

# Statistical Learning: Project Presentation

G. Dunlavey, W. Ren, A. Taqi

# Research Question

The MLB's "luxury tax," implemented in the 2003 Collective Bargaining Agreement, is a rule penalizing franchises whose team payroll for a given year exceeds an agreed threshold. This project attempts to test the tax's effect on quality of play by comparing the rates and numbers of above-replacement retirees (referred to here as "couldabeens") in the pre-rule era (1969-2002) and post-rule eras (2003-2018).

# Theory

Existing literature has established the depressive effect of the post-2003 CBAs on labor share in the MLB (Bradbury, 2019).

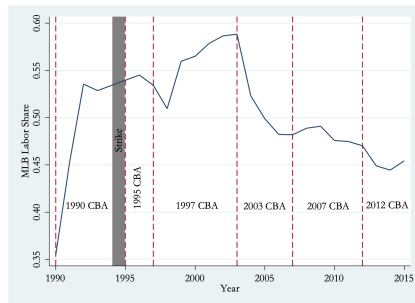


Figure 3. MLB Labor Share by Collective Bargaining Agreement (1990-2015)

How might the “luxury tax” increase the number of above-replacement retirees?

- Players don't gain free agency until six years of MLB service time, making rookies cheaper than veterans.
- Farm teams not counted towards salary threshold, guaranteeing reserve pool of rookies.
- Teams direct limited budget towards retaining a handful of elite veterans, filling out roster with rookies.
- Good-but-not-Mike-Trout veterans replaced with marginally inferior rookies to stay below salary threshold.

# Methods: The Data

We got our data from <https://stathead.com/baseball/> and divided it into four data sets:

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

The research question in this paper hinges on identifying above-replacement retirees. To motivate this, we will first define our classifier for a “couldabeen”.

## Methods: The Couldabeen Classifier

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  $\sigma_Y$ . Then, given a threshold  $t \in \mathbb{R}$ , we construct the corresponding classifier for “couldabeen” status  $C$  of a given retired player  $p$  (from the year  $Y$ ) to be as follows:

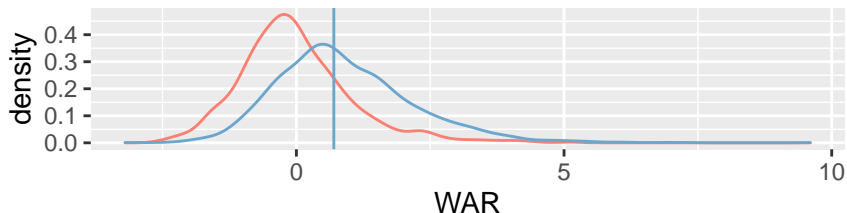
$$C(p) = \begin{cases} \text{True}, & \text{WAR}_p \geq m_Y + t\sigma_Y \\ \text{False}, & \text{WAR}_p < m_Y + t\sigma_Y \end{cases}$$

# Visualization: Couldabeen Classification

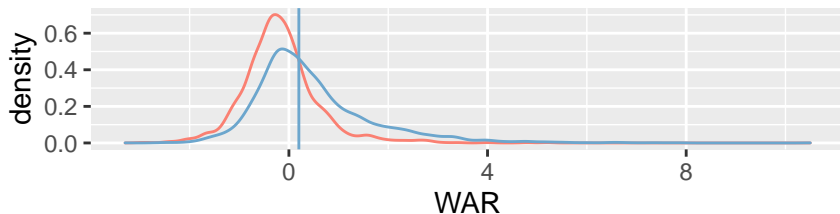
- We can graph the densities of the WAR statistic for each player type (i.e. pitcher or position) to illustrate our premise.
- Although the threshold defining a “couldabeen” is year specific, for the sake of visualization these graphs combine all data from 1969-2018.
- In blue, we have the rookie players (with the median across all years 1969-2018 denoted), and in red, we have the corresponding retired players.

# Visualization: Couldabeen Classification

## Pitchers



## Position



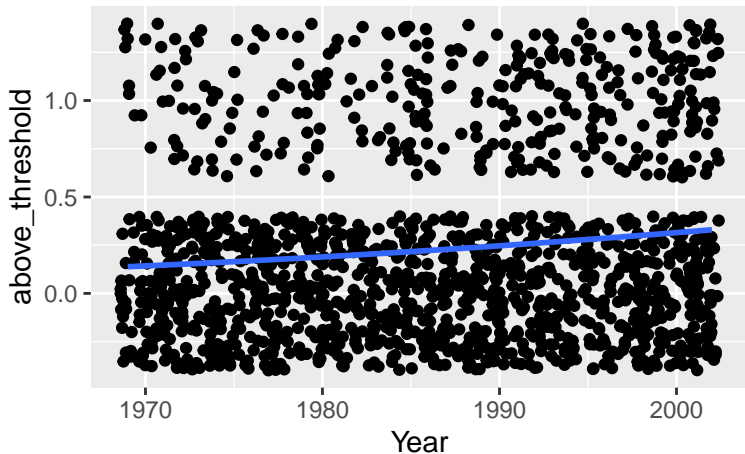


# Methods: Modeling

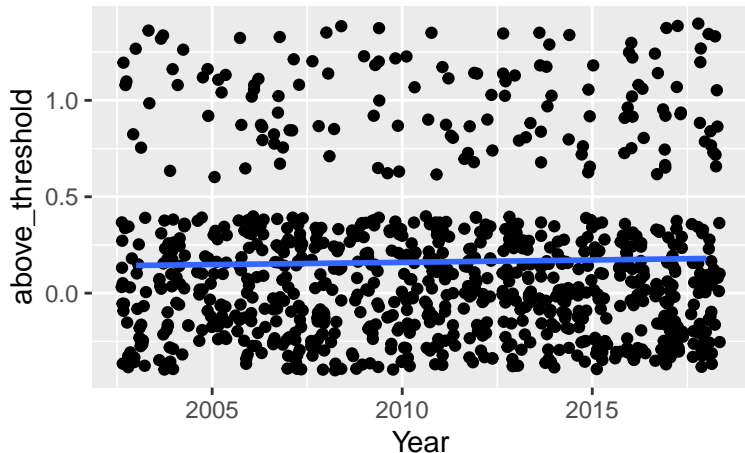
Once every retired player is classified appropriately, we run two primary models:

- ❶ **Logistic Model:** On the retired players dataset, each player has a new column, `exceeds_threshold` which is the result of the classifier  $C(p)$ .
  - This classifier is a boolean, so we run two models on the pre-rule era and the post-rule era and see the effects Year has on `exceeds_threshold`.
  - 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  $\beta_{Year}$ .

# Logistic Model: Couldabeens Retirees (Pre-rule Era: 1969-2002)



# Logistic Model: Couldabeens Retirees (Post-rule Era: 2003-2018)



# Overall Results: Logistic Models

## Post-rule era

- Post-rule era model, we obtain a parameter  $\beta_{Year} = 0.01781$ . So indeed,  $\beta_{Year} > 0$ .

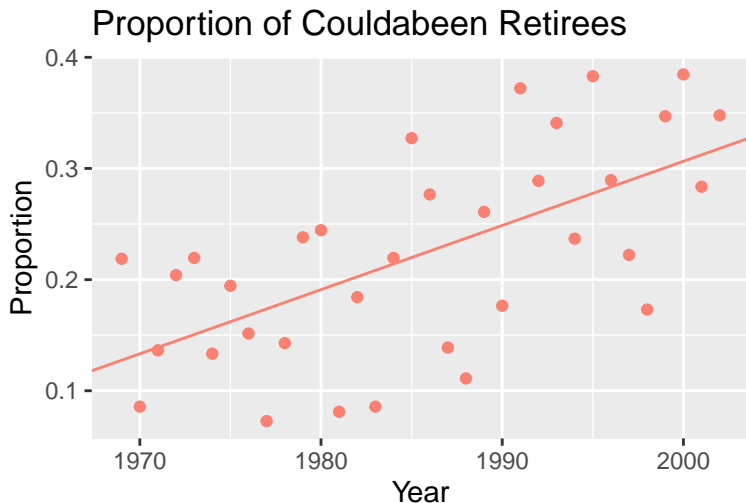
## Pre-rule era

- For the pre-rule era model, we find  $\beta_{Year} = 0.03392$ . So indeed, we also have  $\beta_{Year} > 0$ . So, the effect is positive in both eras (will be discussed later in linear models).
- All in all, we may interpret the positive coefficient in both eras as saying that the probability a couldabeen is classified is greater at the end of an era than at its start for **both eras**.
- Conclude that up until the rule has been implemented, the proportion of couldabeens has been rising, and it rapidly drops in 2003 (year of the rule) only to increase again.

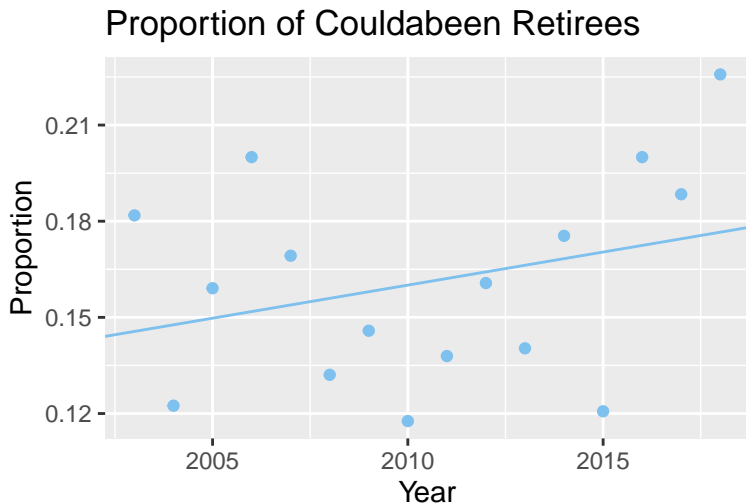
# Methods: Modeling

- ❖ **Linear Model:** 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.
- So, we run a linear model fitting  $\text{Year} \sim \text{prop}$  for the pre-rule era and the post-rule era.
- Again, because there will always be “couldabeens”, we do not expect a large effect size and hence a very significant result.
- Like the logistic models, 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  $\beta_{\text{Year}}$ .

# Linear Model: Couldabeens Retirees (Pre-rule Era: 1969-2002)



# Linear Model: Couldabeens Retirees (Post-rule Era: 2003-2018)



# Overall Results: Linear Models

## Post-rule era

- Post-rule era model:  $\beta_{Year} = 0.002064$ . As such, since  $\beta_{Year} > 0$ .
- There is evidence that the rule has lead to an increase in the proportion of couldabeens.
- Predictions: Initially, the post-rule model predicts that in 2003 we get a proportion of  $\beta_0 + 2003\beta_1 = 0.144$  and in 2018, we get a proportion of  $\beta_0 + 2018\beta_1 = 0.177$ .

## Pre-rule era

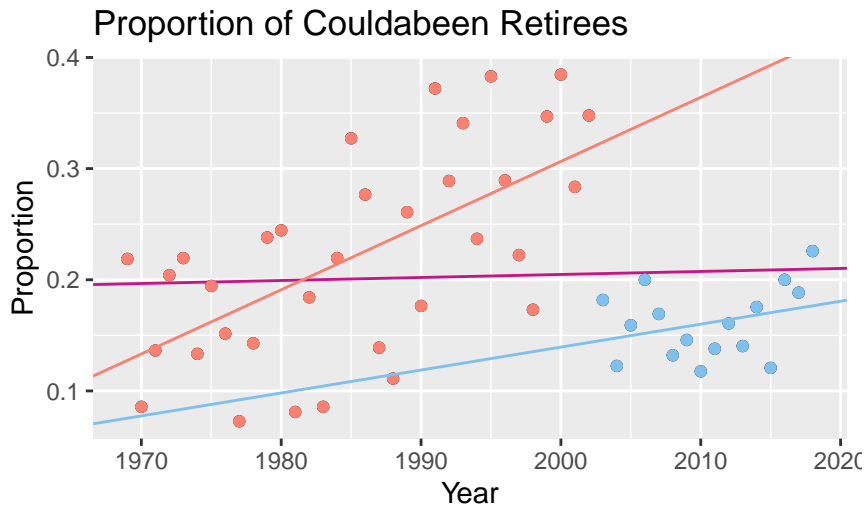
- Pre-rule era, we obtained a parameter  $\beta_{Year} = 0.005773$ . As such, we also have  $\beta_{Year} > 0$ .
- Conclude that up until the rule has been implemented, the proportion of couldabeens has been rising, and it rapidly drops in 2003 (year of the rule) only to increase again.



# Simpson's Paradox

- We chose to partition the dataset into the post-rule and pre-rule eras and fit a linear model  $\text{Year} \sim \text{prop}$ .
- In both partitions, we find that the parameter  $\beta_{\text{Year}} > 0$ .
- However, if we do not make the partition, we find that  $\beta_{\text{Year}} \approx 0$ .
- This raises some questions regarding the role our partition plays in our inference and modeling choices.
- In fact, this is *Simpson's Paradox*.

# Simpson's Paradox



# Threshold Stability

Now, we will analyze the effects of the threshold  $t$  on the impact of our  $\beta_{\text{Year}}$  variable in our linear models. It is important to see what the effects are since our inference relies on  $\beta_{\text{Year}}$  being positive for any  $t \in \mathbb{R}$ . Recall the definition of the Couldabeen classifier:

## The Couldabeen Classifier

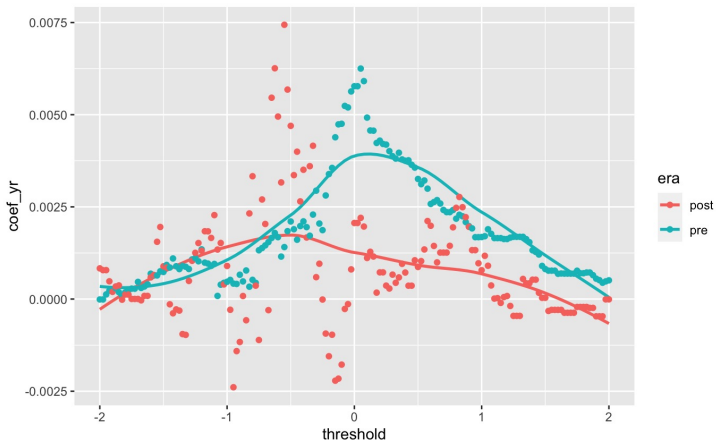
Given a threshold  $t \in \mathbb{R}$ , we construct the corresponding classifier for “couldabeen” status  $C$  of a given retired player  $p$  (from the year  $Y$ ) to be as follows:

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

# Threshold Stability

- Obtained a positive result on the slope of  $\beta_{Year}$  for a particular threshold in the post-rule era.
- Varying the threshold tells another story.
- Run a linear model with  $Year \sim prop$  against a varying threshold  $t$ .
- Then, we found that  $\beta_{Year}$  is quite unstable.
- We obtained the following graph on the next slide.

# Threshold Stability



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

# Threshold Stability: Possible Adjustments

- Thresholds every year were calculated as a function of *only* that year's data.
- So, the sample from which we obtain the threshold by each level (corrospounding to each year) is very small meaning small changes to the threshold cause high levels of noise in our final model.
- To treat this problem, we may consider “smoothing” the threshold out by taking the data of that year and adjacent years. For instance, instead of considering just 2018 data, we may take 2017-2019 data for a “window length of 1”.
- 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.

# 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/>