Qualifying exam questions

Instructions

This file contains two questions for the take-home portion of the 2024 Statistics qualifying exam at the University of Illinois Urbana-Champaign. This part of the exam is assigned at noon on June 27 and is due at 6 PM on June 28.

You are required to submit your answers to the questions in this document as a pdf file. Your submission should be a reproducible technical report and should include all relevant code necessary to answer the two qualifying exam questions. You can submit your finalized pdf file to Aaron Thompson at aaron5@illinois.edu.

DO NOT ENTER YOUR NAME; ENTER THE ANONYMOUS CODE THAT YOU WERE GIVEN.

You will need to load in the following packages:

```
library(tidyverse)
library(Lahman)
```

Question 1 [10 points]

In this question you will be tasked with creating a data set named `bat`. This data set will be analyzed in the next question. This data set will be constructed from the two data sets described below:

- `dat_chadwick`: A data set containing batting information on all balls in play from 2017-2022 with the 2020 season excluded. In particular we have the launch angle (angle of ball off bat) and the exit velocity (velocity of ball off bat) for each ball in play. This data set includes batter's name, the team they played for, the year. It also includes a unique batter identifier called `key_bbref`.
- `batters_fulltime`: A data set containing batting outcome information for full-time players from 2017-2022 with the 2020 season excluded. These outcomes are aggregates over a season. In particular we have the number of home runs hit by a player and the number of at bats (at bats are chances to hit a home run). This data set includes batter's name, the team they played for, the year. It also includes a unique batter identifier called `playerID`. This unique identifier is the same as the `key_bbref` identifier in the `dat_chadwick` data set.

These data sets are loaded in below:

```
dat_chadwick = read_csv("dat_chadwick.csv")
#head(dat_chadwick)
#View(dat_chadwick)

batters_fulltime = Batting %>%
  filter(yearID >= 2017, AB >= 550, yearID != 2020) %>%
  ## definition of full-time players
  filter(AB >= 550) %>%
  left_join(People %>% dplyr::select(playerID, nameFirst, nameLast)) %>%
  mutate(name = paste(nameFirst, nameLast)) %>%
  dplyr::select(name, playerID, teamID, yearID, HR, AB) %>%
  rename(year = yearID)
```

```
#head(batters_fulltime)
#View(batters_fulltime)
```

You need to transform the `dat_chadwick` data set to create features of the exit velocity and launch angle distributions for each player season. This transformed data set will be joined with the `batters_fulltime`. The goal of the analysis in the next question is to use the features of the exit velocity and launch angle distributions that you construct in this question to model home runs. You may restrict attention to complete cases, although that is neither required and it might not be necessary.

```
bat = bat[complete.cases(bat), ]
```

Note that home runs are relatively rare events. They require for a ball to be hit very hard (high exit velocity) and within a range of launch angles. Thus, average exit velocities and average launch angles may be of limited utility for the modeling of home runs. You will need to create additional features of the exit velocity and launch angle distributions beyond simple averages, and you will need to provide justification for these choices [3 points].

For clarity, note that a successfully created `bat` data set is worth 7 points. Inclusion of justified features of the exit velocity and launch angle distributions is worth 3 points.

Display the the first 6 rows of your `bat` data set:

Question 2 [10 points]

In this question you are to model home runs. Your model should include year and features of the exit velocity and launch angle distributions that you constructed in the previous question. You should also add a term that isolates the Colorado Rockies (teamID: COL) because the high altitude of the stadium that this team plays its games allows for batters to hit home runs more often. You should also consider adding a term that isolates the 2019 season because it has been demonstrated that there was an uptick in home run hitting due to changes made to the baseball itself.

You need to justify your model using a combination of model selection criteria [1 points], diagnostics [2 points], out-of-sample prediction [2 points]. You also need to discuss strengths and weaknesses of your final model [1 points]. You do not need to report every model that you fit. However, you do need to show comparisons involving at least two carefully constructed candidate models. Your model should perform well as judged by these criteria [4 points].

Note that there is code in the `yang2021supp.R` file that implements the "quasi-empirical residual distribution function" model diagnostic methodology for Poisson regression. Inclusion of this file does not require you to implement Poisson regression.