Number of outs a pithcer recorded in a given year
- FIP: Fielding Independent Pitching Metric
- TEAMFIP: A given team's average Fielding Independent Pitching in 2013

Salaries

- yearID: Year of Record
- teamID: Team of Record
- lgID: League Identifier
- playerID: A players unique I.D.
- salary: Salary compensation of an individual player on a roster
- Payroll: Total amount of money paid to field a team for a given season in Millions $

Teams

- yearID: Year of Record
- teamID: Team of Record
- lgID: League Identifier
- W: Number of Wins a team recorded in a season
- L: Number of Loses a team recorded in a season
- R: Number of Runs scored FOR a team recorded in a season
- RA: Number of Runs scored AGAINST a team in a season
- EWP: Expected Win Percentage
- EWins: Expected WINS (Forecast)
- AWP: Actual Win Percents

Creating Data Frames

Sean Lahman's Major League Baseball Statistic Descriptions

Lahman's "Salaries" data frame highlights and summarizes the payroll of each team in 2013 along with the individual amount players made in 2013. For this analysis, we were not interested in individual pay, only the sum of what each team spent in 2013. Therefore, we will remove players names, IDs, and individual salaries from the table.

Lahman's "Teams" data frame highlights and summaries each team's wins, losses, runs scored, runs allowed, and playoff win statistics. These columns will be used to help us predict future win totals and see if our model is statistically significant and accurate at predicting playoff teams.

Lahman's "Batting" data frame highlights and summarizes the the team batting statistics of each team in 2013. For this analysis, we were not interested in individual pitching performance, only the sum of how each team performed in 2013.

Lahman's "Pitching" data frame highlights and summarizes the the team pitching statistics of each team in 2013. For this analysis, we were not interested in individual pitching performance, only the sum of how each team performed in 2013.

Summary Statisitics

teamID	lgID	Payroll	W	L	R	RA	TEAMwOBA	TEAMFIP	EWP	EWins	AWP	WinPctDiff
ARI	NL	$90	81	81	685	695	37.26%	4.155110	49.42%	80	50.00%	0.58%
ATL	NL	$88	96	66	688	548	37.85%	3.619982	59.00%	96	59.26%	0.26%
BAL	AL	$84	85	77	745	709	38.28%	4.305974	51.98%	84	52.47%	0.49%
BOS	AL	$152	97	65	853	656	42.00%	4.047866	60.35%	98	59.88%	-0.48%
CHA	AL	$120	63	99	598	723	35.47%	3.987054	42.47%	69	38.89%	-3.58%
CHN	NL	$101	66	96	602	689	37.94%	4.640854	44.62%	72	40.74%	-3.88%
CIN	NL	$106	90	72	698	589	36.66%	3.760333	56.75%	92	55.56%	-1.19%
CLE	AL	$76	92	70	745	662	37.02%	4.423223	54.71%	89	56.79%	2.08%
COL	NL	$74	74	88	706	760	39.43%	3.985013	47.06%	76	45.68%	-1.38%
DET	AL	$146	93	69	796	624	39.43%	3.329263	59.62%	97	57.41%	-2.21%
HOU	AL	$18	51	111	610	848	35.50%	4.893713	37.12%	60	31.48%	-5.64%
KCA	AL	$80	86	76	648	601	33.53%	3.828672	53.01%	86	53.09%	0.08%
LAA	AL	$124	78	84	733	737	38.70%	4.070884	49.78%	81	48.15%	-1.63%
LAN	NL	$223	92	70	649	582	37.87%	3.795892	54.35%	88	56.79%	2.44%
MIA	NL	$34	62	100	513	646	32.33%	3.662555	40.88%	66	38.27%	-2.61%
MIL	NL	$77	74	88	640	687	37.57%	4.381636	47.17%	76	45.68%	-1.49%
MIN	AL	$75	66	96	614	788	35.46%	4.206027	40.15%	65	40.74%	0.59%
NYA	AL	$232	85	77	650	671	35.04%	3.959896	48.73%	79	52.47%	3.74%
NYN	NL	$49	74	88	619	684	35.07%	3.586913	46.01%	75	45.68%	-0.34%
OAK	AL	$60	96	66	767	625	38.54%	3.964019	58.12%	94	59.26%	1.14%
PHI	NL	$170	73	89	610	749	35.71%	4.286271	41.86%	68	45.06%	3.20%
PIT	NL	$77	94	68	634	577	38.25%	3.492685	53.76%	87	58.02%	4.26%
SDN	NL	$66	76	86	618	700	35.61%	4.251065	45.03%	73	46.91%	1.88%
SEA	AL	$74	71	91	624	754	37.34%	4.067213	42.49%	69	43.83%	1.34%
SFN	NL	$140	76	86	629	691	37.48%	3.742470	46.25%	75	46.91%	0.67%
SLN	NL	$92	97	65	783	596	38.36%	3.517016	60.75%	98	59.88%	-0.87%
TBA	AL	$53	92	71	700	646	36.67%	3.740951	53.21%	86	56.79%	3.58%
TEX	AL	$113	91	72	730	636	37.89%	3.812511	55.49%	90	56.17%	0.68%
TOR	AL	$126	74	88	712	756	37.07%	3.920583	47.60%	77	45.68%	-1.92%
WAS	NL	$114	86	76	656	626	35.62%	3.754192	51.87%	84	53.09%	1.21%

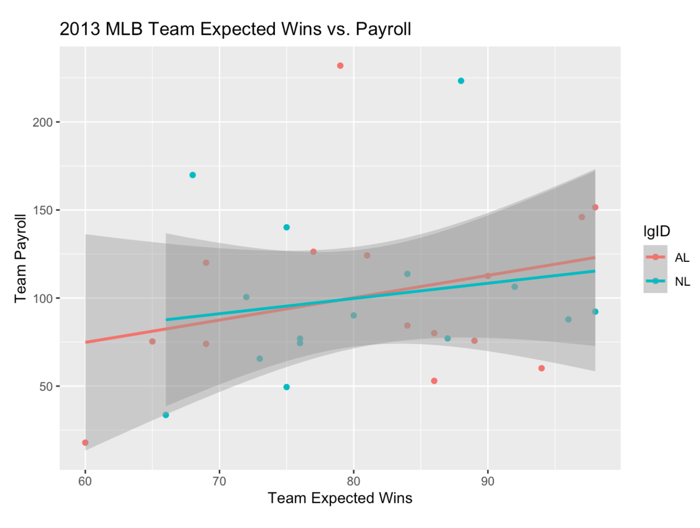
Testing Assumptions

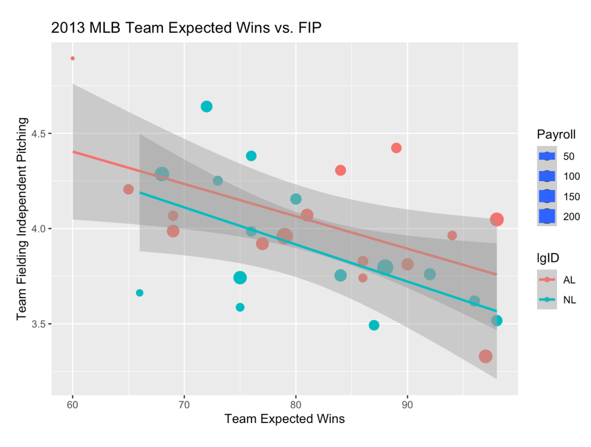
```
Original Proposed Model
Expected Wins ~ Team WOBA + Team FIP + Payroll + error
```

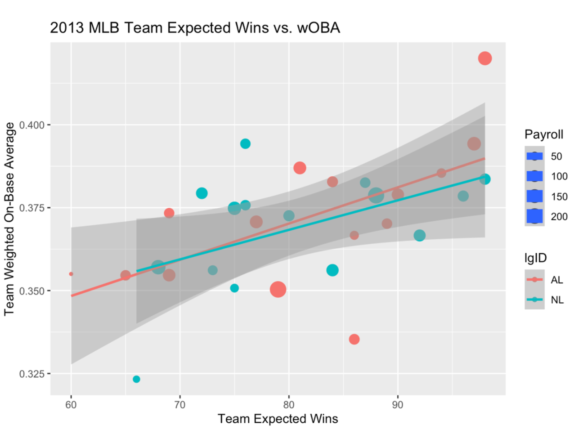
- Assumption 1: There is a linear relationship between the predictors (x) and the outcome (y)
- Assumption 2: Residual Errors are independent from each other and predictors (x)
- Assumption 3: Predictors (x) are independent and observed with negligible error
- Assumption 4: Residual Errors have constant variance
- Assumption 5: The observations are independent

Assumption #1: Liner Relationship

Now that we have gathered all of our data and made the proper modifications, lets dive into some analysis and confirm the five assumptions of multivariate regression. First, we will start with confirming that there exists a linear relationship between each of the independent variables and the dependent variable. We can confirm this by plotting each of three scenarios. From the graphs below, it appears that there is a some what linear relationship between each of the exploratory variables and expected wins. We can move onto the next assumption.







Assumption #2: No Multicollinearity

Note that their is a linear relationship with the exploratory variables and the dependent variable, expected wins, we can test if there is multicollinearity present between the exploratory variables. To begin this, we want to review the summary results running our 'Wins' model listed above.

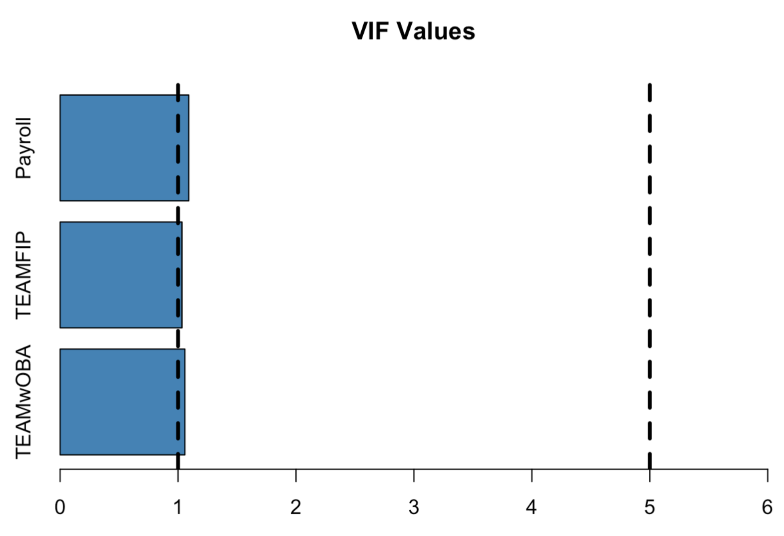
Regession Results

Independent Variables	Estimate	Std. Error	T-Value	P-Vale
(Intercept)	29.9094609	31.1478520	0.9602415	0.3457792
TEAMwOBA	302.2009623	73.0446942	4.1372062	0.0003266
TEAMFIP	-15.4419183	3.9202042	-3.9390597	0.0005479
Payroll	0.0052564	0.0289201	0.1817559	0.8571839

From these regression results we can observe a few things. First, we see that both wOBA and FIP are statistically significant at a 99% confidence interval with Payroll being statistical insignificance. Additionally, we can observe that the exploratory variables explain 58.23 percent of the variation in expected wins. However, we still need to test multicollinearity first! So lets do that and see if we can remove certain variables from our initial model.

For us to test for multicollinearity, we will us a variance inflation factor (VIF) in order to detect if multicollinearity exists. The value for VIF starts at 1 and has no upper limit. A general rule of thumb for interpreting VIFs is as follows:

- A value of 1 indicates there is no correlation between a given predictor variable and any other predictor variables in the model.
- A value between 1 and 5 indicates moderate correlation between a given predictor variable and other predictor variables in the model, but this is often not severe enough to require attention.
- A value greater than 5 indicates potentially severe correlation between a given predictor variable and other predictor variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.



After looking observing our graph, we can see that there is minimal amounts of multicollinearity present within the model. However, it is not severe enough for us to remove or change any of the variables present within the model. We will continue with the next assumption and test if residual values of the exploratory variables have a mean of zero.

Note: Multicollinearity is when exploratory variables (x) are independent and observed with negligible error. Meaning these variables have little to no relationship with each other, giving us the most accurate model and explanation of the dependent variable as possible.

Assumption #3: Independence

Next we want to test if our explanatory variables are independent. Meaning that the exploratory variables (x) are independent and observed with negligible error of each other. We can test this through a what is called a Durbin-Watson Test. A model with a test static output between 1.21 and 1.65 at a 95 percent confidence interval for a Durbin-Watson test is considered to be statistically significant. Meaning that the residuals of exploratory variables within the model are not correlated. A violation of this assumption would underestimate the standard errors within the model giving a false positive of statistical significance.

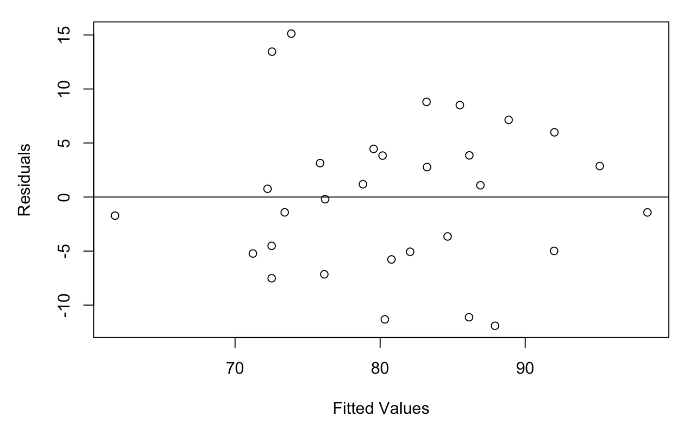
Durbin-Watson Residual Autocorrelation Test

statistic	p.value	autocorrelation	method	alternative
2.203622	0.554	-0.1076603	Durbin-Watson Test	two.sided

In this test we see that our model fails the Durbin-Watson Test and that there is some kind of serial correlation of the residuals within our explanatory variables. Meaning that there will be over estimated coefficients of our models. After doing a couple of variations of this model, there wasn't a single combination of these variables, other than by themselves, where the Durbin-Watson test passed. Meaning that these variables are indeed correlated to some extent. For exploratory reason, let's continue with our testing.

Assumption #4: Homoskedasticity

The next test we want to cover is testing if our residual errors are constant. This will mean that a model is considered to be homoskedastic. Homoskedasticity is considered to be important because the alternative will increase the coefficient estimates and may provide another false positive. For this test we will utilize a Breusch-Pagan test for our model that we already know to be faulty, but lets explore!



Breusch-Pagan Homoskedasticity Test

statistic	p.value	parameter	method
1.342987	0.7189516	3	studentized Breusch-Pagan test

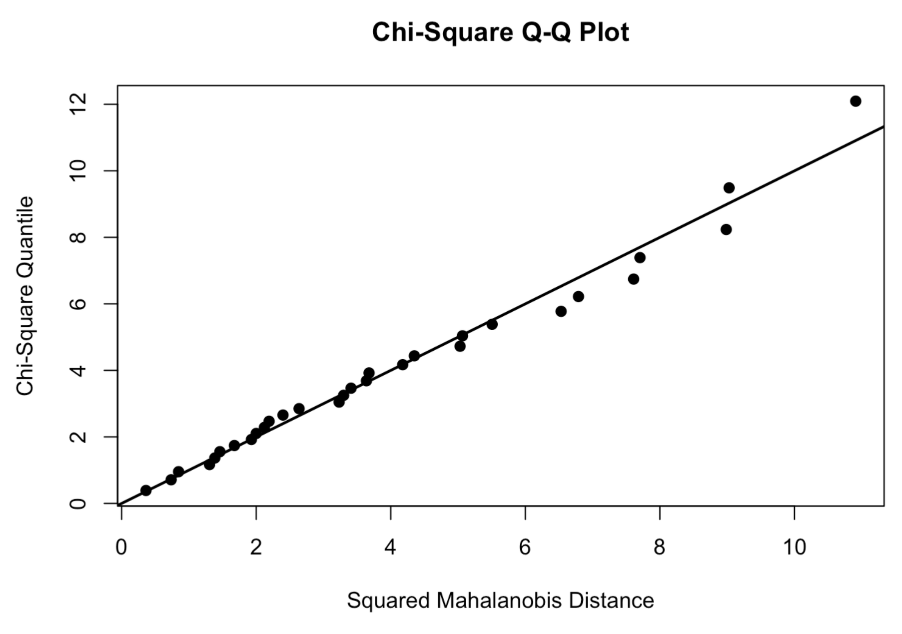
After examining the results, we see that our residuals are not considered to be constant and that heteroskedasticiy exists within the model. Which will cause our model to have overestimates and provide us with variables that appear to be statistically significant but are not in actuality.

Assumption #5: Multivariate Normality

The final assumption that needs to be fulfilled is to determine if each variable has multivariate normality. Which flows a Gaussian normal distribution that shows the major of a data occurs in the middle with two "tails" at the ends. We can achieve this by utilizing a Quantile-Quantile Plot along with an Anderson-Darling Test from the MVN package. This test tries to identify if each individual variable within the model is considered to have normal distribution and if a value is considered to be statistically significant, then they will be considered to not have normality and should be removed from the model.

Multivariate Normaility Testing

Test	Variable	Statistic	p value	Normality
Anderson-Darling	EWins	0.2515	0.7168	YES
Anderson-Darling	TEAMwOBA	0.4490	0.2592	YES
Anderson-Darling	TEAMFIP	0.2403	0.7548	YES
Anderson-Darling	Payroll	0.7694	0.0404	NO



As we can observe from the table and graph above, the majority of the variables are considered to have a normal distribution. Unlike a team's payroll, that appears not be a good variable to include for our model. For now, we will keep Payroll and see what other tests, if any it fails but will remove it from the final model we create. Now that we have finished all of our assumptions, let's wrap up our analysis and review how our model did.

Conclusion

We conclude that our variables of FIP, wOBA, and Payroll are not considered to be variables that can predict a team's expected wins total for a given season. Due to high levels of serial correlation between the explanatory variables and the heteroskedasticity of the residual errors within the model. For future exploration, a user may want to look into utilizing a time-series analysis or bring in more variables to explore the variation present within the model.

Reference

- Hayley Jang. (2019) Salary Distribution and Team Outcome: The Comparison of MLB and KBO. Journal of Global Sport Management 4:2, pages 149-163.
- Markdown Guide: `https://www.markdownguide.org/extended-syntax/#definition-lists`
- Robert Breunig, Bronwyn Garrett-Rumba, Mathieu Jardin, Yvon Rocaboy. (2014) Wage dispersion and team performance: a theoretical model and evidence from baseball. Applied Economics 46:3, pages 271-281.
- Sean Lahman Database: `https://www.seanlahman.com/baseball-archive/statistics/`
- Toward Data Science : `https://towardsdatascience.com/all-the-statistical-tests-you-must-do-for-a-good-linear-regression-6ec1ac15e5d4`
- Wooldridge, Jeffrey M. Introductory Econometrics: A Modern Approach. Australia: Cengage, 2020.