Statistical Learning: Project Presentation

G. Dunlavey, W. Ren, A. Taqi

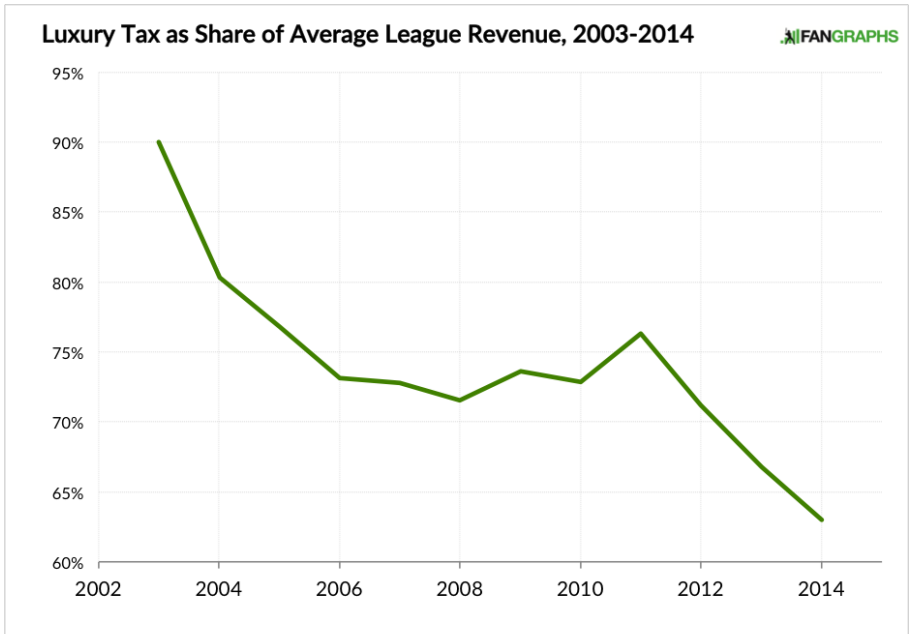
Research Question

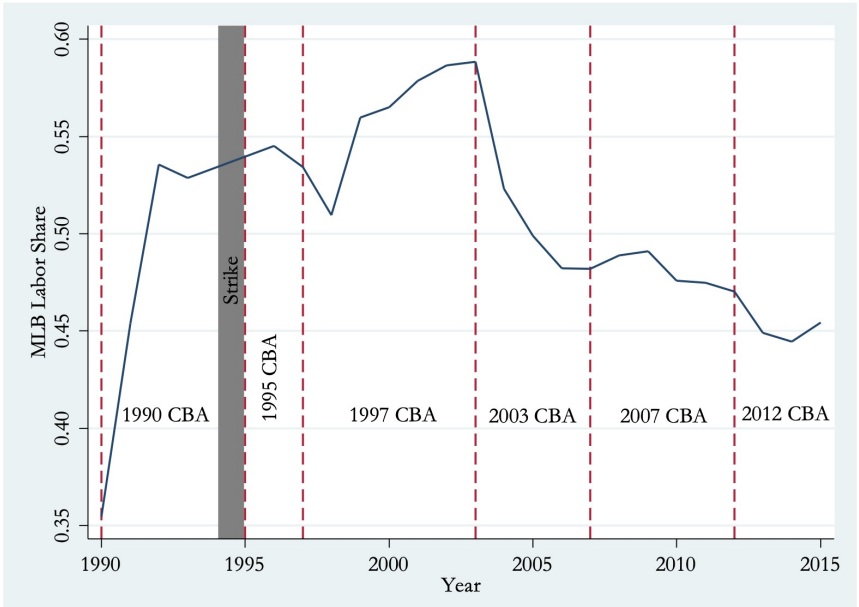
The MLB's "luxury tax," implemented in the 2003 Collective Bargaining Agreement, is a rule penalizing franchises whose team payroll for a given year exceeds an agreed threshold. This project attempts to test the tax's effect on quality of play by studying the number of above-rookie retirees (referred to here as "couldabeens") as a share of total retirements.

Why does the luxury tax matter?

Although originally pitched as a way to "even the playing field," the luxury tax has increasingly functioned as a salary cap. Existing literature has established a continuing decline in labor share in the MLB since the 2003 CBA (Bradbury, 2019).

$$\textit{Labor Share} = \frac{\textit{Total MLB Revenue}}{\textit{Total MLB Player Payroll}}$$





Theory

How might the "luxury tax" increase the number of above-rookie 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.
- Hypothesis: 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

Methods: WAR

Wins Above Replacement, or WAR, is a baseball statistic which seeks to measure a player's total contribution to his team. A WAR of 0.3 means the player's team will win 0.3 more games per season than if he had been substituted for a replacement-level player.

Position WAR:

$$WAR = \frac{(Player\ Runs - Average\ Runs) + (Average\ Runs - Replacement\ Run}{Game\ Runs\ to\ Wins\ Estimator}$$

where

$$Player\,Runs = Batting\,Runs + Baserunning\,Runs + DoublePlay\,Runs + Fielding\,R$$

Pitcher WAR:

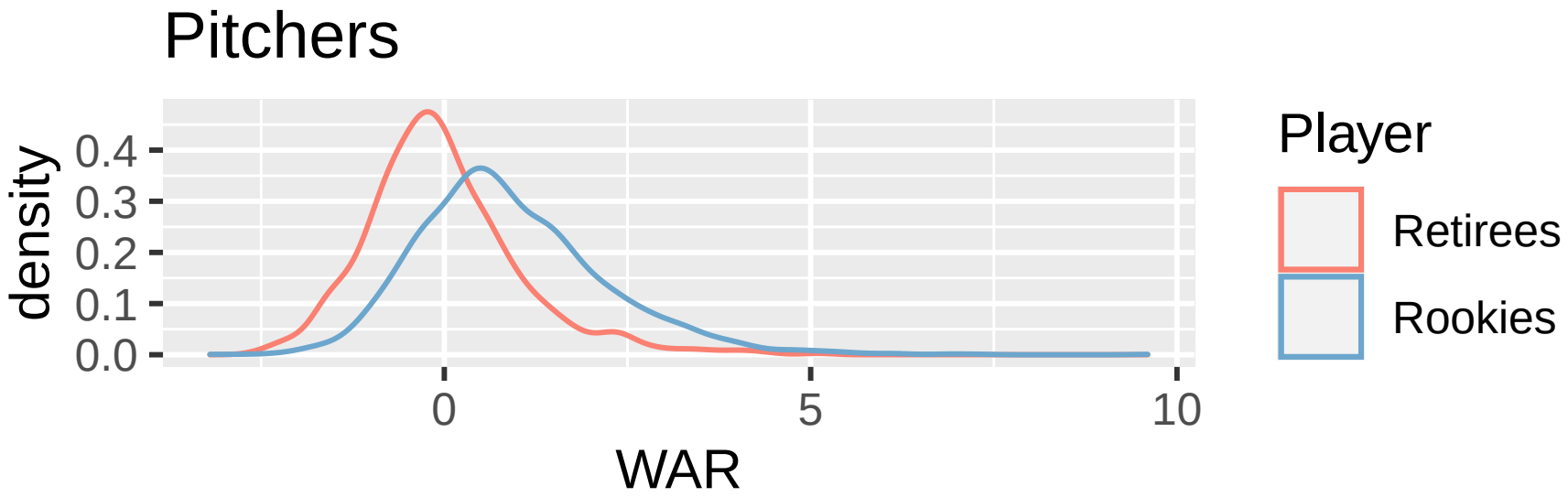
$$(Adjusted\ Average\ Runs\ Allowed - Adjusted\ Player\ Runs\ Allowed) +$$

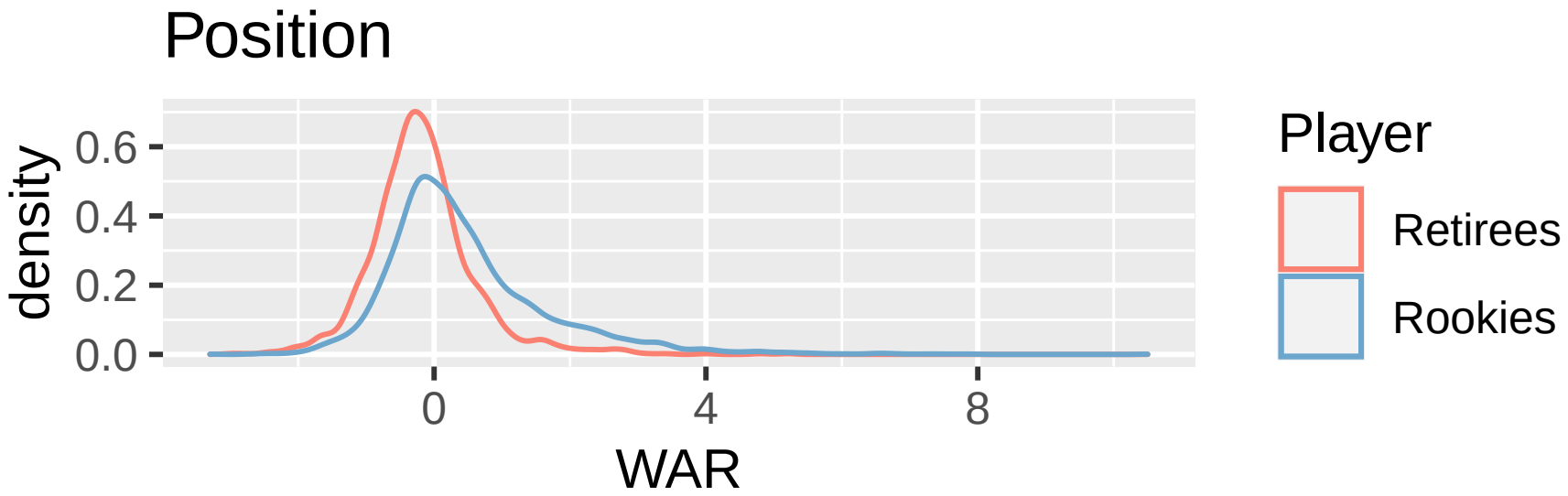
Methods: The Couldabeen Classifier

We want to calculate whether a given retiring player is better than the average rookie replacing him. For a given year Y, we first compute the mean rookie's WAR, call it $Rookie_Y$. Then, we construct the corresponding classifier for "couldabeen" status C of a given retired player p (from the year Y) to be as follows:

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

Visualization: Couldabeen Classification





(No Partition)

A Confounding Variable: *Moneyball* and the Sabermetric Revolution

"Sabermetrics is the search for objective knowledge about baseball through analysis of the statistical record." - from the Society for American Baseball Research, or SABR

A Confounding Variable: *Moneyball* and the Sabermetric Revolution

Timeline:

- 97 Bill James, inventor of term "sabermetrics," publishes first "book": *1977 Baseball Abstract*. It sells 75 copies.
- 97 Billy Beane promoted to general manager of Oakland Athletics. He's ready every *Baseball Abstract* ever published.

(October 2002) Athletics finish season with MLB's best record and second-lowest budget.

(November 2002) Beane declines $12.5 million offer from Boston Red Sox. Boston hires Bill James instead.

(June 2003) Michael Lewis publishes *Moneyball: The Art of Winning an Unfair Game*.

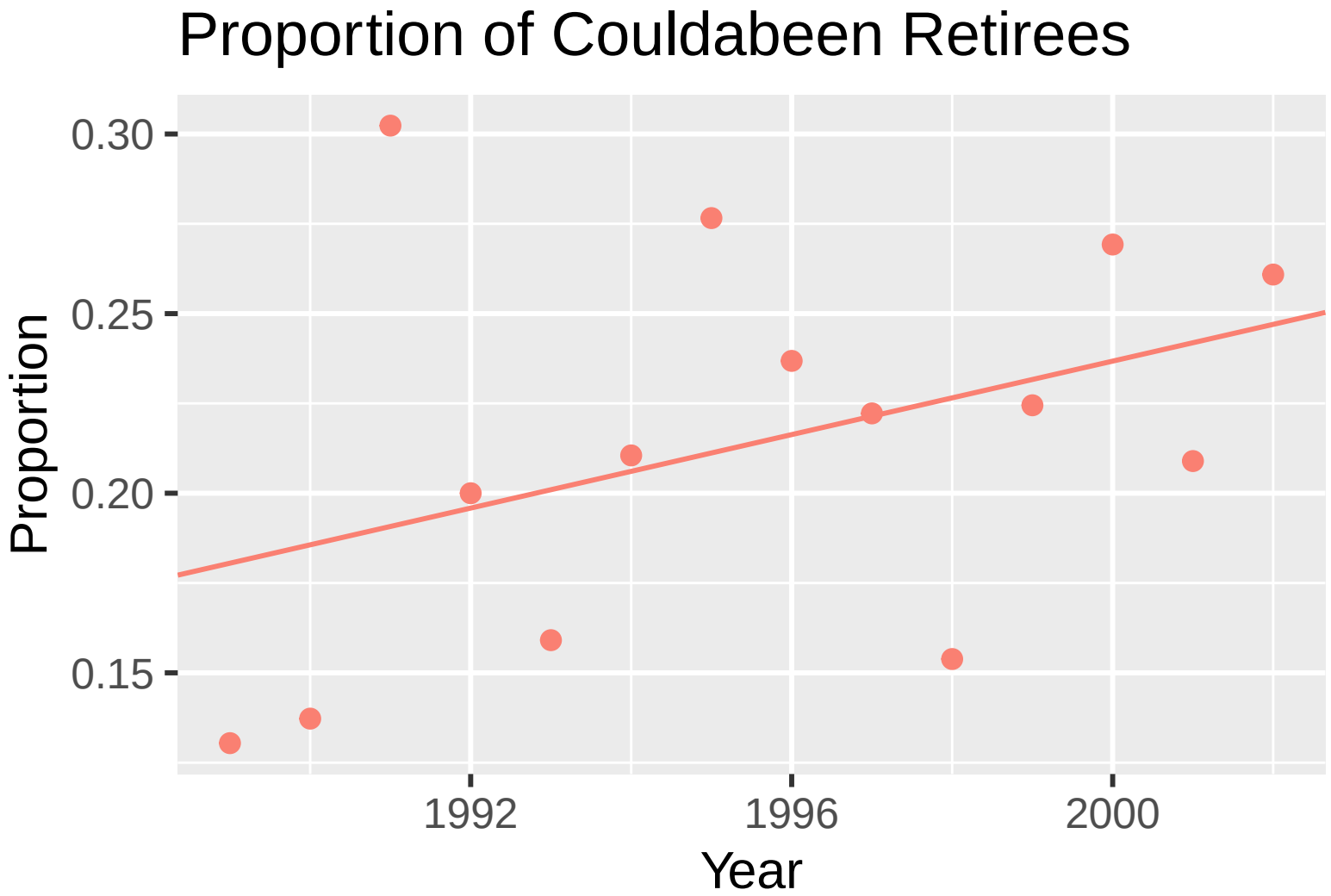
(October 2004) Boston wins their first World Series since 1918.

Methods: Modeling

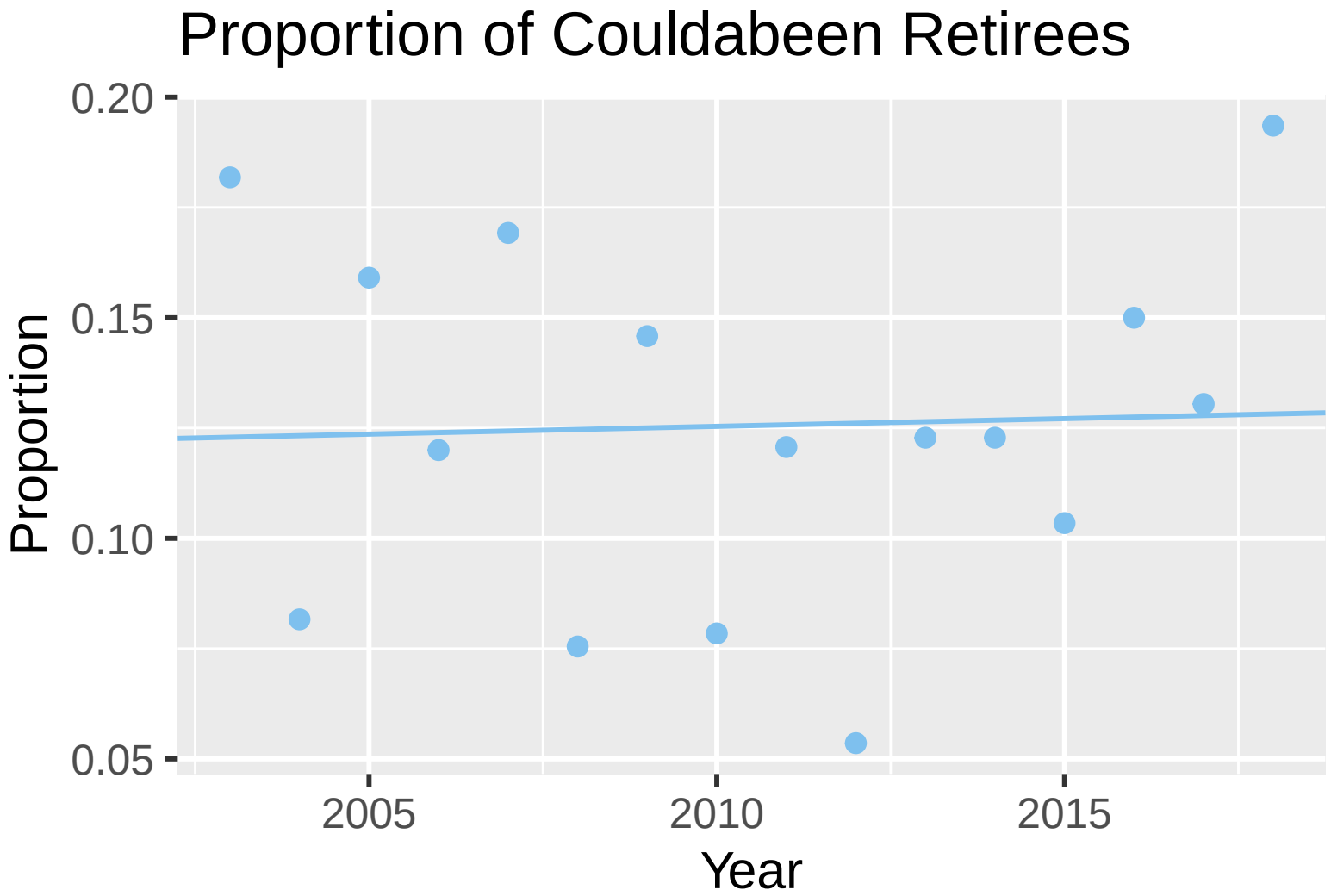
Linear Model: After classifying all retired players, get proportion of "couldbaeen" retirees and call this `prop`.

- As such, we now have 50 data points (for each year), so we run a linear model fitting `Year ~ prop`.
- Because there will always be "couldabeens", we do not expect a large effect size and hence a very significant result.
- If our research hypothesis is correct (that there is an effect), we expect to see a positive coefficient for β_{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.

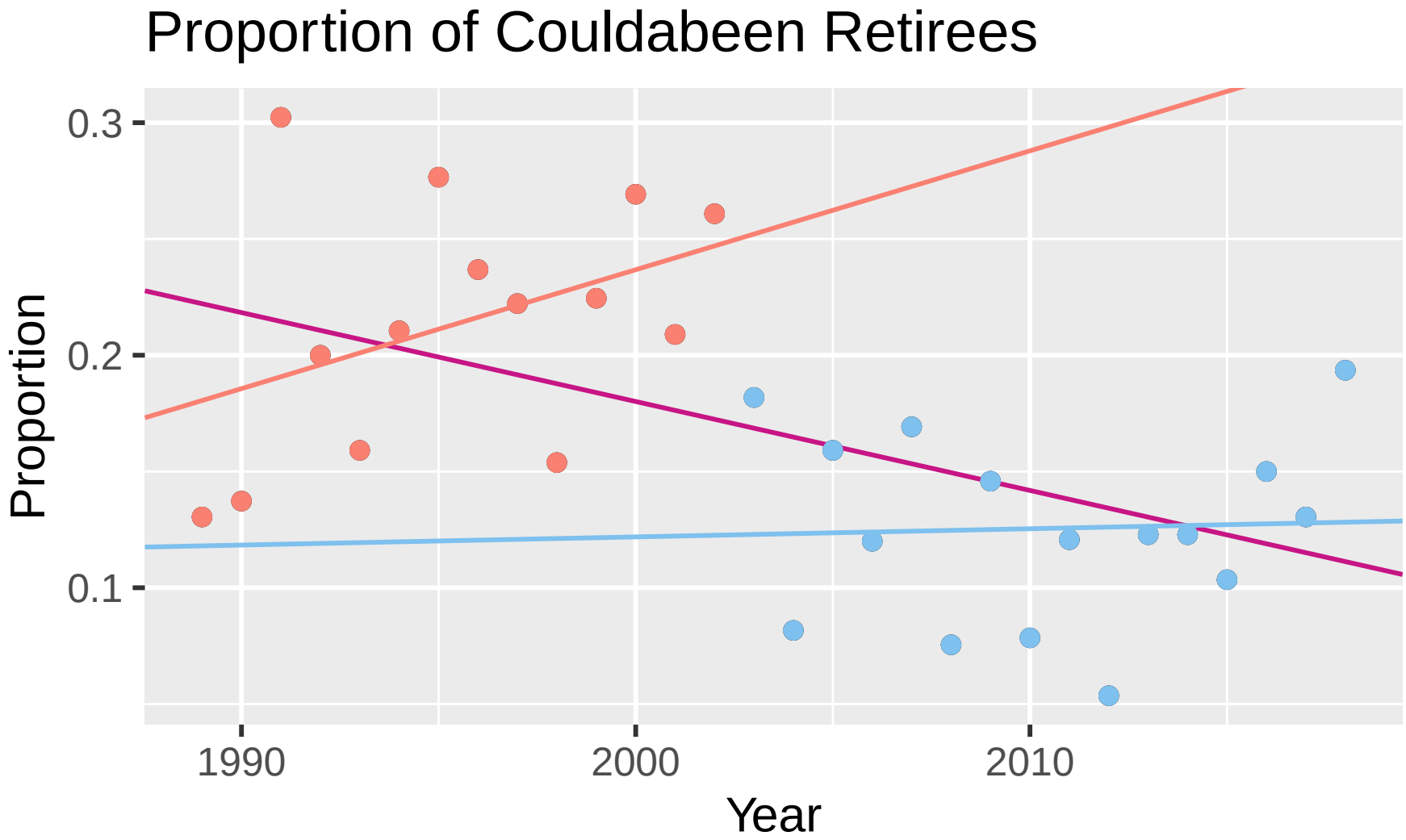
Pre-rule era

- 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 `Year ~ prop`.
- 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.
- In fact, this is *Simpson's Paradox*.

Simpson's Paradox



Threshold Stability

Recall the definition of the Couldabeen classifier:

The General Couldabeen Classifier

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:

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

Threshold Stability

- Analyze the effects of the threshold t on the impact of our β_{Year} fitting the linear models.
- Our inference relies on β_{Year} being positive for any $t \in \mathbb{R}$.
- Obtained a supporting result for β_{Year} for only a particular threshold ($t = 0$) in the post-rule era.
- Varying the threshold tells another story.
- Run a linear model with `Year ~ prop` against a varying threshold $t \in \mathbb{R}$.
- Found that β_{Year} is quite unstable.

Threshold Stability

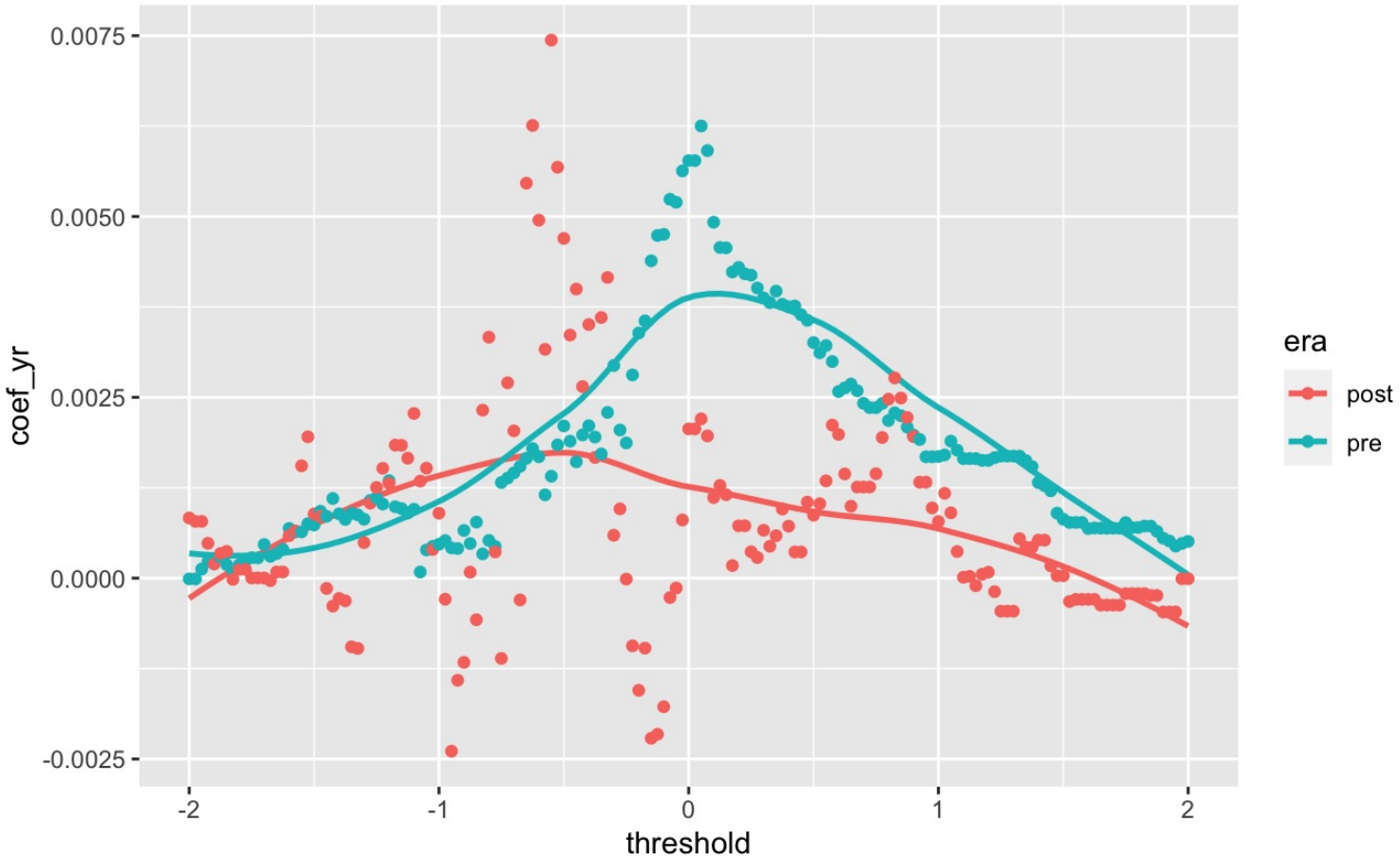


Figure 1: Parameters of the Linear Model against varying threshold for couldabeen status

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 (corrosponding 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.

Conclusion

To summarize: - No correlation between year and proportion of couldabeens across entire 50-year span. - Positive but non-significant relationship between year and proportion of couldabeens when partitioned into pre- and post-*Moneyball* eras. - Strongly significant negative relationship between publication of *Moneyball* and proportion of couldabeens. - Somewhat significant relationship between labor share and proportion of couldabeens in post-*Moneyball* era.

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/