Pitching and Batting Statistics:

$\hat{WinPercentage} = 49.35 - 10.63(\hat{ARAPG\ center}) + 0.22(\hat{SV\ center}) + 3.69(\hat{AvgRunPerGame\ center}) - 0.01(\hat{BB\ center}) - 0.05(\hat{HR\ center}) + 0.01(\hat{BBAllowed\ center}) + 158.18(\hat{SLG\ center}) + 2.85(\hat{HR9\ center})$

The combined model initially selected by the algorithm had several insignificant terms. These insignificant terms were *HR9 center*, *BBAllowed center*, and *BB center*. To ensure the overall effectiveness of the model, we dropped them one at a time in the order that they appeared in the stepwise algorithm output whilst checking to see if the remaining predictors became insignificant. Fortunately, upon dropping *HR*9, *BBallowed*, and *BB* from the model, all other predictors remain significant. This left us with:

$\hat{WinPercentage} = 49.22 - 9.21(\hat{ARAPG\ center}) + 0.20(\hat{SV\ center}) + 3.60(\hat{AvgRunPerGame\ center}) - 0.04(\hat{HR\ center}) + 158.61(\hat{SLG\ center})$

Interpretation

Since all of our predictors across all models are numeric only, interpretations will be the same across all predictors. For example, if one wanted to interpret the coefficient of $\hat{ARAPG\ center}$ from the combined model, one would say that ceteris paribus, on average, for every 1 additional run allowed per game above the median (4.07), that corresponds with the Twins' win percentage decreasing by 9.21 percentage points.

Discussion

Models and Implications

Looking at the batting only model, At Bats (AB), Hits (H), and Home Runs (HR) are the best predictors of a Twins win for their offense. Hits and home runs make sense as great predictors since more hits and home runs especially will result in a higher score.

A strange statistic which appeared in our batting model was AB, which does make sense at a second glance. If a team has more at bats, they are cycling through more batters and those runners are getting on base rather than out. However, the coefficient is negative, meaning more at bats would equal a lower chance of winning. Moving onto the pitching only model, Average runs allowed per game (ARAPG), Saves (SV), and Home Runs Allowed (HRAllowed) are the best predictors for a Twins win based solely on defense. Average runs allowed per game being strong makes sense since more allowed runs would mean lower win probability. Saves also check out as each save is directly responsible for a win. However, there was some trouble with Home Runs Allowed. The coefficient is positive, meaning more home runs allowed contributes to a higher win percentage, which doesn't make logical sense in the context of baseball. Lastly, the combined model tells us that the best predictors are Average runs allowed per game (ARAPG), Saves (SV), Average runs per game (ARPG), Home Runs (HR), and Slugging Percentage (SLG). Average runs allowed per game and saves both have positive coefficients just as they did in the other two models. Average runs per game and slugging percentage both have positive coefficients. This makes sense as an increase in either one means that more runners are getting on base for the Twins. Once again, there is an odd predictor, home runs. The coefficient for home runs is negative, meaning the more home runs the Twins hit, the lower their win percentage becomes, which again does not make sense in baseball. In hindsight, this is possibly due to the drawbacks of the stepwise selection algorithm.

Confounding Variables

Upon testing for multicollinearity by looking at variance inflation factors (VIF), we found that *SLG* had a VIF score of 15. This abnormally high, but when other predictors in the combined model are considered, they are also directly related to the Twins putting runners on base, specifically Home Runs and Average runs per game.