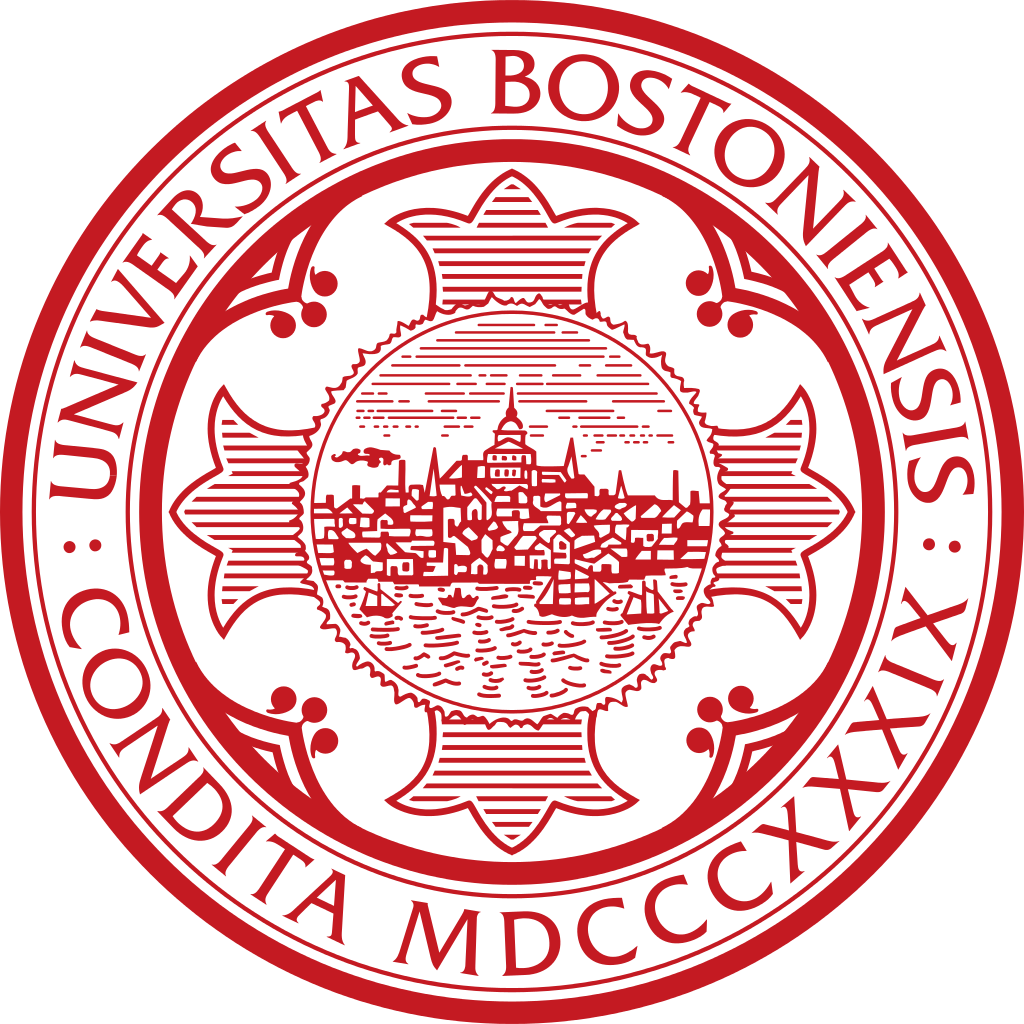
****

**Baseball Statistical Analysis**

*Total Payroll and its Impact on Win Percentage*

Addressed to:

**MLB Team Organizations**

Brendan Rosen

SM222 Professor Kahn

Fall 2015

**Executive Summary**

As America’s past time, Major League Baseball and the thirty organizations within the American and National league have had a lot of weight to carry for their respective cities and states. For this reason, teams should strive to obtain the most up to date and precise data analytics concerning interactions and tendencies between MLB-wide data over the past fourteen years (omitting the current year). Understanding and knowing these certain tendencies and trends will lead to a better success rate for teams. Organization executives and coaches can focus on what is important and make better-informed decisions. This report’s purpose was to investigate whether or not total payroll had any statistically significant relationship in increasing a teams win percentage. The principal finding of the analysis exposed that payroll does in fact play role in affecting win percentage.

Although this is true, there is greater depth to our conclusion. In the course of reaching this verdict, the examination revealed that understanding control variables and understanding the direction of causality between variables were most important in determining the true significance of total payroll. The simple regression between total payroll and win percentage proved that payroll was highly significant in relation to win percentage. But, when statistically relevant factors were added to the original simple regression, home attendance seemed to be the most correlated with win percentage, not total payroll. Further analysis was done to determine the correlation between total payroll and win percentage through a regression between the change in win percentage and the change in total payroll. The result was that total payroll was not statistically significant. To check the direction of causality, another regression was created for total win percentage in the current year against the payroll per team the following year.

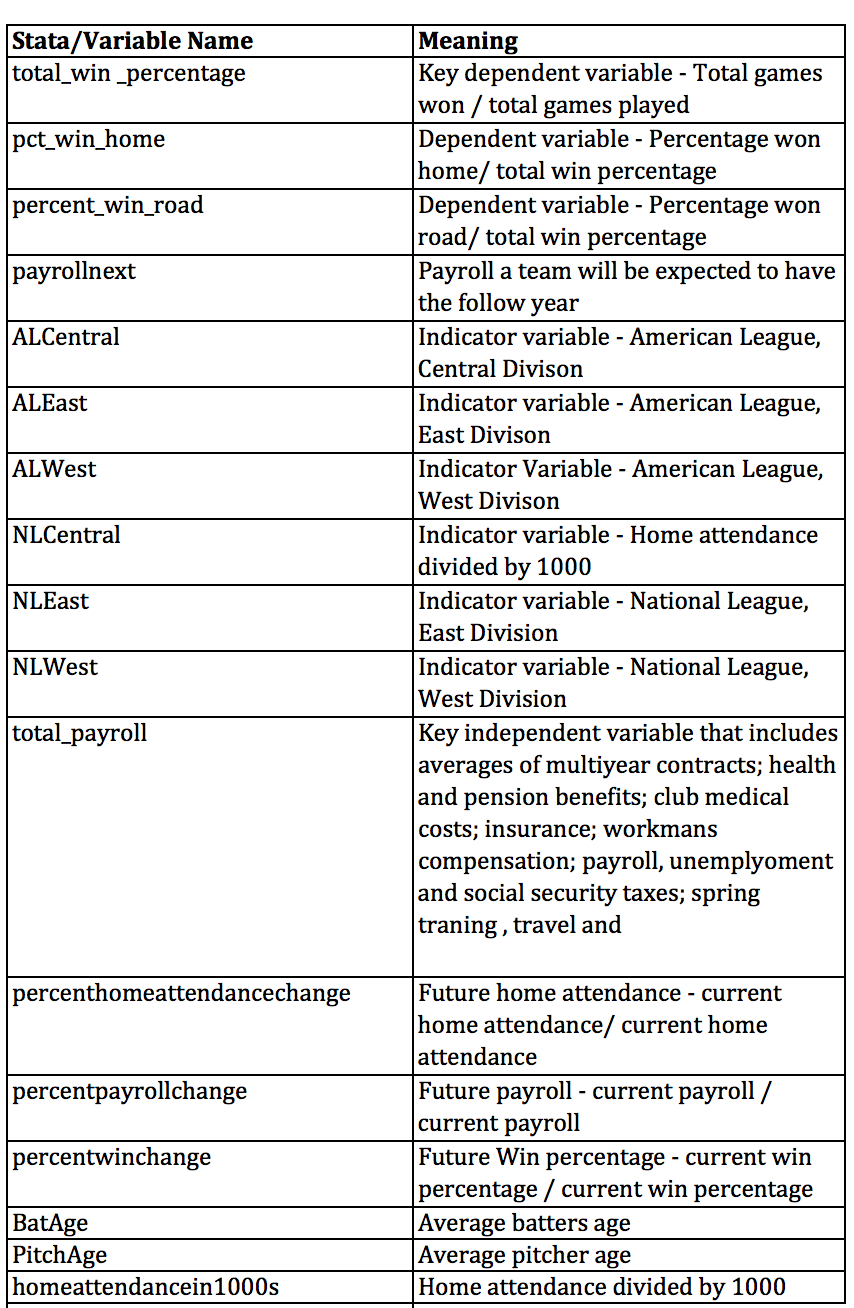
The result was that win percentage is highly significant in relation to a team’s payroll the following year. This comes as no surprise since teams that are winning games can leverage the fact that they are winning, leading to team’s charging higher fees to television networks, and increasing ticket prices for next season since more fans are going to want to go to games where the team is winning.[[1]](#footnote-1)

The next step in the report was to investigate into the direction of causality with home attendance and win percentage. The end result determined the change in wins was likely causing the change in home attendance. This is due to the fact that home attendance has a higher correlation with the percentage of road wins compared to the percentage of home wins. If home attendance was causing a change in win percentage, it should be more significantly related with home win percentage than road win percentage. Since this was not the case, the conclusion was that likely win percentage was causing a change in home attendance, not vice versa.

The final step in the report, in conjunction with the information learned about the direction of causality between total payroll and win percentage, found that in a multiple regression with home attendance controlled for, total payroll is in fact statistically significant in relation to win percentage. The primary finding of this report concludes that wins affects future payrolls, which allows for more wins. The analysis of the data relied heavily on the use of multiple regressions to uncover the true correlation between total payroll and win percentage.

**Introduction**

There is substantial debate as to whether the disparity in a team’s payroll gives an unfair advantage to teams with higher payrolls like the New York Yankees versus teams like the Miami Marlins with lower payrolls. The current thought is that having a larger payroll allows teams to buy better, younger players which in return will positively effect a teams win percentage. This report investigates whether or not total payroll is a significant determinant of win percentage across the MLB over the past fourteen years. The data set[[2]](#footnote-2) used gathered MLB statistics for all thirty teams over the past fourteen years, resulting in each observation being a single season for a specific MLB team in a certain year. This data set contains 450 observations for each variable. This report does not limit itself to determining the correlation between total payroll and win percentage; in addition we investigate the effect of league and division, home attendance, age of players on win rates and on the measured impact of payroll. Below is a table that lists the variables used in the report and their meaning.

****

**Win Percentage**

*Table 1* shows the summary statistics about the variable we predict, the percent of games won. The bottom one percent had a seasonal total win percentage of 34.2% or less, which in a 162 game regular season are approximately 55 wins. The top one percent of teams had a total win percentage of 63.6%, which equates to winning approximately 103 games out of a 162 game regular season. As predicted, the standard deviation is almost exactly 50% because for every win a team has, a loss for the opposing team is the result. Table 1 shows that there is a standard deviation of 7.09% or about 11 games in a 162 game regular season. What this means is that 95% of teams, and their respected seasons, have a win percentage that lies within +/- 22 games (2 empirical standard deviations) of the average of 81 wins (0.5 % win total), or 95% of the observations have been between 59 and 103 % total wins. The histogram in *Figure 1* graphically displays the variable win percentage, and its normal distribution.

**Simple Regression**

The next step in the analysis was to run a simple regression using payroll to predict win percentage to test how correlated total payroll and total win percentage were.

The best fitting linear relationship between win percentage and total payroll is win% = .45043[[3]](#footnote-3) + .000587\*total payroll. In non-statistical terms, each additional dollar increase in total payroll will result in an additional increase of .000587 of win percentage. In *Table 2, Regression 1* shows that the t-stat is 7.31 for total payroll - proving that the variable is statistically relevant. The adjusted R-squared value of 10.47% means that the total payroll (in this regression) explains 10.47% of the total variance of the dependent variable, total win percentage. Although *Table 2, Regression 1* has a high t-stat value, the value of coefficient could be biased since the simple regression does not control for possibly confounding factors. The total payroll variable could be picking up the influence of the omitted variables causing it to be biased. An additional point of interest is the unclear direction of causality between wins and payroll. Because of these questions, the next step in the analysis of win percentage and total payroll is to add possibly confounding factors to *table 2, Regression 1,* which will help determine more accurately the effect of total payroll relative to total win percentage.

**Regression of Potentially Confounding Variables and Total Payroll**

*Table 3 (total payroll in millions)* shows the regression we ran to determine which variables we are able to measure that affect win percentage and are correlated with total payroll, therefore, may bias its coefficient in our win percentage model. Since the purpose of this regression is to determine whether the control variables are correlated with total payroll, total payroll is made the dependent variable. Home attendance was included in the regression because total payroll is generally higher the more people that come to games, since the ticket revenues will be increased.

This relationship between the two variables can work in both directions, on one hand; an increase in home attendance will have a positive affect on total payroll but a higher total payroll could also lead to higher home attendance since teams could buy better players or make improvements to their stadiums to increase fan attendance. We consider this issue in more depth later on in this report.

*Table 3 (total payroll in millions)* also includes indicator variables for both the American and National league teams with their respective divisions of east, west, and central. In order to truly know the effect total payroll has on win percentage, the analysis needed to consider the control variables including leagues and divisions and their effect on total payroll.The average opening day payroll for the American League teams is$97.7 million, compared to $88.6 million for National League teams; the analysis explains whether these variables are possibly confounding factors with total payroll.[[4]](#footnote-4) Lastly, the regression includes average batter ages and pitching ages for teams. Since the most prevalent belief across the MLB is that buying the best young players will increase your chance of winning, the regression accounts for players age[[5]](#footnote-5) to test whether or not these are control variables to consider.

As *Table 3 (total payroll in millions)* shows, home attendance is highly correlated with total payroll, showing that every fan increase in home attendance leads to an increase in $40.34 of total payroll. Furthermore, NLWest, NLCentral, ALEast, and PitchAge have relatively high t-stats, proving these variables have a statistical significant effect on total payroll. *Table (total payroll in millions)* verified the assumption that total payroll in *Table 2, Regression 1* could have been picking up the significance of these correlated variables, thus making its coefficient biased. Further analysis and regressions were needed to determine the true significance of total payroll and the possibly confounding factors.

**Total Payroll and Possibly Confounding Variables vs. Total Win Percentage**

Given the correlation between total payroll and these possibly confounding variables, two forms of regression analysis were run to help determine if payroll in fact had any effect on holding these variables constant. With the addition of these control variables, the t-stat of total payroll dropped from 7.31 in *Table 2, Regression 1* to -1.61 in *Table 2, Regression 2,* muchbelow the significant relevant range (of 1.96). There was a positive bias in the simple regression of *Table 2, Regression 1* most likely due to the fact that total payroll and home attendance are positively correlated, and in *Table 2, Regression 2* home attendance increases win percentage. After inputting the data into IBM’s Watson analytics tool we can graphically see the positive correlation between total payroll and home attendance in *Figure 2*.

Using IBM’s Watson analytics tool again in *Figure 3,* we vividly see the correlation between win percentage and league and division. Control variables, league and division seemed to have relatively similar impacts on win percentage besides for variables NLCentral and ALWest. NLCentral’s coefficient and t-stat in win related regressions were consistently the most statistically relevant in regards to percent win total as the graph demonstrates. On the other hand, ALWest had the lowest statistical relevance in respected regressions, as we can see in *Figure 3*. To our surprise the remaining control variables of average batting age and pitching age did not have a meaningful impact on win percentage. Although, in some cases, average pitchers age had a somewhat statistically relevant affect with a t-stat value above 1.96, we concluded that the control variables of average batters age and pitching age did not demonstrate a relevant impact on win percentage. With control variables accounted for, we took our analysis a step further. Although the level of payroll did not explain win percentage; perhaps changes in payroll would affect changes in winning ratios.

**Change in Win Percentage vs. Change in Payroll**

Two variables were generated, percentwinchange and percentpayrollchange to test the correlation between payroll and win percentage in *Table 4*. The reasoning behind the drop in observations from 450 to 414 is that the calculation for the delta percentages requires a previous year and for year 2000 we do not have the 1999 information in the data set. The analysis proves that even in a simple regression there seems to be no relationship; with a t-value that is much below the relevant range (of 1.96). Both *Table 2, Regression 2* and *Table 4* prove that total payroll might not have the statistically relevant affect in explaining the variance in win percentage as we saw in *Table 2, Regression 1.*

The ambiguity in the direction of causality between payroll and win as stated earlier in the analysis questioned whether wins was causing an increase in total payroll or whether it was vice versa. Our next analysis tested this through the generation of the variable payroll of next year to see whether win percentage had any positive correlation with the payroll a team has the following year.

**Correlation Between Win Percentage and Payroll Next Year**

The analysis in *Table 3 (payroll next year)* shows that an increase in win percentage of the current year is highly correlated with the total payroll of the following year for a certain team with a t-stat of 9.45. Although there is no correlation in the change in win percentage and the change in total payroll, total win percentage does have a significant effect on the payroll a team would have the following year. The win percentage variable is highly significant due to the fact that a team winning more games generally will result in teams charging higher fees to television networks and increasing ticket prices for the next season, thus increasing total payroll because more fans are going to want to go to games where the team is winning. With the direction of causality between wins and payroll accounted for, the next step in the analysis was to investigate the variable of home attendance and determine its true relationship with win percentage.

**Home Attendance and Win Percentage**

*Table 3 (Total payroll in millions)* proved that home attendance was the factor most highly correlated with total payroll and the most significant factor to explain the variance in total win percentage. Further analysis was done to determine the significance of home attendance in relation to total win percentage. In *Table 4* we test the percent change in win rate and the percent change in home attendance to see whether or not there was a correlation between the two variables. The coefficient for percenthomeattendchange is .3058, meaning that for every change in home attendance win percentage also changes by 30.58%. The t-value of 6.35 indicates that this relationship is highly significant.

The problem with this regression is that there is no way of telling which direction it runs. Does the change in win percentage lead to a change in home attendance, or vice versa?

A multiple regression was used to determine the effect home attendance and its impact on the percentage of home wins and road wins separately, including controls for league, division, and average batting and pitching age. The results in *Table 2 (Percent Win Home)* show that home attendance has a highly significant effect on win percentage at home with a t-value of 9.26. The point of concern is that *Table 2 (Percent Win Road)* runs the same regression with percentage of road wins versus home attendance and controls; in *Table 2 (Percent Win Road),* home attendance has a larger coefficient and its t-value is 11.37, 2.11 higher than its regression on home wins. Because *Table 2 (Percent Win Home)* and *Table 2 (Percent Win Road)* prove that home attendance does not effect home win percentage more than it does with road win percentage, it is likely that changes in win percentage cause a change in home attendance rather than vice versa in *Table 2, Regression 2.* With this knowledge, a new and final regression was run to test the correlation of total payroll, win percentage, and relevant control variables without home attendance controlled for.

**Total Payroll, Win Percentage, and Relevant Control Variables**

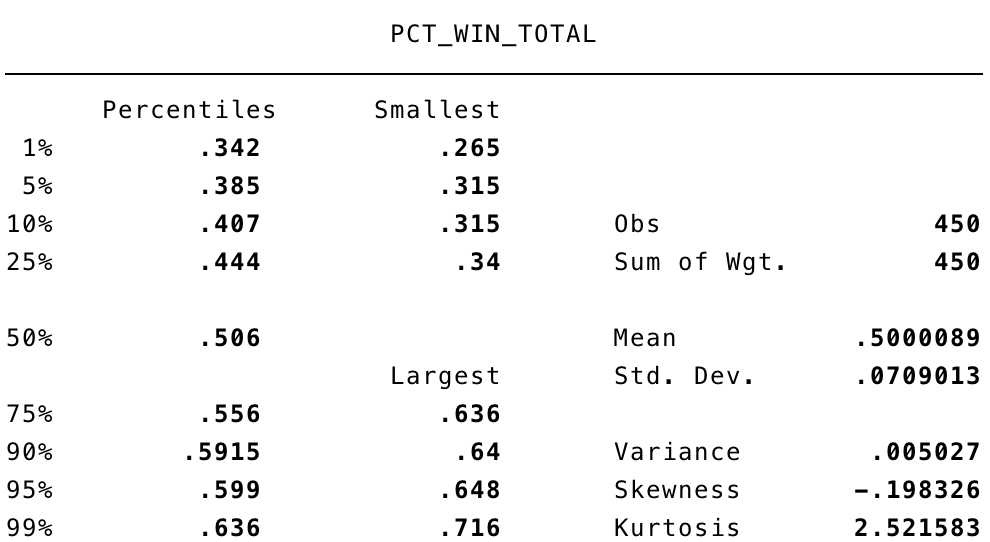
With the determination that win percentage was likely causing the change in home attendance, the variable home attendance was dropped from the final regression. In *Table 2, Regression 3* we see that total payroll does in fact have a highly significant relationship with win percentage with a t-stat of 6.75. In non-statistical terms, each additional dollar increase in total payroll will result in an additional increase of .000564 of win percentage. Other than NLCentral, no other indicator variables used in the final regression in *Table 2, Regression 3* showed a meaningful impact on win percentage. Together with *Table 3 (Total payroll next year),* the conclusion would be that win affects future payrolls, which allows for more wins.

**Conclusion**

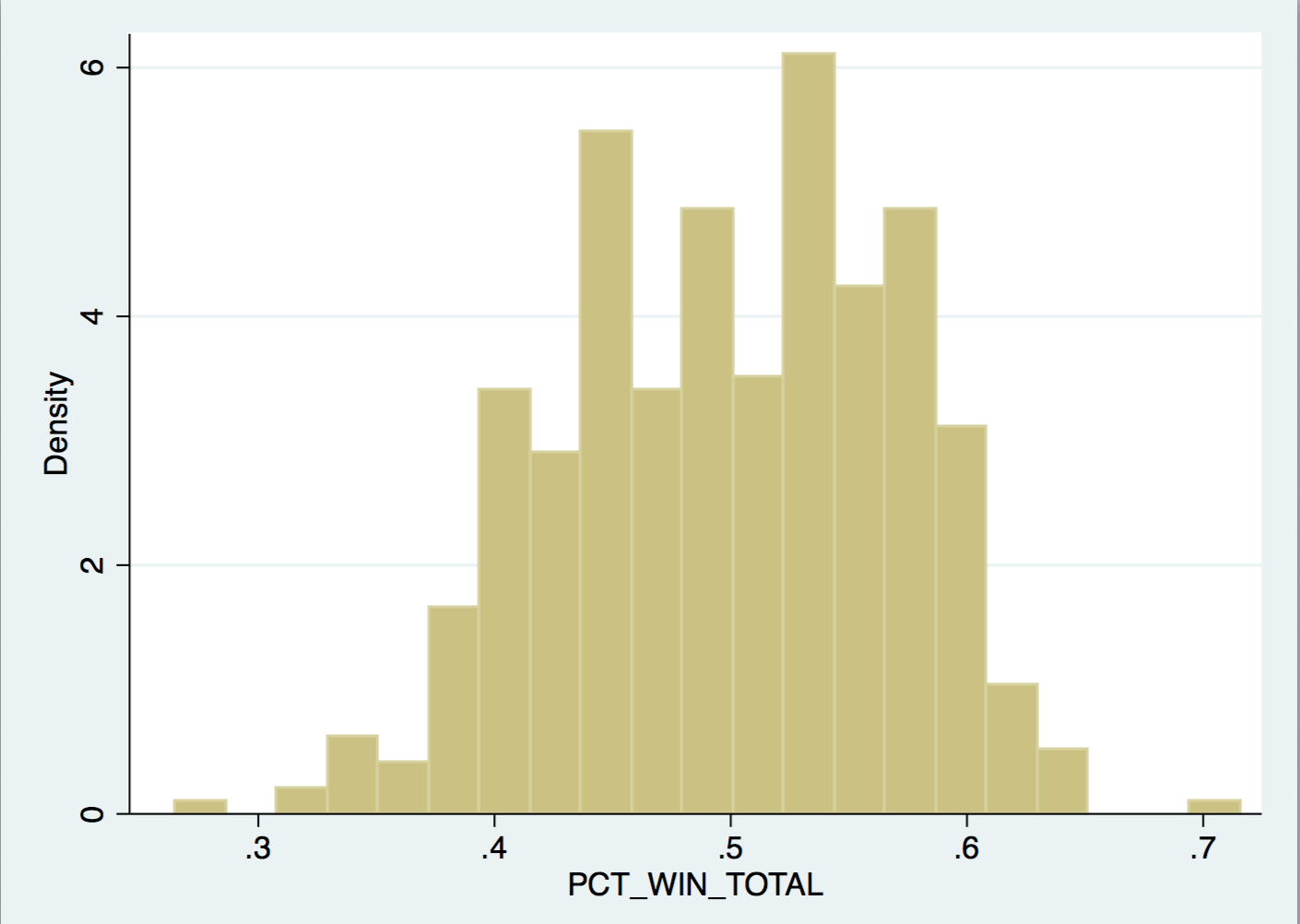
The goal of this report was to investigate whether total payroll affects win percentage. This report displays how easy it is to misinterpret information, when either significant factors are not controlled for or if it is unclear the causality between two variables. The initial simple regression in *Table 2, Regression 1* showed total payroll as being significantly correlated with win percentage. The unclear direction of causality between win percentage and payroll led to the next regression in *Table 3 (Payroll for Next Year)* where we determined that win percentage does have an effect on the total payroll of a team for the next year. The next step was running a multiple regression in *Table 3 (Total Payroll Millions*) where we accounted for any possibly confounding variables with total payroll. This regression determined that total payroll was picking up the influence of omitted variables causing it to be biased in the simple regression. The most correlated variable with total payroll was home attendance, further analysis was done to determine the relationship between home attendance and win percentage.

We first ran a simple regression of the percent change in win rate and the percent change in home attendance which concluded that the two variables were in fact highly correlated. Still, the question of the direction of causality remained; whether or not wins affected home attendance or vice versa. To determine the answer, two multiple regressions were run, *Table 2 (Percent Win Home)* and *Table 2 (Percent Win Road).* The conclusion was that home attendance had a larger coefficient and t-stat in the in *Table 2 (Percent Win Road),* proving that it is likely that win percentage causes a change in home attendance. With this information, *Table 2, Regression 3* was run, without controlling for home attendance, to see the true relationship between payroll and win percentage. With all controls and relevant factors accounted for, the conclusion would be that wins affects future payrolls, which allows for more wins.

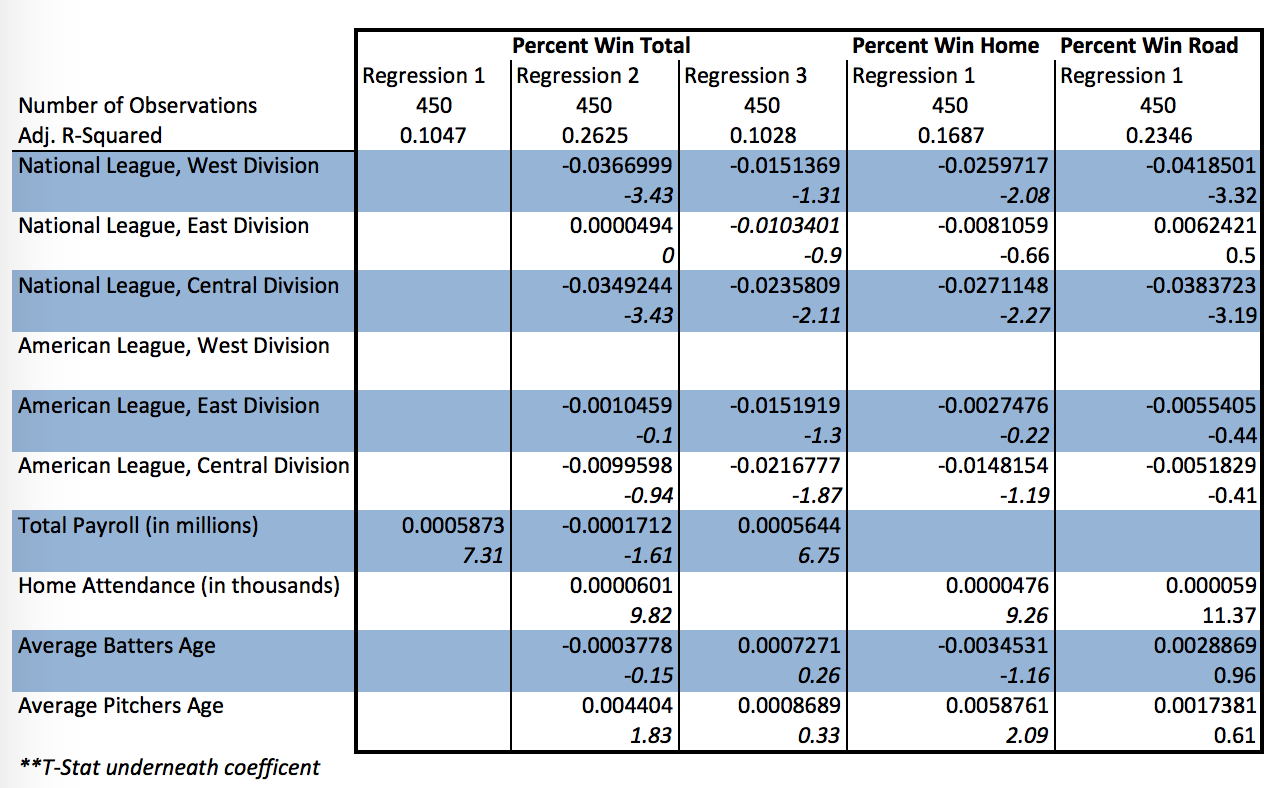
**Appendix**



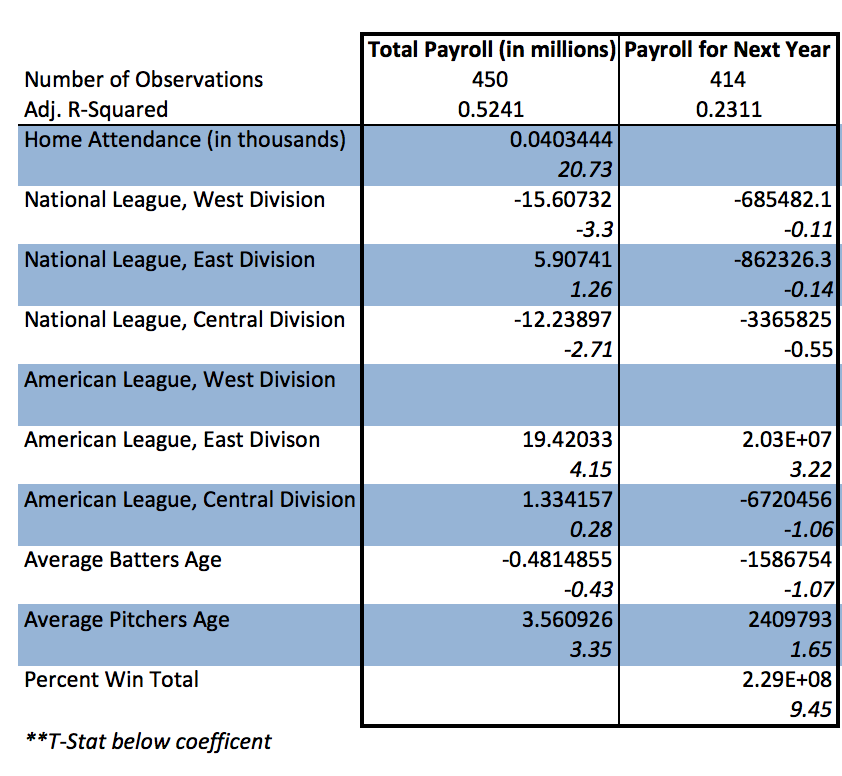
**Detailed Summary, Table 1**



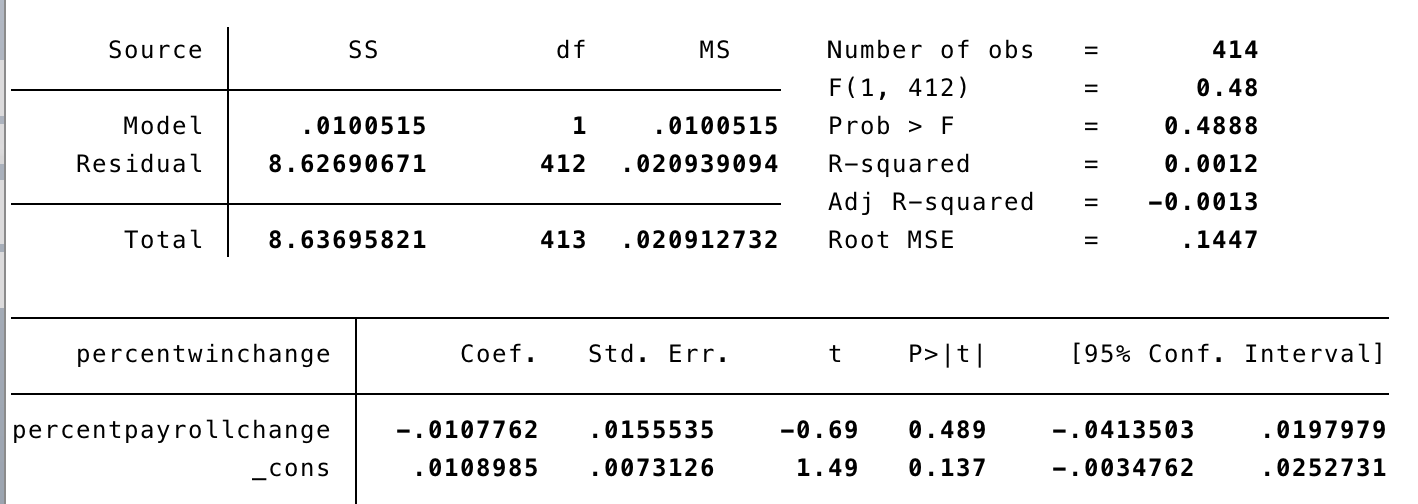
**Histogram, Figure 1**

****

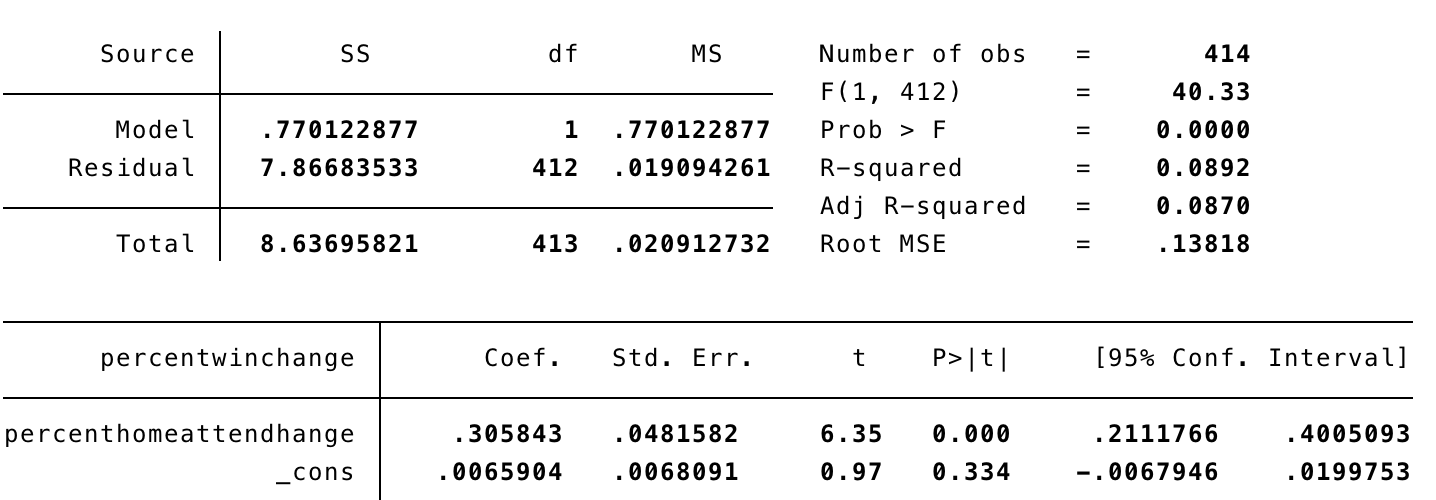
**All win related regressions, Table 2**

****

**All payroll related regressions, Table 3**

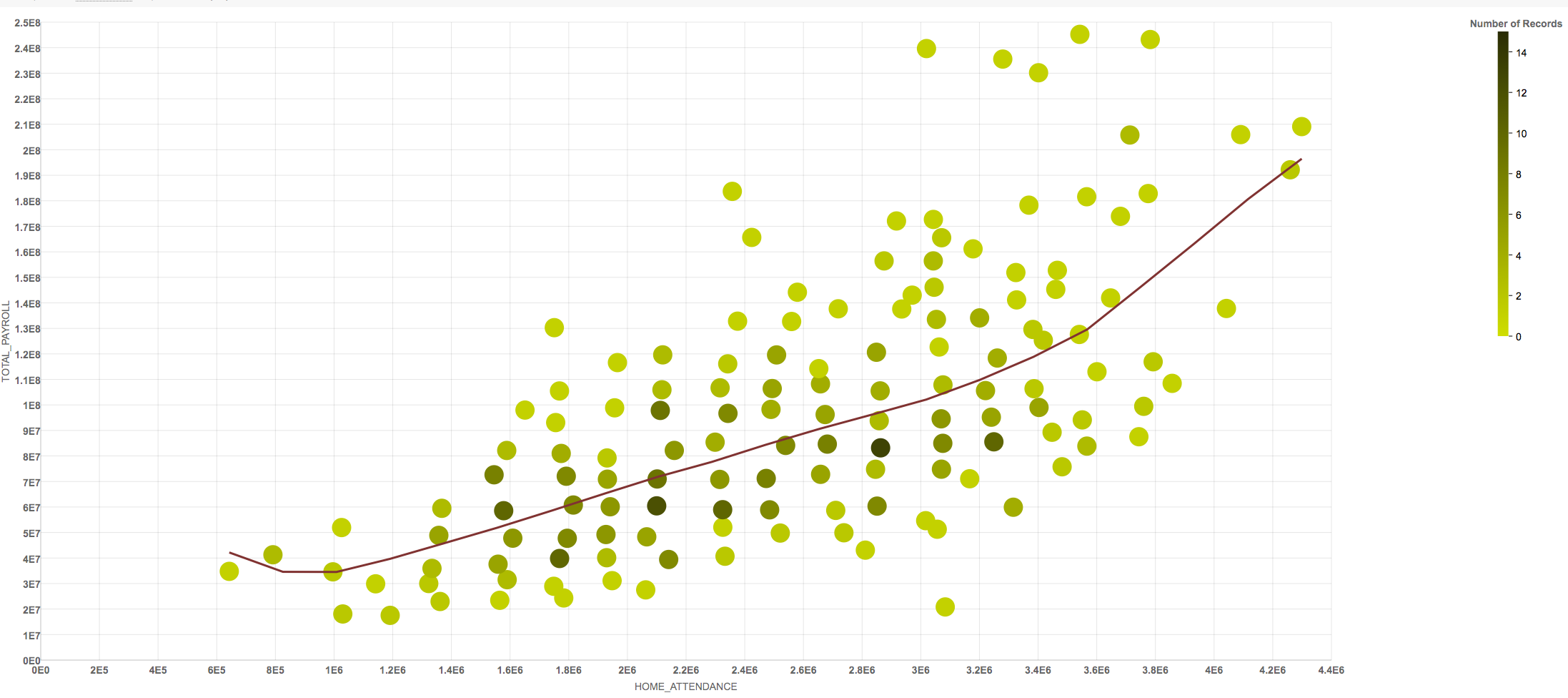
****

**Table 4**

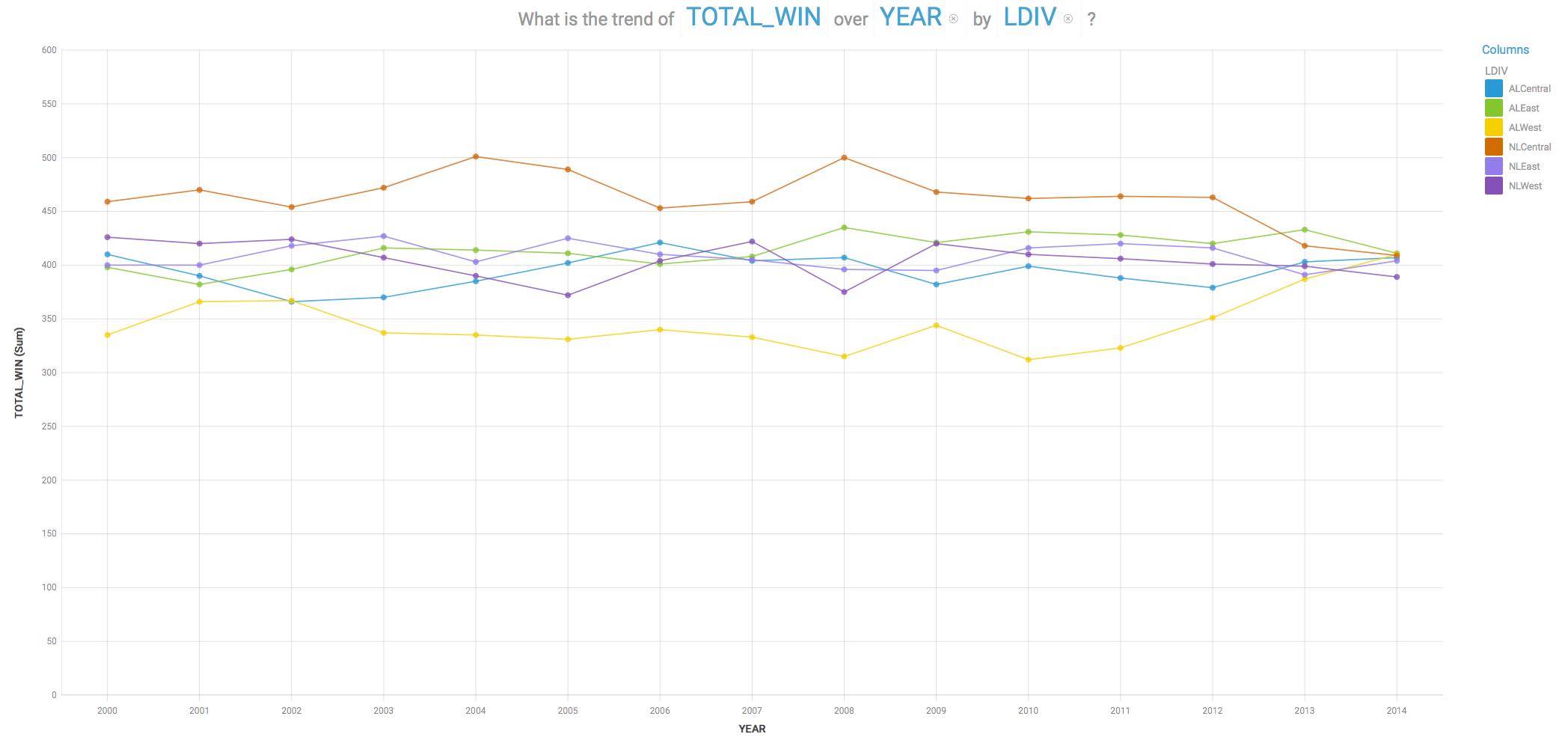
****

**Table 5**

*Correlation between Total Payroll and Home Attendance*



**Figure 2**



**Figure 3**

**Works Cited**

Data Set Material/References:

--MLB Teams and Baseball Encyclopedia

[Baseball-Reference.com](http://baseball-reference.com/) <[http://baseball-reference.com](http://baseball-reference.com/)>

--World Series Winners, Stats, and Results and Postseason Series

<http://www.baseball-reference.com/postseason/>

--Major League Baseball Team Payrolls 1998-2015

<http://www.stevetheump.com/Payrolls.htm>

--Regular Season Standings

<http://mlb.mlb.com/mlb/standings/?ymd=20130930#20101003>

--MLB Payroll Tracker | Spotrac

<http://www.spotrac.com/mlb/payroll/2011/>

--Sortable team Stats

[http://boston.redsox.mlb.com/stats/sortable.jsp?c\_id=bos#elem=[object+Object]&tab\_level=child&click\_text=Sortable+Team+hitting&game\_type=%27R%27&season=2015&season\_type=ANY&league\_code=%27MLB%27&sectionType=st&statType=hitting&page=1&ts=1448815035399](http://boston.redsox.mlb.com/stats/sortable.jsp?c_id=bos#elem=%5Bobject+Object%5D&tab_level=child&click_text=Sortable+Team+hitting&game_type=%27R%27&season=2015&season_type=ANY&league_code=%27MLB%27&sectionType=st&statType=hitting&page=1&ts=1448815035399) <<http://boston.redsox.mlb.com/stats/sortable.jsp?c_id=bos#elem=%5Bobject+Object%5D&tab_level=child&click_text=Sortable+Team+hitting&game_type=%27R%27&season=2015&season_type=ANY&league_code=%27MLB%27&sectionType=st&statType=hitting&page=1&ts=1448815035399>>

--The Analytics Edge -

<https://github.com/pedrosan/TheAnalyticsEdge/blob/master/data/baseball.csv>

References made in paper:

--How are wins, attendance and payroll all related?

<http://www.hardballtimes.com/how-are-wins-attendance-and-payroll-all-related/>

--Analysis of impact of Payroll on regular and post season

<http://digitalcommons.iwu.edu/cgi/viewcontent.cgi?article=1046&context=uer>

1. Swartz, Matt. "Predicting Major League Baseball Salary Growth." The Hardball Times. March 28, 2014. Accessed November 28, 2015. http://www.hardballtimes.com/predicting-major-league-baseball-salary-growth/. [↑](#footnote-ref-1)
2. Citations listed in references [↑](#footnote-ref-2)
3. From stata regression \_cons=.4504259 [↑](#footnote-ref-3)
4. Gleeman, Aaron. "AL Out-spends NL by $10 Million per Team, AL East Out-spends Other Divisions by Wide Margin." HardballTalk. April 4, 2011. Accessed November 29, 2015. http://mlb.nbcsports.com/2011/04/04/al-out-spends-nl-by-10-million-per-team-al-east-out-spends-other-divisions-by-wide-margin/. [↑](#footnote-ref-4)
5. Batters age (positional players) and pitchers age [↑](#footnote-ref-5)