WSABI Summer Research Lab **Summer 2025**

Lab 5: Confounding

Instructor: Jonathan Pipping *Authors: Ryan Brill*

5.1 Park Effects

We wish to estimate a park effect of each MLB ballpark, or the expected runs scored in a half-inning at that ballpark above that of an average ballpark, ceteris paribus (all else equal).

5.1.1 Simplest Version of the Problem

To begin, simply compute the mean runs scored in a half-inning at each ballpark. This is what they did in the OpenWAR paper [BJM]. Is there anything wrong with this? Are there any confounders? What might they be?

The most egregious confounders are **offensive and defensive quality**. For instance, consider the Yankees in 2021, who were in a great division (2021 AL East). In baseball, teams play the other teams in their division most, meaning that the Yankees, Red Sox, Blue Jays, and Rays (great offensive teams in 2021) played there a lot. This means that the runs scored in a half-inning at Yankee Stadium would be higher than average, but maybe not for the effect we're looking for! How might we solve this? The answer is by using a richer dataset.

5.1.2 Data

We are given data of half-innings, with each row representing one half-inning. Each row contains the following variables:

- i: index of the i^{th} half-inning
- y_i: runs scored in the i^{th} half-inning
- o_i: the offensive team-season corresponding to the i^{th} half-inning
- d_i: the defensive team-season corresponding to the i^{th} half-inning
- p_i: the park the i^{th} half-inning was played at

5.1.3 Your Task:

1. Devise a model that estimates park effects while adjusting for offensive and defensive quality.
2. Compare your refined estimates to the original naive estimates by visualization.
3. Test the two estimates against each other using out-of-sample predictive performance. Did we improve?
4. Which park effects differ the most between the two estimates?

References

[BJM] Baumer, B. S., Jensen, S. T., and Matthews, G. J. (2015). openWAR: An open source system for evaluating overall player performance in major league baseball. Journal of Quantitative Analysis in Sports, 11(2):69–84.