Baseball Statistic Analysis

Andrew Norman

2025-05-01

1. Motivation, load the packages and datasets, view the structures, and skim them.

The topic of interest is predictive baseball analysis relating to next years team win ratio with the teams end of year statistics from the previous season. I am interested in predictive baseball analysis because of the film *Moneyball* in which the General Manager and assistant GM try to use data analytics to remove bias from their player scouting and selection to get winning team statistics from overlooked players due to many reasons on a limited budget. I do not have the same scouting and player level statistics that scouts and managers do, but using team level batting statistics I would like to try to model team's win ratios based off of end of season team batting statistics and payroll from the previous year since if you cannot hit and drive in runs then you cannot win games.

2. Merge all the datasets into one and skim to check for missing values.

Table 1: Data summary

Name	mlb_data
Number of rows	775
Number of columns	36
Column type frequency:	
character	2
numeric	34
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
team	0	1.00	12	21	0	35	0
est_payroll	6	0.99	11	12	0	769	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
year	0	1	2012.00	7.22	2000.00	2006.00	2012.00	2018.00	2024.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
batters	0	1	48.08	6.17	34.00	44.00	47.00	52.00	70.00	
batter_age	0	1	28.63	1.29	25.40	27.80	28.50	29.40	33.50	
runs_per_game	0	1	4.56	0.50	3.13	4.22	4.54	4.88	6.04	
games	0	1	157.88	20.02	58.00	162.00	162.00	162.00	163.00	
plate_apps	0	1	6031.21	788.10	2011.00	6092.00	6179.00	6268.50	6537.00	
at_bats	0	1	5381.93	702.37	1752.00	5462.00	5520.00	5574.00	5770.00	
runs	0	1	719.13	120.52	219.00	673.00	730.00	786.00	978.00	
hits	0	1	1382.66	204.20	390.00	1349.00	1414.00	1476.00	1667.00	
doubles	0	1	275.90	46.56	73.00	263.00	280.00	300.00	376.00	
triples	0	1	27.68	9.72	3.00	21.00	28.00	34.00	61.00	
hr	0	1	171.82	40.59	51.00	148.00	171.00	199.00	307.00	
rbi	0	1	685.44	115.85	204.00	641.50	695.00	750.00	932.00	
sb	0	1	90.55	32.94	14.00	67.50	88.00	111.00	223.00	
cs	0	1	34.12	11.84	3.00	26.00	33.00	42.00	74.00	
bb	0	1	507.62	91.30	147.00	466.50	514.00	561.00	775.00	
so	0	1	1173.68	216.83	440.00	1048.50	1183.00	1327.50	1654.00	
ba	0	1	0.26	0.01	0.21	0.25	0.26	0.27	0.29	
obp	0	1	0.32	0.01	0.28	0.32	0.32	0.34	0.37	
slg	0	1	0.41	0.03	0.34	0.40	0.41	0.43	0.50	
ops	0	1	0.74	0.04	0.62	0.71	0.74	0.76	0.85	
ops_plus	0	1	97.30	8.53	73.00	91.00	97.00	102.00	126.00	
tb	0	1	2229.38	332.05	650.00	2155.00	2276.00	2393.00	2832.00	
gdp	0	1	119.43	22.63	28.00	111.00	122.00	133.00	170.00	
hbp	0	1	57.81	15.92	10.00	48.00	56.00	67.00	116.00	
sh	0	1	40.70	23.34	0.00	22.00	38.00	58.00	119.00	
sf	0	1	41.96	10.17	7.00	36.00	42.00	48.00	75.00	
ibb	0	1	34.11	16.45	1.00	22.00	34.00	44.00	153.00	
lob	0	1	1096.45	153.95	337.00	1080.00	1121.00	1162.50	1301.00	
world_series_win	0	1	0.03	0.18	0.00	0.00	0.00	0.00	1.00	
wins	0	1	78.93	15.40	19.00	71.00	81.00	90.00	116.00	
losses	0	1	78.93	15.38	17.00	71.00	80.00	89.00	121.00	
win_ratio	0	1	0.50	0.07	0.25	0.45	0.50	0.56	0.72	
inflation_multiplier	0	1	1.44	0.23	1.03	1.28	1.40	1.60	1.88	

3. Clean the data.

Table 4: Data summary

Name	mlb_data
Number of rows	775
Number of columns	37
Column type frequency:	
character	1
numeric	36
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
team	0	1	12	21	0	31	0

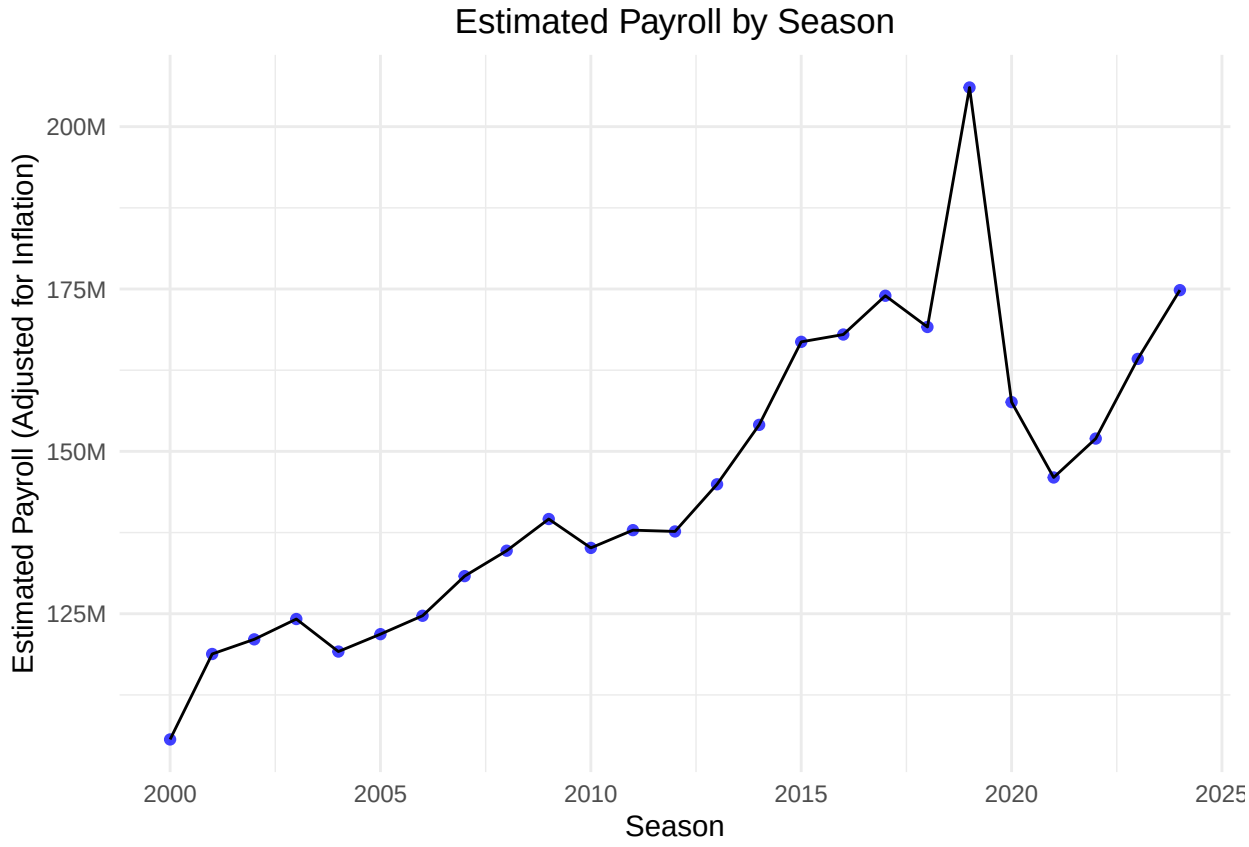
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
year	0	1	2012.00	7.22	2000.00	2006.00	2012.00	2018.00	2024.00	
batters	0	1	48.08	6.17	34.00	44.00	47.00	52.00	70.00	
batter_age	0	1	28.63	1.29	25.40	27.80	28.50	29.40	33.50	
runs_per_game	0	1	4.56	0.50	3.13	4.22	4.54	4.88	6.04	
games	0	1	157.88	20.02	58.00	162.00	162.00	162.00	163.00	
plate_apps	0	1	6031.21	788.10	2011.00	6092.00	6179.00	6268.50	6537.00	
at_bats	0	1	5381.93	702.37	1752.00	5462.00	5520.00	5574.00	5770.00	
runs	0	1	719.13	120.52	219.00	673.00	730.00	786.00	978.00	
hits	0	1	1382.66	204.20	390.00	1349.00	1414.00	1476.00	1667.00	
doubles	0	1	275.90	46.56	73.00	263.00	280.00	300.00	376.00	
triples	0	1	27.68	9.72	3.00	21.00	28.00	34.00	61.00	
hr	0	1	171.82	40.59	51.00	148.00	171.00	199.00	307.00	
rbi	0	1	685.44	115.85	204.00	641.50	695.00	750.00	932.00	
sb	0	1	90.55	32.94	14.00	67.50	88.00	111.00	223.00	
cs	0	1	34.12	11.84	3.00	26.00	33.00	42.00	74.00	
bb	0	1	507.62	91.30	147.00	466.50	514.00	561.00	775.00	
so	0	1	1173.68	216.83	440.00	1048.50	1183.00	1327.50	1654.00	
ba	0	1	0.26	0.01	0.21	0.25	0.26	0.27	0.29	
obp	0	1	0.32	0.01	0.28	0.32	0.32	0.34	0.37	
slg	0	1	0.41	0.03	0.34	0.40	0.41	0.43	0.50	
ops	0	1	0.74	0.04	0.62	0.71	0.74	0.76	0.85	
ops_plus	0	1	97.30	8.53	73.00	91.00	97.00	102.00	126.00	
tb	0	1	2229.38	332.05	650.00	2155.00	2276.00	2393.00	2832.00	
gdp	0	1	119.43	22.63	28.00	111.00	122.00	133.00	170.00	
hbp	0	1	57.81	15.92	10.00	48.00	56.00	67.00	116.00	
sh	0	1	40.70	23.34	0.00	22.00	38.00	58.00	119.00	
sf	0	1	41.96	10.17	7.00	36.00	42.00	48.00	75.00	
ibb	0	1	34.11	16.45	1.00	22.00	34.00	44.00	153.00	
lob	0	1	1096.45	153.95	337.00	1080.00	1121.00	1162.50	1301.00	
world_series_win	0	1	0.03	0.18	0.00	0.00	0.00	0.00	1.00	
wins	0	1	78.93	15.40	19.00	71.00	81.00	90.00	116.00	
losses	0	1	78.93	15.38	17.00	71.00	80.00	89.00	121.00	
win_ratio	0	1	0.50	0.07	0.25	0.45	0.50	0.56	0.72	
est_payroll	0	1	104229685	27189418.14	671500.67	136833.92	310000.00	130693200	291463084.00	
inflation_multiplier	0	1	1.44	0.23	1.03	1.28	1.40	1.60	1.88	
adjusted_payroll	0	1	143648776	40861848.29	247774.99	746626.33	580752318	176514704	402590983.44	

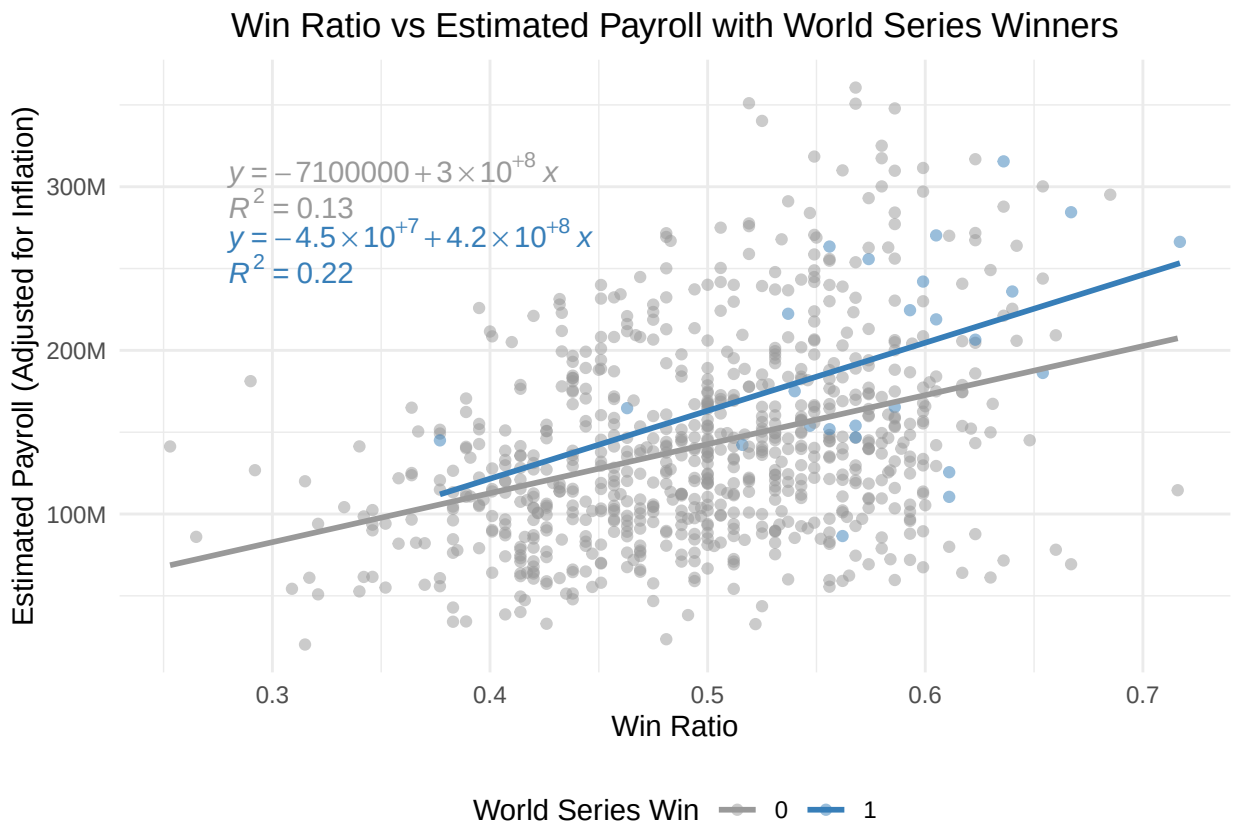
The data set contains 38 variables covering batting statistics, win statistics, and payroll information. Focusing on key batting statistics over the last 25 years, the average number of runs per season is 719, with an average of 171 home runs. The mean batting average is .256, calculated by dividing total hits by total at-bats. The average on-base percentage (OBP) is .325, while the average slugging percentage (SLG) is .414, calculated as total bases divided by at-bats. On-base plus slugging (OPS) averages .740, and OPS+ averages 97.1. In addition to batting metrics, the average estimated team payroll was $140 million. These key variables are central for predicting next season's win ratio using the previous season's end-of-year statistics.

All data were collected from baseball-reference.com for free use. Setting up the data involved creating multiple .csv files containing the relevant statistics. Separate files were constructed for batting statistics, payroll estimates, win statistics, and inflation multipliers to ensure payroll data is accurately adjusted and comparable. A dummy variable for the World Series winner each year was also added. After merging the four data sets by year and team name, missing values in the estimated payroll category were manually addressed, and team names were updated to current names to account for teams that moved cities over the last 25 years, ensuring consistency across the data.

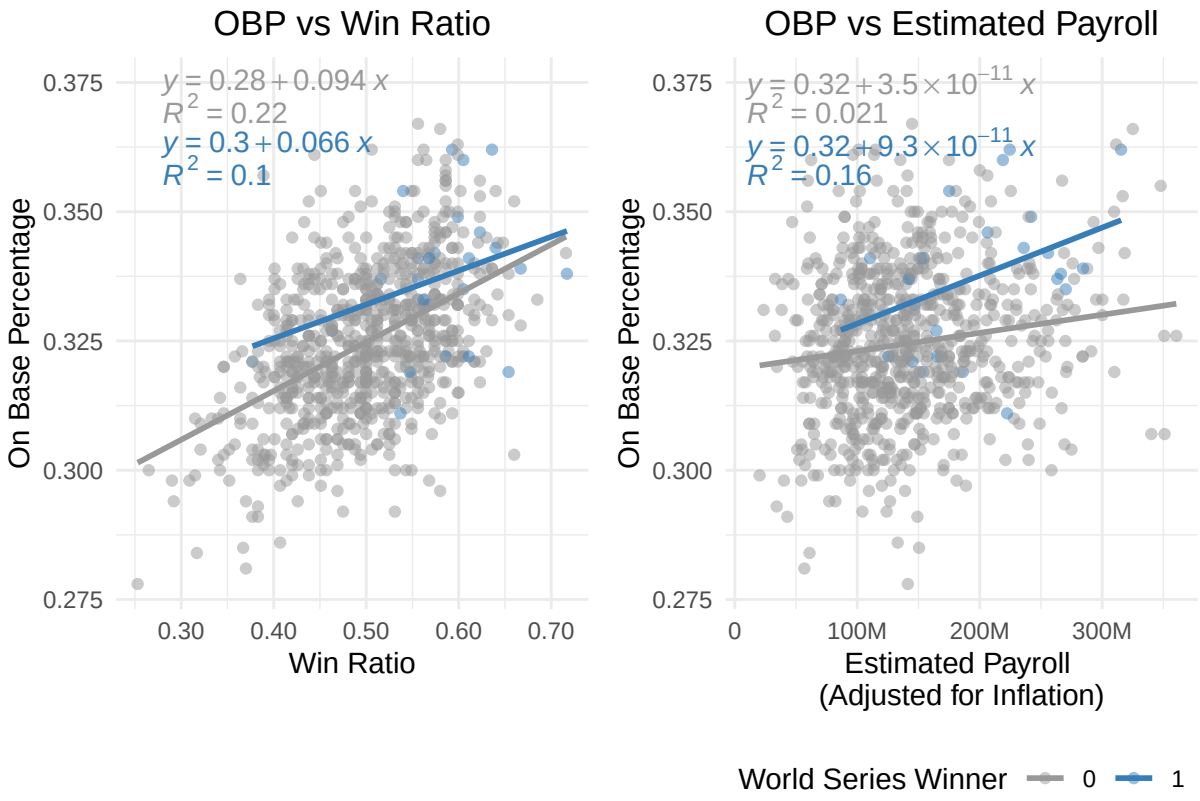
4. Data Exploration



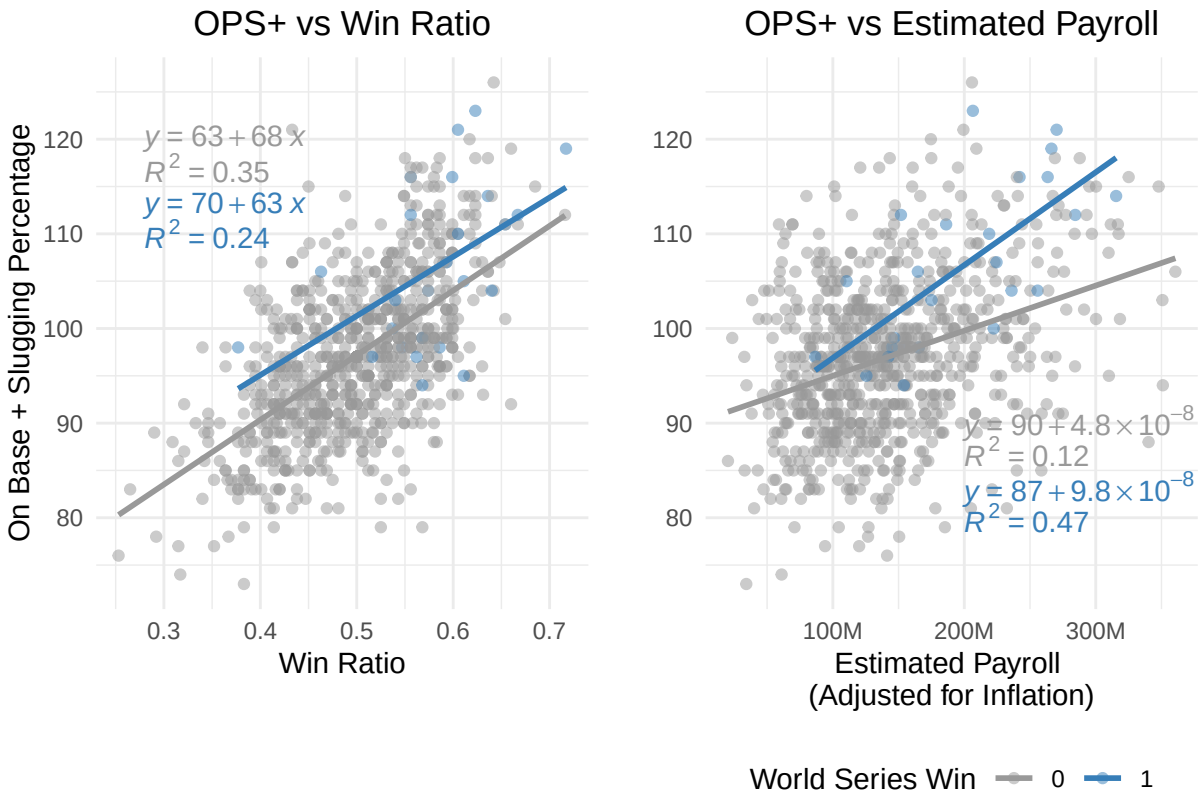
It appears that league payroll has been trending upwards since 2000 with intermittent decreases until reaching a sharp decline rapidly into the 2020 season and currently trending towards recovery to pre-2020 season levels. The overall increase in payroll is expected since rising popularity levels of the sport generate more revenue than previous years and players will want a proportional, if not larger, piece of the pie. There is a likely explanation for the large dip in 2020 most likely due to COVID-19. The 2020 MLB season was massively delayed due to COVID-19 and shortened from 162 games to 60 games total. Due to this player salaries were massively reduced for this year and players were only paid a portion of their contracted earnings for the year. However, this does not detract from the overall trend increase in payroll each year for the prior 20 years.



There is a clear relationship between estimated payroll and win ratio. As teams spend more money they are more likely to increase their win ratio. Notice how in the last 25 years there are only 4 teams that have spent more than \$250 million and had a win ratio lower than .500. All 4 teams that this is true for as well were very close to a .500 win ratio as well. There are teams that have achieved this for way less, but according to the statistics, wins can be bought to a degree. There are two outcomes that both result in success from the view of an owner that may differ from the views of fans. The first outcome is to win as many games as possible and have a good win ratio. As the team wins more that generates more interest from fans, which increases game attendance and therefore revenue. The second outcome is winning the World Series, which has the same effects as winning as many games as possible in the regular season, but to a far greater degree. The relationship for non-World Series winning teams has an R^2 of .13 and a strong correlation coefficient of $3 * 10^8$. This means that 13% of the variation in win ratio can be explained by the variation in explained payroll, and an increase in win ratio by .1 is associated with an increase in estimated payroll by \$30 million. Comparing these values to World Series winning teams, 22% of the variation in win ratio is explained by the variation in estimated payroll and an increase in win ratio by .1 is associated with an increase in estimated payroll by \$42 million. Based off of this, World Series winning teams have to pay more for their wins than non-winning teams, which makes sense since they are presumably paying more money for better players on their team to increase their odds of winning the World Series.



Shifting the focus to On Base Percentage (OBP) and separating the observations once again by World Series winners and non-winners the difference shows. In the film *Moneyball* the managers focus on picking players that get on base, expressed through OBP. Based off of the plot with Win Ratio and OBP it appears this may work since there is quite a strong relationship between the two variables for non-winning teams with an R^2 of .22 and a correlation coefficient of .094, which the Oakland Athletics did not win the World Series in 2002 when the movie is based. For non-winning teams a .1 increase in win ratio is associated with a .0094 increase in OBP. Therefore, picking the players for a team based around OBP may have been an actually viable strategy at that point in time. Shifting to World Series winning teams and a different idea emerges, OBP actually matters less for World Series winning teams. Less of the variation in Win Ratio can be explained by OBP with winners retaining an R^2 of .1, which is still very large for only including Win Ratio to explain OBP, but is much lower than the non-winners. The winners also have a correlation coefficient of .066 which is also noticeably lower since a .1 increase in win ratio only has an associated increase of .0066 in OBP, but their average OBP is also higher than non-winners. Moving over to look at how OBP is effected by estimated payroll it shows more of the same as how win ratio was effected by payroll. The relationship between OBP and payroll is more than two times more important for World Series winning teams than non-winning teams, with winning teams achieving a coefficient more than two times as large and an R^2 more than seven times as large. This is expected since for World Series winning teams getting on base is not enough, hitting and driving in runs is more important to winning as many games as teams need to to get to and win the World Series.

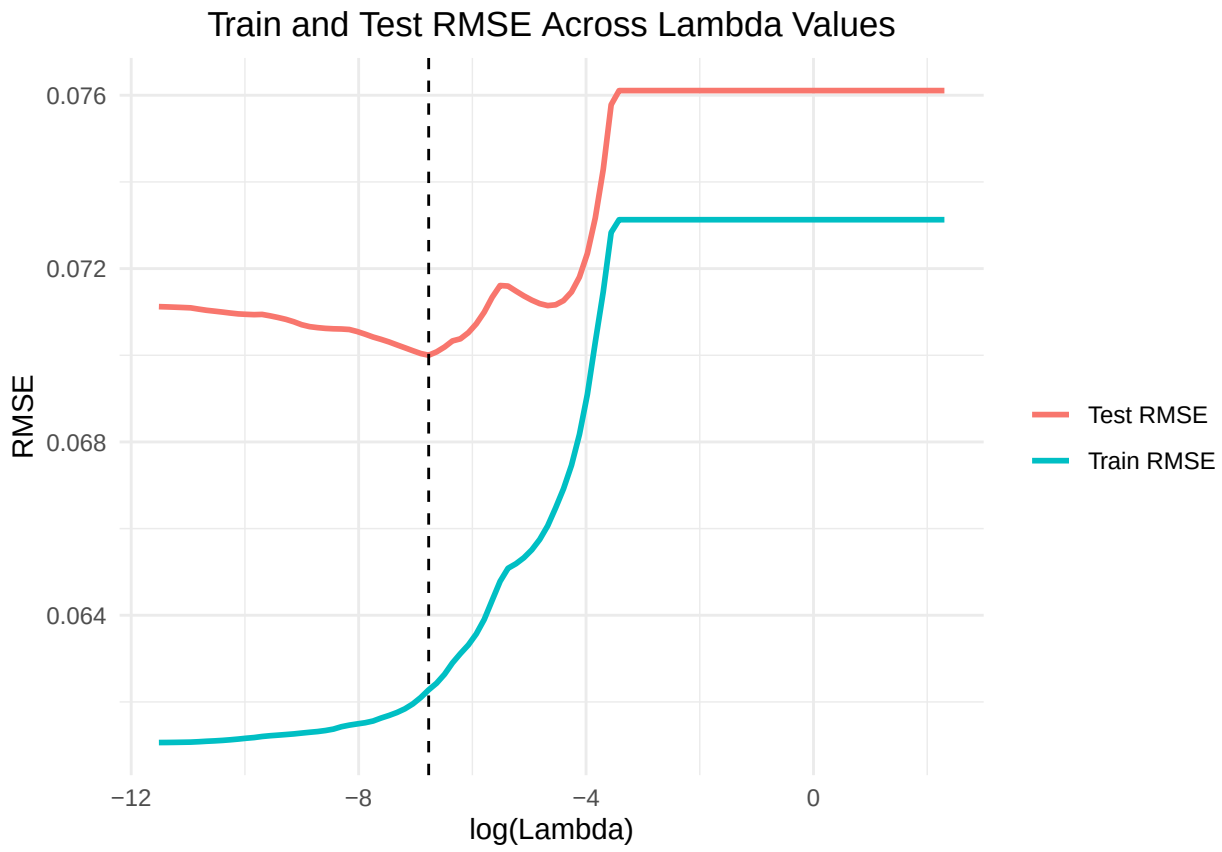


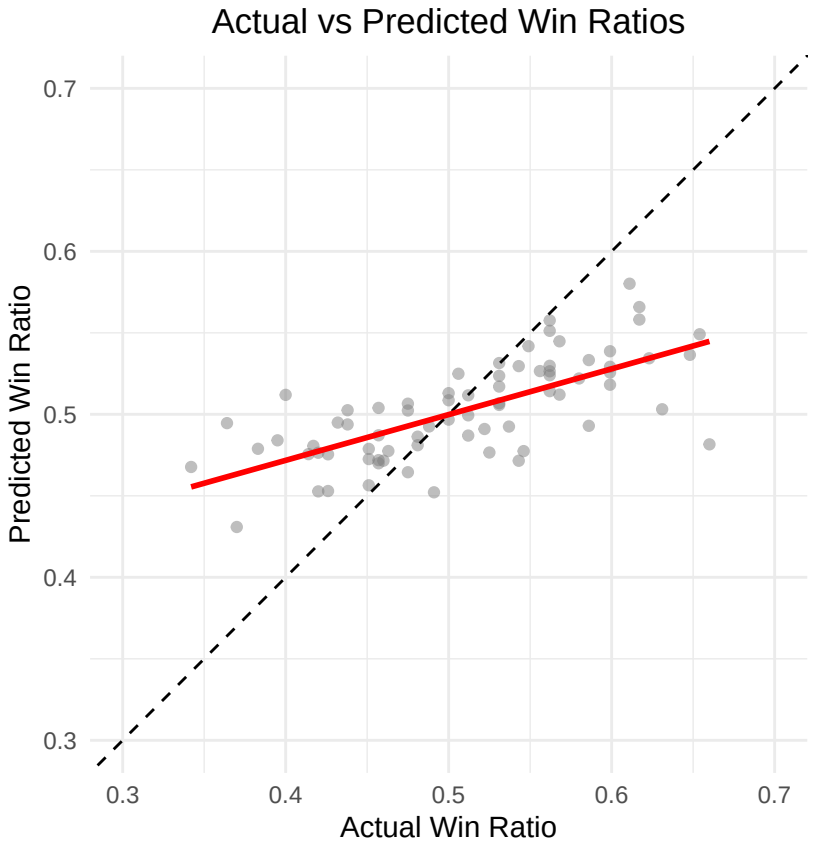
OPS+ is a statistic that combines on base percentage and slugging percentage and then normalizes them over the entire league. This normalization is to account for differences in lengths of ballparks and other variables that change by ballpark. This is statistic is vitally important for winning games since you cannot win games if you do not get on base, and then slugging percentage is calculated by $Slugging = \frac{(1*Singles)+(2*Doubles)+(3*Triples)+(4*HomeRuns)}{(AtBats)}$. Slugging weighs the importance of the hits against at bats. In recent years MLB teams have been shifting their focus to OPS+ since it is vitally important to winning games since low slugging percentages mean teams not getting impactful hits when needed and scoring. It appears that win ratio has roughly the same effect on OPS+ for World Series winning and non-winning teams. World Series winning teams have a noticeably lower correlation coefficient decreasing from 67 to 59 meaning that a .1 increase in win ratio is associated with an 6.7 unit increase in OPS+ for non-winning teams and a 5.9 unit increase for World Series winning teams. The R^2 values reflect that 35% of the variation in win ratio can be explained by OPS+ for non-winning teams and 24% for World Series winning teams, which are quite high for only explaining OPS+ with win ratio. Shifting to the effect of payroll on OPS+ and there is a huge difference between winners and non-winners. For World Series winning teams the R^2 is .47 which is astronomically large for only explaining OPS+ with estimated payroll. This can be interpreted as World Series winning teams are able to more efficiently convert payroll to offensive performance. This is best represented by the current Los Angeles Dodgers roster. The Dodgers spend the most money out of any team in the league in payroll and purchase the best batters on the market, spending $70 million a year on just one player with remarkable batting skills named Shohei Ohtani. For non-winning teams the R^2 value of the effect of payroll on OPS+ is .12 which is large in its own right, but a huge step down from winning teams. It is less likely that OPS+ is associated with payroll for non-winning teams since they may have more elite level rookies or mediocre veteran players that have a good season and produce good offensive numbers for the team, but are not being paid much at the time of their elite performances.

5. Predictive Model

Table 7: Train and Test RMSEs for LASSO Across Lambda Values

lambda	train_rmse	test_rmse
0.000	0.061	0.071
0.000	0.061	0.071
0.000	0.062	0.070
0.000	0.062	0.070
0.000	0.062	0.070
0.001	0.062	0.070
0.001	0.062	0.070
0.001	0.062	0.070
0.001	0.062	0.070
0.001	0.062	0.070
0.001	0.062	0.070
0.001	0.062	0.070
0.002	0.063	0.070
0.002	0.063	0.070
0.002	0.063	0.070
0.002	0.063	0.071
0.003	0.064	0.071
0.003	0.064	0.071
0.004	0.064	0.071
0.004	0.065	0.072
0.005	0.065	0.072





The LASSO model seeks to predict the next season's win ratio based off the end of season batting statistics from the previous season including end of season payroll as well. A complete success would be the predicted values having very small residuals from the dashed black line with a slope of one. The model did not achieve perfect results, but still shows a strong positive relationship in predicting the next season's win ratio. The difference of prediction vs actual most likely come from seasonal roster changes that happen due to trades, drafting, and moving players up and down from organizational minor league teams. The model cannot predict these shocks due to them being unpredictable in nature with only using year-end batting statistics. If this was attempted with individual player level statistics with multiple observations for each player over a year, it may be predictable that decreasing performance may lead to a trade but it would be unpredictable to where a player would go and who they would be replaced by. The variables that the LASSO model chose as the most important were: batters, batter age, average runs per game, runs, number of triples, home runs, batting average, stolen bases, OBP, OPS+, and estimated payroll. The three variables with the highest coefficients are runs, stolen bases, and OPS+. Estimated payroll and OBP have the next highest coefficients and are very important to the model too. This suggests that all the variables that were explored closer are impactful to a team's future success. Based on the RMSE values and final model, it looks like the LASSO model is underfitting since the slope of the predicted values vs actual values is less than 1. To improve the fitting on this model more defensive statistics could be added to the data and increase the amount of seasons in the data frame to give more data to test and train on.