Econometric Analysis of Pitch Dynamics and Batter Stance in Baseball

SABERMETRICS, SCOUTING AND THE SCIENCE OF BASEBALL

Olubayode Ebenezer (Master's Program in Sports Data Analytics)[1]
Advisor: Dr. Daniel Larson, Phd.[2]

The UNIVERSITY of OKLAHOMA

INTRODUCTION

- Imagine a tense moment in the World Series, where the batter faces a full count. The crowd holds its breath as the pitcher winds up.
- What factors do you think are running through the batter's mind when deciding whether to swing at the pitch?
- Understanding these factors is critical for developing effective strategies in baseball.
- Traditional statistics lead to false cause fallacy, mistaking correlation for causation for those factors.
- Rohrer use graphical causal models to distinguish correlation from causation and enhance causal inference (Rohrer, 2018, p. 3).

INTRODUCTION

Why Causal Models ?

Identify Cause-and-Effect Relationships

- identifying & quantifying the strength & direction of causal relationships
- effective decision-making (Rohrer, 2018, p. 3).

Avoiding the False Cause Fallacy

- Observing two trends together does not imply one causes the other.
- Rigorous testing for causality is essential to avoid misleading results.

Bayesian Causal Models

- distinguishing correlation from causation using prior knowledge, essential for reliable data analysis (Rohrer, 2018, p. 8).

Causal Inference

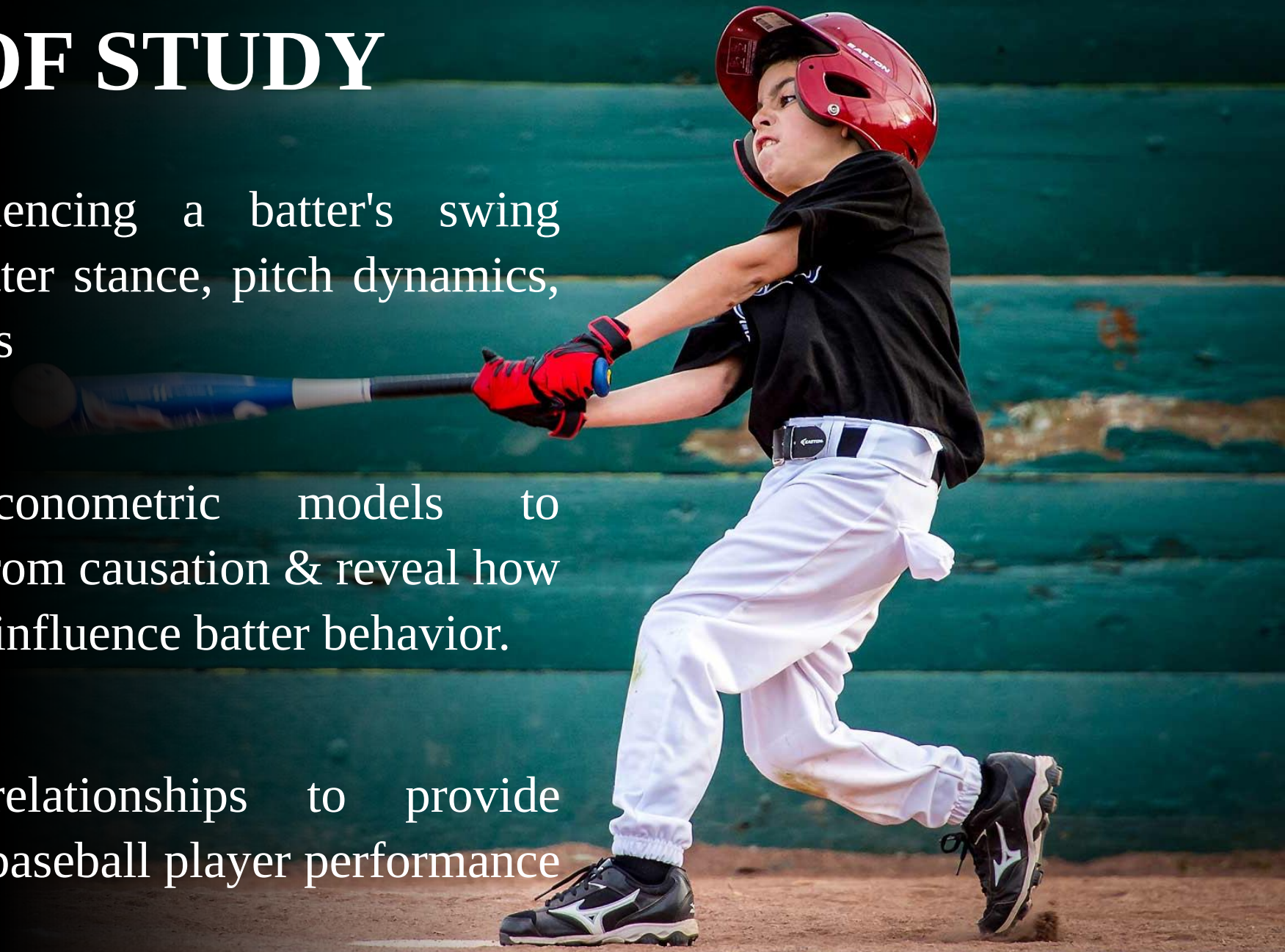
- for modeling & testing hypotheses about causal impacts, which is critical for effective policy and decision-making (Pearl, 2009).

PURPOSE OF STUDY

- Analyze factors influencing a batter's swing decision, focusing on batter stance, pitch dynamics, and confounding variables

- Using Bayesian econometric models to differentiate correlation from causation & reveal how pitchers can strategically influence batter behavior.

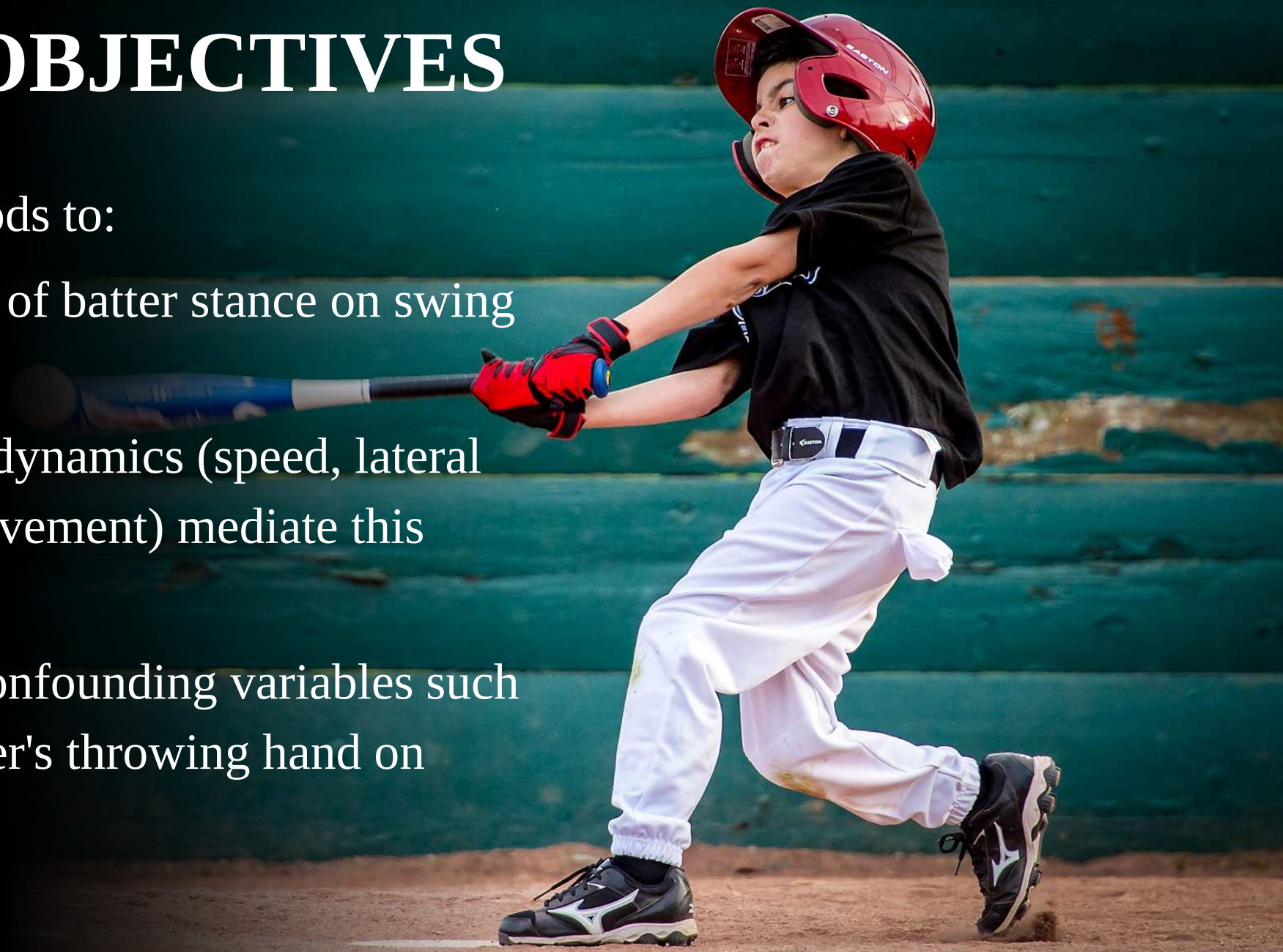
- Understand causal relationships to provide strategies for optimizing baseball player performance and decision-making.



SPECIFIC OBJECTIVES

Bayesian statistical methods to:

1. Isolate the direct effect of batter stance on swing likelihood.
2. Understand how pitch dynamics (speed, lateral movement, vertical movement) mediate this relationship.
3. Assess the impact of confounding variables such as pitch type and pitcher's throwing hand on these relationships.



VARIABLES

Dependent Variables:

- **Swing Likelihood (Y)**

Independent Variables:

- **Batter Stance (Stand)**
- **Release Speed**
- **Lateral Movement (pfx_x)**
- **Vertical Movement (pfx_z)**
- **Pitch Type**
- **Pitcher's Throwing Hand (p_throws)**

RESEARCH QUESTIONS

Main Research Question:

- How does the batter's stance (left or right) influence the likelihood of swinging, and how is this mediated by the release speed and pitch trajectory?

Specific Research Questions:

1. Direct effect of the batter’s stance on decision to swing?
2. How does the lateral movement (pfx_x) of the pitch, influenced by batter stance, affect swing likelihood, considering the pitch type as a confounding variable?
3. How does the vertical movement (pfx_z) of the pitch, influenced by batter stance, affect swing likelihood, considering the pitch type as a confounding variable?
4. Is there an interaction effect between pitch type and batter stance on swing likelihood, influenced by the pitcher's throwing hand?
5. How do pitch type, release speed, and batter stance interact to influence the likelihood of a swing in baseball?

LITERATURE REVIEW

Recent Bayesian Studies in Baseball Decision-Making

- Studies shows batters integrate prior information with observational data to navigate pitch uncertainty (Brantley & Kording, 2022).

- Bayesian framework developed to enhance swing decision-making, accounting for contextual influences (Yee & Deshpande, 2023)

RESEARCH GAPS IDENTIFIED

- need for more robust analytical techniques that differentiate between correlation and causation, particularly in the context of swinging in baseball.

- lack of comprehensive causal models that account for various confounding factors & interactions affecting swing likelihood in baseball.

Justification for the Study:

Improving Strategic Decision-Making:

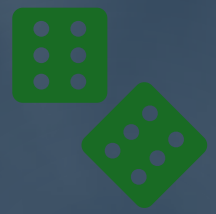
- aid coaches and players in developing better strategies
- enhancing performance on the field.

RESEARCH METHOD

1. Define Generative Model:

Identify all relevant variables affecting the swing decision.

Formulate the logistic regression model to represent the probability of a swing.



2. Define Estimands:

Describe association btw the factors

Swing probability given the input variables.

3. Design Estimator:

Define the likelihood function based on the log. Regres. model.

Choose appropriate priors for the coefficients to produce estimands

.

RESEARCH METHOD

4. Test Estimator:

- Generate synthetic data using the generative model to test the estimator
- Fit the Bayesian model to the synthetic data and evaluate the estimation accuracy.

5. Analyze and Summarize:

- Fit the Bayesian model to the actual swing data.
- Use (Markov Chain Monte Carlo) to sample from the posterior distributions in estimating the distribution of the model's parameters.
- Summarize the results and interpret the impact of each predictor on the swing decision.

DATA COLLECTION

Data Collection Procedures

- Source: From the Miami Marlins, spanning three seasons (2 million pitches)

Mitigating Bias in Data Analysis

Balanced Sample Sizes:

- Adjust dataset to have equal numbers of right and left stances.

Control for Confounding Variables:

- Include pitcher's hand (left or right) as a covariate to account for pitcher-batter matchups.

DATA ANALYSIS

Statistical Software & Libraries: Python, PyMC4, Pandas etc.

Analytical Techniques:

- Use **Bayesian models** to estimate probabilities and infer causal relationships between variables.
- **Logistic regression** to model swing likelihood as a function of batter stance, pitch dynamics, and other covariates.
- **Causal Inference:** Directed Acyclic Graphs **(DAGs)** to understand the causal structure of the data.

Generative Model

- Outcome: Swing decision (0 for no swing, 1 for swing)
- variables :
  - X1: Pitch speed
  - X2: Pitch type (encoded numerically)
  - X3: pfx_x
  - X4: pfx_z
  - X5: p_throws
- The generative model for the swing decision Y defined using a logistic regression framework:
- $P(Y = 1 \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5)}}$
- $P(Y = 1 \mid X)$ = Prob of swing decision given predictors X
- Right side = Log func. mapping linear predictor to prob.
- β is impact of each predictor on swing likelihood.

Generative Model

The logistic function $\frac{1}{1+e^{-z}}$

Large Positive Predictor

- e^{-z} approaches zero.
- Probability approaches 1 (high likelihood of swing)

Large Negative Predictor:

- e^{-z} becomes very large.
- Probability approaches 0 (low likelihood of swing)

Predictor Near Zero:

- Logistic function yields probabilities around 0.5.
- Indicates uncertainty in swing decision

Define Estimands & Estimator

Estimands:

- Posterior distributions of the coefficients $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$
- The probability of a swing decision given the input variables $P(Y = 1 \mid X)$

Design an Estimator:

- estimate the coefficients using BLR . The priors for the coefficients chosen based on expert knowledge.
- $\beta_i \sim (0,10)$for $i = 0,1,2,3,4,5$

Likelihood:

- $P(Y \mid \beta, X) = \left(\frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5)}}\right)$

Posterior:

- $P(\beta \mid Y, X) \propto P(Y \mid \beta, X)P(\beta)$

Posterior Distribution Basics

Posterior Distribution:

- Represents updated beliefs about parameters after observing data.
- Derived using Bayes' theorem:
- $P(\beta \mid Y, X) \propto P(Y \mid \beta, X)P(\beta)$

Components:

Prior $P(\beta)$: Initial beliefs about parameters; e.g., βi∼Normal(0,10).

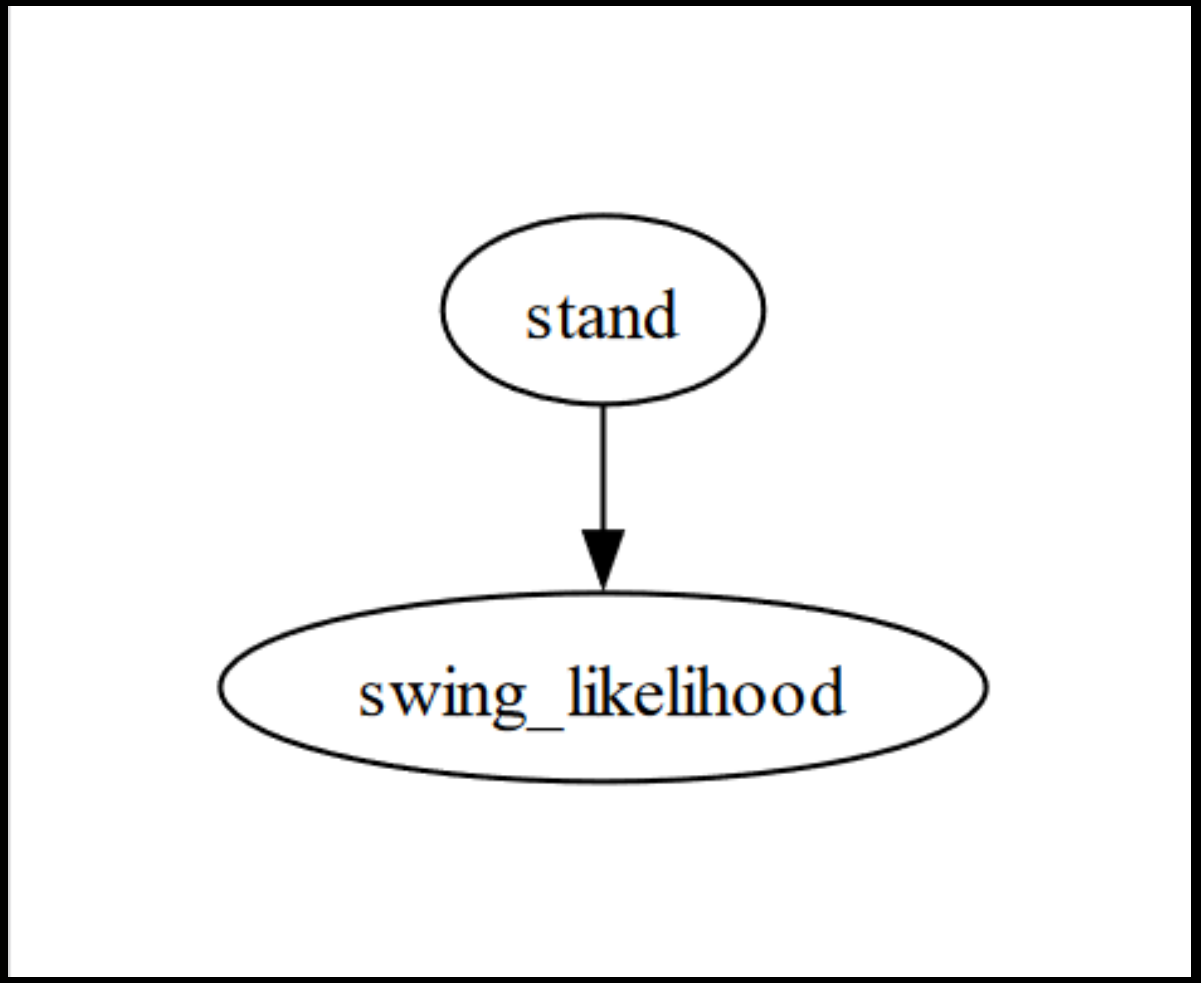
Probability of data given parameters.

Likelihood Expression:

- For logistic regression;
  - $P(Y_i = 1 \mid X_i, \beta) = \frac{1}{1+\exp(-X_i\beta)}$

ANALYSIS AND RESULTS:

1. Direct effect of the batter's stance on decision to swing?



Drawing the DAG

Result

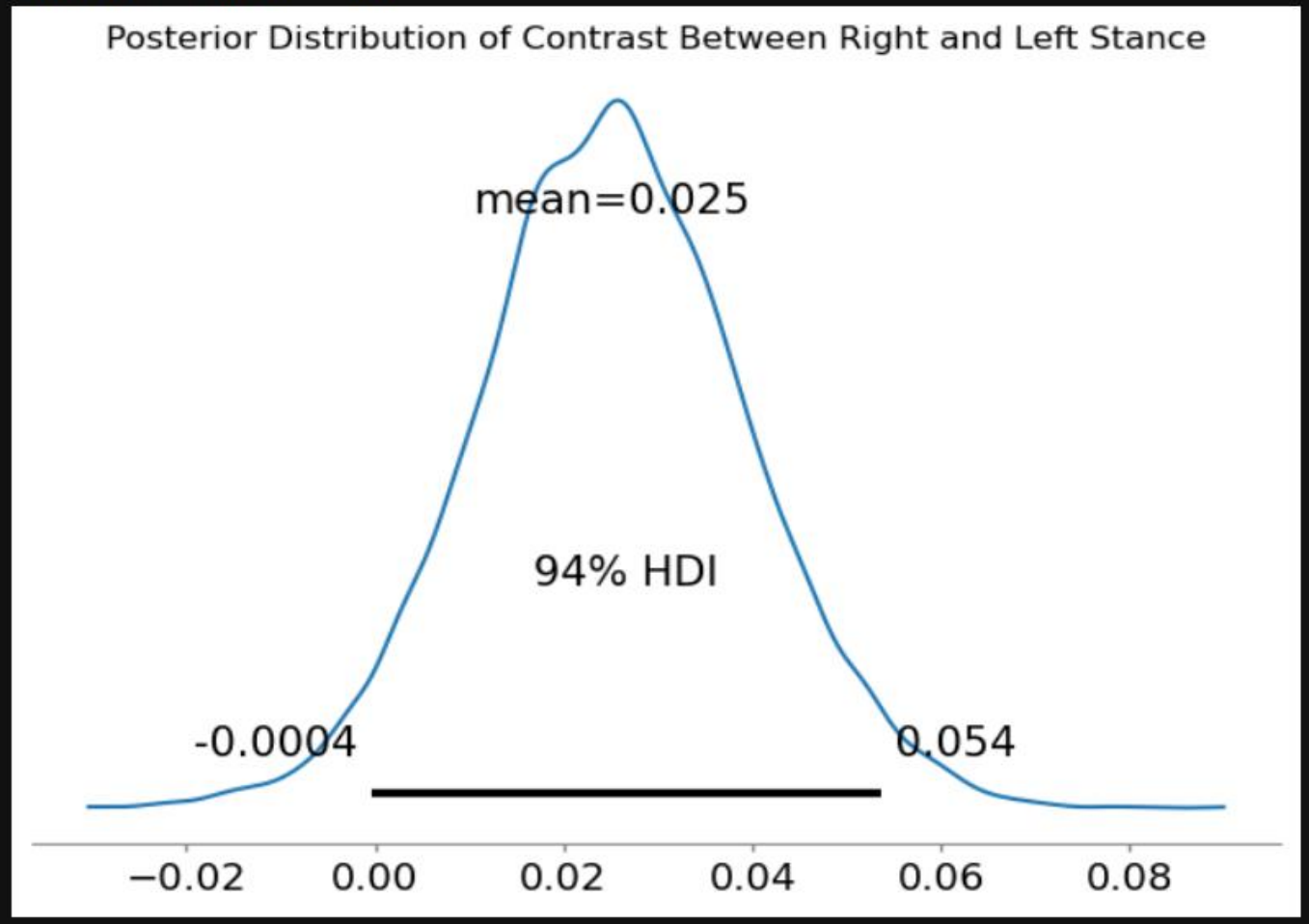
Mean is 0.025

Right Stance shows 2.5% Swing Probability

HDI ranges from -0.0 to 0.054 suggesting small variability

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	\
contrast	0.025	0.014	-0.0	0.054	0.0	0.0	8991.0	

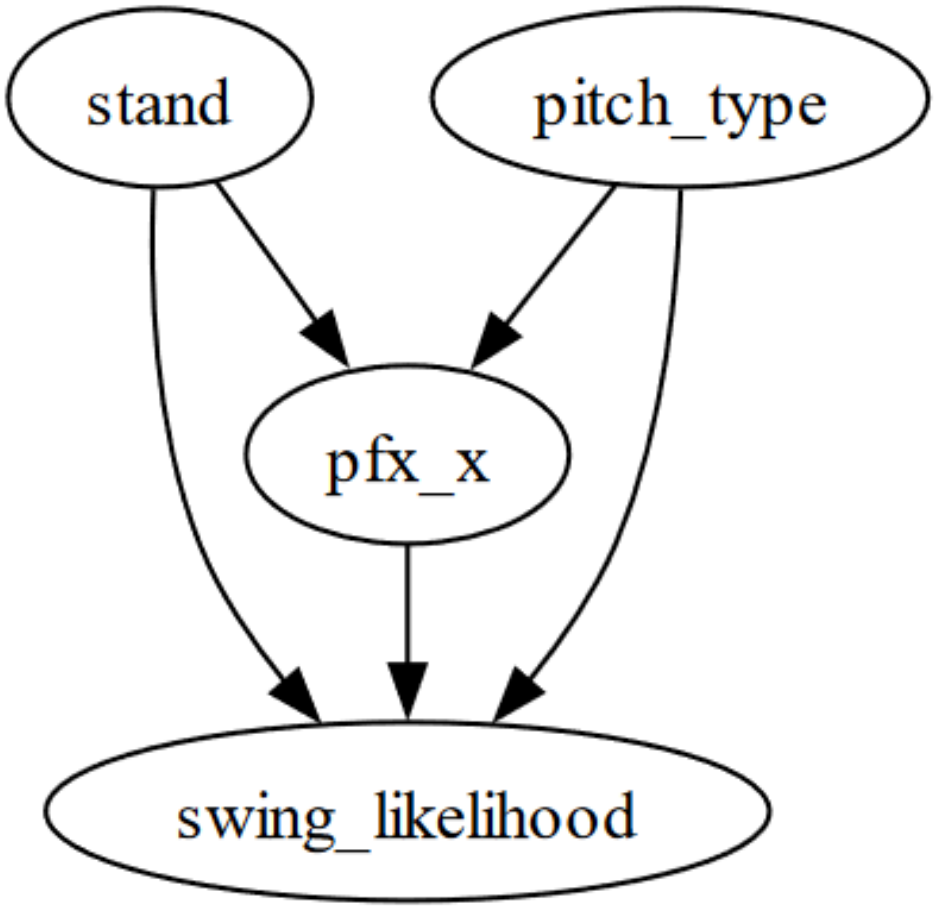
	ess_tail	r_hat
contrast	7065.0	NaN



Probability Distribution by Stance

ANALYSIS AND RESULTS:

- 2. How does the lateral movement (pfx_x) of the pitch, influenced by batter stance, affect swing likelihood, considering the pitch type as a confounding variable?



Drawing the DAG

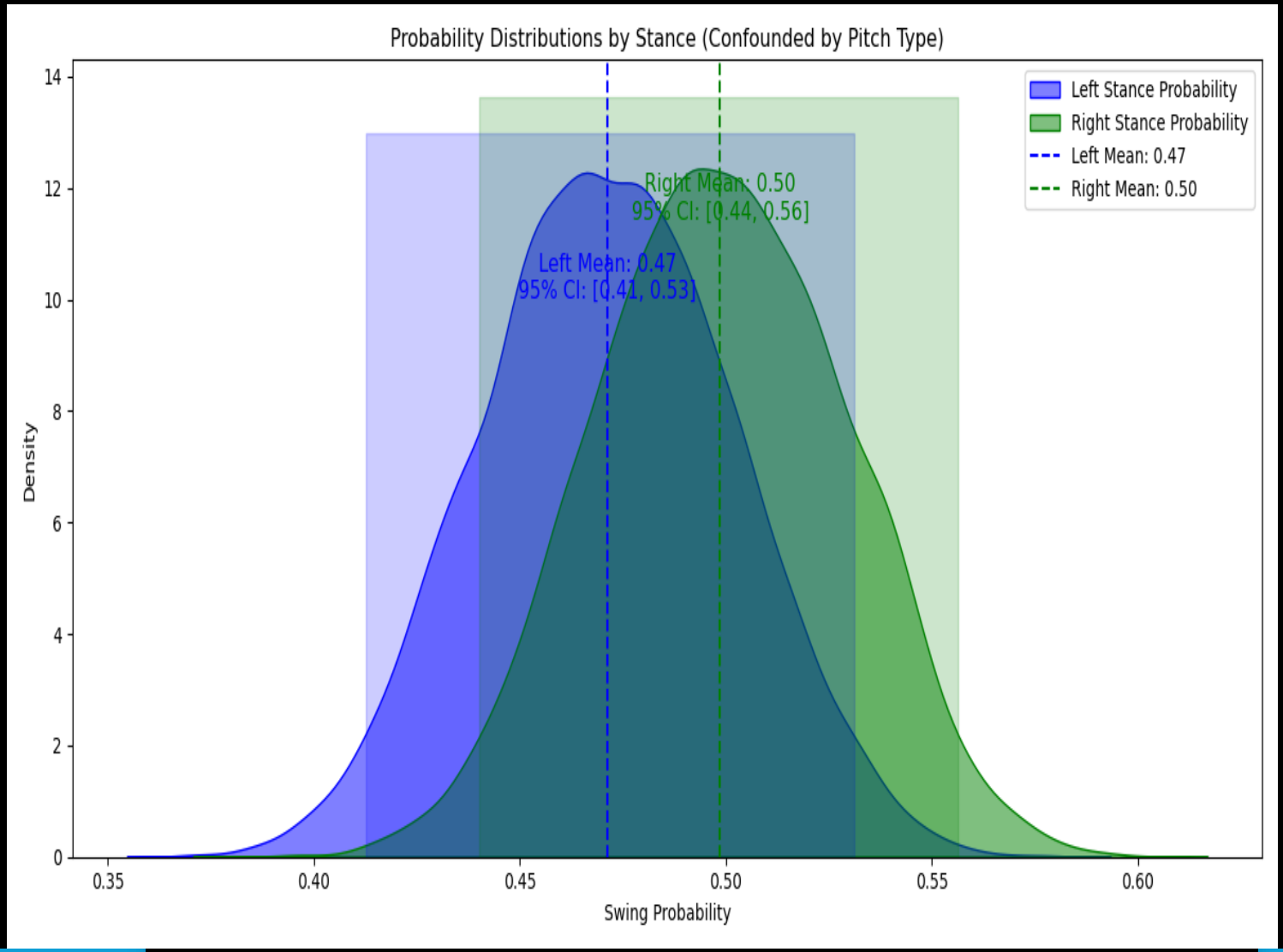
Result

Contrast mean of 0.03

HDI – hdi_3 is 0.001, hdi_97 is 0.053

SD is 0.014 (low variability)

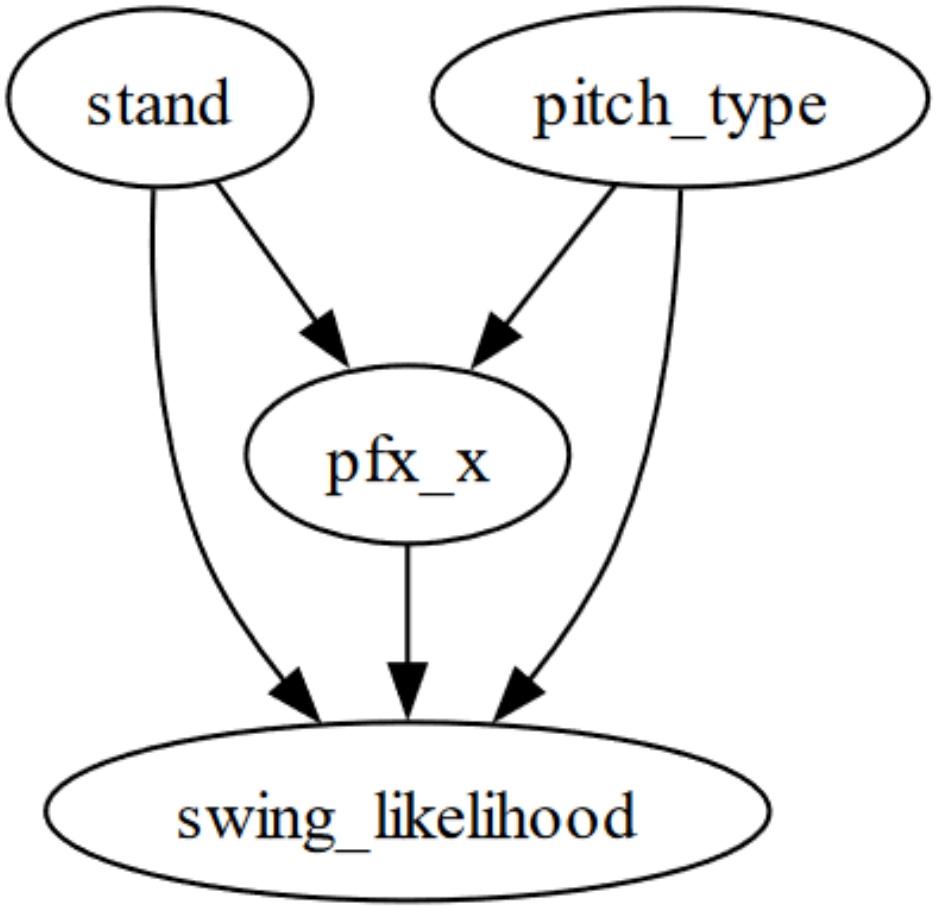
HDI does not include 0 , statistically significance



Probability Distribution by Stance

ANALYSIS AND RESULTS:

- 3. How does the vertical movement (pfx_z) of the pitch, influenced by batter stance, affect swing likelihood, considering the pitch type as a confounding variable?



Drawing the DAG

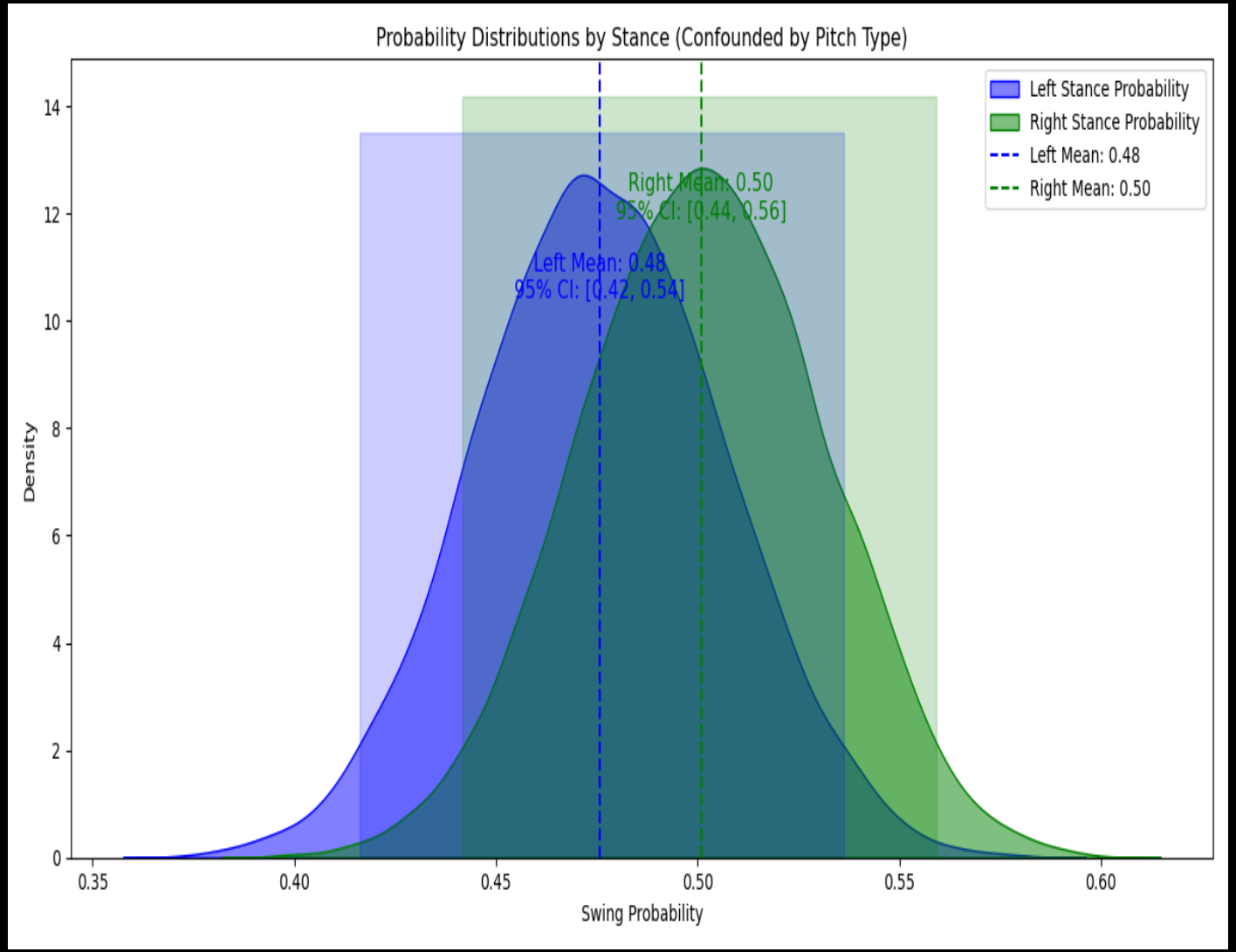
Result

Contrast mean of 0.025

HDI – hdi_3 is -0.002, hdi_97 is 0.05

SD is 0.014 (low variability)

HDI does include 0 , not statistically significance, uncertainty whether effect is genuinely positive

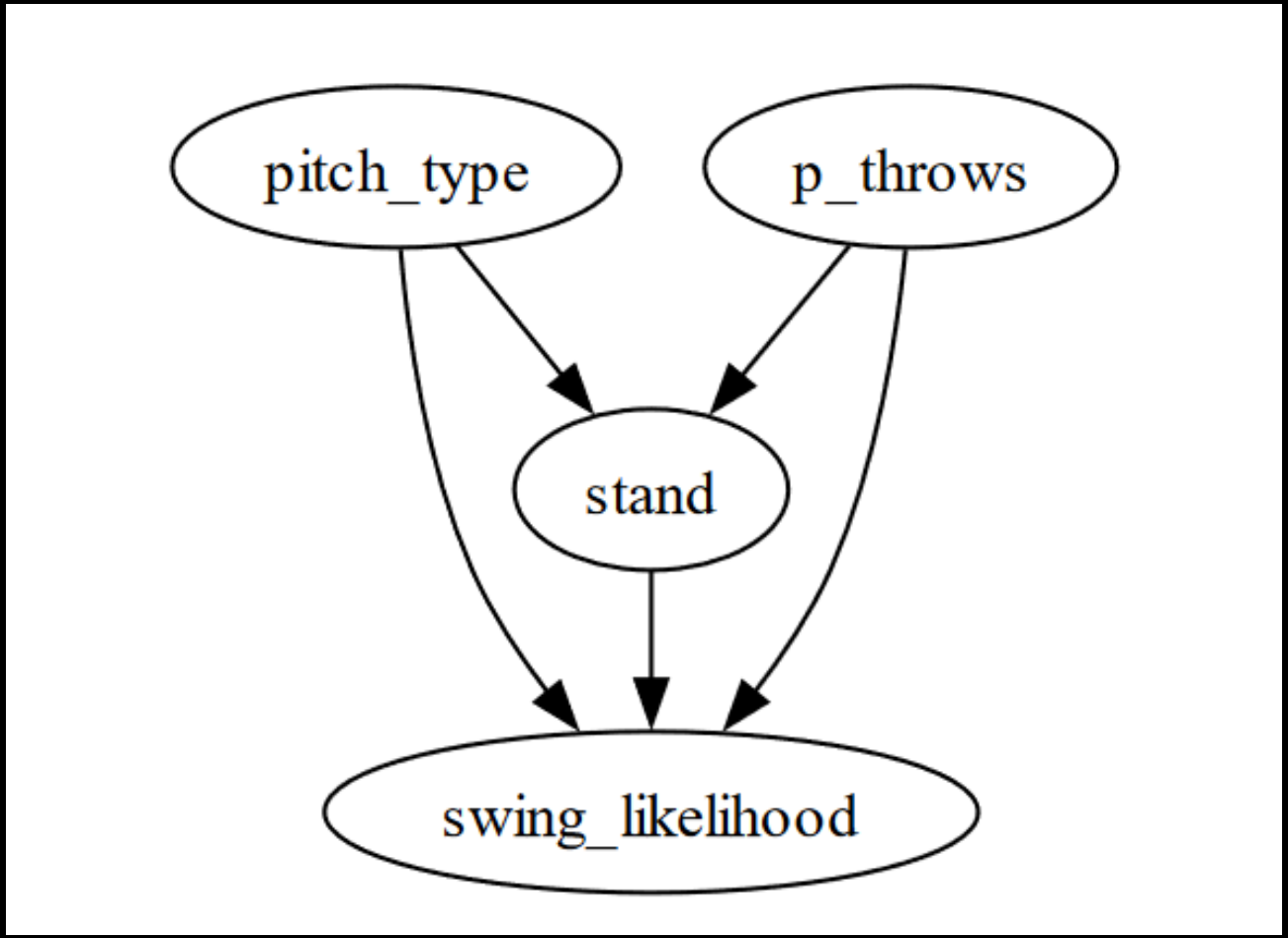


Probability Distribution by Stance

ANALYSIS AND RESULTS:

- 4. Is there an interaction effect between pitch type and batter stance on swing likelihood, influenced by the pitcher's throwing hand?

Confounder: Pitcher's throwing hand (p_throws).



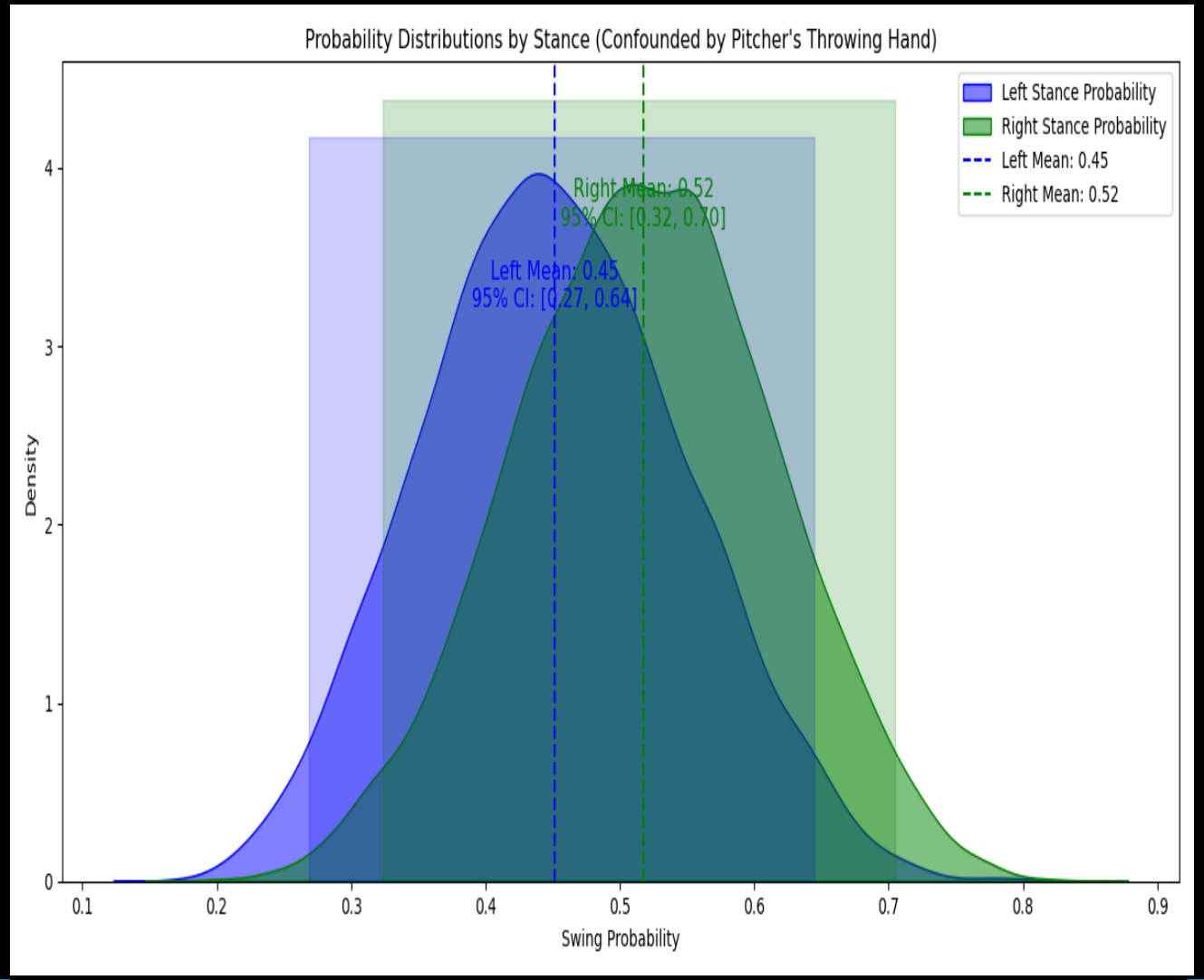
Drawing the DAG

Result

Contrast mean of 0.067

SD is 0.137 (high variability due to inconsistent across samples e.g., diff pitch type)

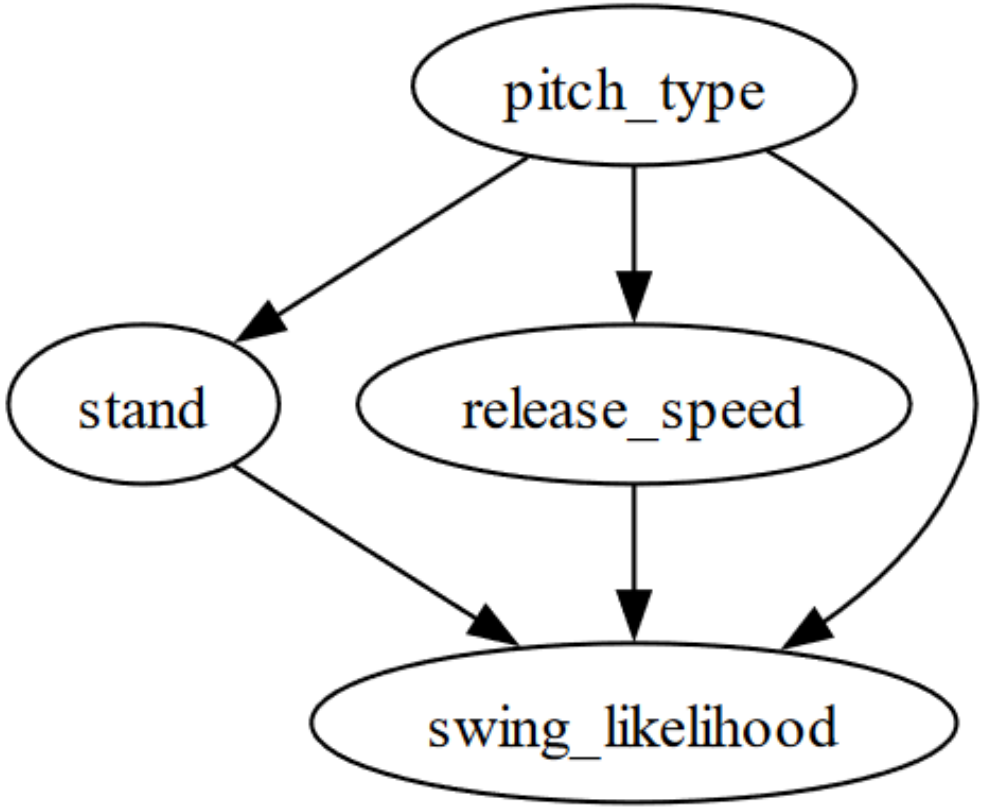
- Right stance shows a 6.7% higher swing probability than left, with variability.
- HDI range from -0.191 to 0.318 indicates uncertainty in effect size.



Probability Distribution by Stance

ANALYSIS AND RESULTS:

- 5. How do pitch type, release speed, and batter stance interact to influence the likelihood of a swing in baseball?



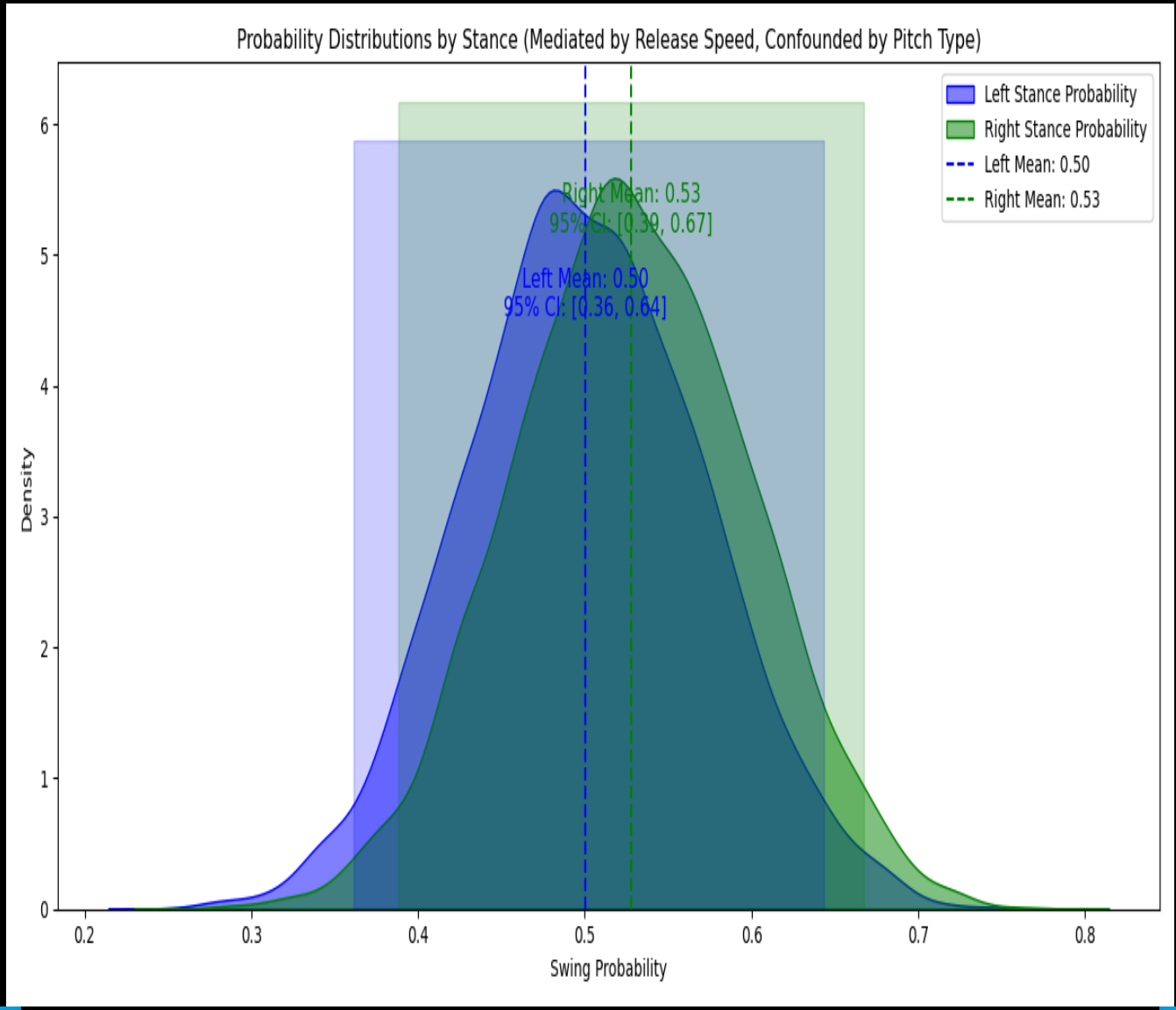
Drawing the DAG

Result

Contrast mean is 0.027

SD is 0.014 (low variability)

- HDI range from 0.001 to 0.055 suggests a small but significant effect (a consistent positive interaction effect across samples.)



Probability Distribution by Stance

FINDINGS

Analysis	Contrast Mean	HDI Range	SD (Variability)	Key Interpretation
Direct Effect of Batter's Stance	0.025	-0.0 to 0.054	Small	The right stance shows a small ↑ in swing prob. compared to the left. However, the effect is not consistently statistically significant because the HDI includes zero. The variability in the effect size suggests that the influence of the right stance on swing prob. may depend on interactions with other factors.
Lateral Movement considering Stance, pitch type on Swing Likelihood	0.03	0.001 to 0.053	Low	Lateral pitch movement significantly ↓ swing prob., but the right stance still shows ↑ swing prob. Statistically significant
Vertical Movement considering Stance, pitch type on Swing Likelihood	0.025	-0.002 to 0.05	Low	The vertical movement has a less consistent effect on swing likelihood, as the HDI includes zero, indicating no statistically significant impact. Further investigation needed.
Interaction Between Pitch Type, Stance & Pitcher's Throwing Hand (p_throws)	0.067	-0.191 to 0.318	High	The interaction shows variability in swing probability. High variability highlights complex interdependencies and the need for further analysis.
Pitch Type, Release Speed, and Stance Interaction	0.027	0.001 to 0.055	Low	Right stance shows a 2.7% ↑ swing prob. > left. Has a small but consistent positive interaction effect suggesting that the combination of pitch type, speed, and stance influences swing likelihood.

FINDINGS & CONCLUSION

Optimize Batters' Stance:

- The right stance generally increases swing probability.
- Batters can focus on refining their right stance, especially in situations where they anticipate pitches with significant lateral movement.

Pitching Strategy:

- Pitch type and lateral movement significantly influence swing decisions.
- Pitchers can strategically adjust pitch types and lateral movements to exploit weaknesses in batters' stances, particularly targeting scenarios where batters' stances are less effective.

Coaching and Training:

Coaches can design tailored training programs that focus on improving batters’ adaptability to different pitch movements, and guide pitchers in developing pitches that exploit batter stance tendencies.

REFERENCES

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. Advances in Methods and Practices in Psychological Science, 1(1), 27–42. https://doi.org/10.1177/2515245917745629

Pearl, J. (2009). Causal inference in statistics: An overview. Statistics Surveys, 3, 96–146. https://doi.org/10.1214/09-SS057

Kruschke, J. K. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan (2nd ed.). Academic Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.