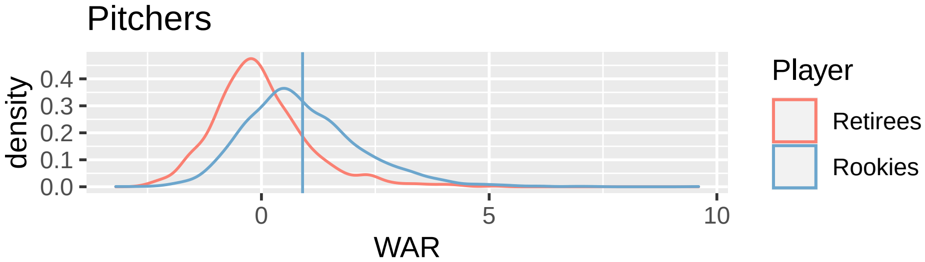
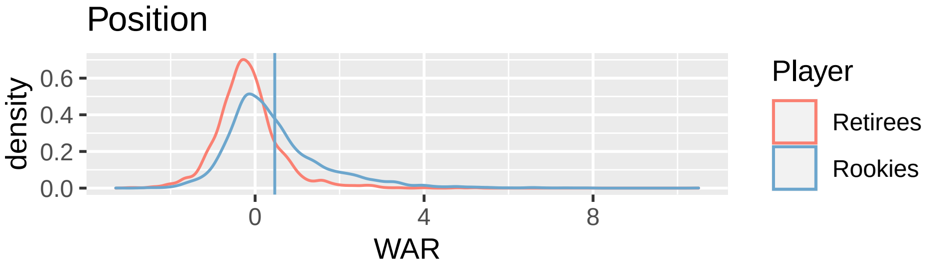
Statistical Learning: Project Presentation

G. Dunlavey, W. Ren, A. Taqi

Visualization: Couldabeen Classification



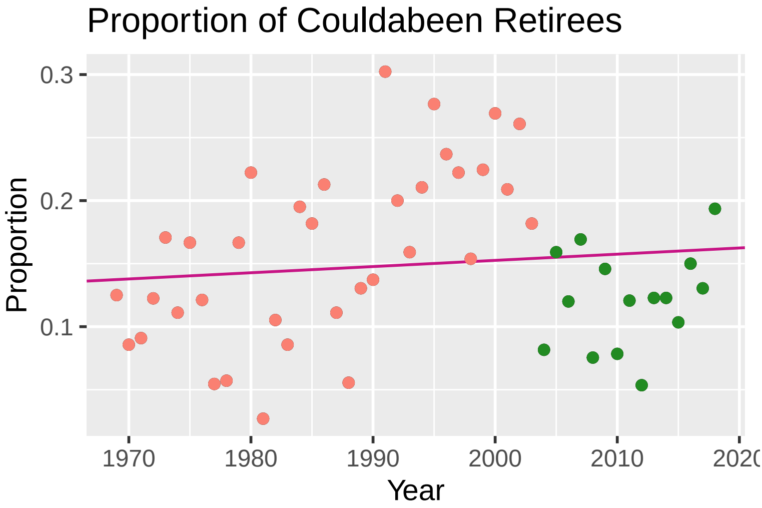


Methods: Modeling

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: Year



Hypothesis Test: Why Split the Data?

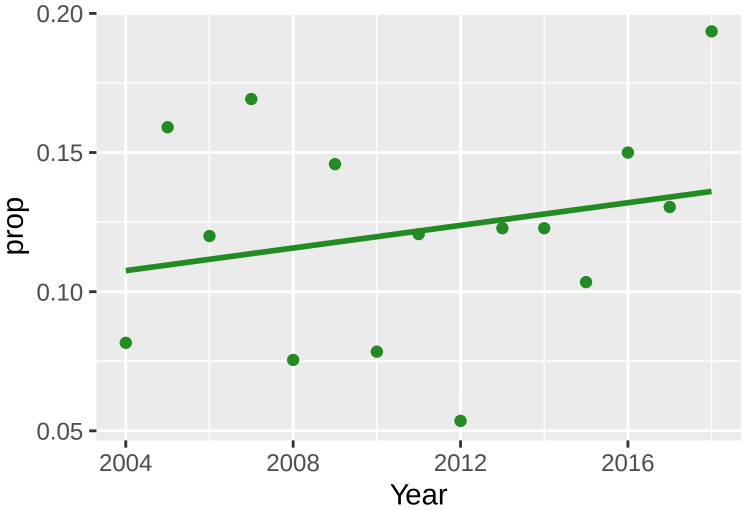
Since the release of Moneyball in 2003 seems to be an important **confounding variable**, we perform a hypothesis test on the `postMoneyball` classifier of a year Y, defined as follows:

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

- 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 on Year

Post-Moneyball era

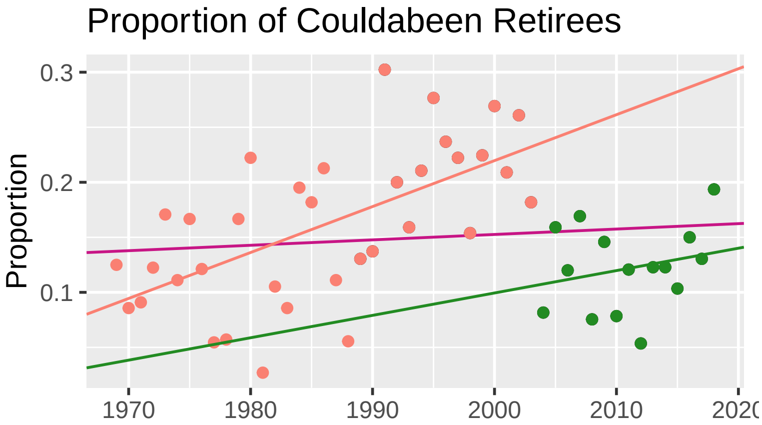
- $\beta_{Year} = 0.002034$.
- $\beta_{Year} > 0$ supports the hypothesis that there is an increasing rate of couldabeens since the luxury tax.
- β_{Year} is not statistically with a high p-value of 0.398.

Pre-Moneyball era

- $\beta_{Year} = 0.003357$.
- β_{Year} is not statistically significant with a high p-value of $p = 0.300$.

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 in fact *Simpson's Paradox*.



What next?

So far, we have fit some linear models seeing the effect sizes of `Year` on the response `prop`. But `Year` is no mighty predictor...

- 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...
- Resample, or bring in new data.

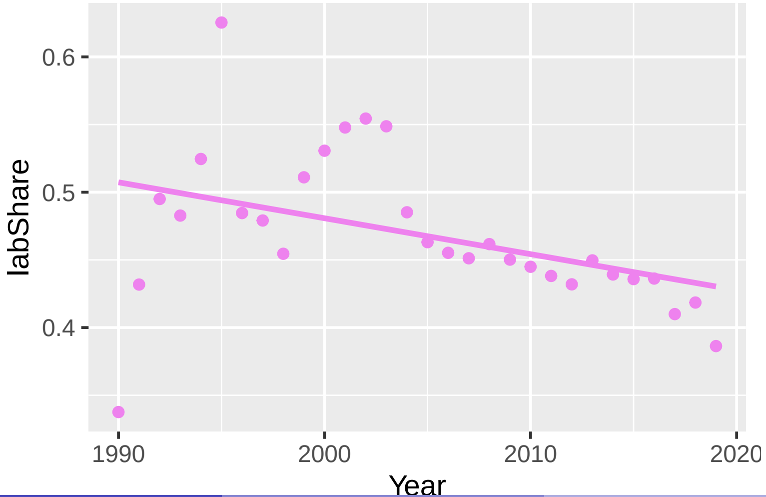
A New Approach: Labor Share

Since our couldabeen proportions are sorted by year, we needed year-sorted data. So, we used the Total Revenue and Total Payroll.

$$laborShare(Y) = \frac{totalPayroll(Y)}{totalRevenue(Y)}$$

A New Approach: Labor Share

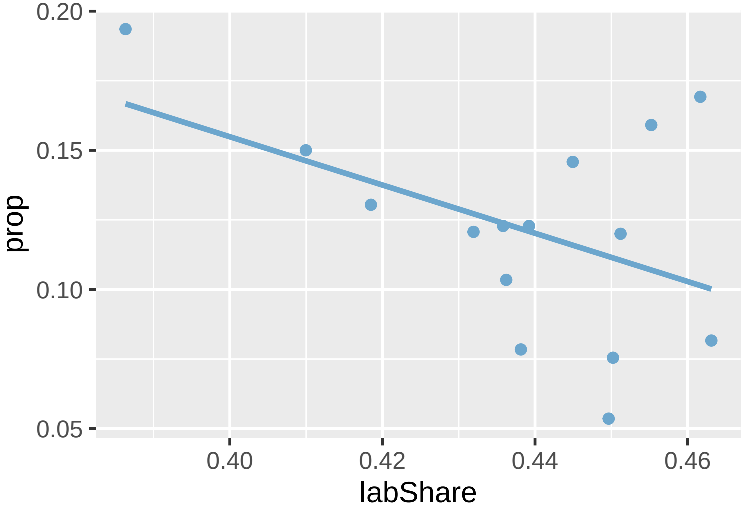
- Revenues are increasing, but fraction of revenue that goes to payroll has been decreasing. . .



A New Approach: Labor Share

- In fact, study by [Bradbury] shows labor share is decreasing due to the rule. . .
- The rule means more veterans are playing, and Sabermetrics means choosing rookies is cheaper.
- What is left is to analyze `prop ~ laborShare`.

Linear Model: Labor Share



Linear Model: Labor Share

- Alas, we find $\beta_{LabShare} = -0.867$ with a p-value of $p = 0.084$.
- Labor share is indeed **negatively** correlated to proportion of couldabeens.
- Research paper by [Bradbury] conveys that labor share is indeed diminshing due to the *luxury tax*.

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

Labor Share: Conclusions

- Luxury Tax $\implies$ Declining Labor Shares [Bradbury]
- Declining Labor Shares $\implies$ Increasing Couldabeens [Linear Model]
- Luxury Tax $\implies$ Increasing Couldabeens [Conclusion]

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/