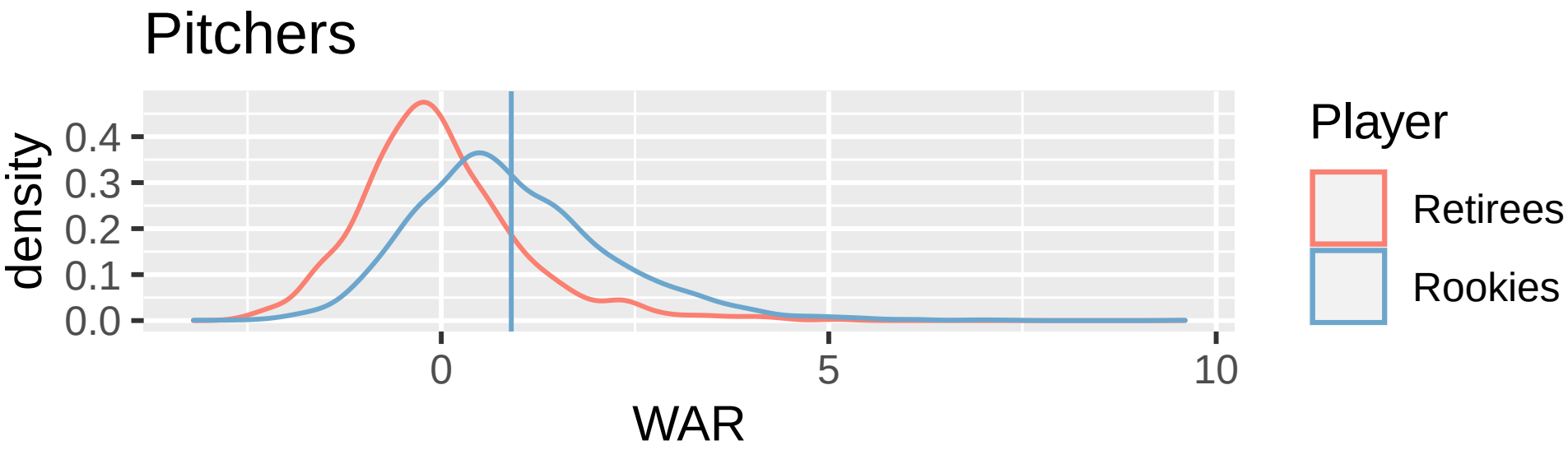
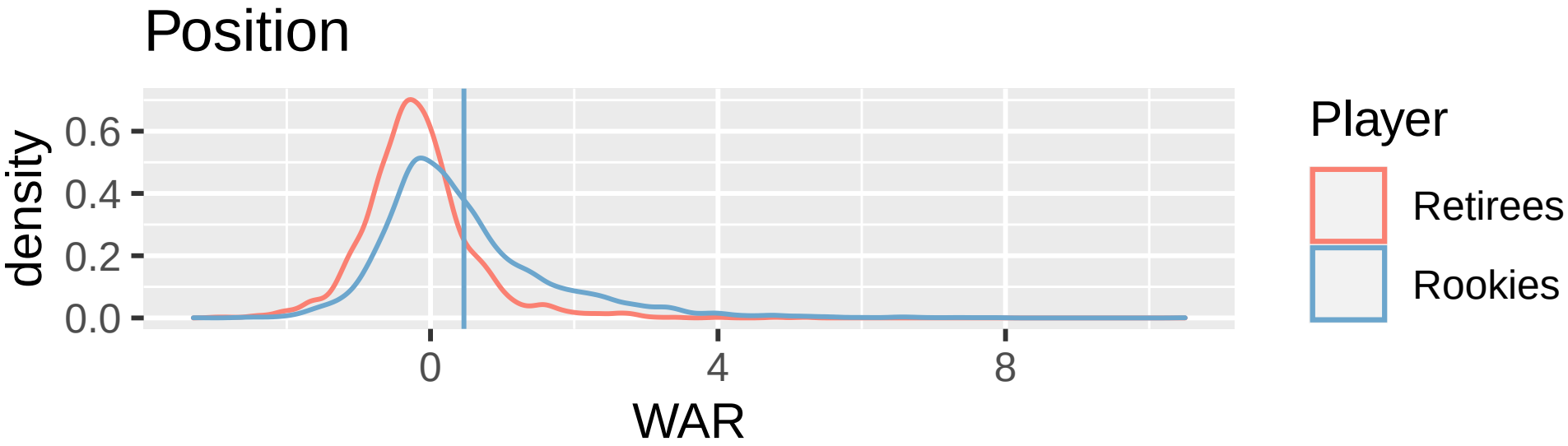
Statistical Learning: Project Presentation

G. Dunlavey, W. Ren, A. Taqi

Visualization: Couldabeen Classification



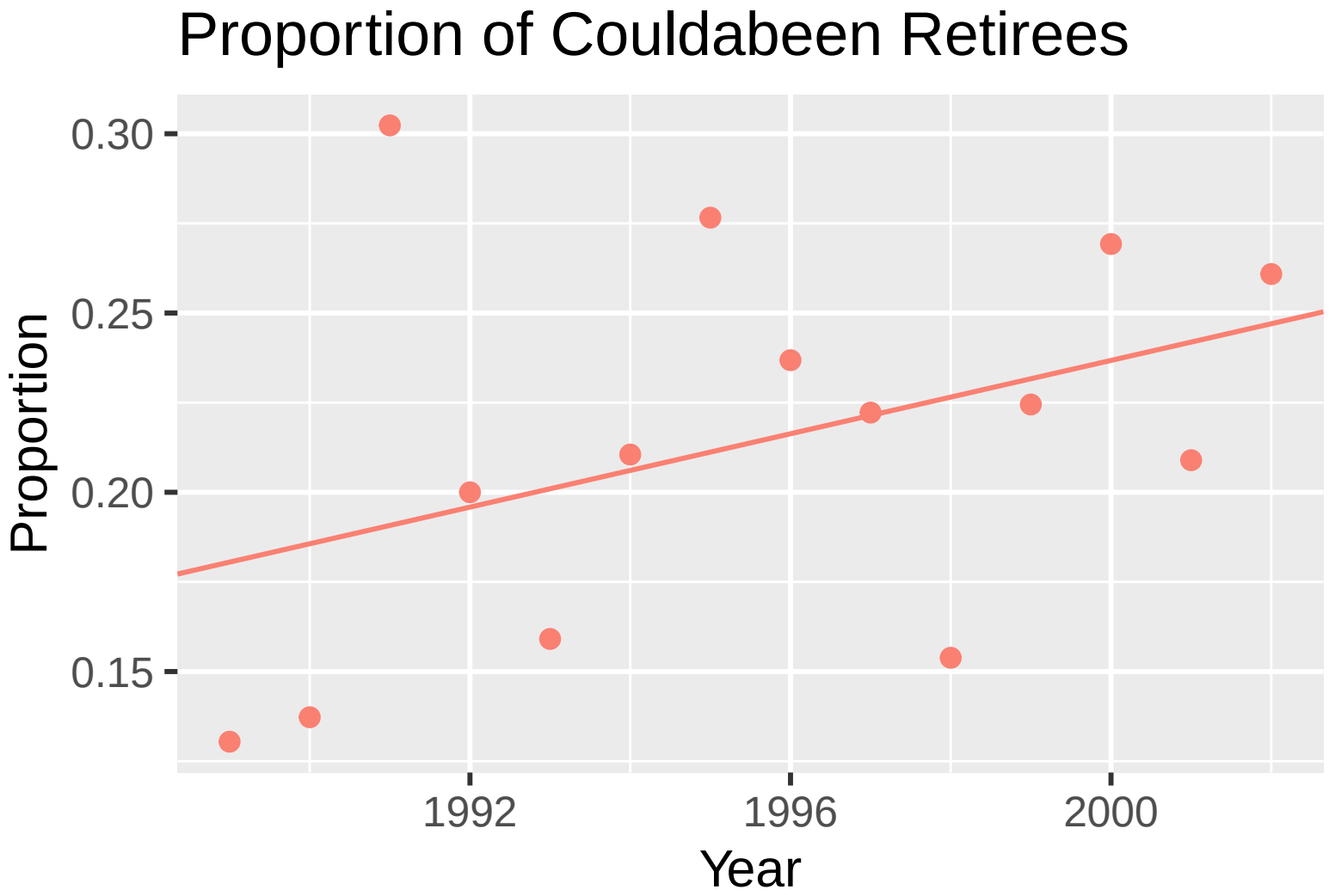


Methods: Modeling

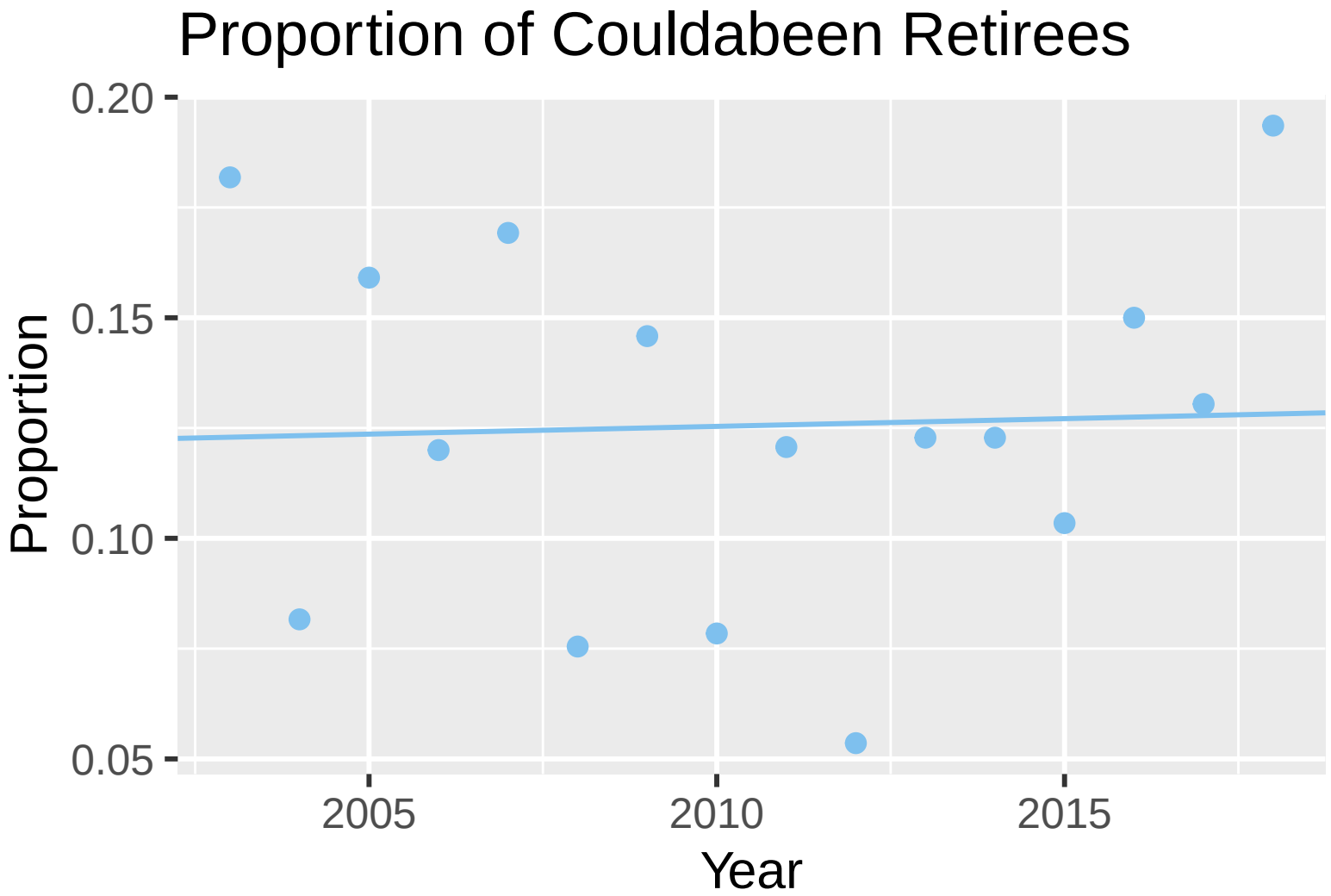
Linear Model: After classifying all retired players, get proportion of "couldbaeen" retirees and call this `prop`.

- 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: 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.

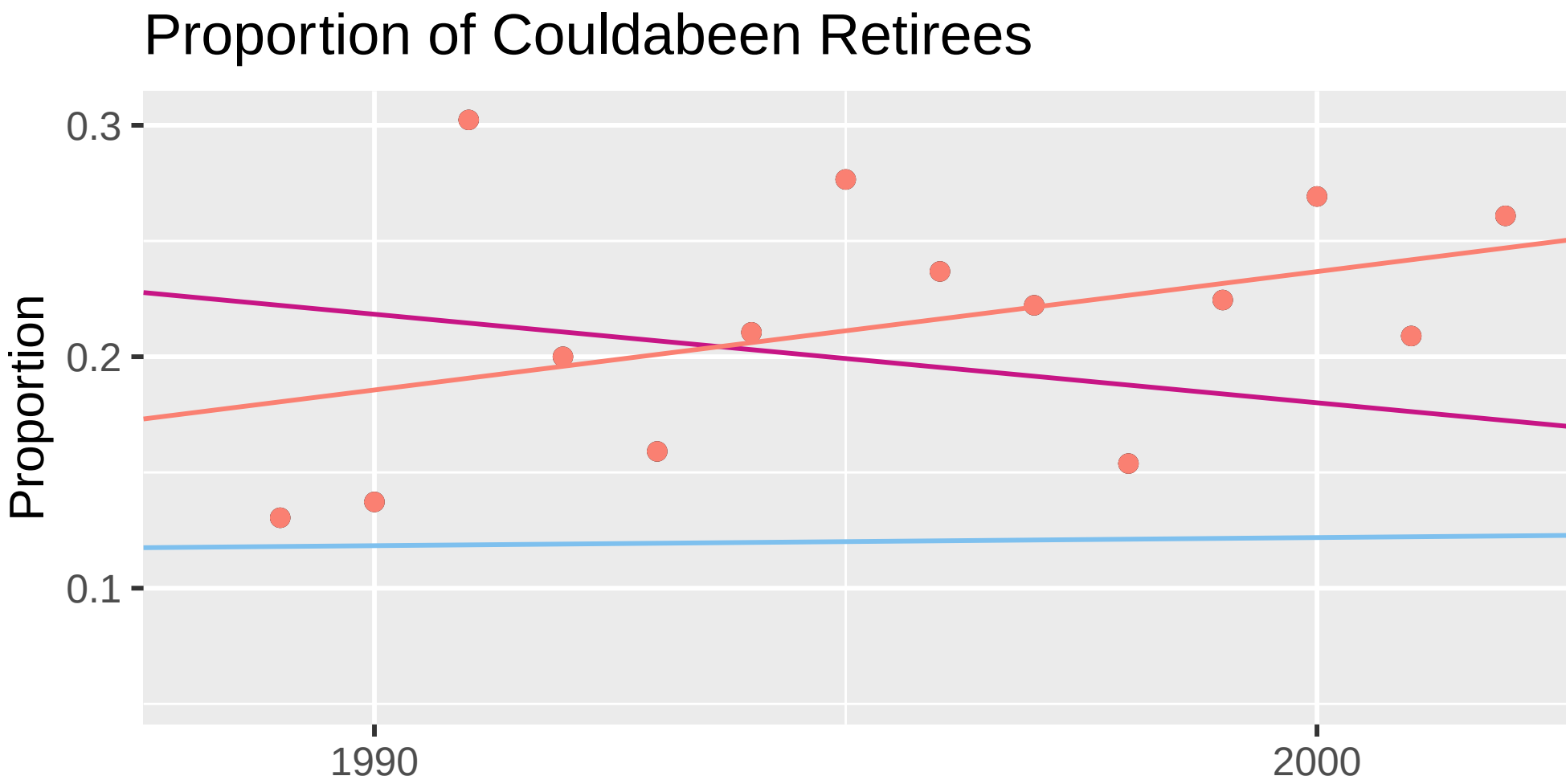
Pre-rule era

- 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

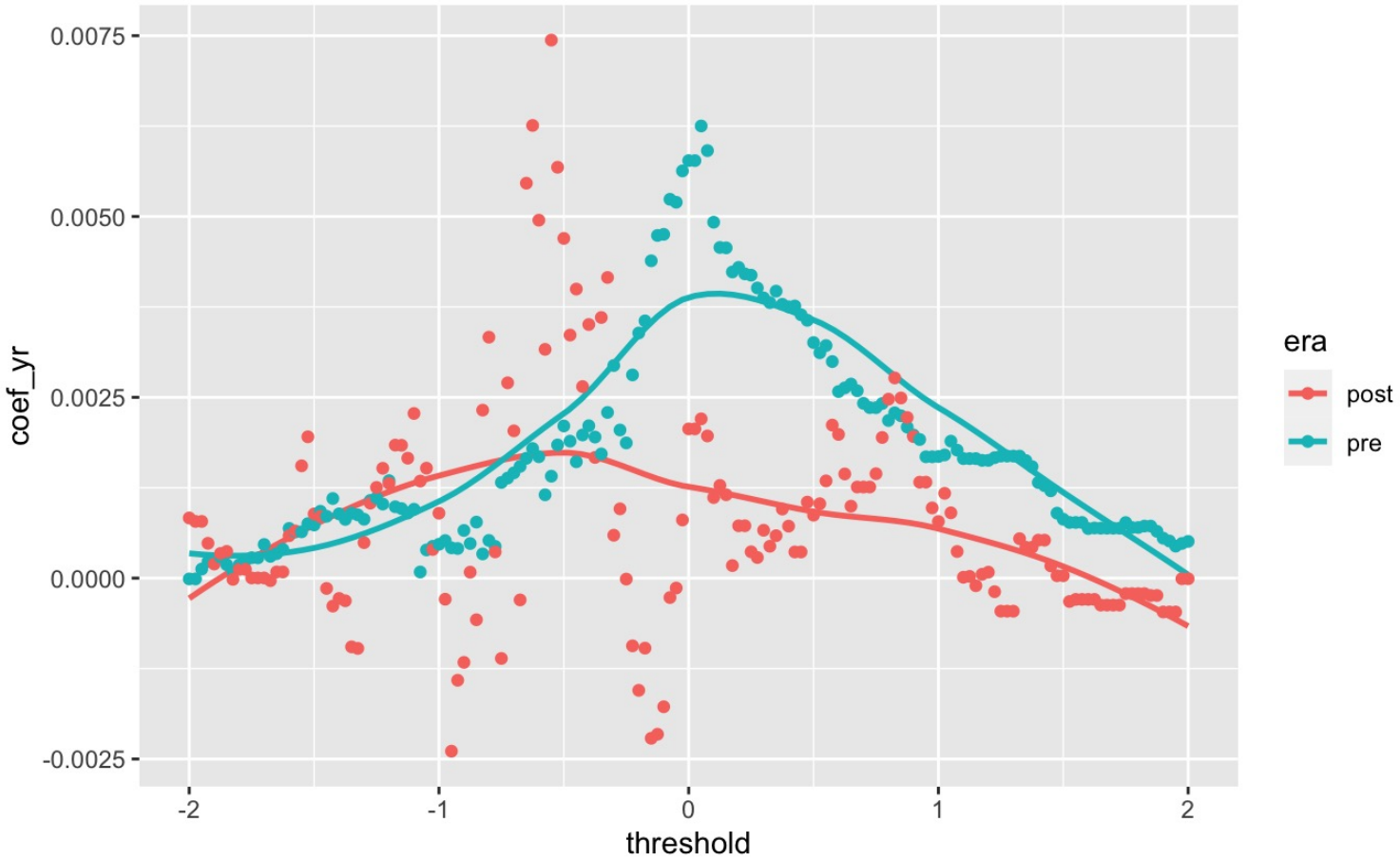


Figure 1: Parameters of the Linear Model against varying threshold for couldabeen status

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.

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/