Statistical Learning: Project Presentation

G. Dunlavey, W. Ren, A. Taqi

Research Question

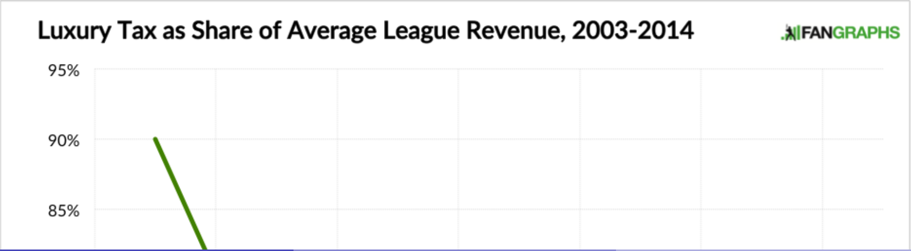
The MLB's "luxury tax," implemented in the 2003 Collective Bargaining Agreement (CBA), is a rule penalizing franchises whose team payroll for a given year exceeds an agreed threshold. This project attempts to test the tax's effect on quality of play by studying the number of premature retirees (referred 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 linked the continuing decline in MLB labor share to the 2003 CBA (Bradbury, 2019).

$$Labor\ Share\,(Y) = \frac{Total\ MLB\ Revenue\,(Y)}{Total\ MLB\ Player\ Payroll\,(Y)}$$

::: columns :::: column



Theory

How might the "luxury tax" increase the number of premature 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: Teams will replace good-but-not-Mike-Trout veterans with marginally inferior rookies to stay below salary threshold. This will lead to a rise in premature retirements.

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.

The WAR statistic is calculated differently for position players and pitchers.

Position WAR

Position WAR:

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

Where:

$$\begin{aligned} Player\ Runs = Batting\ Runs + Baserunning\ Runs + Double\ Play\ Runs + \\ Fielding\ Runs + Positional\ Adjustment \end{aligned}$$

Pitcher WAR

Pitcher WAR:

Abrev.	Meaning
AARA	Adjusted Average Runs Allowed
APRA	Adjusted Player Runs Allowed
ARRA	Adjusted Replacement Runs Allowed

Table 1: Key

$$WAR = \frac{(AARA - APRA) + (ARRA - AARA)}{Game\ Runs\ to\ Wins\ Estimator}$$

WAR

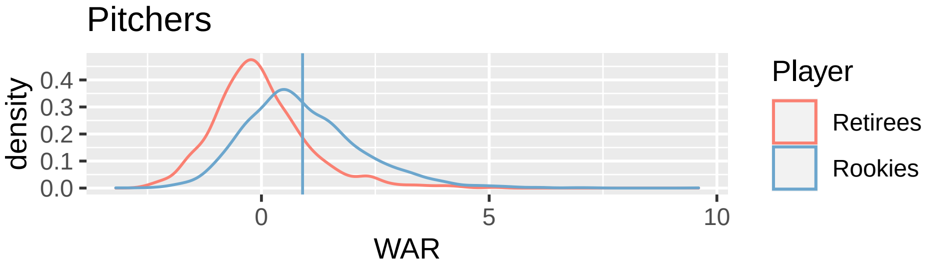
However, the "replacement-level player" used in WAR is an estimate for the average *midseason* replacement. Being better than a midseason replacement does not necessarily make you better than the generation of rookies *actually* replacing you. If we want to argue that a given retiree, on merit, *should have kept playing*, we require a higher standard.

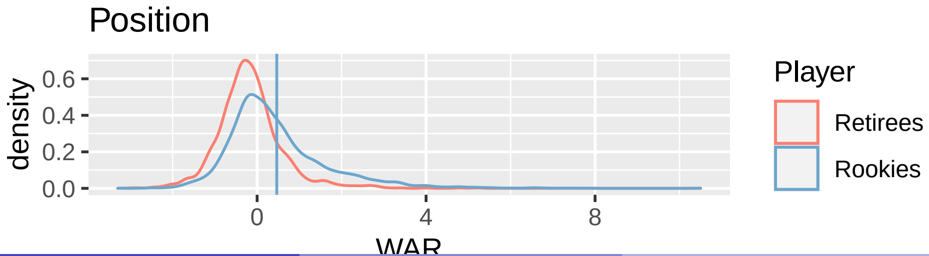
Methods: The Couldabeen Classifier

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:

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

Visualization: Couldabeen Classification

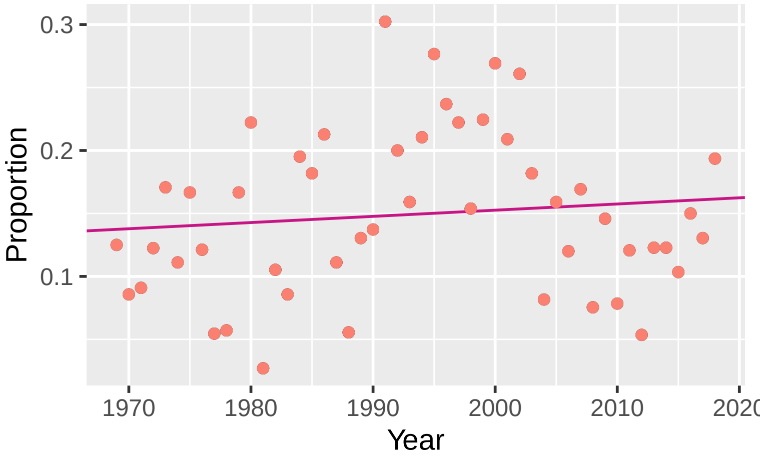




Methods: Modeling

Linear Model: After classifying all retired players, we divide the number of "couldabeen" retirees by the total number of retirements for each year ('prop'). We then run a linear model fitting `Year ~ prop`.

Linear Model: Year



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

The Sabermetric Revolution

1977: Bill James, inventor of term "sabermetrics," publishes first "book": *1977 Baseball Abstract*. It sells 75 copies.

1997: Billy Beane promoted to general manager of Oakland Athletics. He has 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. Red Sox hire Bill James instead.

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

October 2004: Red Sox win their first World Series since 1918.

2006: *Time* lists Bill James among "100 Most Influential People in the World." Nearly every MLB franchise employs a sabermetrics team.

The Sabermetric Revolution

> *"[Presidential politics] reminded me of baseball, when you see the same recycled clichés and conventional wisdoms over and over again, some of which isn't very wise." - Former *Baseball Prospectus* partner Nate Silver, on why he founded *538**

Moneyball and the “sabermetric revolution” changed the face of statistics as a profession. But much more importantly, it changed baseball. We acknowledge this with a binary classifier, stating whether a given observation occurred before or after the publication of *Moneyball*.

Testing the Significance of *Moneyball*

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, defined as follows:

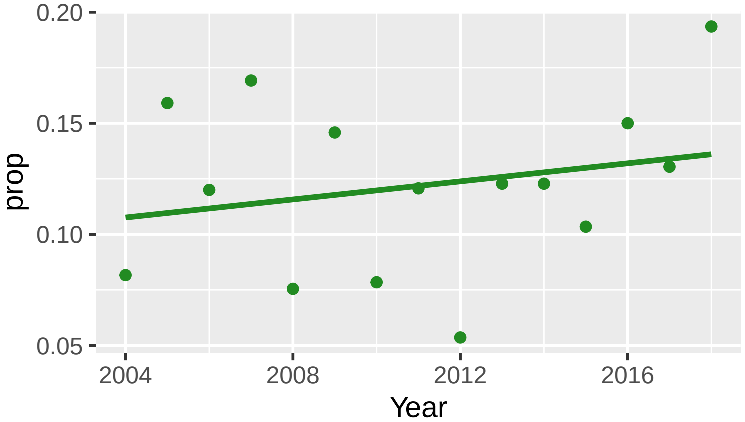
$$\text{postMoneyball}(Y) = \begin{cases} \textit{True}, Y > 2003 \\ \textit{False}, Y \leq 2003 \end{cases}$$

- We fit a model `Year ~ postMoneyball` and perform *LSS*.

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

Linear Model on Year (Post-*Moneyball*)

Linear Model: After partitioning the dataset, we run another linear model fitting `Year ~ prop`.



Linear Model on Year

Pre-*Moneyball* era

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

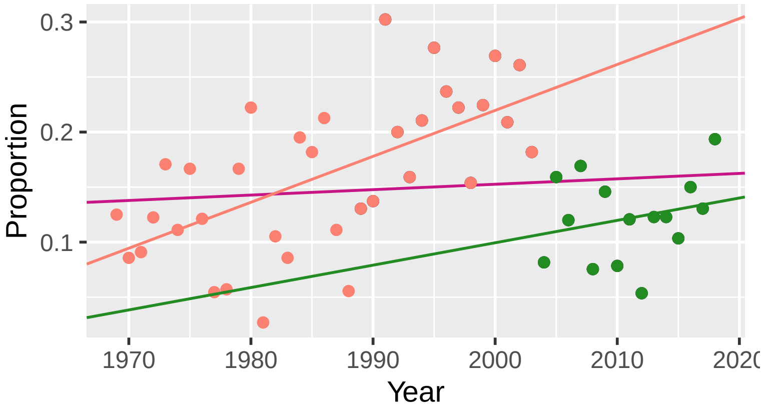
Post-*Moneyball* era

- $\beta_{Year} = 0.002034$.
- $\beta_{Year} > 0$ 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.

Simpson's Paradox

- Partitioning and fitting linear model with `Year ~ prop` yields $\beta_{Year} > 0$ in both partitions.
- However, if we do not make the partition, we find that $\beta_{Year} \approx 0$.
- This is evidence of *Simpson's Paradox*.



What next?

So far, we have fit some linear models seeing the effect sizes of `Year` on the response `prop`.

- We know that we must partition the dataset into pre-*Moneyball* and post-*Moneyball* due to the confounding variable.
- From the linear model on the pre-*Moneyball* era, we find that $\beta_{Year} > 0$, suggesting that it is *certainly possible* that the rule had an effect on the game.

Nonetheless, $p = 0.3$ is statistically insignificant. We need a new approach.

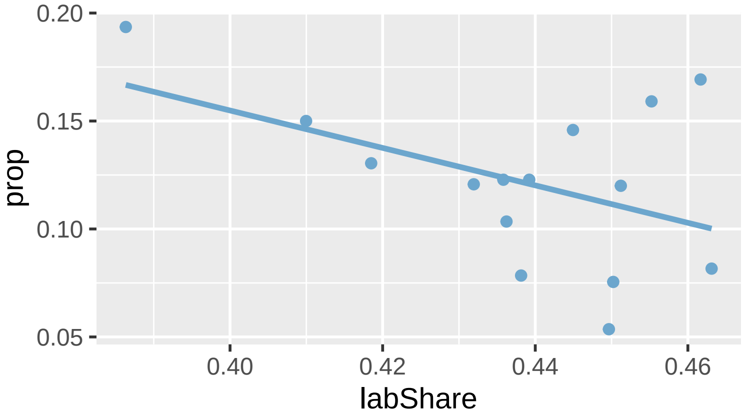
A New Approach: Labor Share

We know that labor share has declined steadily since the 2003 CBA, and Bradbury (2019) has established a causal relationship. Recall:

$$\textit{Labor Share}\,(Y) = \frac{\textit{Total MLB Revenue}\,(Y)}{\textit{Total MLB Player Payroll}\,(Y)}$$

Linear Model: Labor Share

So how is `laborShare` affecting `prop`? We lag 'laborShare' by one year (since failing to sign a couldabeen will show up in *next* year's payroll) and then fit `prop ~ laborShare`:



Results

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	0.502	0.204	2.464	0.028	0.062	0.941
labShare	-0.867	0.464	-1.868	0.084	-1.870	0.136

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

Conclusion

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.

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*.