Technical Report

Group 6

Abstract

The competitive balance tax (implemented in 2002) is a rule implementing salary caps for baseball players in a given team. This paper attempts to uncover if there is any underlying difference in the number of lost potential players (called "couldabeens") in the pre-rule (1968-2002) and post-rule eras (2002-2018).

Introduction

Major League Baseball's "competitive balance tax," first implemented in 2002, has become a favorite subject of outrage among players and fans alike in recent years. On paper, the policy was meant to make the sport more competitive and boost salaries for players on lower-revenue teams. The owners and the players' union would negotiate the largest reasonable amount a team could spend on its roster, and any team that wanted to spend beyond that cap would pay a "tax" (a share of their excess payroll) to be redistributed to the poorer teams. In practice the policy has effectively become a salary cap, particularly after the 2016 renegotiations. In 2018 only two MLB teams out of thirty went over the salary threshold, and only one of those by any significant margin (the team that did so, the Boston Red Sox, unsurprisingly went on to win the World Series). With leaguewide revenue growth far outpacing growth in the revenue cap and most younger players effectively locked out of salary negotiations by free agency rules, players have understandably chafed at the limitation on their salaries, with mutterings of a player strike unless the rule is altered.

But while the depressive effect on players' salaries is not in dispute, the question of whether the rule *hurts the game* (a far graver sin among baseball fans than merely conspiring to lower salaries) is more open. At least in the conventional wisdom, the tax encourages teams to cultivate a massive pool of recruits and then pay the best of them the minimum permitted salary for up to seven years before they age into free agency. Once these players enter free agency, the narrative goes, the team dumps them for younger players whom they can pay the minimum salary. The remainder of their salary cap goes towards retaining a few older superstars with the name recognition to bring in fans, with many good-but-not-Mike-Trout players pushed into early retirement. The implication, then, is that the salary policy "hurts the game" because the MLB is less likely to be putting the best 30 shortstops in the world on the field at any given time. I would like to test this anecdotal observation empirically.

So, the working hypothesis states that since the advent of the competitive balance tax, the number of better players forced into retirement annually will have increased. A "better player forced into retirement" will be defined as a player under the age of 32 who a) left baseball due to their contract not being renewed, rather than injury or personal circumstances, and b) was outperforming the median rookie in their position at the time of their retirement. In other words, we might expect to find the number of couldabeens to increase after the implementation of the rule.

Methods

WAR Data

Our data comes from https://stathead.com/baseball/ and we mainly have four sets of data.

1. Rookie pitchers
2. Rookie position players
3. Retired pitchers
4. Retired position players

The research question in this paper hinges on classifying retired players of lost potential. To motivate this, we will define our classifier for a "couldabeen" below.

Payroll data

We also have data on the revenues and payrolls in the MLB for a given year.

1. MLB yearly payroll
2. MLB yearly total revenue

The Couldabeen Classifier

For a given year Y, we first compute the median rookie's WAR, call it μ_Y. Additionally, compute the standard deviation of that data and call it σ_Y. Then, given a threshold $t \in \mathbb{R}$, we construct the corrosponding classifier for "couldabeen" status C of a given retired player p (from the year Y) to be as follows:

$$C(p) = \begin{cases} True, \text{ WAR}_p \geq \mu_Y + t\sigma_Y \\ False, \text{ WAR}_p < \mu_Y + t\sigma_Y \end{cases}$$

The Moneyball Classifier

To demonstrate that the release of *Moneyball* in 2003 is a fair place to partition the data, we perform a hypothesis test on the `postMoneyball` classifier of a year Y, which we define as follows:

$$\text{postMoneyball}(Y) = \begin{cases} True, Y > 2003 \\ False, Y \leq 2003 \end{cases}$$

Once every retired player is classified appopriately, we run a few models and a hypothesis test:

(i) Linear Model (`prop ~ Year`): After aggregating all the retired players and their classifications, we summarize the data by year to obtain the proportion of retired players that were couldabeens that year, called `prop`. As such, we now have 50 data points (for each year), and a response variable being the proportion of couldabeens. As such, we run a linear model fitting `prop ~ Year` for the pre-rule era and the post-rule era, and on the unpartitioned dataset. Because there will always be "couldabeens", we do not expect a large effect size and hence a very significant result, however, the **sign** of our coefficient will be essential for our inference. If our research hypothesis is correct (that there is an effect), we expect to see a positive coefficient for β_{Year} on the post-rule era partition.

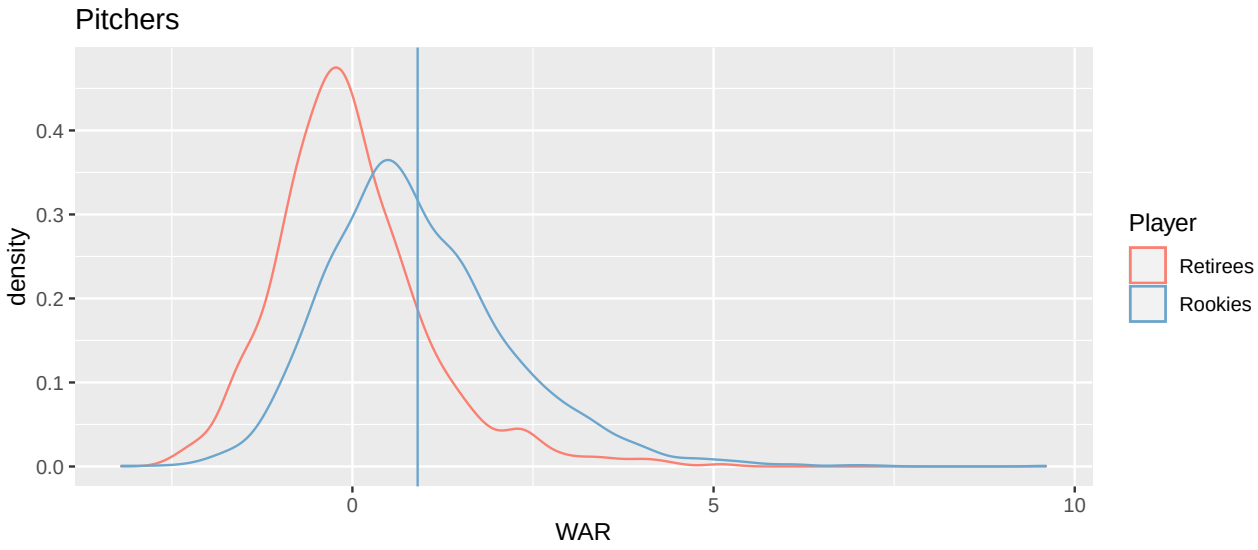
(ii) Hypothesis Test (`prop ~ postMoneyball`): With the dummy variable `postMoneyball` at hand, we perform a hypothesis test to see if the mean proportion of couldabeens in the pre-Moneyball era is different than that of the post-Moneyball era. If the result of our hypothesis test is significant, we proceed to partition our dataset since it affirms our belief that the publication of 'Moneyball' is indeed a confounding variable.

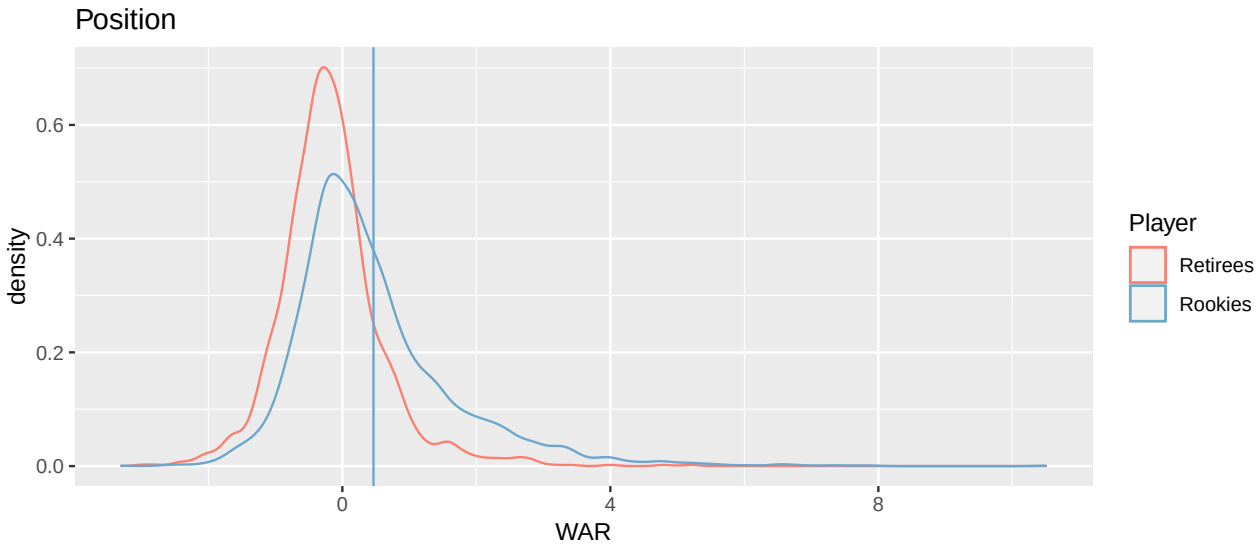
(iii) Linear Model (`prop ~ laborShare`): Lastly, we use the payroll data to obtain the labor share in the MLB for a given year, and fit that data to `prop`.

Exploratory Data Analysis

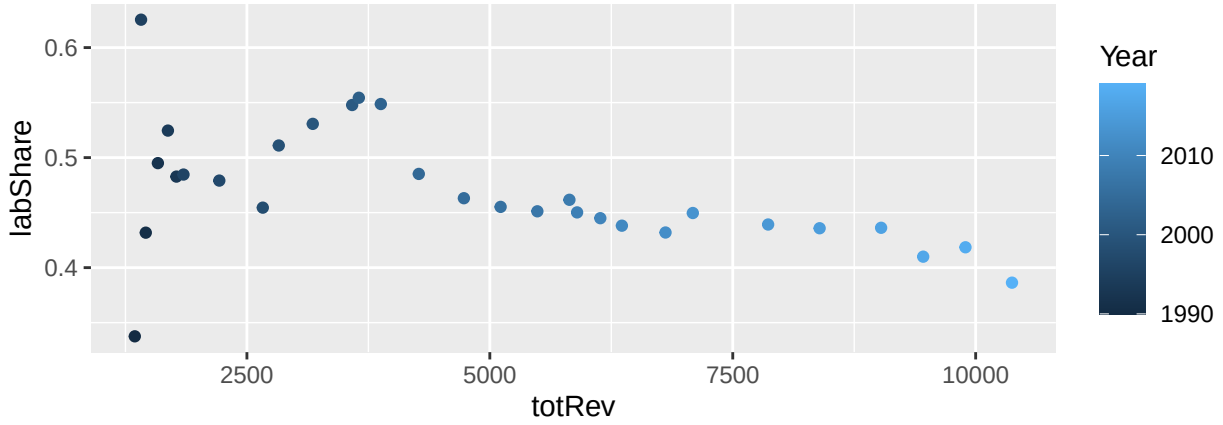
WAR Density Plots

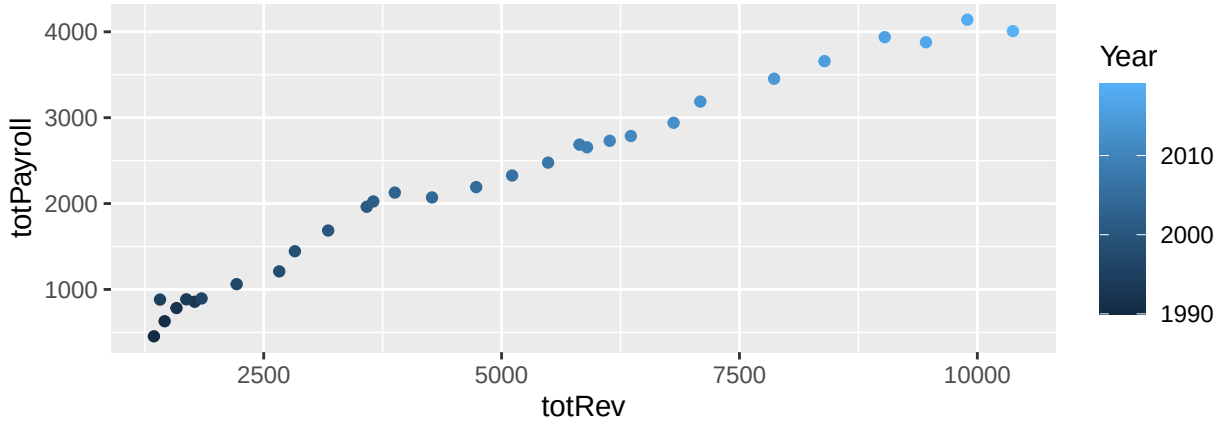
Below, we can visualize the densities of the WAR statistic for each type of player. In blue, we have the rookie players (with the median line denoted), and in red, we have the corrosponding retired players. As such, we can visualize the "couldabeens" as the retired players to the left of the median line (if the threshold $t = 0$), and we correspondingly count the number of couldabeens based on year, from which we make our inference on the impact of Year.

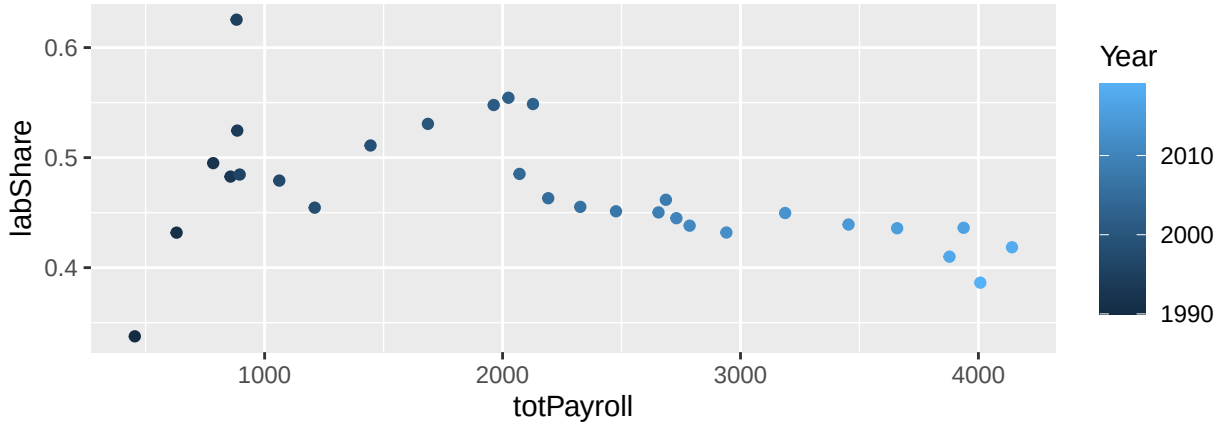




Payroll Predictor Data

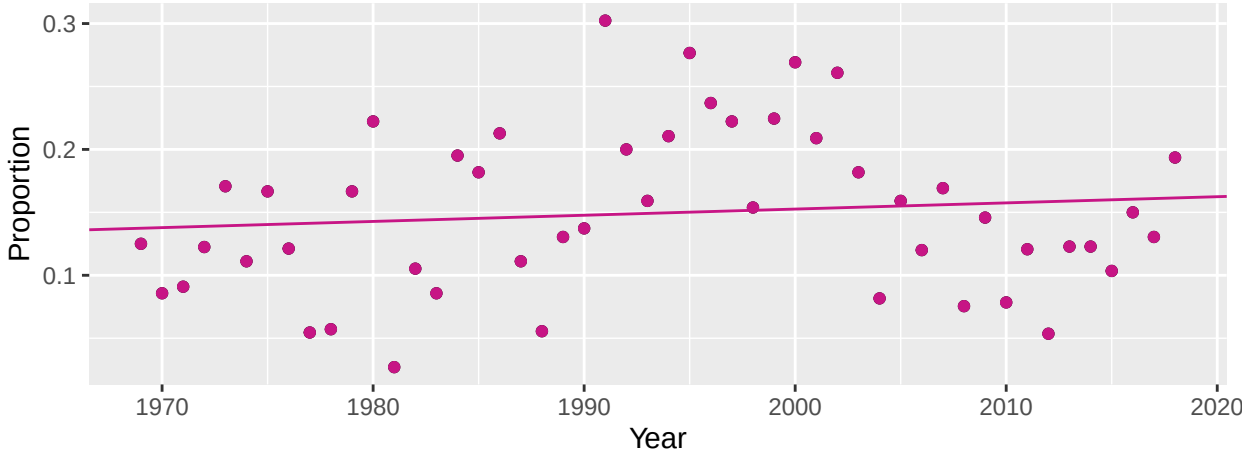






Linear Models

Initial Model: Year



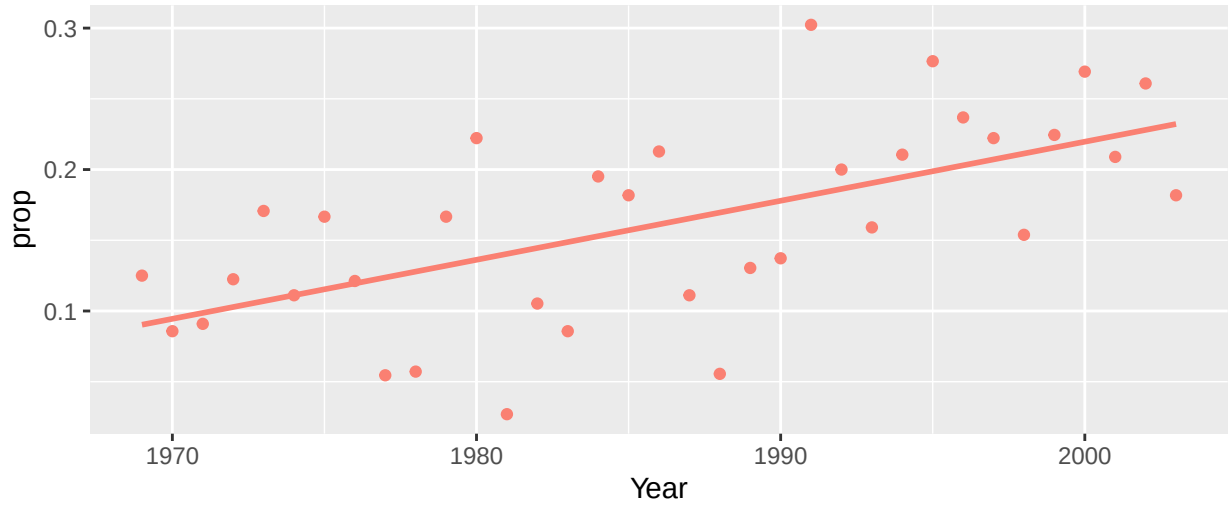

```
##
## Call:
## lm(formula = prop ~ Year, data = couldabeens_t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11623 -0.03820 -0.01083  0.04694  0.15415
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.8305575  1.2597256  -0.659    0.513
## Year         0.0004916  0.0006319   0.778    0.440
##
## Residual standard error: 0.06448 on 48 degrees of freedom
## Multiple R-squared:  0.01245,    Adjusted R-squared:  -0.008123
## F-statistic: 0.6052 on 1 and 48 DF,  p-value: 0.4404
```

Hypothesis Test: Why Split the Data?

```
##
## Call:
## lm(formula = prop ~ postMoneyball, data = couldabeens_t)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.134209 -0.042568  0.001007  0.045264  0.141090
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.16124    0.01052  15.334   <2e-16 ***
## postMoneyball -0.03944    0.01920  -2.054   0.0454 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06221 on 48 degrees of freedom
## Multiple R-squared:  0.08081,    Adjusted R-squared:  0.06166
## F-statistic:  4.22 on 1 and 48 DF,  p-value: 0.04543
```

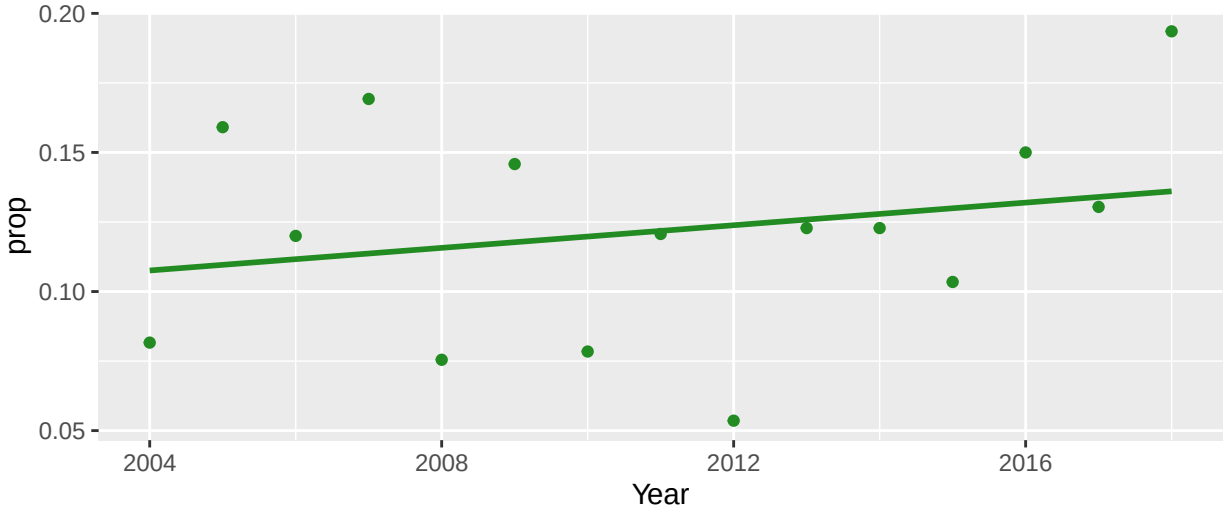
Splitting the Data

Pre-Moneyball Model



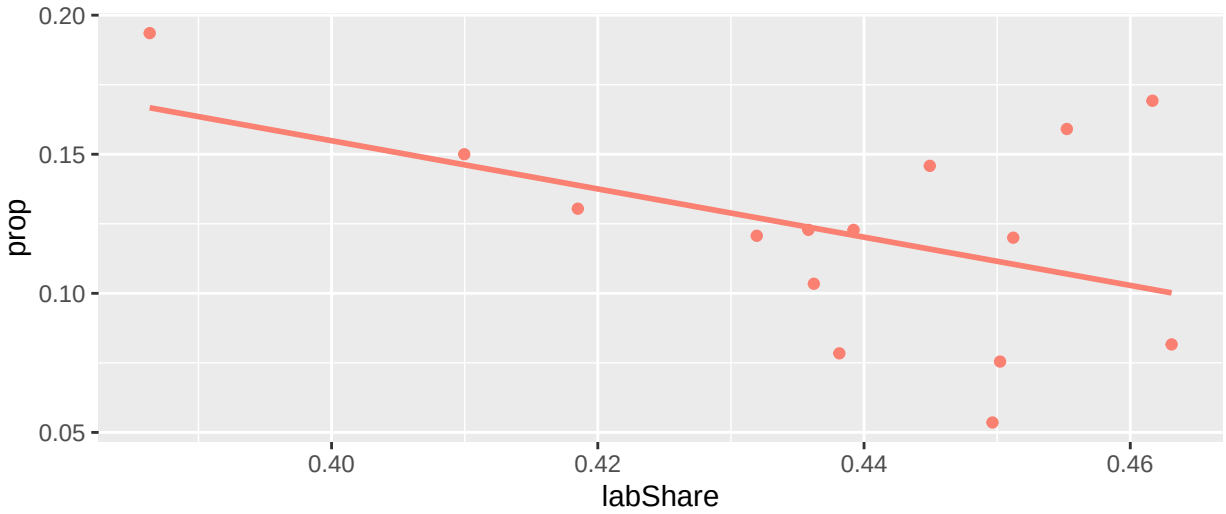

```
##
## Call:
## lm(formula = prop ~ Year, data = couldabeens_pre)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.114028 -0.042000  0.008993  0.034685  0.120220
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.1282589  1.8541569  -4.384 0.000112 ***
## Year         0.0041740  0.0009336   4.471 8.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05578 on 33 degrees of freedom
## Multiple R-squared:  0.3772, Adjusted R-squared:  0.3583
## F-statistic: 19.99 on 1 and 33 DF,  p-value: 8.69e-05
```

Post-Moneyball Model




```
##
## Call:
## lm(formula = prop ~ Year, data = couldabeens_post)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07026 -0.02621 -0.00306  0.02307  0.05751
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.967977   4.680332  -0.848    0.412
## Year         0.002034   0.002327   0.874    0.398
##
## Residual standard error: 0.03894 on 13 degrees of freedom
## Multiple R-squared:  0.05548,    Adjusted R-squared:  -0.01718
## F-statistic: 0.7636 on 1 and 13 DF,  p-value: 0.3981
```

Labor Share Model




```
##
## Call:
## lm(formula = prop ~ labShare, data = couldabeens_post)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.058253 -0.019268 -0.001008  0.018178  0.067824
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5017     0.2036   2.464   0.0284 *
## labShare     -0.8670     0.4642  -1.868   0.0845 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03558 on 13 degrees of freedom
## Multiple R-squared:  0.2116, Adjusted R-squared:  0.151
## F-statistic:  3.49 on 1 and 13 DF,  p-value: 0.08446
```

Results

Linear Model on Year

Pre-*Moneyball* era

- $\beta_{Year} = 0.004174$.
- β_{Year} is statistically significant with a very low p-value of $p \approx 0$.

Post-*Moneyball* era

- $\beta_{Year} = 0.002034$.
- $\beta_{Year} > 0$ supports the hypothesis that there is an increasing rate of couldabeens since the luxury tax.
- β_{Year} is not statistically with a high p-value of 0.398.

Since the pre-*Moneyball* era had more data points, this may explain the lower p-value.

Hypothesis Test

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.161	0.011	15.334	0.000	0.140	0.182
postMoneyball	-0.039	0.019	-2.054	0.045	-0.078	-0.001

Simpson's Paradox

- Partitioning and fitting linear model with `prop ~ Year` yields $\beta_{Year} > 0$ in both partitions.
- However, if we do not make the partition, we find that $\beta_{Year} \approx 0$.

Linear Model on Labor Share

- We find $\beta_{LabShare} = -0.867$ with a p-value of $p = 0.084$.
- Labor share is indeed **negatively** correlated with proportion of couldabeens.

Discussion

Simpson's Paradox

Interestingly, when we choose not to partition the dataset into the post-rule and pre-rule eras and fit a linear model `Year ~ prop`, we find that a very resounding instance of Simpson's Paradox. In both partitions, we find that the parameter $\beta_{Year} > 0$. However, if we do not make the partition, we find that $\beta_{Year} \approx 0$. This raises some questions regarding the role our partition plays in our inference and modeling choices.

Firstly, this begs the question: should we partition the data? For the sake of valid statistical inference, if we assume `Year` is a valid predictor, it initially seems as though it would be errenous practice to do so. If we partition the data, we essentially say that our inferential result is **conditioned** on only that era's data (since our parameter is fit to only that data). As such, (assuming no confounding variable) it would be better statistical practice to not partition the data and only make inference on a β_{Year} parameter trained on the entire dataset.

However, as we can see, if we do not partition the data, our models suffer from Simpson's paradox and there seemingly is no trend along the years. As it suggests, there is indeed a possible explanation for why this partition is necessary. Some confounding variable(s) that the model failed to include (because we did not find and import the data yet) is any salary data (adjusted for inflation) or labor share data in a given year. The reason why this data might be particularly helpful to include in further renditions of this model is due to the fact that in the end, budgets are finite and they are what determines if a player continues playing or not regardless of their ability.

As such, we believe that because the post-rule era and pre-rule era are fundamentally different in terms of team salary budgets (and related variables), it is valid and actually more statistically preferable to inform our inference in this way. So until such data is incorporated into the model in some form or another, the

inference will be on the comparison of the β_{Year} parameters from the two disjoint partitions.

Threshold Stability

Recall, that our couldabeen classifier was constructed as follows:

For a given year Y, we first compute the median rookie's WAR, call it m_Y. Additionally, compute the standard deviation of that data and call it σ_Y. Then, given a threshold $t \in \mathbb{R}$, we construct the corrosponding classifier for "couldabeen" status C of a given retired player p (from the year Y) to be as follows:

$$C(p) = \begin{cases} True, \text{ WAR}_p \geq m_Y + t\sigma_Y \\ False, \text{ WAR}_p < m_Y + t\sigma_Y \end{cases}$$

All in all, we can see there are quite a few problems with our model. Although we have obtained a positive result on the slope of β_{Year} for a particular threshold in the post-rule era, varying the threshold tells another story. In particular, if we run a linear model on the proportion of couldabeens against year i.e. `Year ~ prop` against a varying threshold (in standard deviations of rookie WAR), we found that β_{Year} is quite unstable. Consider the graph below:

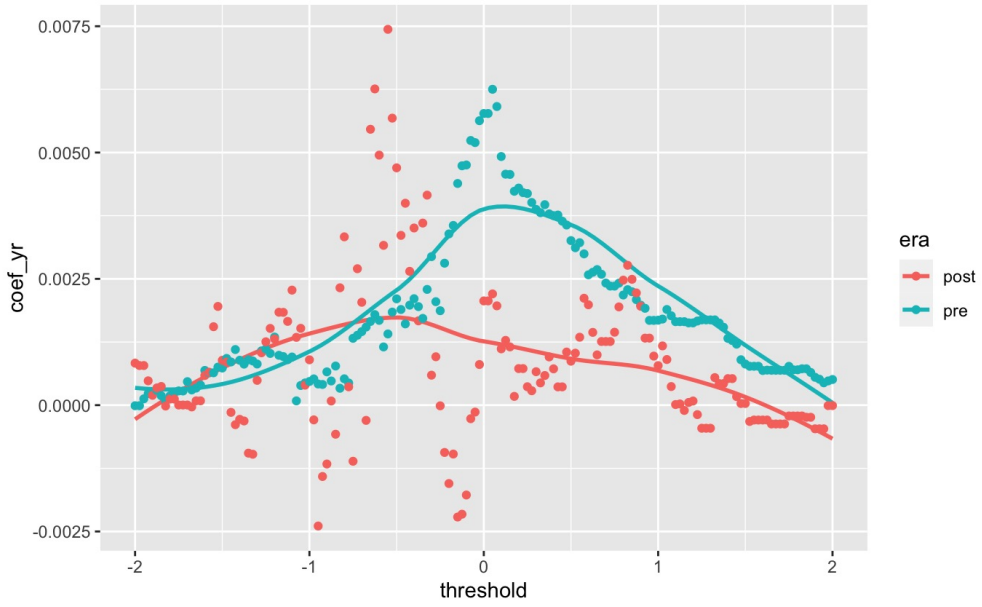


Figure 1: Performance of the Linear Model against varying threshold for couldabeen status

One reason this instability may arise could be due to the fact that our thresholds for every year were calculated as a function of *only* that year's data. As such, the sample from which we obtain the threshold by each level (corrosponding to each year) is very small compared to the whole dataset; as such, small changes to the threshold cause high levels of noise in our final model.

To treat this problem, we may reconsider our method of computing the threshold. Instead of strictly using just the data for a given year, it may be beneficial to "smooth" the threshold out by taking the data of that year and adjacent years (like a window). For instance, instead of considering just 2018 data, we may take 2017-2019 data for a "window length of 1". Hopefully, once this adjustment is made, our results will be more stable and less noisy with respect to the threshold, which should be just an arbitrary value rather than a value that carries so much weight in the entire inferential process.

Furthermore, the addition of some supplementary yearly salary/budgets data may be useful as another predictor since `Year` alone does not seem to have good explanatory power.

Resampling Results

Conclusions

To summarize:

- No correlation between year and proportion of couldabeens in full data set ($\beta_{Year} \approx 0$).
- Positive but non-significant relationship between year and proportion of couldabeens when partitioned into pre- and post-*Moneyball* eras.
- Significant negative relationship between publication of *Moneyball* and proportion of couldabeens. ($p = 0.045$)
- Somewhat significant negative relationship between labor share and proportion of couldabeens in post-*Moneyball* era. ($p = 0.084$)

Previous literature has established that the 2003 CBA, which implemented the luxury tax, led to a decline in labor share. We find *some* evidence for a link between a lower labor share and a higher proportion of "couldabeens" retiring. However, we cannot definitively conclude that the MLB luxury tax has increased the proportion of couldabeens.

Code Appendix

Importing the Datasets

```
# Load rookies datasets
df_pit_rkes <- read_csv("../data/rookie-pitcher.csv")
df_pos_rkes <- read_csv("../data/rookie-position.csv")
# Load retirees datasets
df_pit_ret <- read_csv("../data/retirees-pitcher.csv")
df_pos_ret <- read_csv("../data/retirees-position.csv")
```

Wrangling the Datasets

```
# select appopriate columns from player datasets
wrangle_init <- function(dataset){
  colnames(dataset)[3] <- "WAR"
  dataset %>% select(WAR, Year)
}
```

```
# Obtain wrangled datasets
pit_rkes <- wrangle_init(df_pit_rkes)
pit_ret <- wrangle_init(df_pit_ret)
pos_rkes <- wrangle_init(df_pos_rkes)
pos_ret <- wrangle_init(df_pos_ret)
```

Finding the Couldabeen Thresholds

```
# obtain summary of WAR: median and variance
find_thresholds <- function(dataset, sds = 0){
  dataset %>%
    group_by(Year) %>%
    summarize(mean_WAR = median(WAR), sd_WAR = sqrt(var(WAR))) %>%
    mutate(threshold = mean_WAR + sds*sd_WAR) %>%
    select(Year, threshold)
}
```

```
# Get thresholds in each year
pit_thresholds <- find_thresholds(pit_rkes)
pos_thresholds <- find_thresholds(pos_rkes)
```

```
head(pit_thresholds)
```

Classifying Retiree Couldabeens

```
# Appends above_threshold column to retirees dataset (checks if a retiree exceeds a threshold)
compare_thresholds <- function(dataset, summary_dataset){
  above_threshold <- rep(NA, nrow(dataset))
  for(i in 1:nrow(dataset)){
    year <- as.numeric(dataset[i,2])
    above_threshold[i] <- (dataset[i,1] > summary_dataset[year - 1968, 2])
  }
  dataset <- cbind(dataset,above_threshold)
  colnames(dataset)[3] <- "above_threshold"
  dataset
}
```

```
# See and record which players cross that year's adjusted threshold from rookie players
pit_ret <- compare_thresholds(pit_ret, pit_thresholds)
pos_ret <- compare_thresholds(pos_ret, pos_thresholds)
# Get all retired couldabeens by combining the rows
retirees <- rbind(pit_ret,pos_ret)
```

```
head(retirees)
```

Counting Couldabeens by Year

```
# Counts couldabeens in a given year
count_cbns <- function(dataset){
  dataset %>%
    group_by(Year) %>%
    summarize(cbns = sum(above_threshold))
}
```

```
# Count the retiree couldabeens by year
couldabeens <- count_cbns(retirees)
```

```
#head(couldabeens)
```

Counting Retirees in a Given Year

```
# Yields the sum of retirees in two datasets (pitchers, position commonly)
total_retirees_by_yr <- function(pitchers, position){
  year_count_pitchers <- retirees_by_yr(pitchers)$retirees
  year_count_position <- retirees_by_yr(position)$retirees
  data.frame(Year = 1969:2018, retirees = year_count_position + year_count_pitchers)
}
# Counts the retirees in a single dataset
retirees_by_yr <- function(dataset){
  year_count <- dataset %>%
    group_by(Year) %>%
    summarize(retirees = n())
}
```

```
# Find number of retirees by year
num_retirees <- total_retirees_by_yr(df_pit_ret, df_pos_ret)
num_retirees <- data.frame(retirees = num_retirees)
```

```
#head(num_retirees)
```

Retiree Proportion of Couldabeens

```
# Append number of retirees that year
couldabeens <- cbind(couldabeens, num_retirees)
# Find and append proportion of couldabeens : retirees
couldabeens <- couldabeens %>% mutate(prop = cbns/retirees)
```

```
#head(couldabeens)
```

Modeling Helper Functions

```
# Filters year to be in the post-rule era
postrule <- function(dataset, rule_year = 2002){
  dataset %>% filter(Year > rule_year)
}
# Filters year to be in the pre-rule era
prerule <- function(dataset, rule_year = 2002){
  dataset %>% filter(Year <= rule_year)
}
# Prep the boolean data for a logistic model
prep_booleans <- function(dataset){
  bools <- as.numeric(dataset$above_threshold)
  dataset[,3] <- bools
  dataset
}
```

Model Functions

```
# Logisitic regression model on exceeding the threshold
logistic_model <- function(dataset){
  logistic_model <- glm(formula = above_threshold ~ Year, data = dataset, family = "binomial")
  logistic_model
}
# Linear model predicting proportion of couldabeens based on the year
linear_model <- function(dataset){
  linear_model <- lm(formula = prop ~ I(Year), data = dataset)
  linear_model
}
```

Logistic Models

```
# Partition dataset into years before and after rule
retirees_pre <- prerule(retirees)
retirees_post <- postrule(retirees)
# On the retirees pre-rule dataset
model_log_pre <- logistic_model(retirees_pre)
model_log_post <- logistic_model(retirees_post)
```

Linear Models

```
# Partition dataset into years before and after rule
couldabeens_pre <- prerule(couldabeens)
couldabeens_post <- postrule(couldabeens)
# Obtain linear model for pre-rule years
model_pre <- lm(formula = prop ~ I(Year), data = couldabeens_pre)
coefs_pre <- model_pre$coefficients
# Obtain linear model for post-rule years
model_post <- lm(formula = prop ~ I(Year), data = couldabeens_post)
coefs_post <- model_post$coefficients
```

References

(1) https://stathead.com/baseball/
(2) Bradbury, John Charles. "What Explains Labor's Declining Share of Revenue in Major League Baseball?" (2019).
(3) https://blogs.fangraphs.com/mlbs-evolving-luxury-tax/
(4) Lewis, Michael. *Moneyball: The Art of Winning an Unfair Game*. New York: Norton, 2003.
(5) Hayes, Hannah. "What will Nate Silver do next?" *uchicago.edu*.
(6) Birnbaum, Phil. "Asking the Right Qustions." *SABR.org*.