Reds Hackathon

Luke Hamm, Jadon Hsu, Jake Snyder, Harrison Stanton, Tyler Weiner

February 2024

1 Model

1.1 Choosing Stats to Consider

In order to determine which characteristics were important for starters and relievers, and in turn who should switch roles, the first thing we needed to do was pick which stat would be used to judge a pitcher's success. We initially were considering ERA estimators, such as xFIP and SIERA, and WAR. There were good characteristics about ERA estimators and WAR: however, we decided on looking more into ERA estimators. There were a couple of reasons for this decision. When researching more about WAR, one important thing that came up was that the small changes in WAR do not tell you much about the difference between two players. FanGraphs glossary page for WAR says that if player A has a 6.4 WAR and player B has a 6.1 WAR for the season, that this is not a significant enough difference to tell you whether or not player A was better than player B. So, while it would be nice to use WAR because it can better describe someone's worth across roles (for example, determining whether someone is more valuable as a starter or a reliever), we decided on an ERA estimator because we were able to distinguish more when players' numbers were close together.

From there, we settled on SIERA over xFIP. The reason for this is that SIERA is supposed to be slightly more predictive than xFIP: additionally, SIERA takes into account all balls in play, not just homeruns. This means that we learn more about how and why a pitcher succeeds[1]

1.2 Creating the Regressions

Since the goal of our model is to differentiate what leads to success for starting pitchers versus relievers, the first step we took was to filter the FanGraphs data frame by position. We then filtered out relievers who had thrown less than 250 pitches on the season, and starters who had thrown less than 400 pitches. This is because some of the variables that we ended up looking at, like Stuff+ and

[1] FanGraphs Sabermetrics for xFIP, SIERRA, and WAR was used as a source

Location+, become more predictive around those ranges. Additionally, with a bigger sample size of pitches thrown, our data contains fewer outliers.

Then, we decided to make two different multiple linear regressions, one for the starter's SIERA and one for the reliever's SIERA, in order to compare and contrast. We went through and picked variables that were not directly related to the inputs of SIERA, and ones that made sense in the context of our regression. This meant leaving out any stat that had BB, GB rate, SO, or PA, and leaving behind any cumulative stat or other ERA estimators. After creating our initial multiple linear regressions, we wanted to make sure that a linear regression was a viable technique to use with this data. To check this, we created scatter plots of the residual values vs. the fitted values and histograms of the residuals. As you can see in Figure 1 and Figure 2, our scatter plots are randomly distributed above and below the dashed line at $y = 0$, and the histograms both show that our residuals are normally distributed, meaning that the linear regression is viable technique to represent the pitchers' SIERA.

In order to filter our multiple linear regressions down from what they were initially at, we used $\alpha = .05$ to determine which variables were statistically significant in our regression. One by one, we took out the variables with the highest p-values and reran the regressions, until we were left with only statistically significant results, as shown in Figures 3 and 4.

Once we had our multiple linear regressions finalized, we took the reliever regression, and added a column to the starting pitcher data frame calculating their projected SIERA in a relief pitcher role. We did the same thing for the relief pitchers using the starting pitcher regression. An additional step that was needed to make our results accurate was to subtract .37 from the projected SIERA values of pitchers moving from a starting role to the bullpen, and adding .37 to the SIERA values of pitchers moving from the bullpen to the rotation. This change is part of the formula for SIERA, and occurs because pitchers who move to the bullpen typically gain velocity because of a smaller workload.

2 Characteristics

One of the biggest differences between the starting pitcher and relief pitcher SIERA was that the relief pitchers had O-swing% and Z-swing% in their regression. This is likely because relief pitchers face fewer batters than starters do. Thus, preventing balls in play is more important for a reliever than it is for a starter, meaning that getting swings on pitches outside the zone would be crucial, as well as limiting swings on pitches in the zone. Additionally, the relievers had RE24 factor into their regression, while starters had LOB%. This is likely due to the nature of each statistic. Relievers are expected to come in with runners on and limit damage, while starters are expected to not let their own base runners score. Conversely, Z-contact%, exit velocity, barrel percentage, and best fastball grade were roughly twice as important for starters as relievers. This is likely due to the nature of their roles. Starters face more batters over the course of a season, meaning that they can pitch to contact more effectively because they have a much bigger sample size to work with. If they are consistently allowing weak contact, over time that will result in more outs, whereas a reliever is more likely to struggle from bad variance. Along the same lines, barrel percentage and exit velocity are more likely to affect a starting pitcher than a reliever for the same reasons. Consistently allowing hard hit balls over the course of many starts will lead to damage being done, whereas the smaller sample size that a reliever has can lead to them not fully realizing the negative effects of giving up hard hit balls. Similarly, a higher fastball grade is more important for starters because they face hitters multiple times in the order, not just once or being used situationally. This information tells us that having swing and miss stuff that limits balls in play is more important for a reliever, whereas consistently getting weak contact is more important for a starter.

3 The Final Decisions

Using the model, we narrow candidates to ones who are both in the top 100 for change in SIERA and appear on the list for multiple seasons. Out of these candidates, we focus on shifting from a starting role to a relief role, mainly because the majority of candidates who are currently relievers and would be better suited to be a starter are out of the league or a free agent. This leaves three starting pitchers: Tyler Anderson, Marco Gonzales, and Kyle Hendricks. Tyler Anderson has, according to our model, been better suited as a reliever the past three seasons and as an older player (Figure 5). A transition to a less pitch heavy role would dually serve to prolong his career as well as make him a more efficient player. Marco Gonzales is currently projected to be the number 3 starting pitcher in the rotation and, although this position is valuable, he had one of the highest differences in projected SIERA from our model vs. actual SIERA in 2023. Kyle Hendricks is projected as the 4th starting pitcher for the Cubs and will be an unrestricted free agent after this upcoming season. At 34 years old, a switch to reliever could make him both a more desirable free agent as well as extending his career. Although considered as a finalist, ultimately we decided Wade Miley was too valuable in his current role to switch. Although he has had one of the larger decreases in projected SIERA in the 2023 season, as well as having consecutive years where he would have been better suited in the reliever role, he is projected as the number 2 starting pitcher for the Brewers and has an inning pitched clause in his contract which makes his switch to reliever impractical for him. We also factored innings pitched when looking at these final candidates. Obviously Anderson with having three seasons will have the most, but Hendricks had over 100 innings pitched in both seasons the model projected him as a better reliever, and Gonzales had over 200 total innings pitched over the two seasons.

4 Appendix

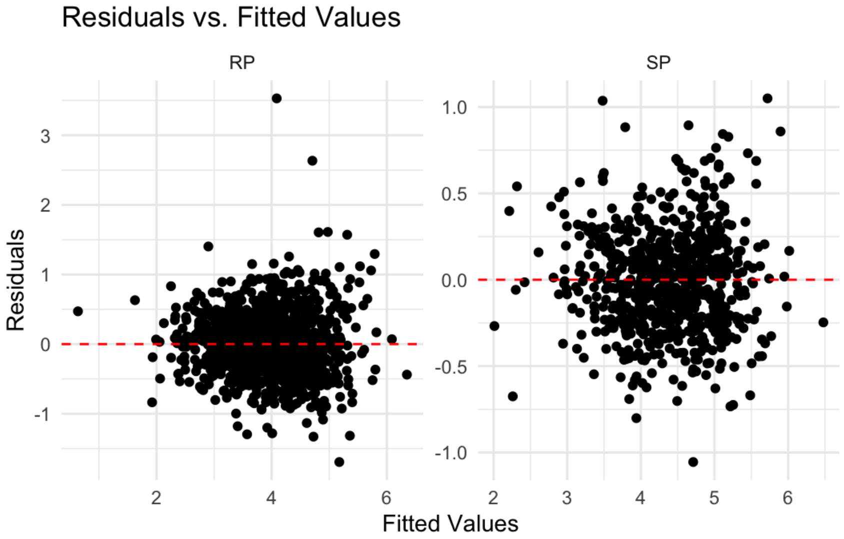


Figure 1: Scatter plot of the Residuals vs. the Fitted values for both the relief pitcher and starting pitcher regression

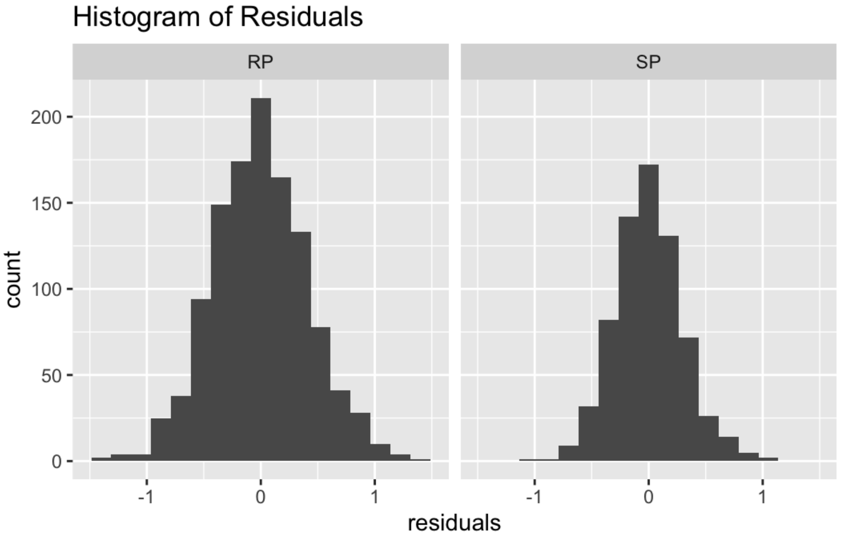
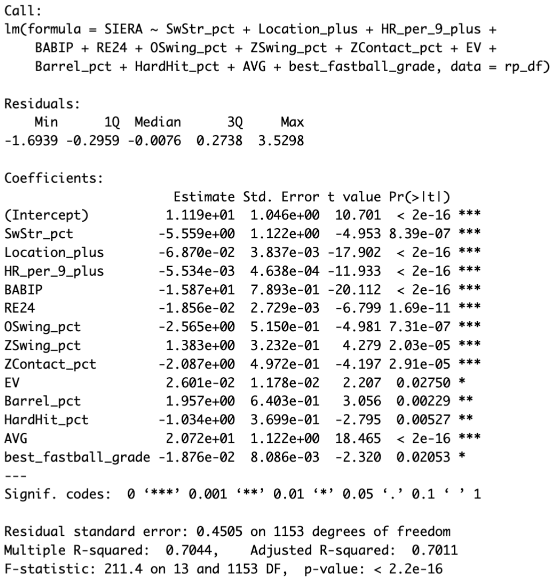


Figure 2: Histogram of the residuals for both the relief pitcher and starting pitcher regressions



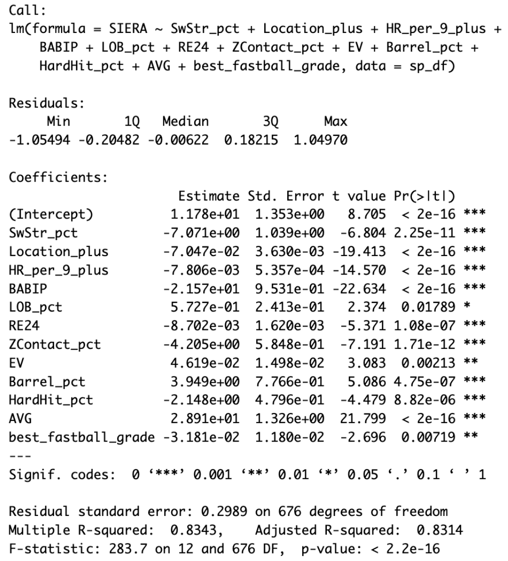
```
Call:
lm(formula = SIERA ~ SwStr_pct + Location_plus + HR_per_9_plus + 
    BABIP + RE24 + OSwing_pct + ZSwing_pct + ZContact_pct + EV + 
    Barrel_pct + HardHit_pct + AVG + best_fastball_grade, data = rp_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6939 -0.2959 -0.0076  0.2738  3.5298 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1.119e+01  1.046e+00  10.701  < 2e-16 ***
SwStr_pct           -5.559e+00  1.122e+00  -4.953 8.39e-07 ***
Location_plus       -6.870e-02  3.837e-03 -17.902  < 2e-16 ***
HR_per_9_plus       -5.534e-03  4.638e-04 -11.933  < 2e-16 ***
BABIP               -1.587e+01  7.893e-01 -20.112  < 2e-16 ***
RE24                -1.856e-02  2.729e-03  -6.799 1.69e-11 ***
OSwing_pct          -2.565e+00  5.150e-01  -4.981 7.31e-07 ***
ZSwing_pct           1.383e+00  3.232e-01   4.279 2.03e-05 ***
ZContact_pct        -2.087e+00  4.972e-01  -4.197 2.91e-05 ***
EV                   2.601e-02  1.178e-02   2.207  0.02750 *  
Barrel_pct           1.957e+00  6.403e-01   3.056  0.00229 ** 
HardHit_pct         -1.034e+00  3.699e-01  -2.795  0.00527 ** 
AVG                  2.072e+01  1.122e+00  18.465  < 2e-16 ***
best_fastball_grade -1.876e-02  8.086e-03  -2.320  0.02053 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4505 on 1153 degrees of freedom
Multiple R-squared:  0.7044,	Adjusted R-squared:  0.7011 
F-statistic: 211.4 on 13 and 1153 DF,  p-value: < 2.2e-16
```

Figure 3: Summary output from R of the relief pitcher regression with only statistically significant variables left



```
Call:
lm(formula = SIERA ~ SwStr_pct + Location_plus + HR_per_9_plus + 
    BABIP + LOB_pct + RE24 + ZContact_pct + EV + Barrel_pct + 
    HardHit_pct + AVG + best_fastball_grade, data = sp_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.05494 -0.20482 -0.00622  0.18215  1.04970 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          1.178e+01  1.353e+00   8.705  < 2e-16 ***
SwStr_pct           -7.071e+00  1.039e+00  -6.804 2.25e-11 ***
Location_plus       -7.047e-02  3.630e-03 -19.413  < 2e-16 ***
HR_per_9_plus       -7.806e-03  5.357e-04 -14.570  < 2e-16 ***
BABIP               -2.157e+01  9.531e-01 -22.634  < 2e-16 ***
LOB_pct              5.727e-01  2.413e-01   2.374  0.01789 *  
RE24                -8.702e-03  1.620e-03  -5.371 1.08e-07 ***
ZContact_pct        -4.205e+00  5.848e-01  -7.191 1.71e-12 ***
EV                   4.619e-02  1.498e-02   3.083  0.00213 ** 
Barrel_pct           3.949e+00  7.766e-01   5.086 4.75e-07 ***
HardHit_pct         -2.148e+00  4.796e-01  -4.479 8.82e-06 ***
AVG                  2.891e+01  1.326e+00  21.799  < 2e-16 ***
best_fastball_grade -3.181e-02  1.180e-02  -2.696  0.00719 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2989 on 676 degrees of freedom
Multiple R-squared:  0.8343,	Adjusted R-squared:  0.8314 
F-statistic: 283.7 on 12 and 676 DF,  p-value: < 2.2e-16
```

Figure 4: Summary output from R of the starting pitcher regression with only statistically significant variables left

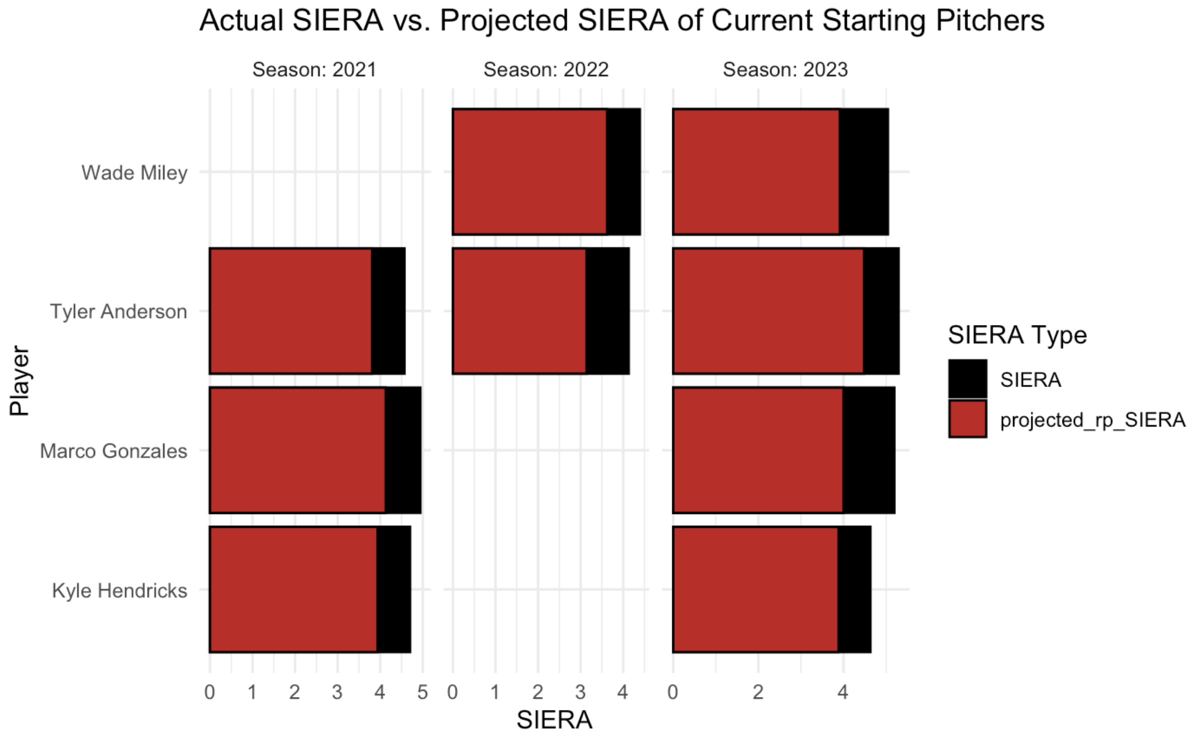


Figure 5: Actual SIERA of current starting pitchers vs. projected SIERA in a relief role, sorted by the repeat cases in top 100

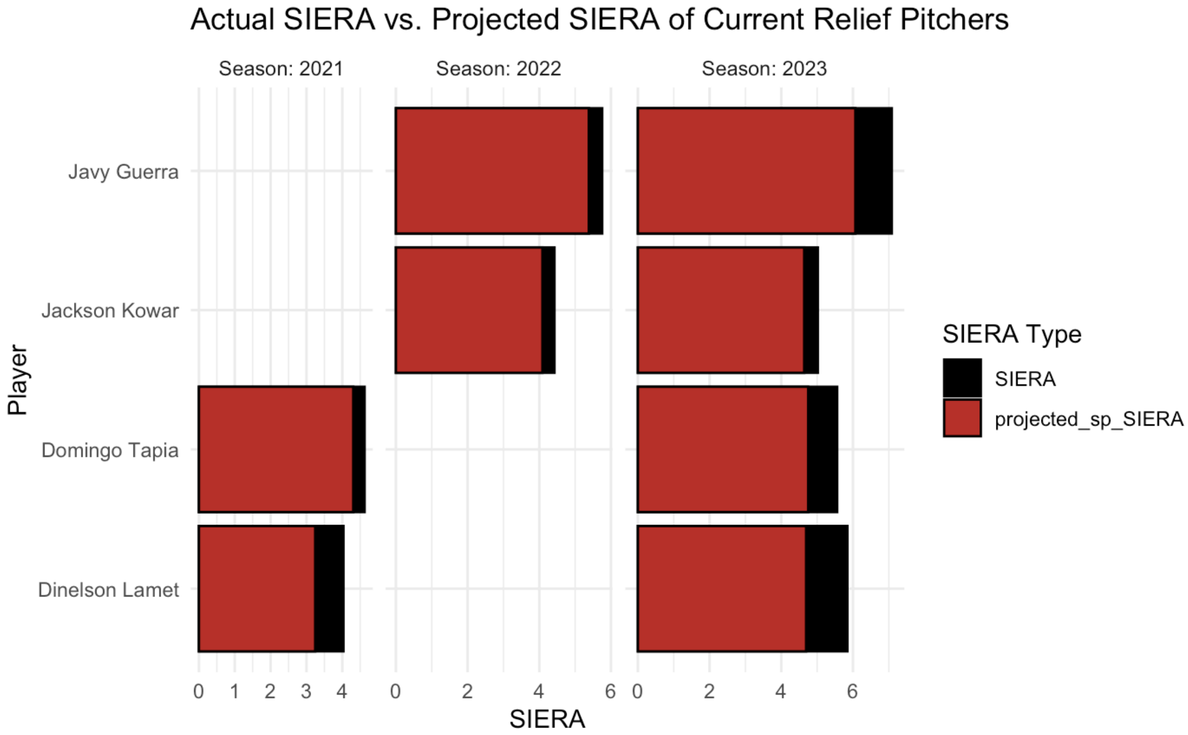


Figure 6: Actual SIERA of current relief pitchers vs. projected SIERA in a starting role, sorted by the repeat cases in top 100