Supplementary Files for Comparing baseball players across eras via the novel Full House Model

Shen Yan[1], Adrian Burgos Jr.[2], Christopher Kinson[1], Brandon Niedert[3],
and Daniel J. Eck[1]

1. Department of Statistics, University of Illinois Urbana-Champaign
2. Department of History, University of Illinois Urbana-Champaign
3. FuboTV Inc.

1 Data Collection for Batting Statistics

Baseball-Reference Win Above Replacement(bWAR) is collected from the website: https://www.baseball-reference.com/data/war_daily_bat.txt, which is stored in batters_bWAR.csv.

Fangraphs Win Above Replacement(fWAR) is scraped from the website: https://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=al&qual=0&type=c,4,5,6,23,58&season=2021&month=0&season1=2021&ind=0&team=0&rost=&age=&filter=&players=&startdate=&enddate=, which is stored in batters_fWAR.csv

Hits, Home Runs, and Walks are collected from Chadwick Baseball Bureau http://www.chadwick-bureau.com/ and they provide data to a number of high-profile clients across the industry, including being a primary provider of historical statistical data for Baseball-Reference for major, minor, international, collegiate, and summer collegiate competitions. The datasets we use are listed on Github: https://github.com/chadwickbureau/baseballdatabank/tree/master/core and stored in the raw_batter.csv.

Hits, Runs, and Home Runs in a team's home and road games are collected from Chadwick Baseball Bureau http://www.chadwick-bureau.com/ that is mentioned above. The datasets we use are listed on Github: https://github.com/chadwickbureau/retrosplits/tree/master/daybyday and stored in the team-xx.csv.

The unique player IDs for Baseball-Reference, and Fangraphs are collected from Chadwick Baseball Bureau http://www.chadwick-bureau.com/ that is mentioned above. The unique player IDs for Baseball-Reference, and Chadwick are also collected from Chadwick Baseball Bureau http://www.chadwick-bureau.com/. The datasets we use are stored in people_bb_f.csv and people_park_bb.csv.

2 Data Preprocess for Batting Statistics

merge_pop.R combines the bWAR, fWAR, and other statistics when a player is traded in midseason. Also, it adds the corresponding MLB eligible population to the dataset. The results are stored in the batters_combined_b.csv and batters_combined_f.csv

park_factor_chad.R applies the park-factor adjustment to the Hits, Runs, and Home Runs. Also, it combines the Hits, Home Runs, and other statistics when a player is traded in midseason and adds the corresponding MLB eligible population to the dataset. The results are stored in the batter_park_factor.csv.

clean_batters.R combines all the important batting statistics based on the players' unique IDs in Baseball-Reference, Fangraphs, and Chadwick. The results are stored in the batters_all.csv.

3 Data Collection for Pitching Statistics

Baseball-Reference Win Above Replacement(bWAR) is collected from the website: https://www.baseball-reference.com/data/war_daily_pitch.txt, which is stored in pitchers_bWAR.csv.

Fangraphs Win Above Replacement(fWAR) is scraped from the website: https://www.fangraphs.com/leaders.aspx?pos=all&stats=pit&lg=al&qual=y&type=8&season=2021&month=0&season1=2021&ind=0&team=0&rost=0&age=0&filter=&players=0&startdate=&enddate=, which is stored in pitchers_fWAR.csv

Earned Run Average(ERA), Strikeouts(SO), and other statistics are collected from Chadwick Baseball Bureau http://www.chadwick-bureau.com/ that is mentioned above. The datasets we use are listed on Github: https://github.com/chadwickbureau/baseballdatabank/tree/master/core and stored in Pitching.csv.

4 Data Preprocess for Pitching Statistics

merge_pop_p.R combines the bWAR, fWAR, and other statistics when a player is traded in midseason. Also, it adds the corresponding MLB eligible population to the dataset. The results are stored in the pitchers_combined_b.csv and pitchers_combined_f.csv

clean_batters.R applies the park-factor adjustment to the Runs. Also, it combines all the important pitching statistics based on the players' unique IDs in Baseball-Reference, Fangraphs, and Chadwick. The results are stored in the pitchers_all.csv.

5 Summary

All the datasets and codes mentioned above can be found on Github: https://github.com/yanshenlmx/Full_House_model_data and our data processing steps are reproducible in Tech_Report.html that is included with this submission.