Lab: Confounding

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).

To begin, simply compute the mean runs scored in a half-inning at each ballpark (that's what they did in the OpenWAR paper (Baumer et al 2015)).

Before going to the next page:

Is there anything wrong with this?
Are there any confounders?

The most egregious confounders are offensive & defensive quality.

For instance, consider the Yankees in 2021, who were in a great division (2021 AL East). The mean runs scored in a half inning at Yankees stadium will be larger because the Yankees, Red Sox, Blue Jays, and Rays play there a lot and are great offensive teams! We need to disentangle the park effect from the quality of the offense and defense.

We are given a dataset of half-innings,

- i = index of i^{th} half inning
- y_i = Runs scored in i^{th} half inning
- O_i = the offensive team-season
- D_i = the defensive team-season
- P_i = park

Devise a model that estimates park effects and adjusts for offensive and defensive quality.

Compare to the original naive estimates by visualization & out-of-sample predictive performance.

Which park effects are most different?