Exploratory Data Analysis

We were chiefly interested in exploring the relationship between win percentage and any numeric variables that appeared to have a strong ($r > 0.5$) correlation coefficient. To do this, we created two separate correlation matrices, one for batting statistics only **(Fig. 2)** and one for pitching statistics only **(Fig. 3)**. In **Fig. 2**, we notice that the correlations with $r > 0.5$ are RBI, R, AvgRunPerGame, SLG, and OPS. In **Fig. 3**, the strongest correlations are ARAPG, RAllowed, WHIP, SV, and HAllowed. These were a part of our decision even though we used stepwise algorithms to see which combinations of variables were the best for final models.

Methods

With preliminary models, our intercept was very strange. Essentially, the intercept could only be interpreted if a team were to allow zero home runs, get zero hits, have a slugging statistic of zero, etc. None of these scenarios are realistic, so we instead centered each possible predictor variable at its median to allow our intercept to be meaningful.

Additionally, when considering all combinations of predictor variables, there are 2^{32} possible models, or almost 4.3 billion. This immediately forced us to use a stepwise algorithm to find the "best" combinations. The drawback of this method is that we don't get to examine the best combinations for models that have 1, 2, ..., 32 total predictors; we instead just get the "best" model overall.

Results

We chose to split the models into three groups: Batting only, Pitching only, and both Batting and Pitching. This was done because batting is purely offensive, whereas pitching is purely defensive. Finally, we combined both to see overall, what predictors make a Twins win more likely. It is important to note that upon assessing each of our models with a residual plot, a quantile-quantile plot, and a histogram of residuals, no significant issues were found. The stepwise algorithm found the following models:

Batting Statistics ONLY:

$$\hat{WinPercentage} = 49.36 - 0.07\,(\hat{AB\,center}) + 0.09\,(\hat{H\,center}) + 0.10\,(\hat{HR\,center})$$

The initial model selected by the stepwise model for batting statistics contained no insignificant terms, thus we keep this model in its entirety.

Pitching Statistics ONLY:

$$\hat{WinPercentage} = 56.71 - 9.94(\hat{ARAPG\ center}) + 0.48(\hat{SV\ center}) + 0.90(\hat{HRAllowed\ center}) - 127.43(\hat{HR9\ center})$$

The initial model chosen by the stepwise algorithm had two predictors that were statistically insignificant, namely $HR9$ *center* and $HRAllowed$ *center*. It turns out that upon dropping $HR9$ *center*, $HRAllowed$ *center* becomes significant. This somewhat makes intuitive sense, as $HR9$ *center* is simply a per nine inning average of $HRAllowed$ *center*. This leave us with:

$$\hat{WinPercentage} = 49.26 - 10.72(\hat{ARAPG\ center}) + 0.53(\hat{SV\ center}) + 0.10(\hat{HRAllowed\ center}).$$