The Changing Criteria in Performance Based Salaries Across Eras of Major League Baseball

Billy Fryer under the mentorship of Dr. Eric Chi

Abstract

There is a consensus that the past 30 years of Major League Baseball can be split up into 3 major eras: The Free Agency Era (1977-1993), the Long Ball/Steroid Era (1994-2005), and the Post-Steroid Era (2006-present). Each different era defined the worth of non-pitchers to coaches and scouts differently. This worth has affected how managers have paid those players in that time period. Across the eras, the criteria used has changed. Using the Lahman Data Sets, multiple regression analysis was utilized in order to quantify which aspects of a player's performance affect his salary the most for each of these eras.

Libraries Used

```
# Libraries
library(Lahman)
library(ggplot2)
library(tidyverse)
library(coefplot)
library(scales)
library(randomcoloR)
```

Data and Cleaning

Lahman is a package in R that contains many baseball data sets. We will be using the Salary, Batting, Fielding and People datasets.

The unifying variable among data sets is the playerID variable. Every player that has played in MLB has a unique playerID.

For more info about the Lahman Package: http://www.seanlahman.com/files/database/readme2017.txt

Salary Data

There is a consensus that the past 30 years of Major League Baseball can been split up into 3 major eras: The Free Agency Era (1977-1993), the Long Ball/Steroid Era (1994-2005), and the Post-Steroid Era (2006-present). Each different era defined the worth of non-pitchers to coaches and scouts differently. Because of this, the criteria used to evaluate which players deserve higher salaries has changed accordingly. Using data from the Lahman data sets, we used multiple regression analysis to quantify which aspects of a player's performance affects his salary the most for each of these eras between the years 1985 and 2016.

Batting Data

The Batting data set contains hitting statistics statistics from the 1871 season to the 2019 season. Due to the limitations in the Salary data set, we only use data from 1985 to 2016. I did not perform any additional manipulation to the Batting Data Set.

Fielding Data

The Fielding data set contains variables relating to the Fielding productivity of each every individual for seasons between 1871 and 2019. We limited our observations between 1985 and 2016 due because those were the only years that we had Salary data available to us. I did have to perform some manipulation on the Fielding data set.

Originally, the Fielding data set has an individual row for each distinct POS (position) that a player played in a season. Using the following website as a guide, we collapse position into only 1 observation per player per season. For this reason, position is not used in our regression model. I then summed all quantitative variables for each player for each season into one observation as well. The resulting data set had one observation per player per season.

I used these website as a guide:

Making Multiple rows into 1 (for POS variable):

https://markhneedham.com/blog/2015/06/27/r-dplyr-squashing-multiple-rows-per-group-into-one/

I learned about merging Lahman data sets from the following website:

http://lahman.r-forge.r-project.org/doc/Salaries.html

```
New.Fielding <- Fielding %>% group_by(playerID, yearID, teamID) %>%
  # summarize
  summarize(League = paste(lgID),
            # condencing multiple positions down into 1 row
            POS = paste(POS, collapse = ", "),
            #Summing all other varables for where a person is listed on multiple rows
            G = sum(G),
            GS = sum(GS),
            InnOuts = sum(InnOuts),
            PO = sum(PO),
            Assists = sum(A),
            Errors = sum(E),
            DP = sum(DP),
            PB = sum(PB),
            WP = sum(WP),
            SB = sum(SB),
            CS = sum(CS),
            ZR = sum(ZR)
            ) %>%
  # Only keep Distinct rows
  distinct()
```

People Data

The People data set gives biographical information about each player. The only information I used from this data set are nameFirst and nameLast which I then combined together to make a singular variable containing a player's first and last names.

```
bio <- People %>% select(c(playerID, nameFirst, nameLast))

# Put Name into 1 Variable
bio$name <- paste(bio$nameFirst, bio$nameLast)

# Delete variables other than playerID and the new name variable
bio <- bio %>% select(c(playerID, name))
```

Merging Data Sets

I then merged all of the previously mentioned datasets using the playerID and yearID variables. These variables help match the proper data to the proper player and the season from which the data was collected.

```
# Combine Salaries data with Batting data
stats <- merge(Salaries,
              Batting,
              by = c("playerID", "yearID", "teamID"),
              all.x = TRUE) %>%
  # Drop nas
  drop_na() %>%
  # Delete Duplicate Variables
  select(-c("lgID.y")) %>%
  # Rename lg.ID.x to lgID
  rename(lgID = lgID.x)


# Merge again with Fielding Data
stats <- merge(x = stats,
               y = New.Fielding,
               by = c("playerID", "yearID", "teamID"),
               all.x = TRUE) %>%
  # Choose SB and CS from Hitting Data not Fielding
  select(-c(SB.y, CS.y)) %>%
  # Changes names: CS.x to CS and SB.x to SB
  rename(SB = SB.x, CS = CS.x)

# Merge with Bio data
stats <- merge(stats,
               bio,
               by  = "playerID",
               all.x = TRUE)
```

It is at this point that we filter out Pitchers from our Data Set. Pitchers are evaluated on a different criteria than the rest of the position players (on their ability to pitch rather than their hitting and fielding skills).

```
# Take out Pitchers
stats <- subset(stats, POS != "P") %>%
  # Take out a few variables (see Future Work for more Information)
    select(-c(PB, WP, ZR, stint, G.x, G.y, SH, X3B)) %>%
  # Reorder variables
  select(playerID, name, teamID, League, yearID, POS, InnOuts, GS, everything())

# Make year a factor variable
stats$yearID <- factor(stats$yearID)
```

```r
# Intercept column of all 1s for regression
stats$intercept <- 1
```

I then subsetted the data by year.

```r
# Subsetting large data set by year
stats85 <- stats %>% subset(yearID == 1985)
stats86 <- stats %>% subset(yearID == 1986)
stats87 <- stats %>% subset(yearID == 1987)
stats88 <- stats %>% subset(yearID == 1988)
stats89 <- stats %>% subset(yearID == 1989)
stats90 <- stats %>% subset(yearID == 1990)
stats91 <- stats %>% subset(yearID == 1991)
stats92 <- stats %>% subset(yearID == 1992)
stats93 <- stats %>% subset(yearID == 1993)
stats94 <- stats %>% subset(yearID == 1994)
stats95 <- stats %>% subset(yearID == 1995)
stats96 <- stats %>% subset(yearID == 1996)
stats97 <- stats %>% subset(yearID == 1997)
stats98 <- stats %>% subset(yearID == 1998)
stats99 <- stats %>% subset(yearID == 1999)
stats00 <- stats %>% subset(yearID == 2000)
stats01 <- stats %>% subset(yearID == 2001)
stats02 <- stats %>% subset(yearID == 2002)
stats03 <- stats %>% subset(yearID == 2003)
stats04 <- stats %>% subset(yearID == 2004)
stats05 <- stats %>% subset(yearID == 2005)
stats06 <- stats %>% subset(yearID == 2006)
stats07 <- stats %>% subset(yearID == 2007)
stats08 <- stats %>% subset(yearID == 2008)
stats09 <- stats %>% subset(yearID == 2009)
stats10 <- stats %>% subset(yearID == 2010)
stats11 <- stats %>% subset(yearID == 2011)
stats12 <- stats %>% subset(yearID == 2012)
stats13 <- stats %>% subset(yearID == 2013)
stats14 <- stats %>% subset(yearID == 2014)
stats15 <- stats %>% subset(yearID == 2015)
stats16 <- stats %>% subset(yearID == 2016)


# Creates a list of all datasets subsetted by year
dsets <- list(stats85, stats86, stats87, stats88, stats89, stats90, stats91, stats92,
              stats93, stats94, stats95, stats96, stats97, stats98, stats99,
              stats00, stats01, stats02, stats03, stats04, stats05, stats06,
              stats07, stats08, stats09, stats10, stats11, stats12, stats13,
              stats14, stats15, stats16)
```

Modeling

The Following Variables Are Used For Modeling:

- GS - Games Started

Batting Statistics

- AB - At Bats
- R - Runs Scored
- H - Hits
- X2B - Doubles
- HR - Home Runs Scored
- RBI - Runs Batted In
- SB - Stolen Bases
- CS - Caught Stealing
- BB - Walks
- SO - Strike Outs
- IBB - Intentional Walks
- HBP - Hit By Pitch
- SF - Sacrifice Flies
- GIDP - Number of Times Grounded into a Double Play

Fielding Statistics

- InnOuts - Time played in the field expressed as outs
- PO - Put Outs
- Assists - Fielding Assists
- Errors - Errors
- DP - Double Plays

Note: All of the variables mentioned take on discrete positive integers

```r
# First Order Linear Model as a function
fit1_model <-function(data){
  lm(salary~intercept +GS+AB+R+H+X2B+HR+RBI+SB+CS+BB+SO+IBB+HBP+SF+GIDP+InnOuts+PO+Assists+Errors+DP, da
}

# Apply model to all data sets
order_1_models <- lapply(dsets, fit1_model)
```

I then made a data frame of all the coefficients of the models created.

```
# make dataframe of coefficients
coef.df <- data.frame(rbind(
            order_1_models[[1]]$coefficients,
            order_1_models[[2]]$coefficients,
            order_1_models[[3]]$coefficients,
            order_1_models[[4]]$coefficients,
            order_1_models[[5]]$coefficients,
            order_1_models[[6]]$coefficients,
            order_1_models[[7]]$coefficients,
            order_1_models[[8]]$coefficients,
            order_1_models[[9]]$coefficients,
            order_1_models[[10]]$coefficients,
            order_1_models[[11]]$coefficients,
            order_1_models[[12]]$coefficients,
            order_1_models[[13]]$coefficients,
            order_1_models[[14]]$coefficients,
            order_1_models[[15]]$coefficients,
            order_1_models[[16]]$coefficients,
            order_1_models[[17]]$coefficients,
            order_1_models[[18]]$coefficients,
            order_1_models[[19]]$coefficients,
            order_1_models[[20]]$coefficients,
            order_1_models[[21]]$coefficients,
            order_1_models[[22]]$coefficients,
            order_1_models[[23]]$coefficients,
            order_1_models[[24]]$coefficients,
            order_1_models[[25]]$coefficients,
            order_1_models[[26]]$coefficients,
            order_1_models[[27]]$coefficients,
            order_1_models[[28]]$coefficients,
            order_1_models[[29]]$coefficients,
            order_1_models[[30]]$coefficients,
            order_1_models[[31]]$coefficients,
            order_1_models[[32]]$coefficients))

# Add year column to data frame
coef.df <- coef.df  %>%
  mutate(year = c(1985:2016)) %>%
  select(year, everything())
```

Visualizations

Coefficient Bump Plot Data Manipulation

To find out which variables in the model were most important, I took the absolute value of the coefficients for the yearly regression models. I then ranked each variable between 1-20 with a ranking of 1 representing the highest magnitude and highest effect on the model. For this project, I created a bump plot to display the importance of different variables in the model over time.

The first code chunk is pure data manipulation while all code for the actual bump plot is in the second code chunk. The code is adapted from https://www.statology.org/bump-chart-in-r-using-ggplot2/

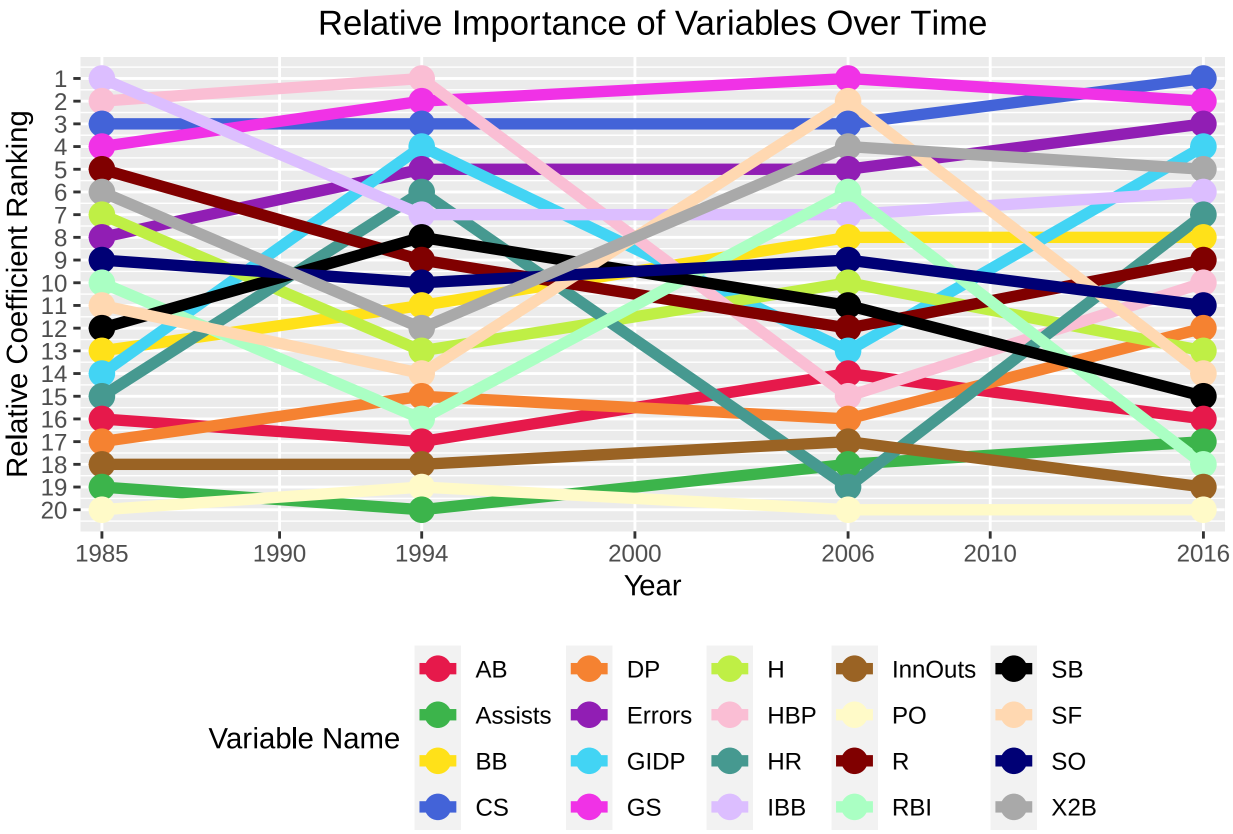
```r
# Data Manipulation

# Abs to get the magnitude of everything in coefficient data frame
rankmat <- data.frame(abs(coef.df)) %>%
  # drop intercept from bump plot
  select(-c(intercept))

rankmat$year <- as.numeric(rankmat$year)

rankmat <- rankmat %>%
  # Only certain values of year
  filter(year %in% c(1985, 1994, 2006, 2016)) %>%
  # group by year
  group_by(year) %>%
  # Make data longer (keep years the same)
  pivot_longer(cols = -c(year)) %>%
  # Get rid of X Intercept
  filter(name != "X.Intercept.")  %>%
  # Arrange
  arrange(year, desc(value)) %>%
  # Assign rank to each variable
  mutate(rank = row_number()) %>%
  # Ungroup
  ungroup()
```

```r
# Graph Code
set.seed(42) # For consistent coloring of variables
ggplot(rankmat, aes(x = year, y = rank, group = name)) +
scale_color_manual(values =c('#e6194B', '#3cb44b', '#ffe119', '#4363d8',
'#f58231', '#911eb4', '#42d4f4', '#f032e6', '#bfef45', '#fabed4', '#469990',
'#dcbeff', '#9A6324', '#fffac8', '#800000', '#aaffc3', '#000000', '#ffd8b1',
'#000075', '#a9a9a9')) +
  geom_line(aes(color = name), size = 2) +
  geom_point(aes(color = name), size = 4) +
  scale_y_reverse(breaks = 1:max(rankmat$rank)) +
  labs(colour = "Variable Name") +
  ylab("Relative Coefficient Ranking") +
  xlab("Year") +
  theme(legend.position = "bottom") +
  scale_x_discrete(name = "Year", limits = c(1985, 1990, 1994, 2000, 2006, 2010, 2016)) +
  ggtitle("Relative Importance of Variables Over Time") +
  theme(plot.title = element_text(hjust = 0.5))
```



Each variable is represented by a different color. The value on the vertical axis is the variable's relative ranking over time for the years listed. The years where rankings are directly present (the circles) were selected because they signify the beginning of a new era.

A major finding was that stats implying good defensive skills, such as assists, tended to matter less in all eras than offensive stats. However, if a player had several errors in the field, this severely impacted his salary.

Another important finding can be shown by following the BB (or Walks) line. The relative importance of walks rose 5 spots from the beginning of our analysis to the end of it. While walks are seemingly random, Billy Beane and the Oakland A's during the Moneyball Movement around 2004 used on base percentage (OBP) to evaluate players. A crucial factor in OBP is walks, hence the increase.

A third and final finding that I would like to mention is the sharp increase in the valuation for Home Runs (HR) between 1985 and 1994. As players such as Barry Bonds began hitting more homers, their salaries increased dramatically. The drop off in importance in 2006 could be either to the steroid scandal being publicized or this could be an anomaly in our data.

Parameter Weight Over Time for Every Variable

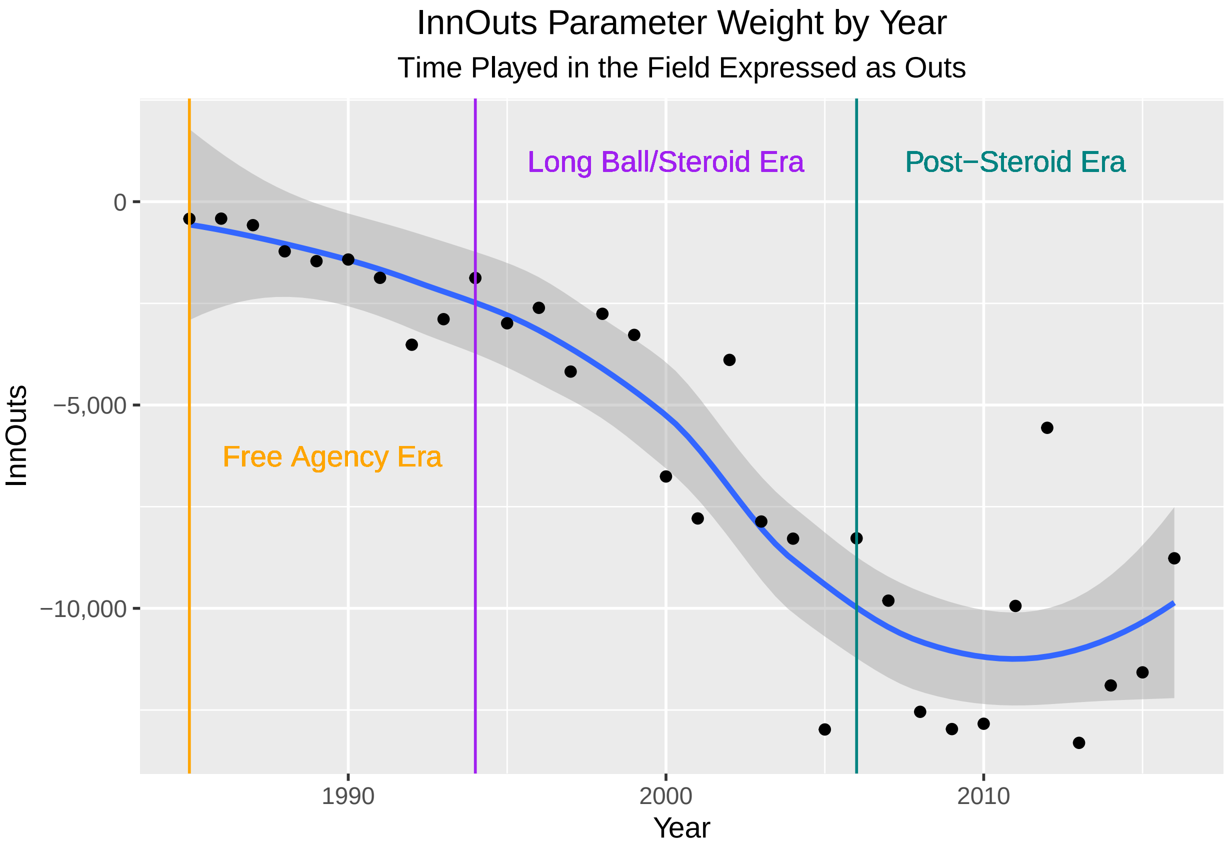
The following graphs display how the valuation of player statistics in relation to the salary change over time. This coefficient represents how much a change in that statistic would affect the player's salary, given that all other variables are held constant. The intent of these plots is to generally describe the trend in each era of how that statistic affected a players salary.

The curving blue line was created as a smoother and roughly describes the trend in the data. The vertical lines on these plots mark the beginning of a new Era in baseball and are labeled appropriately. Each data point is the partial slope estimation for the variable given in the title of the graph for a particular year (on the horizontal axis).

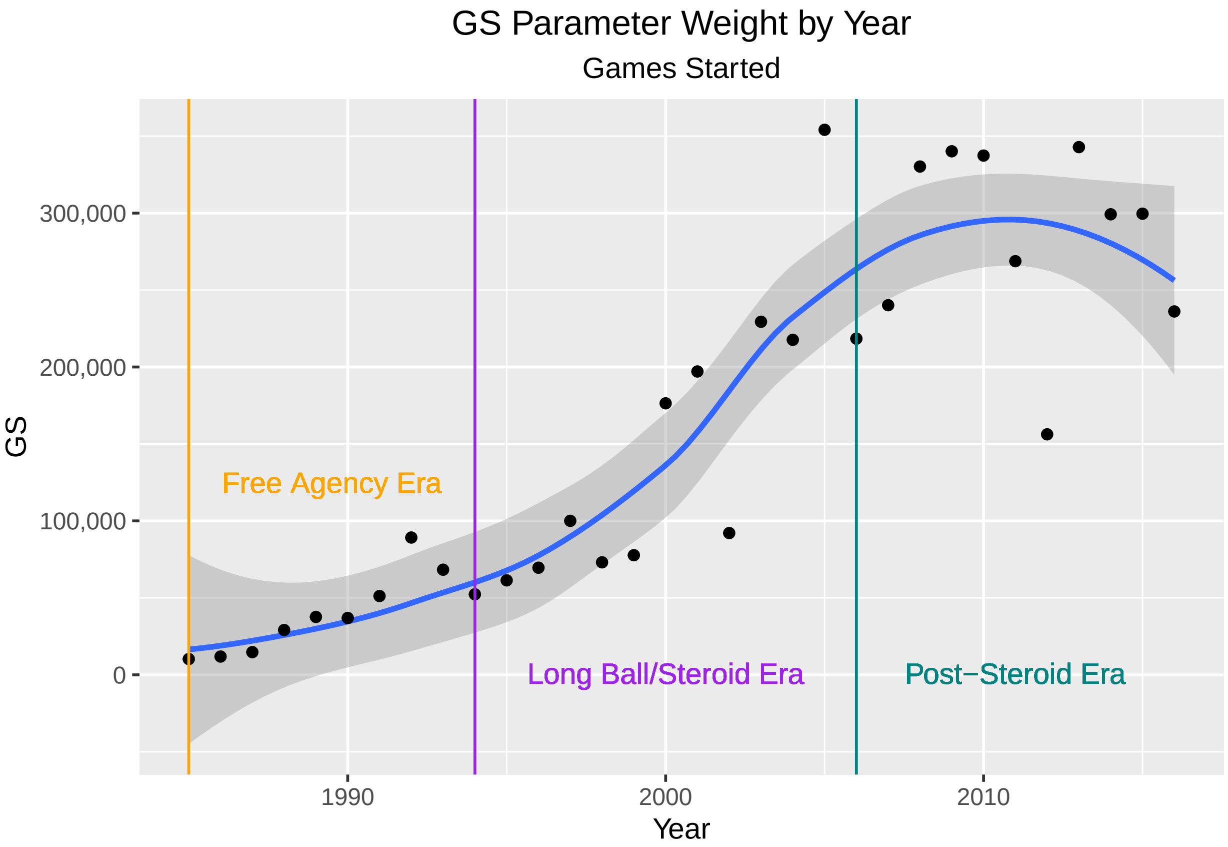
For our purposes, the beginning of the "Free Agency Era" is before 1985, however we do not have enough data from before 1985 to mark the true beginning of the era. As previously mentioned, the beginning of the "Long Ball/Steroid Era" was 1994 and the beginning of the "Post-Steroid Era" was 2006. Also as a reminder, the last year represented in our data set is 2016.

The code for the first plot is shown. The rest are similar in nature.

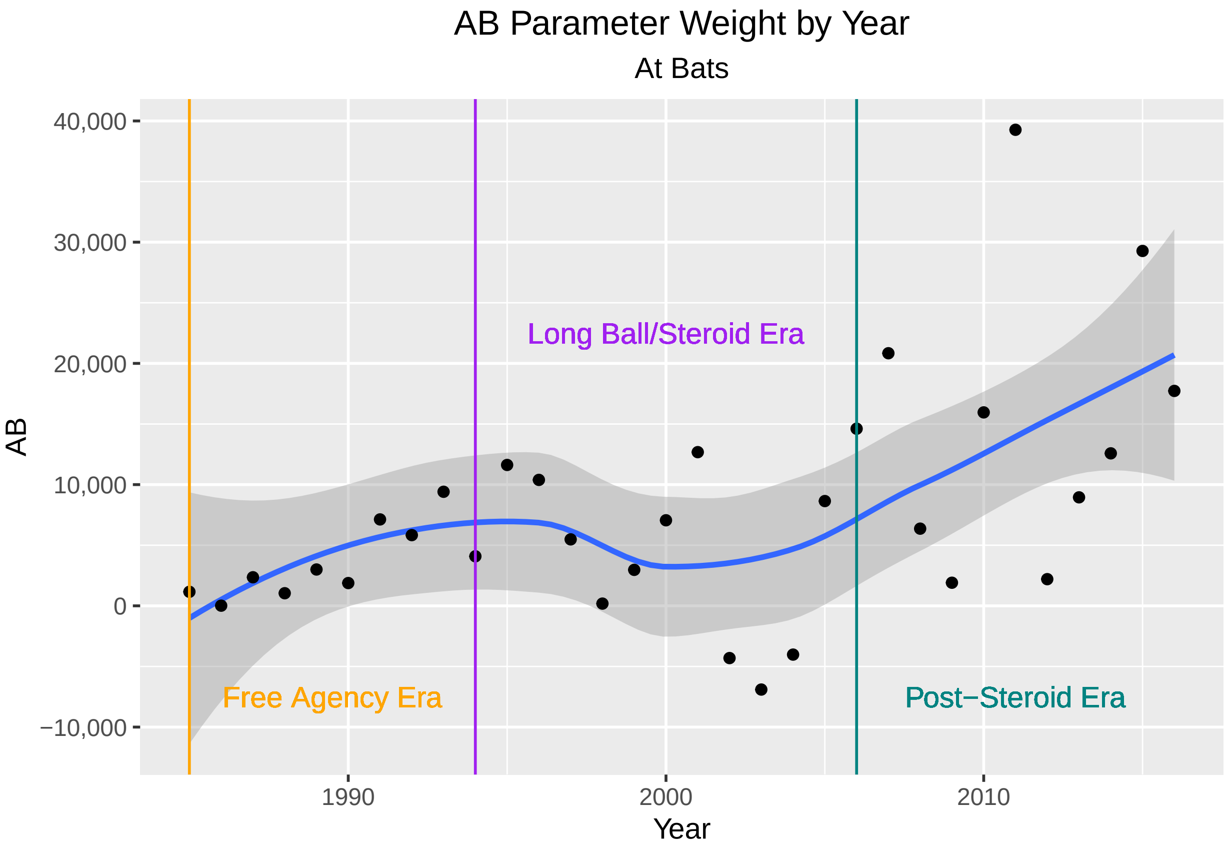
```r
# Example of how InnOuts parameter has changed over years
ggplot(coef.df, aes(x = year, group = 1, y = InnOuts)) +
  geom_smooth() +
  geom_point() +
  ggtitle("InnOuts Parameter Weight by Year", subtitle = "Time Played in the Field Expressed as Outs") +
  scale_y_continuous(labels = scales::comma) +
  # Green Line and Text
  geom_vline(xintercept = 2006, color = "#008381") +
  geom_text(x = 2011, y = 1000,label ="Post-Steroid Era", color = "#008381") +
  # Purple Line and Text
  geom_vline(xintercept = 1994, color = "purple") +
  geom_text(x = 2000, y = 1000,label ="Long Ball/Steroid Era", color = "purple") +
  # Orange Line and Text
  geom_vline(xintercept = 1985, color = "orange") +
  geom_text(x = 1989.5, y = -6250,label ="Free Agency Era", color = "orange") +
  xlab("Year")+
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```



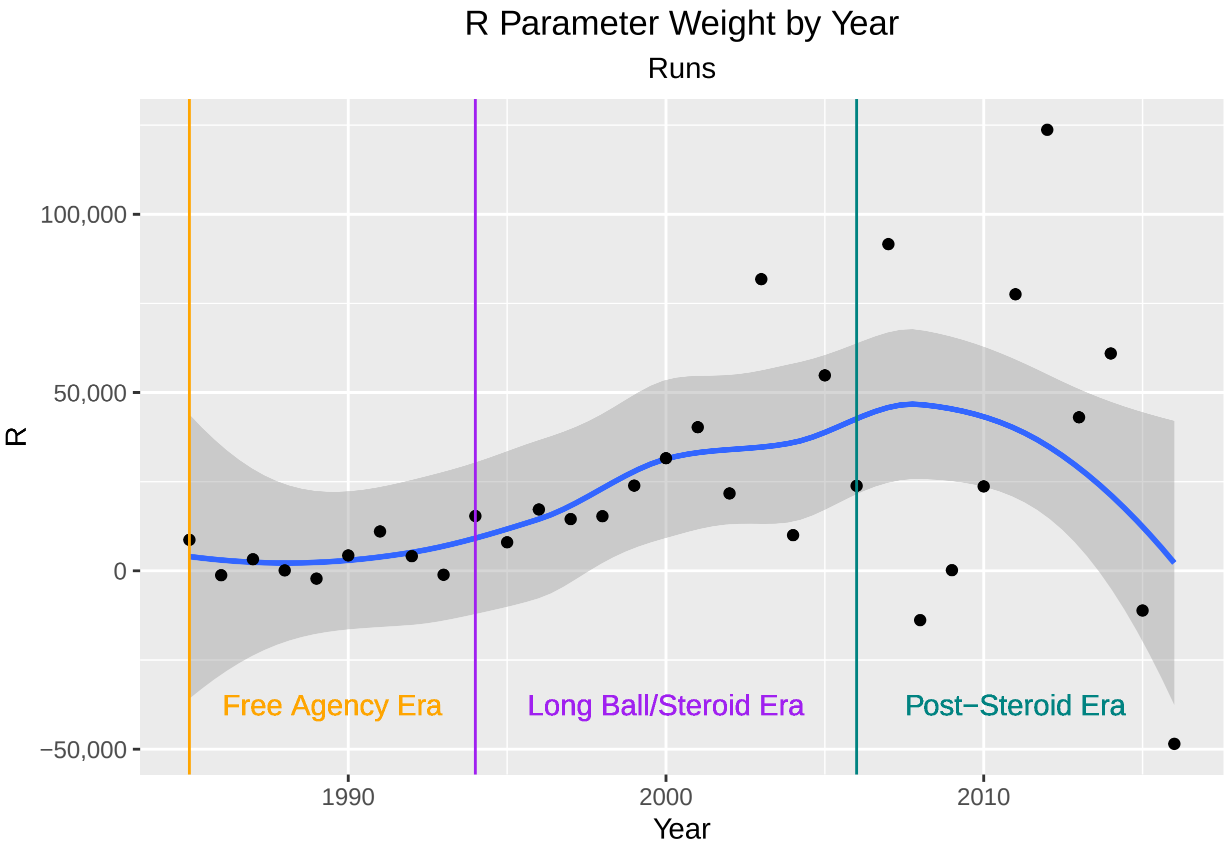
The relative importance of the time played in the field expressed as outs decreased in importance during the Long Ball Era. This is likely because this time period marked the beginning of players not playing every single game. Before this time period, players were expected to play every game of the season. However, more recently players are given a particular day of the week off for health reasons and thus are playing less over the full season.



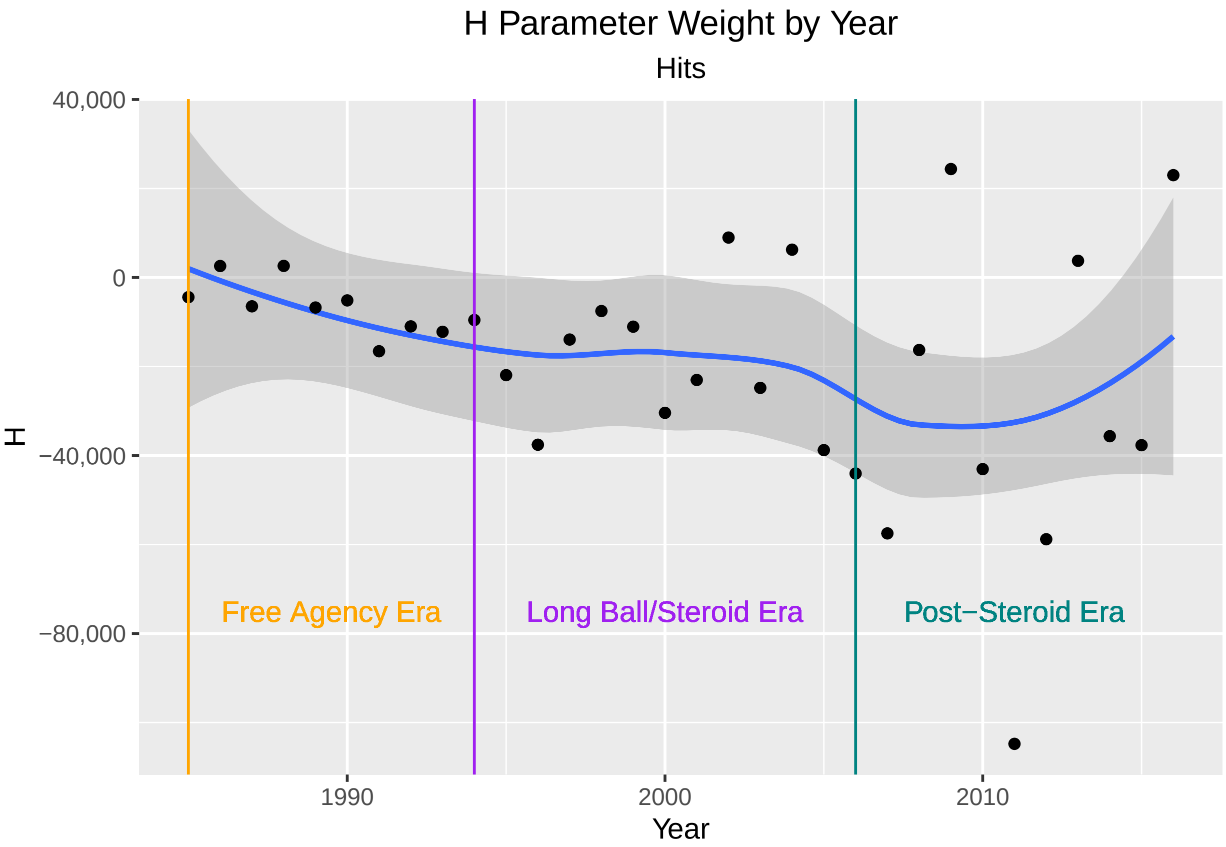
In contrast to InnOuts, GS (Games Started) has very strong positive change over time. As previously mentioned, players have recently began to have play in fewer games in order to prevent injury. This made starting games more important for determining who is a better player. The idea being that if you start the game, you are the best at your position on the team. The scale of this graph is unusually large and needs to be investigated further. In the Post-Steroid Era, the data points seem to level off rather than peak as the smoother suggests.



The At Bats parameter stayed pretty constant throughout the Free Agency and Long Ball Eras. However, in the Post-Steroid Era it seems to be becoming increasingly important. It is important to note that the parameters in the Post-Steroid Era vary greatly per year, so we cannot make a strong conclusion about the relation between At Bats and Salary in the Post-Steroid Era.

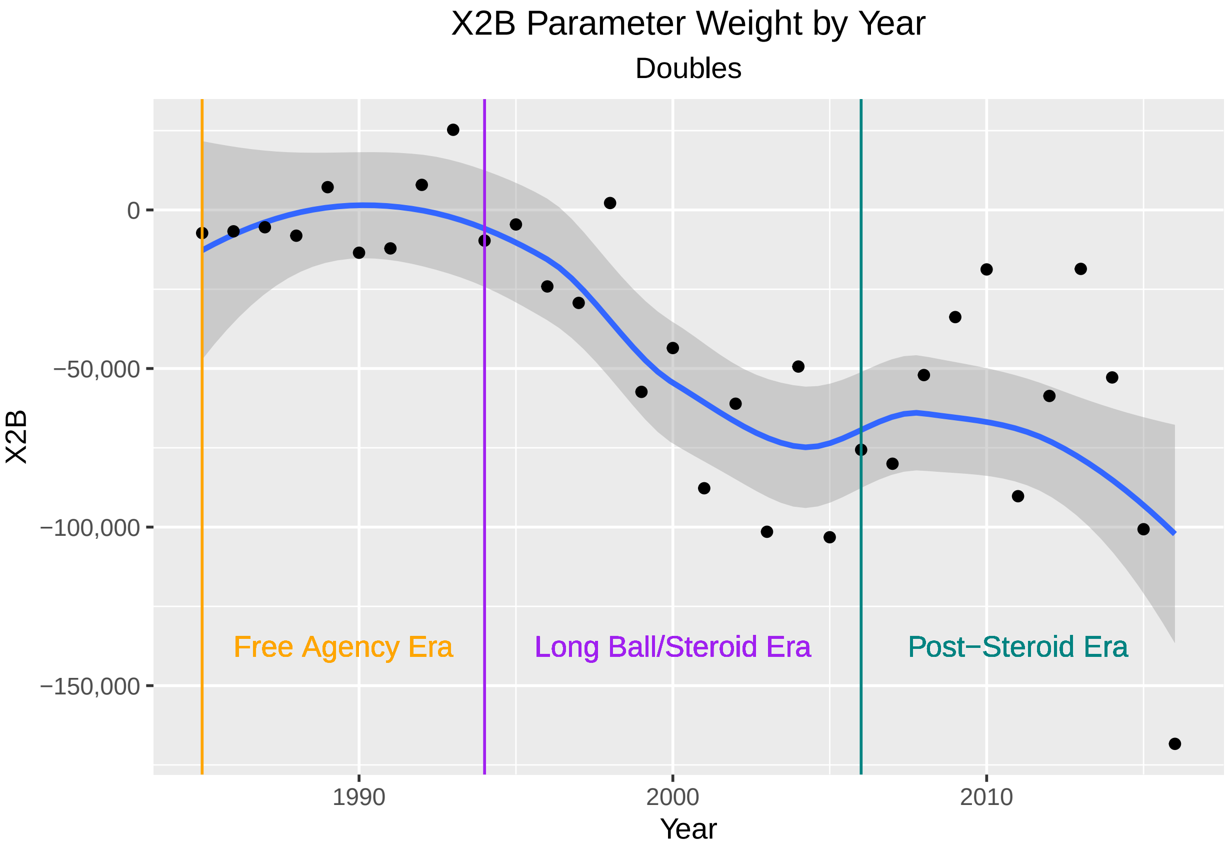


The Runs parameter seemed to be of increasing importance during the Free Agency and Long Ball Eras. However, similar to the At Bats parameter, due to massive variance, we cannot make a strong conclusion about the correlation between Runs and Salary Era Post-Steroid Era either.

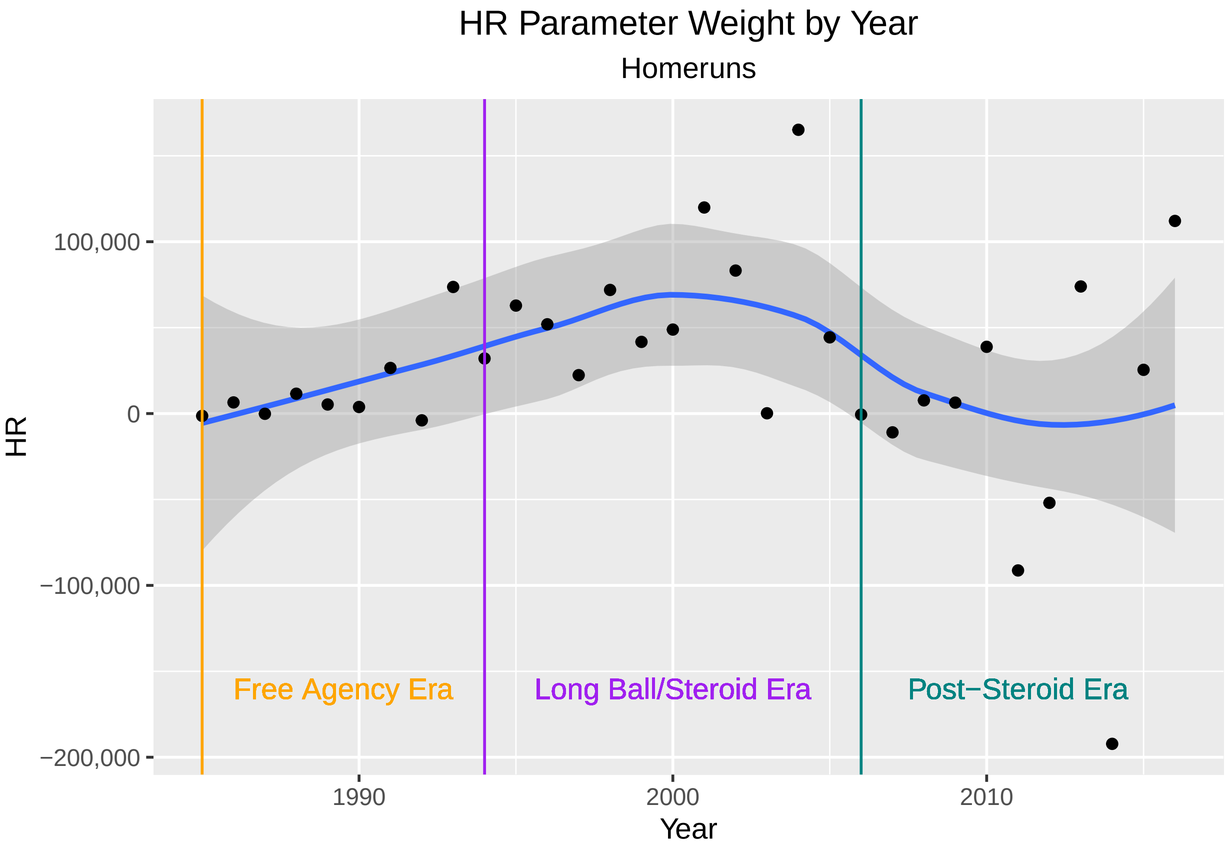


The Hits parameter slightly decreased during the Free Agency and Long Ball eras but again conclusions cannot be made about the Post-Steroid Era due to high variance. There is a strong outlier point in 2011. This would need to be investigated further in the future.

Another interesting topic to explore would be why this parameter is negative. Typically, having more hits is seen as a positive attribute.

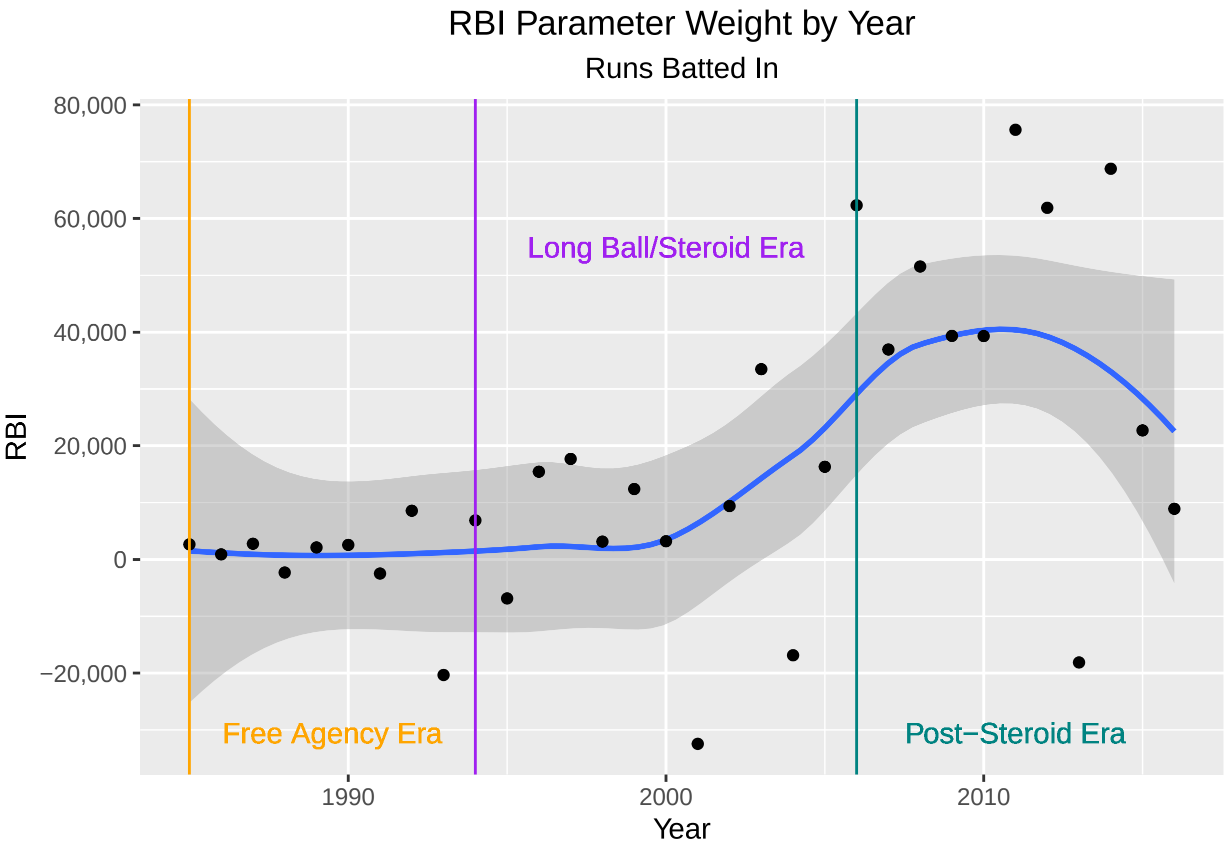


The Doubles parameter seems to be decreasing across all Eras. Again, similar to the Hits parameter, this is odd and would need to be explored more in future work.



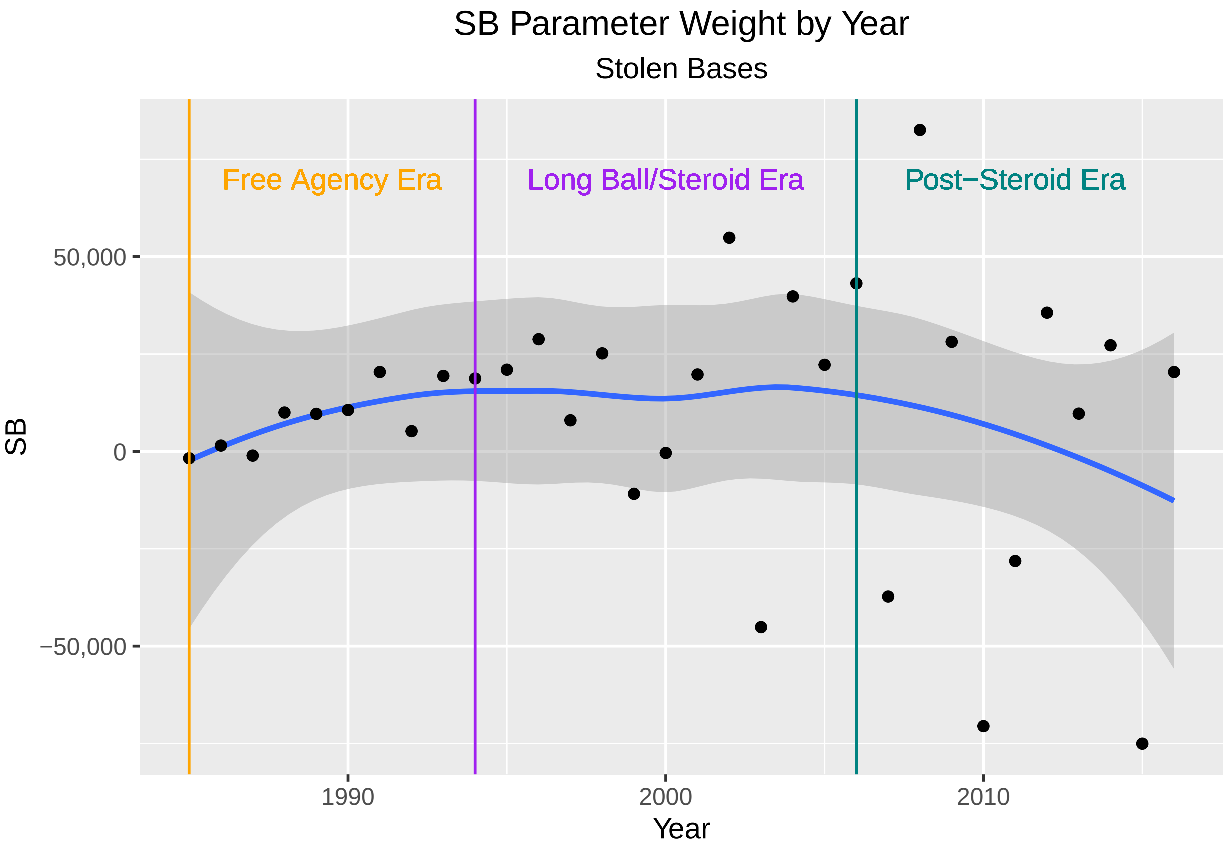
As expected, the Homerun parameter was strongly positive suggesting that people who hit more homeruns have higher salaries. The decline at the end of the Long Ball Era could explained by the steroid era bust which made teams more cautionary towards homerun hitters.

In the Post-Steroid Era, homeruns still seem to have value. Without the outlier of the 2014 season, the smoother would have predicted that the homerun parameter would have returned to the value of around 1995 before the steroid bust. If we had data from the 2017-2019 seasons, this trend probably would have been more evident.

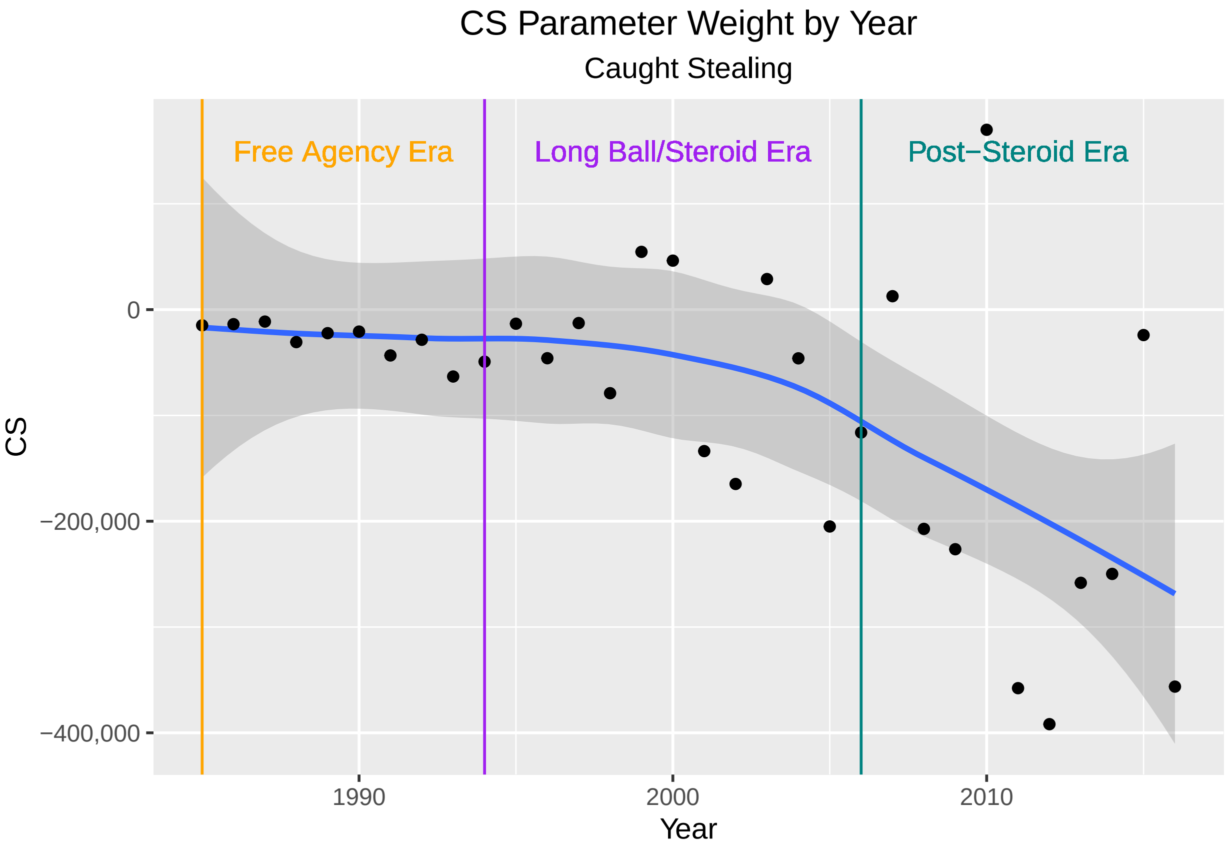


RBIs became increasingly important during the early 2000s. Before, during the Free Agency Era, RBIs were not important at all with slope parameters close to 0. The parameters peaked between 2010 and 2015.

In recent developments, the importance of RBIs has began to decrease. This is because of the advancements in baseball analytics. Analysts have found that RBIs are strongly related to where someone is in hits in the batting order. This means a person who has better hitters in front of them are more likely to have more RBIs than someone with poorer hitters batting before them.

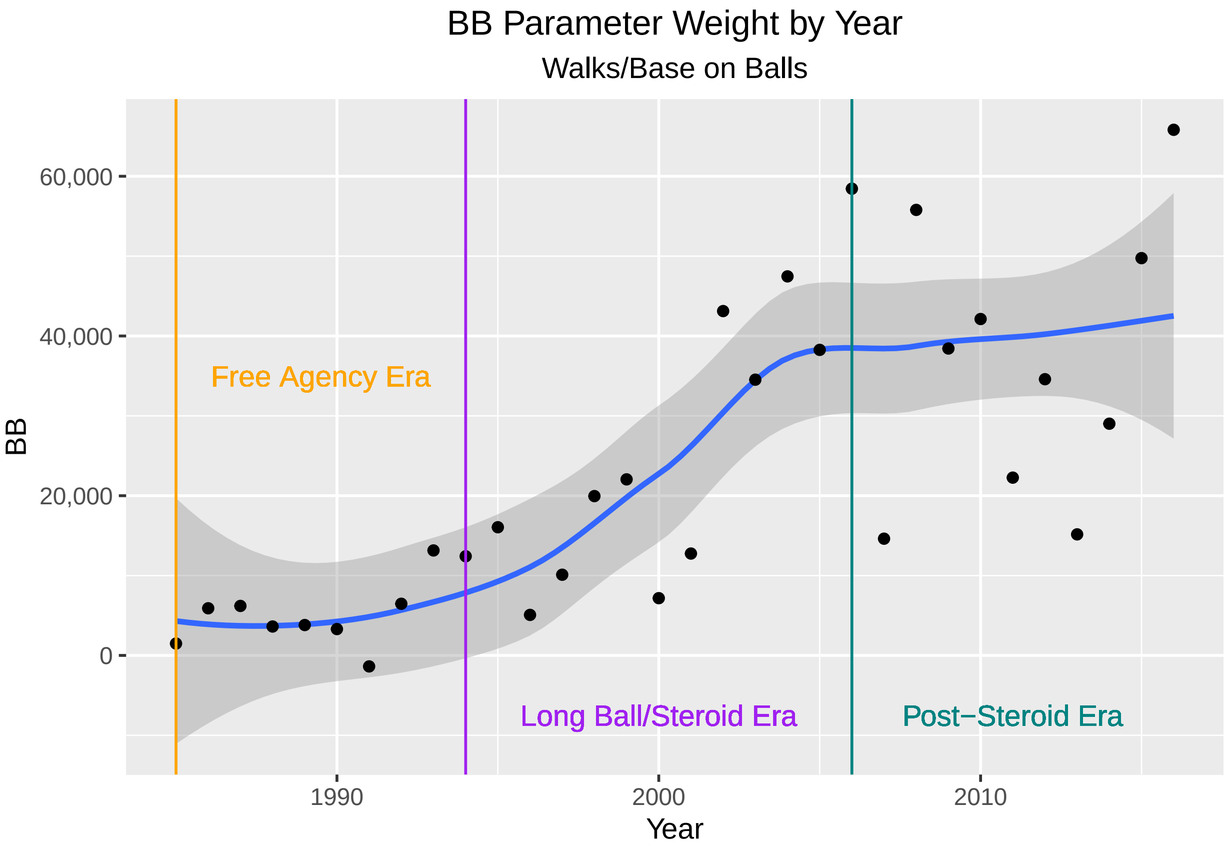


The importance of stealing bases rose slightly during the Free Agency and Long Ball eras. However, during the Post Steroid era, the variance is too large to draw appropriate conclusions.



The importance of not getting caught stealing is of high priority in recent years. This is shown in the graph due to the dramatic downward smoother function and because almost all parameters are less than zero.

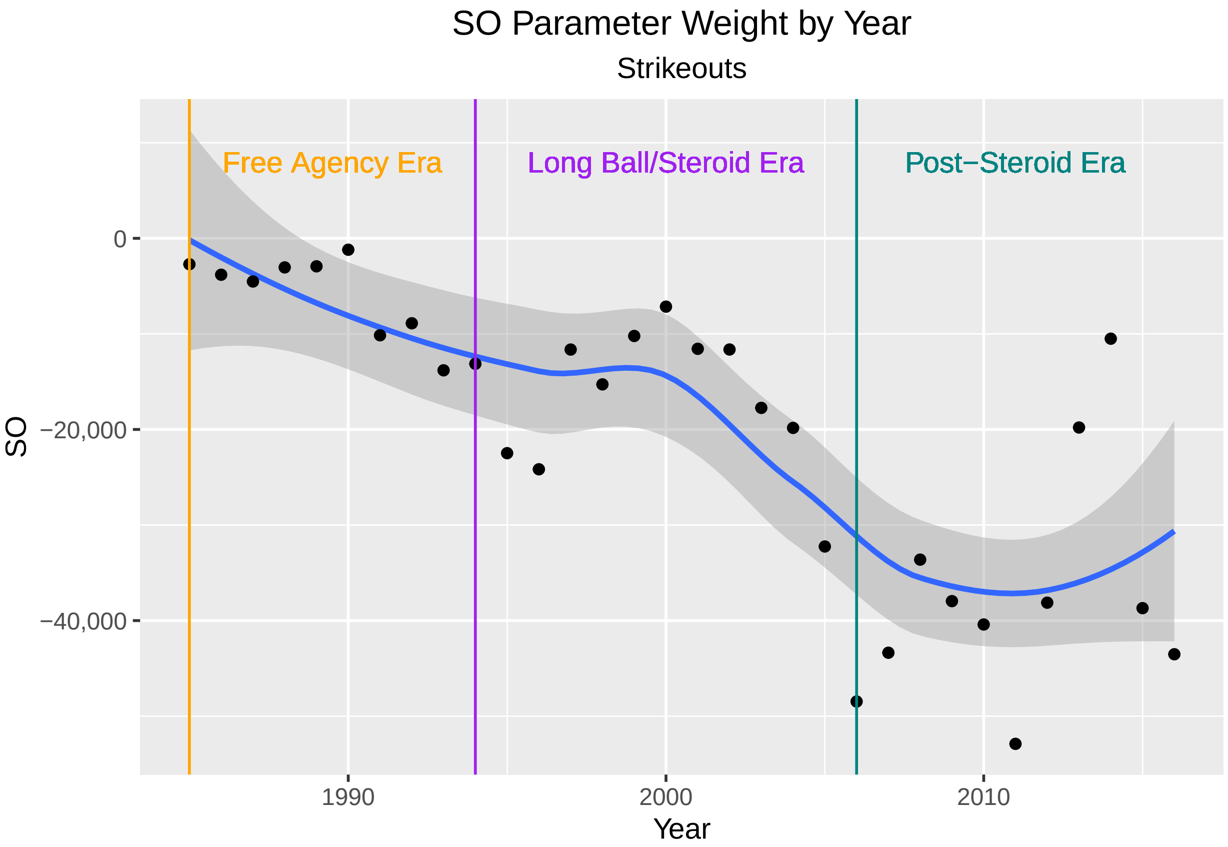
Similar to the Doubles and Games Started parameters, the scale for the Caught Stealing Parameter seems out of proportion. This would need to be explored more carefully in the future.



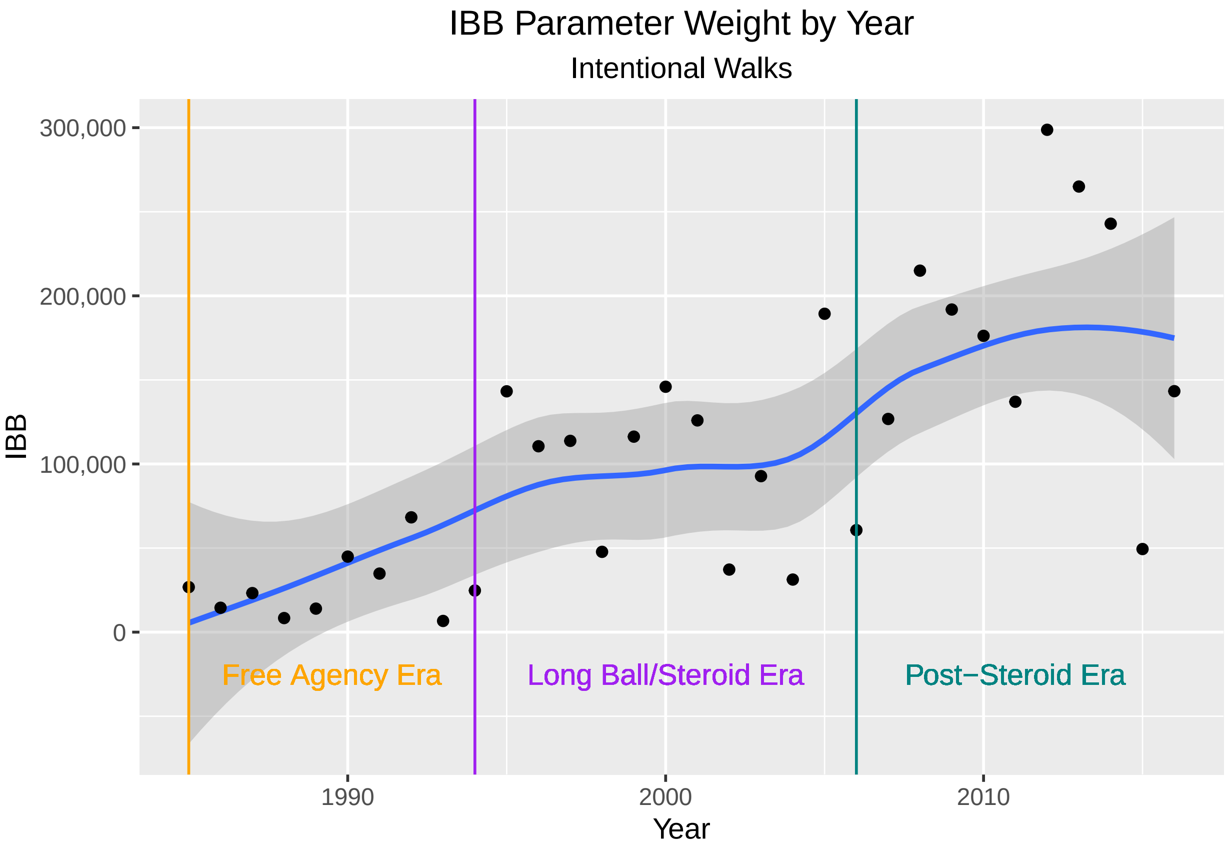
The Walk parameter has a very interesting baseball history. During the Free Agency era, base on balls were relatively unimportant. However, this began to have strong upward trend in the early 2000s and corresponds perfectly with Billy Beane and the Moneyball movement.

Billy Beane and the Oakland Athletics theorized that getting on base by any means necessary was one of the most underrated aspects in baseball. Their thought process was that one cannot score without first getting on base. Beane contracted several players who walked often and due to this strategy was very succesful in the 2004 campaign.

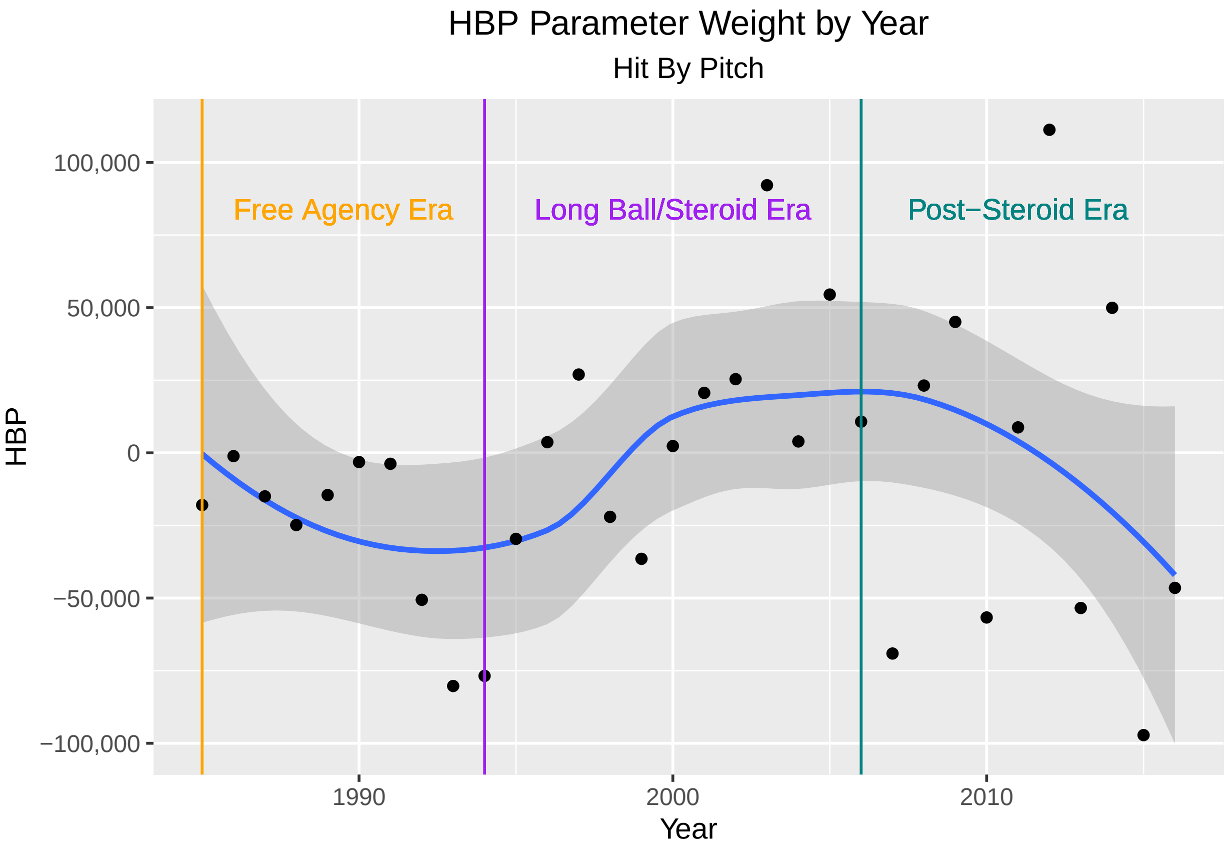
In the Post-Steroid Era, everyone knows about the Moneyball Movement which explains why the slope parameter for Walks is always positive. However, there is large variance between years as to how important Walks are.



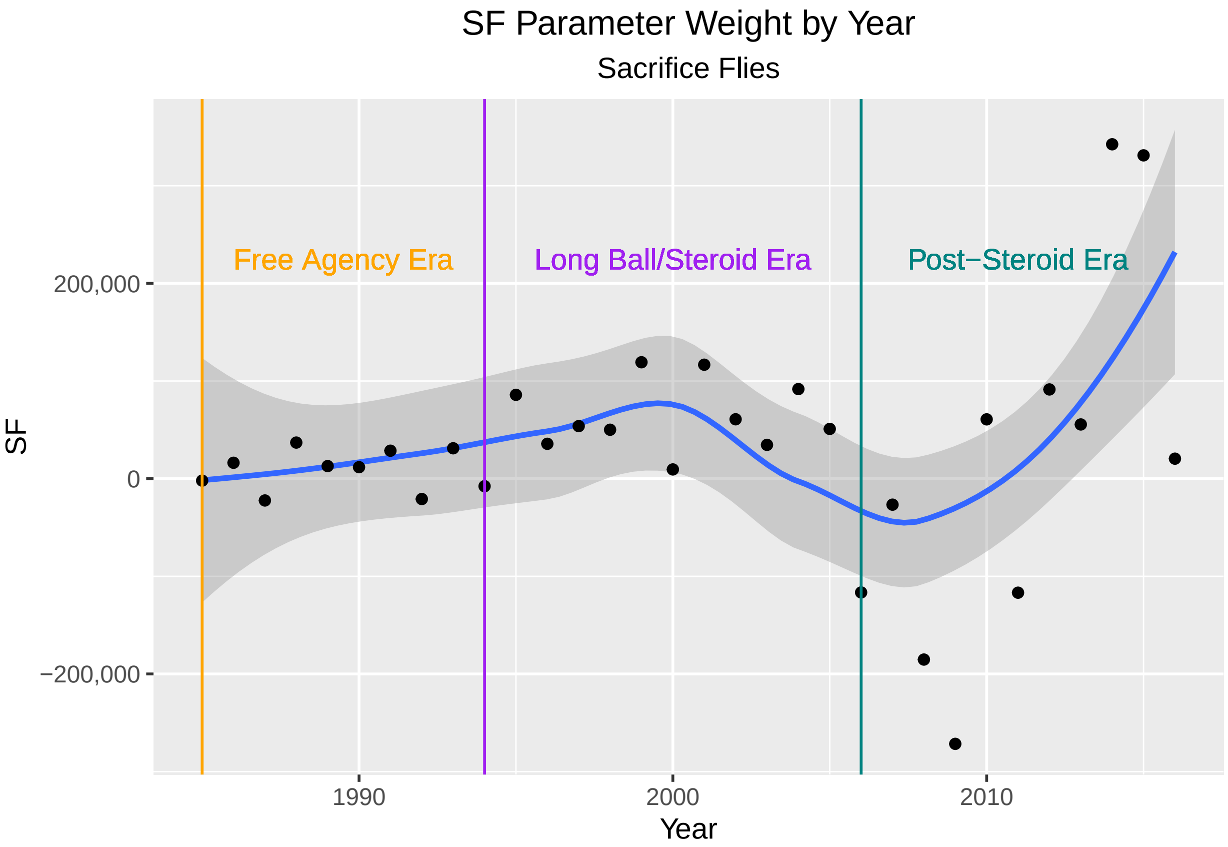
During the Free Agency and the beginning of the Long Ball Era, managers understood that power hitters often struck out. Because this is a sacrifice they were willing to make, the slope parameter for strikeouts was relatively close to zero. However, after the steroid scandal came to light, salaries became more strongly affected by poor strikeout rates. This has not yet recovered in the Post-Steroid era.



Similar to the Walks parameter, the Intentional Walks parameter has seen massive gains in importance in recent years due to Moneyball philosophies. However, this parameter also seems out of scale on the vertical axis and would need to be further explored in the future



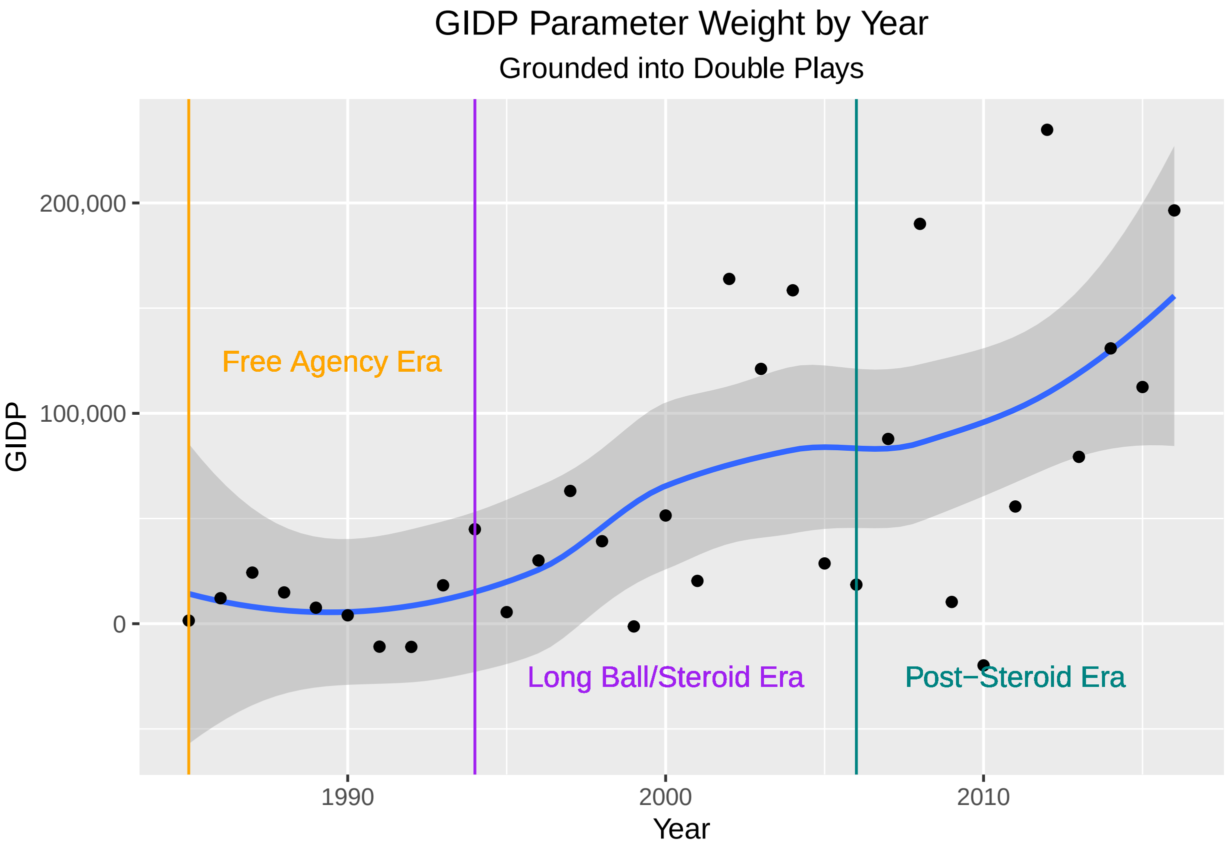
The Hit By Pitch weight seems to vary almost randomly around zero. The smoother attempts to fit a trend, but it rather unsuccessful. This means that being frequently hit by pitches does not really affect a player's salary.



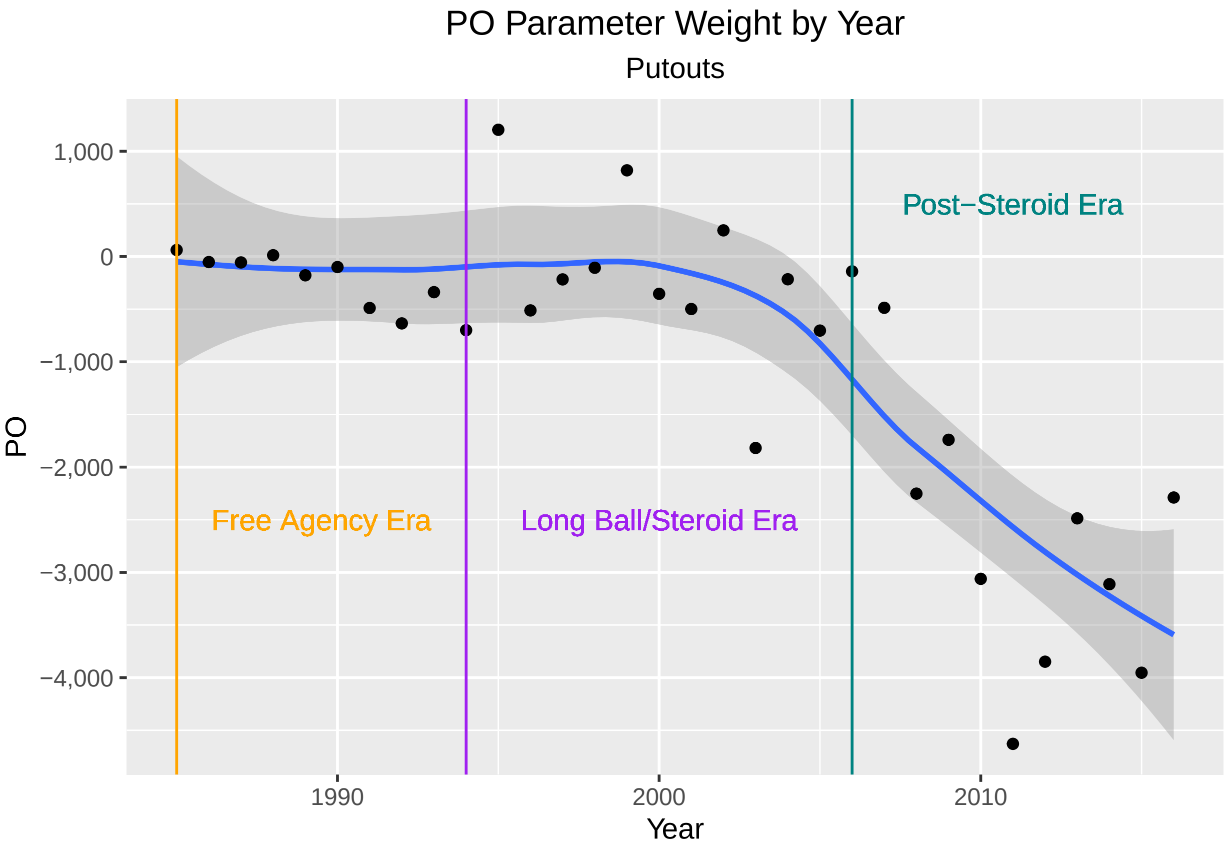
Sacrifice Flies increased in importance during the Free Agency and Long Ball Eras, but has recently taken off into very high importance. This can be attributed to the new data from Statcast about Launch Angle.

The recent thoughts are that hitting in the air more increases the chance of success for getting on base or hitting homeruns. To defend this logic, many hitting coaches claim that "you cannot hit a homerun by hitting the ball on the ground". Because of this philosophy, many players are hitting the ball in the air more which leads to more long fly balls that would score runners from third base (or Sacrifice Flies).

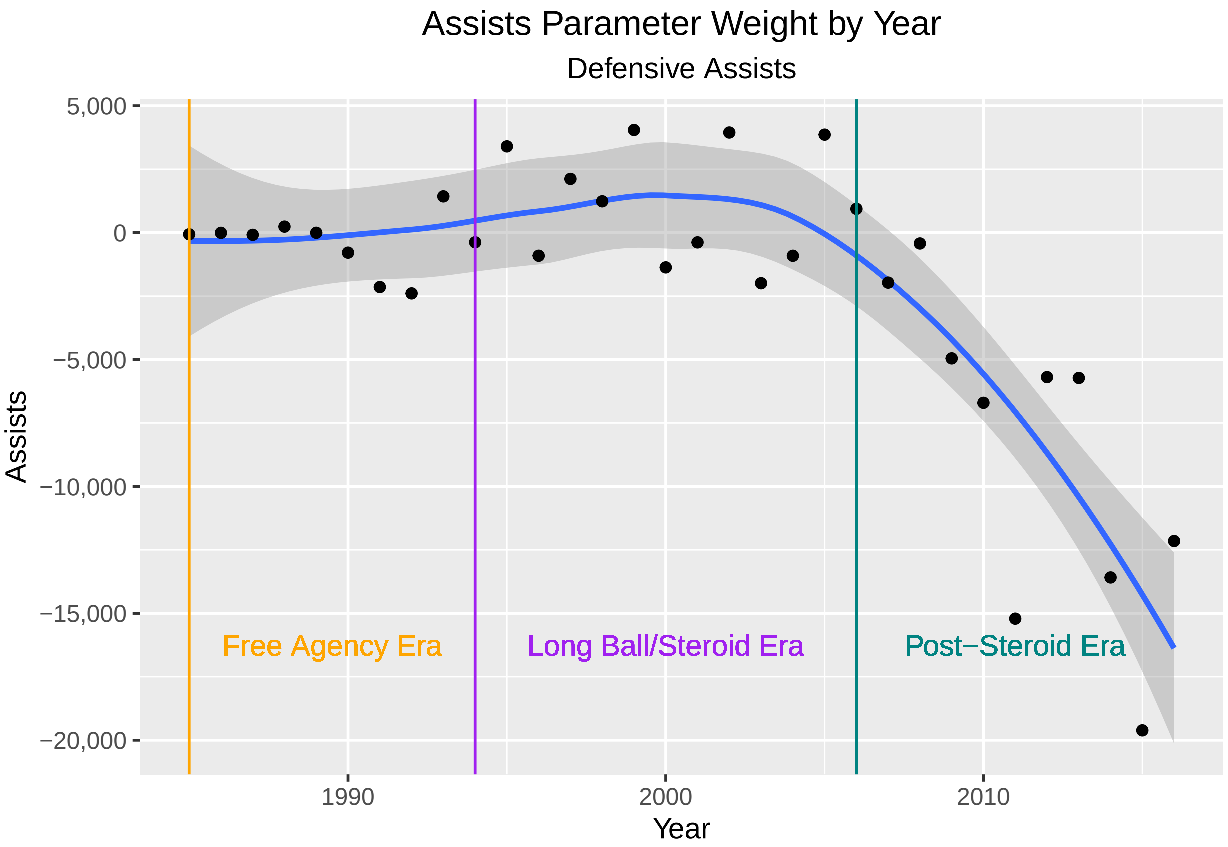
The scale on the vertical axis also appears to be very high, so this would need to be investigated in the future.



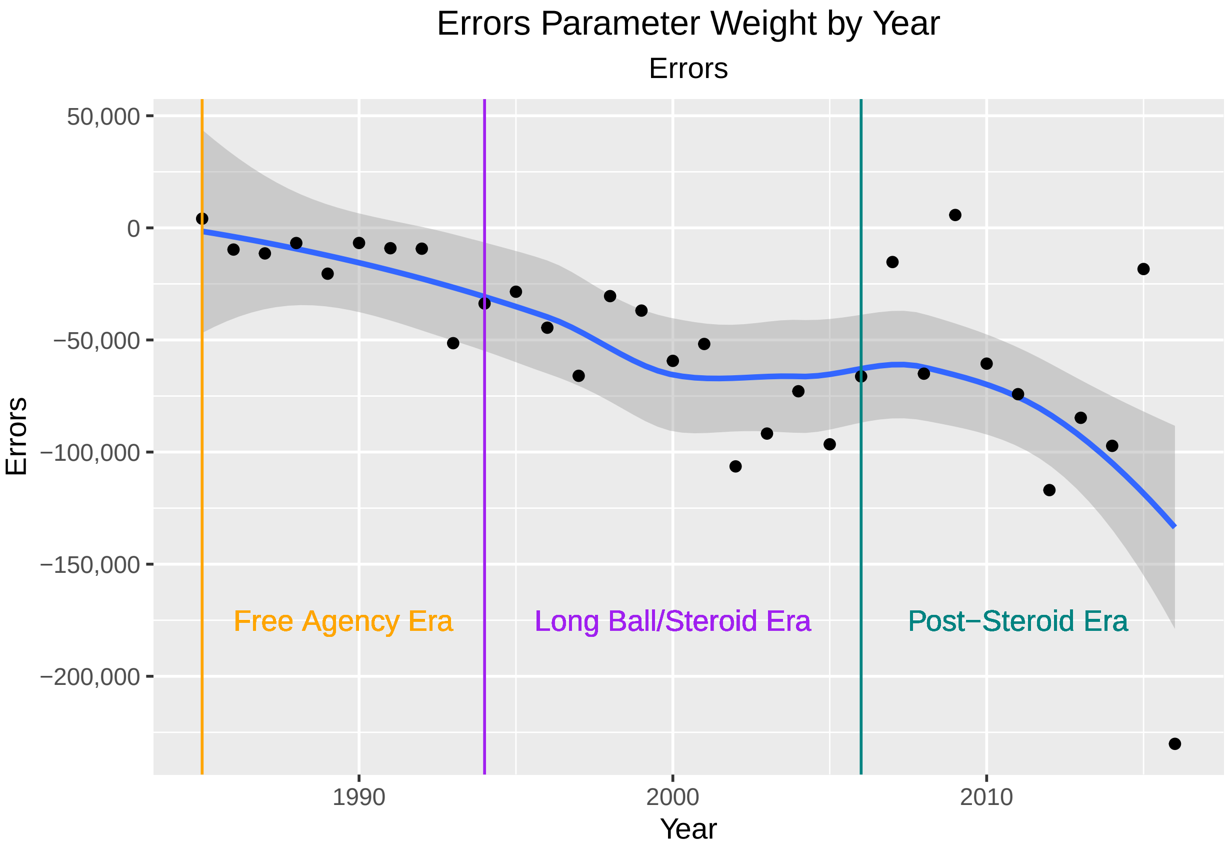
This plot shows that grounding into double plays relates with a high salary. This finding is contrary to intuition and suggests that our model will need to be explored further in the future. Perhaps a higher order term should be considered or the variable should be dropped from the model entirely. For more information, see the Future Work section.



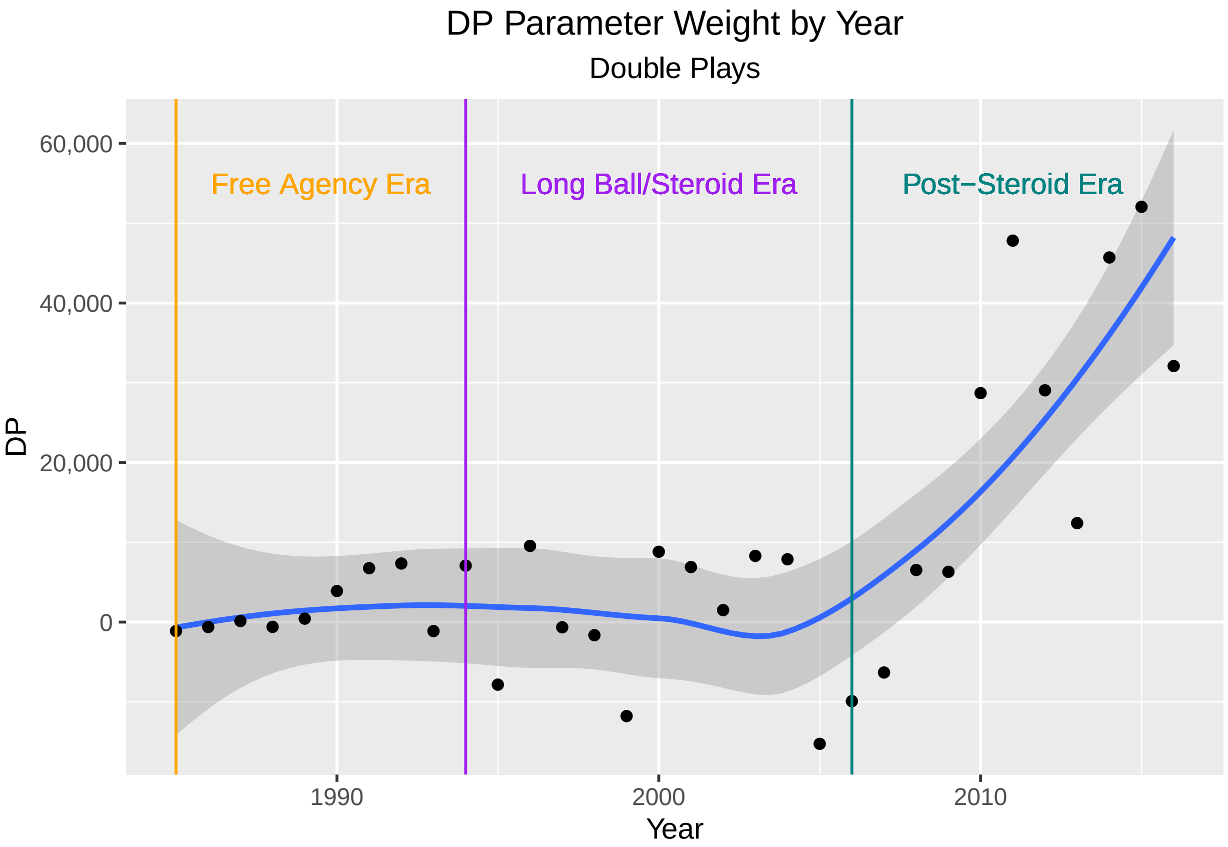
With parameter values only in the thousands (as compared to the 10 thousands), the defensive stat of putouts is the least important variable in our model. This is relatively close to 0 in all three eras.



An explanation similar to that of putouts can be applied to assists. They are of relatively low importance in our model.



Opposite to putouts and assists, Errors measure mistakes that players make in the field. These errors are seen as very costly in regards to a player's salary. This drives home the idea that good defense is an expectation. Having poor defensive skill hurts a player more than having great defensive skill helps a player, with regards to salary.



The Double Play parameter was relatively unimportant during the Free Agency and the Long Ball eras. However, in the Post-Steroid era, it has become more valuable.

Typically, second basemen and shortstops turn double plays most frequently. This increase in the parameter for double plays may be due to higher salaries for shortstops and second basemen for a variety of reasons.

Future Work

In the future, I have a few ideas where to expand on this project.

The most natural expansion would be to expand and perform a similar analysis for Pitchers based off of Pitching criteria. This would allow us to evaluate all players in MLB.

There were also a few curious variables that were not included in the model (both purposefully and unknowingly) or had surprising effects on the models. For variable that had effects that were surprising findings to me, see the Variable Importance over Time Plots beginning on Page 9.

A third way to expand on this project would be to include continuous variables in our analysis such as batting average and on base percentage. I purposefully chose to only include discrete variables for this analysis, however including continuous variables would likely lead to interesting results.

A final way of expanding on this project would be to consider different order models for our equation. For example, considering higher powers or a log transformation on salary would likely lead to better results. Due to time restraints, I was unable to expand on this idea in this research project.