Capstone Project

MLB Payroll vs Winning

Will Pauls

February 18, 2019

## Abstract

Growing up I have always been a big baseball fan. I have been a die-hard Royals fan since the day I was born. Until 2014, I had seen 20 years of Royals seasons end without a playoff birth and not much winning at all. One of my thoughts for this was because the Royals were known as a "cheap" organization. For this project, I wanted to explore whether an MLB teams payroll has a relationship with how well they did that season in terms of winning %. My client for this project could be any MLB teams owner and/or GM, and I would like to report to them whether Payroll has any relationship with winning %.

## Introduction

For this project, I have team and salary data for the 1985 through 2016 seasons via [**Lahman's Baseball Database.**](http://www.seanlahman.com/baseball-archive/statistics/) First I will look at the statistics of teams whose winning increased based on their change in Payroll Rank and change in % of total MLB Payroll they have. Then, I will fit a linear regression model with winning % as the dependent variable, and either rank in team payroll or % of total MLB payroll as the predictor variable based on my findings in the **Probability & Statistics** section. I chose winning % as the dependent variable over Total Win's or playoff appearances due to the fact of strike shortened seasons, and weather delays not made up causing differences in total games played for a few of the seasons, and playoff expansion in 1994 and 2012 causing a difference in total playoff teams for some seasons. After fitting the regression I will do some diagnostic tests to test the fit of the model, and then come to a conclusion.

## Data wrangling

For my Capstone project I will be pulling from the Teams and Salaries datasets from the Lahman package. Performing data wrangling tasks and merging them into a new dataset called baseball

**Steps**  
1. Load in Teams dataset with necessary variables

# make team data frame with necessary variables  
teams = Teams  
teams = subset(teams, select = c(yearID, lgID, teamID, Rank, G, W, L, DivWin, WCWin))

1. Since Salary info dates back only to 1985 I filtered the Teams dataset from 1985 and on to match

# grab teams from 1985 and on to match with years of payroll information  
teams = filter(teams, yearID >=1985)

1. I added a column in teams for winning %

# add winning % column  
teams = mutate(teams, winpercent = round((W/(W+L)), digits = 3))

1. I added a column with a boolean data type (Y/N) of whether that team made the playoffs

# add a playoffs column  
teams$DivWin[is.na(teams$DivWin)] <- "N"  
teams$WCWin[is.na(teams$WCWin)] <- "N"  
teams <- mutate(teams, playoffs = if\_else(DivWin == 'Y' | WCWin == 'Y', 'Y', 'N'))  
teams = subset(teams, select = -c(DivWin, WCWin))

1. I loaded in the Salaries dataset

# add salaries to a data frame  
sals = read\_csv("Salaries.csv")

1. I summed up team payroll by grouping by year and team

# Get team payroll for each year  
sals <- sals %>%   
 group\_by(yearID, teamID) %>%   
 summarise(payroll = sum(salary))

1. I standardized team names in each data set to prepare for merge

#Standardize team names   
sals$teamID[sals$teamID=="CHN"] <- "CHC"  
sals$teamID[sals$teamID=="CHA"] <- "CHW"  
sals$teamID[sals$teamID=="KCA"] <- "KCR"  
sals$teamID[sals$teamID=="LAN"] <- "LAD"  
sals$teamID[sals$teamID=="NYN"] <- "NYM"  
sals$teamID[sals$teamID=="NYA"] <- "NYY"  
sals$teamID[sals$teamID=="SDN"] <- "SDP"  
sals$teamID[sals$teamID=="SFN"] <- "SFG"  
sals$teamID[sals$teamID=="SLN"] <- "STL"  
sals$teamID[sals$teamID=="TBA"] <- "TBR"  
sals$teamID[sals$teamID=="WSN"] <- "WAS"  
  
teams$teamID <- as.character(teams$teamID)  
teams$teamID[teams$teamID=="CHN"] <- "CHC"  
teams$teamID[teams$teamID=="CHN"] <- "CHC"  
teams$teamID[teams$teamID=="CHA"] <- "CHW"  
teams$teamID[teams$teamID=="KCA"] <- "KCR"  
teams$teamID[teams$teamID=="LAN"] <- "LAD"  
teams$teamID[teams$teamID=="NYN"] <- "NYM"  
teams$teamID[teams$teamID=="NYA"] <- "NYY"  
teams$teamID[teams$teamID=="SDN"] <- "SDP"  
teams$teamID[teams$teamID=="SFN"] <- "SFG"  
teams$teamID[teams$teamID=="SLN"] <- "STL"  
teams$teamID[teams$teamID=="TBA"] <- "TBR"  
teams$teamID[teams$teamID=="WSN"] <- "WAS"  
teams$teamID <- as.factor(teams$teamID)

1. I merged the teams and salaries datasets into a new one I named baseball

# merge data into baseball  
baseball <- merge(teams, sals, by=c("yearID", "teamID"))  
  
baseball$teamID[baseball$teamID=="ML4"] <- "MIL"  
baseball$yearID <- as.factor(baseball$yearID)  
baseball$playoffs <- as.factor(baseball$playoffs)

1. I created a column in baseball calculating the amount of payroll per team win (Payroll/W)

# calculate Amount of payroll per team win  
baseball <- baseball %>% mutate(dolperwin = payroll/W)

1. I calculated the average payroll in the MLB per year by summing each teams payroll by year and dividing by the number of teams for that year and added it to a new column in baseball

#calculate total payroll per year for entire MLB  
mlb <- baseball %>%   
 group\_by(yearID) %>%   
 summarise(mlbpayroll = sum(as.numeric(payroll)))  
  
# get number of teams for each year  
 numteams <- baseball %>%   
 group\_by(yearID) %>%  
 summarise(numteams =length(yearID))  
  
mlb$numteams <- numteams$numteams  
  
# calculate average payroll per year  
mlb <- mlb %>%   
 mutate(avgpayroll = mlbpayroll/numteams)  
mlb <- subset(mlb, select = -c(numteams))  
  
# merge into baseball data set  
baseball <- merge(baseball, mlb, by="yearID")

1. I divided each teams payroll by the the total MLB payroll to get a % of payroll each team has compared to entire league and added it into a new column

# calculate percentange of total payroll each team has  
baseball <- baseball %>%   
 mutate(percentofmlb = round((payroll/mlbpayroll), digits = 4))

1. I added a column that calculated the average amount of MLB payroll per each win

#calculate average dollars of mlb payroll per win  
wins <- baseball %>%   
 group\_by(yearID) %>%   
 summarise(mlbwin= (sum(W)))  
  
baseball <- merge(baseball, wins, by="yearID")  
baseball <- baseball %>%   
 mutate(mlbdolperwin = mlbpayroll/mlbwin)  
  
baseball <- baseball %>% group\_by(yearID) %>% mutate(payrank = dense\_rank((desc(payroll))))

1. I added a column with the difference in payroll from the previous year for each team

# calculate teams difference in payroll from previous year  
baseball<- baseball %>%  
 arrange(teamID, yearID) %>%  
 group\_by(teamID) %>%  
 mutate(paydiff = payroll - lag(payroll))

1. I created a % change of that difference in a new column

# +/- % change of payroll from previous year  
baseball<- baseball %>%  
 arrange(teamID, yearID) %>%  
 group\_by(teamID) %>%  
 mutate(percentdiff = paydiff/lag(payroll))   
  
baseball$percentdiff <- round(baseball$percentdiff, 4)

1. I created a column with a difference in winning % from the previous year for each team

## calculate change in winning percentage from previous year  
baseball<- baseball %>%  
 arrange(teamID, yearID) %>%  
 group\_by(teamID) %>%  
 mutate(winpercentdiff = winpercent - lag(winpercent))

1. I created a column with a difference in a teams % of MLB payroll from previous year

## calculate change in percent of mlb from previous year  
baseball<- baseball %>%  
 arrange(teamID, yearID) %>%  
 group\_by(teamID) %>%  
 mutate(percentofmlbdiff = percentofmlb - lag(percentofmlb))

1. I created a column with the differencr in Payroll Rank from the previous year

## calculate change in payroll rank from previous year  
baseball<- baseball %>%  
 arrange(teamID, yearID) %>%  
 group\_by(teamID) %>%  
 mutate(payrankdiff = payrank - lag(payrank))

1. I created a column to test if payroll increased for each team that year

# create column to test if payroll increased for team  
baseball$payincreased[baseball$paydiff<=0]<- "No"  
baseball$payincreased[baseball$paydiff>0]<- "Yes"  
  
baseball$payincreased <- as.factor(baseball$payincreased)

1. I created a column to test if winning % increased for each team that year

# create a column to test if winning % increased for team  
baseball$winincreased[baseball$winpercentdiff<=0]<- "No"  
baseball$winincreased[baseball$winpercentdiff>0]<- "Yes"  
  
baseball$winincreased <- as.factor(baseball$winincreased)

1. I created a column to test if teams % of MLB payroll increased from previous year

# create a column to test if percent of mlb increased for team  
baseball$percentofmlbincreased[baseball$percentofmlbdiff<=0]<- "No"  
baseball$percentofmlbincreased[baseball$percentofmlbdiff>0]<- "Yes"  
  
baseball$percentofmlbincreased <- as.factor(baseball$percentofmlbincreased)

1. I created a column to test if a Teams Payroll rank increased from previous year

# create column to test if payroll increased for team  
baseball$payrankincreased[baseball$payrankdiff<=0]<- "No"  
baseball$payrankincreased[baseball$payrankdiff>0]<- "Yes"  
  
baseball$payrankincreased <- as.factor(baseball$payrankincreased)

### Baseball Data Set

This is the merged and cleaned data set I will be using for my analysis

#display\_output(baseball, out\_type)  
head(baseball)

## # A tibble: 6 x 26  
## # Groups: teamID [1]  
## yearID teamID lgID Rank G W L winpercent playoffs  
## <fctr> <fctr> <fctr> <int> <int> <int> <int> <dbl> <fctr>  
## 1 1997 ANA AL 2 162 84 78 0.519 N  
## 2 1998 ANA AL 2 162 85 77 0.525 N  
## 3 1999 ANA AL 4 162 70 92 0.432 N  
## 4 2000 ANA AL 3 162 82 80 0.506 N  
## 5 2001 ANA AL 3 162 75 87 0.463 N  
## 6 2002 ANA AL 2 162 99 63 0.611 Y  
## # ... with 17 more variables: payroll <int>, dolperwin <dbl>,  
## # mlbpayroll <dbl>, avgpayroll <dbl>, percentofmlb <dbl>, mlbwin <int>,  
## # mlbdolperwin <dbl>, payrank <int>, paydiff <int>, percentdiff <dbl>,  
## # winpercentdiff <dbl>, percentofmlbdiff <dbl>, payrankdiff <int>,  
## # payincreased <fctr>, winincreased <fctr>,  
## # percentofmlbincreased <fctr>, payrankincreased <fctr>

## Probability & Statistics

### Wins and Payroll Rank

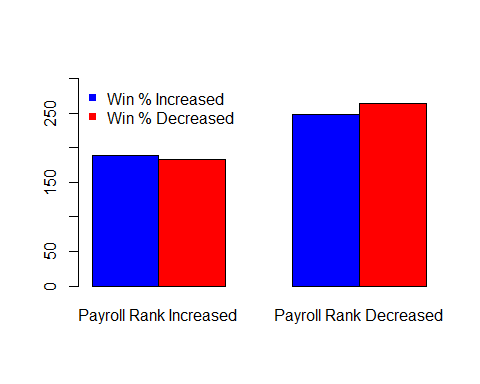
We are looking to see if increasing where you rank in terms of team payroll is correlated with more winning. (This does not necessarily mean the amount a team paid in Payroll went up.)

# summary of teams whose winning increased after payroll rank increased  
summary(baseball$winincreased[baseball$payrankincreased=="Yes"])

## No Yes NA's   
## 183 189 34

# summary of teams whose winning increased after payroll rank decreased  
summary(baseball$winincreased[baseball$payrankincreased=="No"])

## No Yes NA's   
## 264 248 34



#### Hypothesis

: = - The proportion of teams whose wins increased after increasing their Payroll Rank equals the proportion of teams whose winning increased after their payroll rank decreased.  
: - The proportion of teams whose wins increased after increasing their Payroll Rank is greater than The proportion of teams whose winning increased after their payroll rank decreased.

payranktest <- prop.test(x = c(189, 248), n = c((189+183), (248+264)), alternative = "greater")  
# Printing the results  
payranktest

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: c(189, 248) out of c((189 + 183), (248 + 264))  
## X-squared = 0.39359, df = 1, p-value = 0.2652  
## alternative hypothesis: greater  
## 95 percent confidence interval:  
## -0.03464492 1.00000000  
## sample estimates:  
## prop 1 prop 2   
## 0.5080645 0.4843750

Since the p-value is .2652 we do not have enough evidence to reject the null hypothesis that the two proportions are equal to each other.

### Wins and Payroll Amount

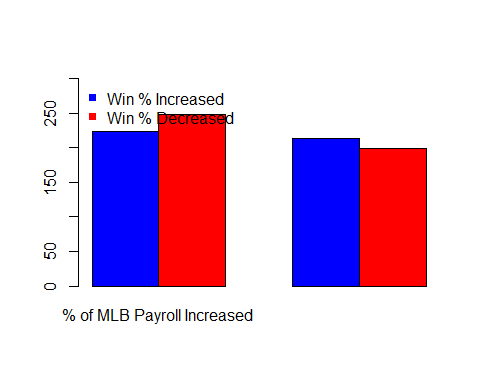
We are looking to see if increasing the amount of money you spend in Payroll in relation to other teams is correlated with winning

# summary of teams whose winning increased after % of mlb pay increased  
summary(baseball$winincreased[baseball$percentofmlbincreased=="Yes"])

## No Yes NA's   
## 248 224 34

# summary of teams whose winning increased after % of mlb pay decreased  
summary(baseball$winincreased[baseball$percentofmlbincreased=="No"])

## No Yes NA's   
## 199 213 34



#### Hypothesis

Since a lower proportion of teams have increased their winning after increasing their payroll than when they decrease in relation to other teams we will test both sides for this hypothesis.

: = - The proportion of teams whose wins increased after increasing their Payroll amount in relation to other MLB teams equals the proportion of teams whose winning increased after their payroll amount decreased in relation to other teams.  
: - The proportion of teams whose wins increased after increasing their Payroll amount in relation to other MLB teams does not equal the proportion of teams whose winning increased after their payroll amount decreased in relation to other teams.

payrolltest <- prop.test(x = c(224, 213), n = c((224+248), (213+199)))  
# Printing the results  
payrolltest

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: c(224, 213) out of c((224 + 248), (213 + 199))  
## X-squared = 1.418, df = 1, p-value = 0.2337  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.11069982 0.02587178  
## sample estimates:  
## prop 1 prop 2   
## 0.4745763 0.5169903

Since the p-value is .2337 we do not have enough evidence to reject the null hypothesis that the two proportions are equal to each other.

### Conclusion

Looking at the proportions of teams whose winning increased after they increased their rank in payroll and after they increased their % of MLB payroll, it is apparent that the amount of money a team puts into their payroll isn't necessarily as important as where they rank in team payroll regardless of how much they spend. For this reason, our linear regrssion model we will use in the Machine Learning section will use Payroll Rank as the Independent variable.

## Machine Learning

### Hypothesis

: = 0 - Teams payroll rank has no linear relationship with their winning %  
: 0 - Teams payroll rank has a linear relationship with their winning %

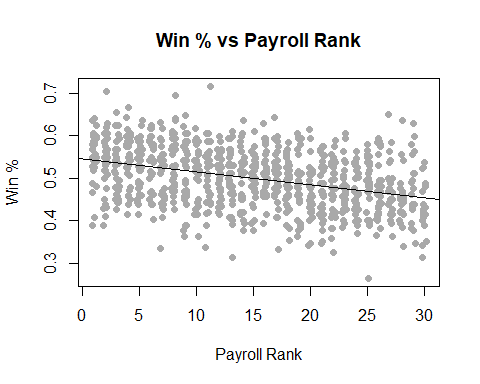
#### Linear regression summary with confidence interval

##   
## Call:  
## lm(formula = baseball$winpercent ~ baseball$payrank)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.204099 -0.046392 0.002225 0.044436 0.204081   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.5455626 0.0043006 126.86 <2e-16 \*\*\*  
## baseball$payrank -0.0030585 0.0002518 -12.15 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06378 on 916 degrees of freedom  
## Multiple R-squared: 0.1387, Adjusted R-squared: 0.1378   
## F-statistic: 147.6 on 1 and 916 DF, p-value: < 2.2e-16

## 2.5 % 97.5 %  
## (Intercept) 0.53712252 0.554002730  
## baseball$payrank -0.00355268 -0.002564411

The P-value is practically 0 so there is a signifcant result and we can reject the null hypothesis and conclude that there is a linear realtionship between a teams winning % and where they rank in Team Payroll.

plot(baseball$winpercent~jitter(baseball$payrank), main = "Win % vs Payroll Rank", xlab = "Payroll Rank", ylab = "Win %", pch = 16, col = "dark grey")  
abline(MLB)



### Testing Linear Regression Assumptions

#### Assessing Outliers

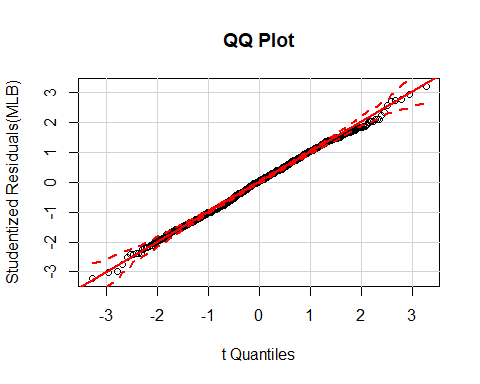
# Assessing Outliers  
  
outlierTest(MLB) # Bonferonni p-value for most extreme obs

##   
## No Studentized residuals with Bonferonni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferonni p  
## 306 -3.220594 0.0013244 NA

There are no outliers that would influence our model in a signifcant way.

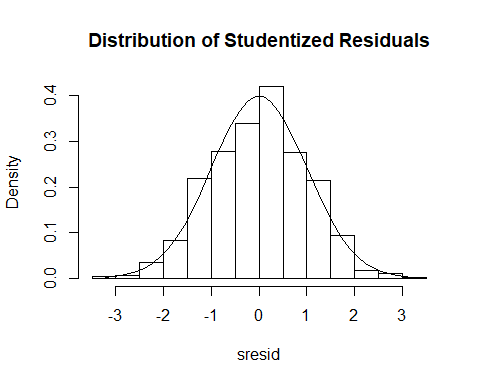
#### Normality of Residuals

# Normality of Residuals  
  
qqPlot(MLB, main="QQ Plot") #qq plot for studentized resid

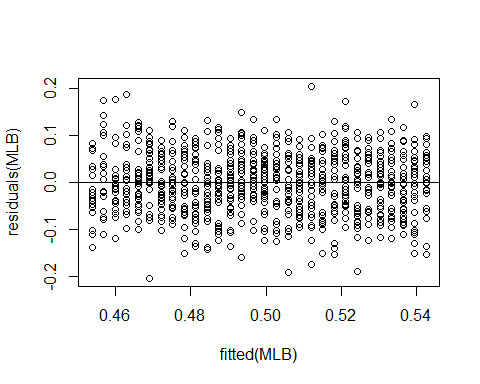


#### Distribution of studentized residuals

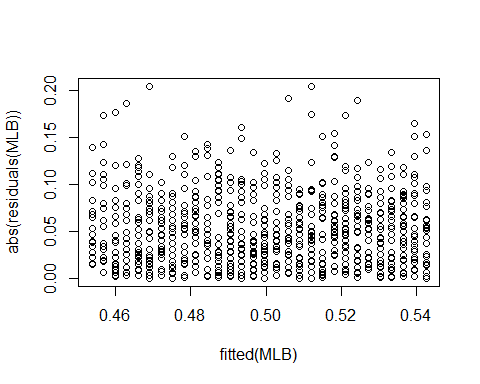
# distribution of studentized residuals  
  
sresid <- studres(MLB)   
hist(sresid, freq=FALSE,   
 main="Distribution of Studentized Residuals")  
xfit<-seq(min(sresid),max(sresid),length=40)   
yfit<-dnorm(xfit)   
lines(xfit, yfit)



# plot studentized residuals vs. fitted values   
plot(fitted(MLB), residuals(MLB))  
abline(h = 0)



plot(fitted(MLB), abs(residuals(MLB)))



summary(lm(abs(residuals(MLB))~fitted(MLB)))

##   
## Call:  
## lm(formula = abs(residuals(MLB)) ~ fitted(MLB))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.051472 -0.029895 -0.005111 0.022959 0.152936   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.046885 0.024313 1.928 0.0541 .  
## fitted(MLB) 0.009121 0.048562 0.188 0.8511   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.03763 on 916 degrees of freedom  
## Multiple R-squared: 3.851e-05, Adjusted R-squared: -0.001053   
## F-statistic: 0.03527 on 1 and 916 DF, p-value: 0.8511

After testing the residuals we can conclude that are data is an approximately normal distribution which is needed to fit are assumptions in our linear model.

#### Evaluation of homoscedasticity

# Evaluate homoscedasticity  
# non-constant error variance test  
ncvTest(MLB)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.03794686 Df = 1 p = 0.8455499

Since the P-value is .84 we can conclude that the standard deviations of the error terms are constant and do not depend on the x-value, which is an assumption needed for our linear model.

#### Global validation of linear model assumptions .

gvmodel <- gvlma(MLB)   
summary(gvmodel)

##   
## Call:  
## lm(formula = baseball$winpercent ~ baseball$payrank)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.204099 -0.046392 0.002225 0.044436 0.204081   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.5455626 0.0043006 126.86 <2e-16 \*\*\*  
## baseball$payrank -0.0030585 0.0002518 -12.15 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06378 on 916 degrees of freedom  
## Multiple R-squared: 0.1387, Adjusted R-squared: 0.1378   
## F-statistic: 147.6 on 1 and 916 DF, p-value: < 2.2e-16  
##   
##   
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS  
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:  
## Level of Significance = 0.05   
##   
## Call:  
## gvlma(x = MLB)   
##   
## Value p-value Decision  
## Global Stat 2.9495 0.5663 Assumptions acceptable.  
## Skewness 0.6193 0.4313 Assumptions acceptable.  
## Kurtosis 1.2533 0.2629 Assumptions acceptable.  
## Link Function 0.1632 0.6862 Assumptions acceptable.  
## Heteroscedasticity 0.9137 0.3391 Assumptions acceptable.

All assumptions to our linear model are acceptable.

## Conclusion

The plot of team payroll rank and winning % showed a clear negative linear relationship. Meaning that as the Teams drop lower in the ranks of team payroll, their winning % is expected to also decrease. After performing all the diagnostic tests, the model passed each one confirming that it is a good model fit. The summary of the model showed a significant relationship between payroll rank and winning % at an = .05 level, therefore, we can reject the null hypothesis and conclude, that there is a relationship between where a team ranks in team payroll and their winning %. According to our model for every spot a team drops in team payroll rank, their winning % is expected to decrease by 0.3%, and over a 162 game season that would be alomost 15 less wins expected for the 30th ranked payroll vs the 1st ranked Payroll.

### Discussion

As a client for an MLB team I would report to them that the amount of money spent in team payroll isn't crucial in increasing your winning % as long as you are spending **more** than your competitors. It matters more to rank ahead of other teams in Payroll regardless of dollar amount spent than it does to just throw money at players.