Using the James-Stein Estimator to Predict Pitcher Performance

Tyler Marshall
Department of Mathematics and Statistics, Amherst College

December 11, 2020

Abstract

There are many statistics that can be used to evaluate pitcher performance in baseball, but the most common one is earned run average (ERA). There is often year-to-year variation in a pitcher's ERA, so it can be challenging to predict how a pitcher will perform in the following season. In this case, pitching data from the 2018 MLB season was used to help make predictions for the 2019 MLB season. Empirical Bayes methods can help with these predictions, and in this case the James-Stein estimator was used. The James-Stein estimator incorporates a shrinkage parameter that pulls all the estimates towards the group mean. The shrinkage parameter helps the James-Stein estimator outperform the maximum likelihood estimator in high-dimensional cases like this one. The James-Stein estimator has a lower average error rate than the maximum likelihood estimator when predicting pitcher ERA for the 2019 MLB season based on 2018 pitching data.

Keywords: baseball, Empirical Bayes, earned-run average, shrinkage

1 Research Question

Does the James-Stein estimator outperform the maximum likelihood estimator when predicting earned run average for Major League Baseball pitchers?

2 Introduction

In this project, I will be using an Empirical Bayes method, specifically the James-Stein estimator, to predict pitcher performance in baseball. The main statistic that this project focuses on is earned run average because it is one of the most common statistics used to evaluate a pitcher. The goal of a pitcher is to prevent the other team from scoring, so earned run average is a statistic that shows how successful a pitcher is at preventing runs. Earned run average is the number of runs a pitcher typically gives up per nine innings pitched without any errors from the defense, and a lower ERA indicates better performance since pitchers want to limit the number of runs allowed.

For example, if a pitcher pitches three innings and gives up one earned run, then his ERA will be 3.00 for that appearance ((1 earned run/ 3 innings pitched) x 9 = 3.00). To calculate a pitcher's ERA for an entire season, we find the total number of earned runs they allowed as well as the total number of innings pitched. To give an example of this, Jacob deGrom allowed 41 earned runs in 217 innings pitched in 2018 which gave him an ERA of 1.70 ((41/217) x 9 = 1.70).

There is often variation in a pitcher's ERA from year-to-year, so I am using an Empirical Bayes method to attempt to accurately predict pitcher ERA in the following season. Since I will be using data from past seasons, I will be able to compare my predictions to the actual results to evaluate the performance of the James-Stein estimator.

I will first provide an overview of the James-Stein estimator before applying the model to data from the 2018 and 2019 Major League Baseball seasons (Willman 2018, Willman (2019)). The data from 2018 will be used to help predict performance for each pitcher in the 2019 season, and the James-Stein estimator will be compared to maximum likelihood estimator to see which performs better. The data from the 2019 season will be used to compare the actual ERA for each pitcher to our predicted values. I am interested to

see whether the James-Stein or maximum likelihood estimator does a better job on predicting pitcher performance. The James-Stein estimator should outperform the maximum likelihood estimator because the James-Stein estimator is biased and contains a shrinkage parameter that brings all estimates towards the group mean (Efron & Hastie 2016). The amount of shrinkage depends on how far away from the league average the pitcher's ERA was. While this does introduce bias into our model, it allows the James-Stein estimator to outperform the maximum likelihood estimator in high-dimensional cases like this one.

I will also provide background on the pitching data that I will be using throughout this report. This is pitching data from the 2018 and 2019 MLB seasons that includes every pitcher who faced at least 150 hitters during both seasons and their respective ERAs each year. Once I have explained the James-Stein estimator as well as the data I am using, I will be able to fit the model and predict pitcher ERA in the 2019 season for all these pitchers. There are 302 pitchers who met the criteria mentioned above in both seasons, so these 302 pitchers are the ones I will try to predict performance for, and then will compare their predictions to the actual results.

To display the results, I created a Shiny App that includes the actual performance for each pitcher in 2018 and 2019 as well as their predicted ERA based on the James-Stein estimator and maximum likelihood estimator. The Shiny App allows users to look at specific players to see how accurate the model predictions were for those players. It also includes a tab that explains how the James-Stein estimator works, so that users can understand where the predictions came from. The Shiny App provides additional insight into the data which helps build on the analysis included in the report.

3 Background

3.1 James-Stein Estimator

Before fitting the model on our data, I want to provide an overview about the James-Stein Estimator, specifically focusing on why it is an ideal model to use in this situation. The James-Stein estimator, which is a type of Empirical Bayes model, is a biased estimator that utilizes a shrinkage parameter to improve overall performance. From the textbook

Computer Age Statistical Inference, "the James-Stein estimator is a data-based rule for compromising between the null hypothesis of no differences and the MLE's tacit assumption of no relationship at all among the μ_i values" (Efron & Hastie 2016). The shrinkage parameter allows for a compromise between two conflicting arguments of assuming no difference between the individual means and assuming no relationship between the individual means. The usage of a shrinkage parameter is why the James-Stein estimator is able to outperform the maximum likelihood estimator. Unlike the James-Stein estimator, the maximum likelihood estimator assumes there is no relationship between the ERA's of different pitchers, so it only cares about the performance of the specific pitcher it is trying to estimate. As a result, the predicted ERA for the following season will be the ERA that the pitcher had in the current season for the maximum likelihood estimator (μ_i is the MLE where μ_i is the ERA from the current season).

For our case with the James-Stein estimator, the shrinkage parameter is impacted by the statistic we are trying to predict, which is ERA. If a pitcher has an ERA near the league average, then the shrinkage will not be very large because the purpose of the James-Stein is to shrink everything back towards the mean, and this is not necessary when the estimate is already close to the mean. On the other hand, if a pitcher has a very high or low ERA, then they will have a larger shrinkage as the estimator brings their predicted ERA back towards the league average. From Computer Age of Statistical Inference page 94 (Efron & Hastie 2016) , the formula for the James-Stein estimate is $\hat{x_i}^{js} = \sigma_0 \hat{\mu_i}^{js}$. In the model fitting section of the report, I will calculate σ_0 and $\hat{\mu_i}^{js}$, and then multiply them together to the get the James-Stein estimate for each pitcher.

The James-Stein estimator is expected to outperform the maximum likelihood estimator in high-dimensional cases which is why I am using it in this context because I am trying to predict pitcher ERA for hundreds of different pitchers. The James-Stein estimator has been used before to predict batting averages (by Brown (2008) and Jiang & Zhang (2010)), but predicting earned run average is a new application of the James-Stein estimator. The main difference between using earned run average compared to batting average is that batting average follows a binomial distribution while earned run average does not. Batting average is easier to predict because it is always between 0 and 1 while earned run average is tougher

because it can range from 0 to infinity. Instead of a binomial distribution like Brown and Jiang used, I used a normal distribution to help predict ERA since ERA roughly follows a normal distribution.

3.2 Pitching Data

Since I am using the James-Stein estimator to predict pitcher performance, I need data from individual MLB pitchers to predict their future performance. This data comes from baseballsavant.com which allows me to download a .csv file that contains statistics (including ERA) for every pitcher who faced at least 150 batters in the 2018 season (Willman 2018). This data will be used to fit the James-Stein estimator, and will allow me to predict pitcher performance in the next season. I also have the data from baseballsavant.com for the 2019 season for all pitchers who faced at least 150 batters, so that I can compare both the James-Stein estimator and the maximum likelihood estimator to the actual ERAs (Willman 2019). There are a total of 302 pitchers that faced at least 150 hitters in both seasons, so these are the pitchers I will be focusing on in my analysis. As mentioned earlier, the James-Stein estimator has been used to predict batting average before, but this is a new application to apply it to pitching statistics. I wanted to explore a new application of Empirical Bayes, so I chose to look at earned run average for pitchers in Major League Baseball. The wrangling code that reads in the data and gets us set up for the analysis can be seen in the appendix. Now that the data is in the proper format, we can complete some preliminary analysis to explore the data.

4 Preliminary Analysis

Before fitting the James-Stein estimator, I want to do some preliminary analysis on the pitching data. I would like to explore both the 2018 and 2019 pitching data before fitting the James-Stein estimator. I will start with an overlay that includes two density lines to compare the distribution of ERA in both of these seasons.

Figure 1 displays the distribution of earned run average for both the 2018 and 2019 MLB seasons. From the overlap of the 2018 and 2019 ERAs, we see that pitcher ERAs

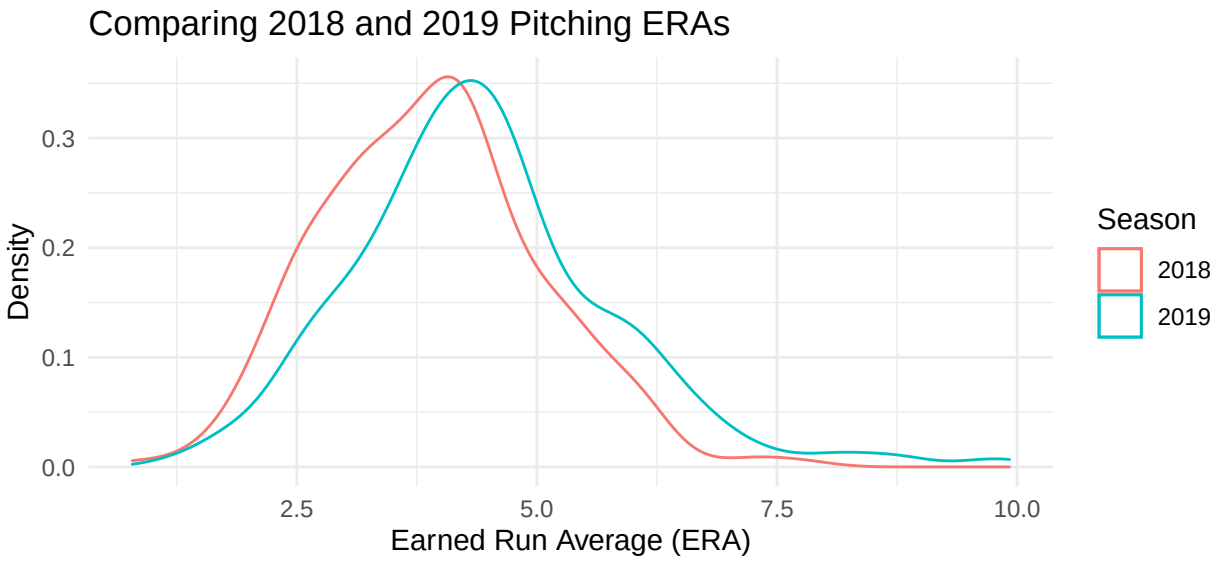

Figure 1: Overlaying Distribution of Pitching ERAs for the 2018 and 2019 MLB Seasons. The red line which represents the distribution of ERA in the 2018 season is further left than the blue line that represents ERA in the 2019 season which demonstrates that ERAs were higher in the 2019 season.

were higher in the 2019 season compared to the 2018. By calculating the average ERA for each of these seasons, we can see that average ERA for qualified pitchers (pitchers who faced at least 150 hitters in both seasons) was .52 runs higher in 2019 (3.9 in 2019 compared to 4.42 in 2018). For those who follow baseball closely, there were concerns about a "juiced" ball in 2019 that led to more home runs for hitters. This could explain why the average ERA increased by half a run in just one season, but I will not dive too deep into that. Since our predictions are based on the 2018 data, it seems likely that a lot of the predictions will be too low since average ERA increased by a decent amount in 2019 (and the James-Stein estimator shrinks everything towards the average ERA in 2018). This will be something to explore once the predictions have been made for each player using the James-Stein estimator.

I would also like to explore the relationship between 2018 and 2019 ERA through a scatter plot. I know that the average ERA increased from 2018 to 2019, but I am interested to see how individual pitchers ERA's vary from one season to the next.

Figure 2 shows that there is variation in a pitcher's ERA from year to year with some

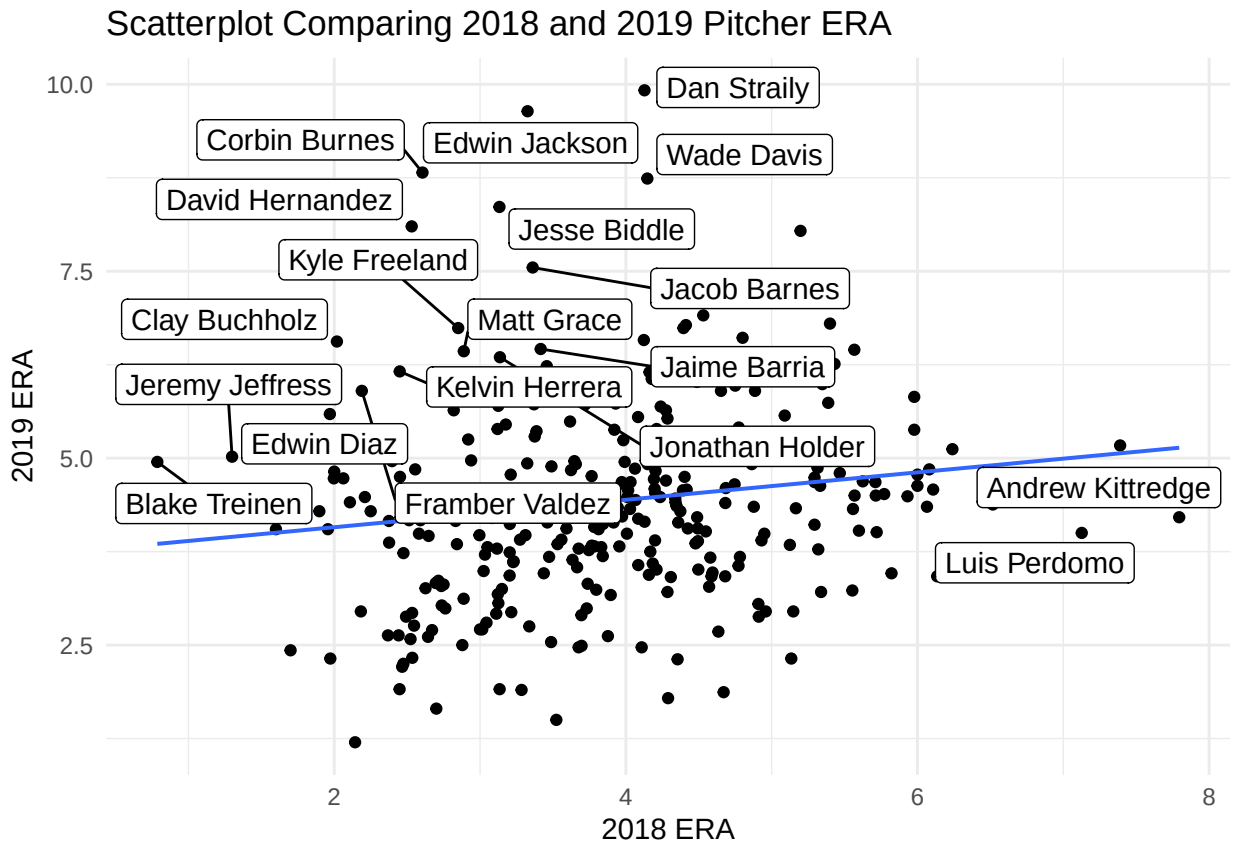

Figure 2: Comparing Individual Pitcher ERA for 2018 and 2019. The players who are labeled had an ERA that differed by at least three runs during the 2018 and 2019 seasons. Most of those players had much higher ERAs in 2019 which is why they can be seen near the top edge of the plot. The two pitchers, Andrew Kittredge and Luis Perdomo, that had much higher ERAs in 2018 are on the right edge of the plot.

pitchers having large differences in ERA between the two seasons. Pitchers whose ERA differed by at least three runs between 2018 and 2019 are labeled in the plot. Pitchers who had a very high ERA in 2018, but an ERA closer to average in 2019 can be seen on the right edge of the plot (this includes Andrew Kittredge and Luis Perdomo). On the flip side, pitchers who had a quality ERA in 2018, but a very high ERA in 2019 can be seen on the top edge of the graph (this includes Edwin Jackson, Dan Straily, Wade Davis, and Corbin Burnes among others). The pitchers in the middle of the plot were more consistent as their ERA was close to average in both seasons observed. The James-Stein estimator will likely be much more accurate for the players that were consistent, and far less accurate for the pitchers who had a large difference in their ERA between the two seasons. Now that I have explored the data, I am ready to fit the James-Stein model on our data.

5 Model Fitting

I am ready to fit the James-Stein estimator to this data. The code below finds the league average ERA for the 2018 season and also tells us how many pitchers are in the data set which we will use in our James-Stein estimator equation.

```
# Define variables for James-Stein estimator (see CASI p. 93)
xbar <- mean(pitchingtotal$MLE) #Average ERA of all pitchers in data for 2018
N <- length(pitchingtotal$MLE) #Number of pitchers in our data set
```

From Computer Age of Statistical Inference page 93 (Efron & Hastie 2016), we have the equation for $\hat{\mu}_i$ which is the predicted value of ERA for each pitcher if we assumed variance was 1. The equation is $\hat{\mu_i}^{js} = \bar{x} + (1 - \frac{N-3}{S})(x_i - \bar{x})$ where S $= \sum_{i=1}^{N}(x_i - \bar{x})^2$. Using our data, x_i is the pitcher's ERA during the 2018 season, $\bar{x}$ is the average ERA of the 2018 season (3.90 which we found above), and N is the number of pitchers in our data (302 found above). However, since we can estimate the variance for our normal distribution, we can use another equation from Computer Age of Statistical Inference to predict ERA (Efron & Hastie 2016). From Computer Age of Statistical Inference page 94 (Efron & Hastie 2016) , $\hat{x_i}^{js} = \sigma_0 \hat{\mu_i}^{js}$, so we need to calculate $\hat{\mu_i}^{js}$ and then multiply it by the square root of the variance.

In the example from Computer Age of Statistical Inference, they are estimating batting average which follows a binomial distribution, so they are using binomial variance (Efron & Hastie 2016). ERA does not follow a binomial distribution, but instead can be estimated by a normal distribution, so using the formula for variance of a normal distribution we can find σ_0^2.

In this case, $\sigma_0^2 = (\sum_{i=0}^{n}(ERA - \bar{x})^2)/N$ where N is the number of pitchers in our data. This comes from the formula for variance of a normal distribution where we have plugged in ERA as x_i, average ERA as $\bar{x}$, and number of pitchers as N into the formula to find variance. This means that all pitchers have the same variance which is a part of the James-Stein estimator.

Now that we have σ_0^2, we can get σ_0 by taking the square root. Then, we can calculate the predicted ERA for each pitcher in the 2019 season by multiplying σ_0 by $\hat{\mu_i}^{js}$. The code below calculates the James-Stein estimate for each player by finding $\hat{\mu_i}^{js}$ and σ_0 and multiplying them together.

```
estimators <- pitchingtotal %>% #Calculate JS estimator
  mutate(
    sigma2 = sum((ERA - xbar)^2)/N, #Variance calculation
    S  = sum((ERA-xbar)^2), #Formula for S from page 92 of CASI
    bhat = 1 - (N - 3)/S, #Formula for bhat from page 92 of CASI
    mu_i = xbar + bhat*(ERA - xbar), #Predicted mean for each pitcher
    JS = round(sqrt(sigma2)*mu_i,2), #JS estimate
    shrink = round(MLE - JS,2), #Shrinkage
    JS_error = round(TRUE_ERA - JS,2) #Error for JS estimator
  ) %>%
  select(player_name, TRUE_ERA, MLE, JS, shrink, JS_error)
```

Table 1 displays the name, ERA, MLE prediction, JS prediction, shrinkage, and JS error for the 15 pitchers with the largest negative shrinkage values. These are the 15 pitchers with the lowest ERAs in the 2018 season which can be seen in the MLE prediction. Almost all of these players regressed in 2019 as 14 of the 15 pitchers on the list had a higher ERA in 2019 than they did 2018. This demonstrates the effectiveness of the shrinkage parameter

Table 1: Glimpse at Pitcher Predictions and Shrinkage

player_name	TRUE_ERA	MLE	JS	shrink	JS_error
Blake Treinen	4.95	0.79	3.62	-2.83	1.33
Jeremy Jeffress	5.02	1.30	3.75	-2.45	1.27
Jose Leclerc	4.35	1.57	3.82	-2.25	0.53
Sean Doolittle	4.05	1.60	3.82	-2.22	0.23
Jacob deGrom	2.43	1.70	3.85	-2.15	-1.42
Blake Snell	4.29	1.90	3.90	-2.00	0.39
Jared Hughes	4.05	1.96	3.92	-1.96	0.13
Hyun Jin Ryu	2.32	1.97	3.92	-1.95	-1.60
Edwin Diaz	5.59	1.97	3.92	-1.95	1.67
Collin McHugh	4.73	2.00	3.93	-1.93	0.80
T.J. McFarland	4.82	2.00	3.93	-1.93	0.89
Clay Buchholz	6.56	2.02	3.93	-1.91	2.63
Matt Strahm	4.73	2.06	3.94	-1.88	0.79
Chris Sale	4.41	2.11	3.95	-1.84	0.46
Kirby Yates	1.20	2.14	3.96	-1.82	-2.76

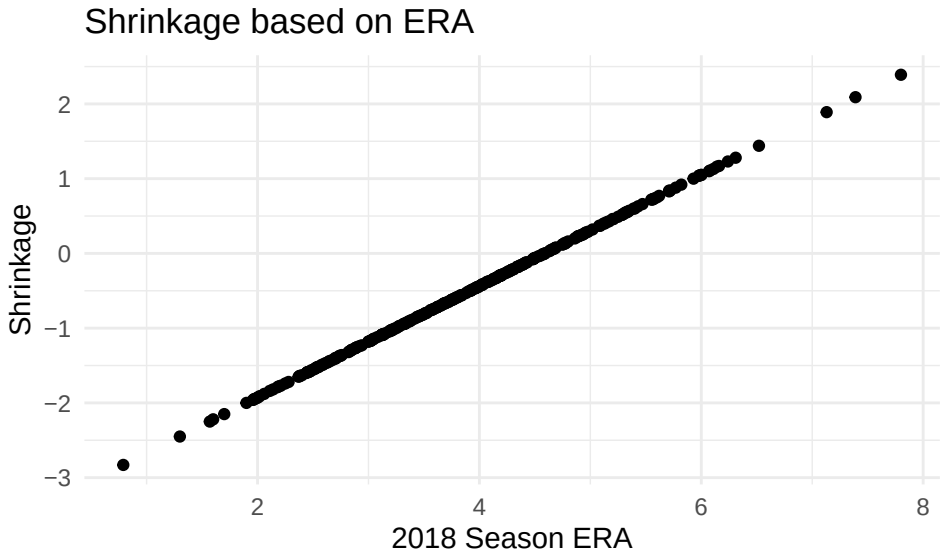

Figure 3: Amount of Shrinkage Based on Pitcher ERA in 2018. The pitchers with the lowest ERAs in 2018 have the most negative shrinkage values while the pitchers with the highest ERAs in 2018 have the largest positive shrinkage.

as the James-Stein estimator expected these players to regress back towards the league average, and almost all of these pitchers did.

Before going further into the analysis, I wanted to look at how the shrinkage parameter is related to ERA. I am expecting the shrinkage to be larger for pitchers who had an ERA farther away from the mean, so I will check to see if this is true.

Figure 3 illustrates the amount of shrinkage for each pitcher's predicted ERA based on their 2018 season ERA. As expected, we see that pitchers with extreme ERA values had much larger shrinkage values than those pitchers who had an ERA near league average. There is a linear relationship between ERA and shrinkage since ERA is the only variable in the data that impacts the amount of shrinkage in the prediction.

It may seem odd that ERA is the only statistic that impacts the amount of shrinkage in the James-Stein estimator. It seems reasonable to assume other statistics like innings pitched could also impact the amount of shrinkage that occurs. Pitchers who pitched more innings likely have more stable ERAs, so they will likely not have as much variation from year to year compared to pitchers who threw fewer innings. Innings pitched would play a role if we were using Bayes, but it does not play a role here since we are using the James-

Table 2: Average Error for JS and MLE Estimators

JS_Prediction_Error	MLE_Prediction_Error
1.803836	2.923653

Stein estimator which is an Empirical Bayes method. Since the James-Stein estimator is a balance between Bayesian and Frequentist statistics, and slightly more influenced by Frequentist, the James-Stein estimator only relies on the statistic of interest, which is ERA in this case, to impact the shrinkage parameter.

6 Analysis

I would like to start by comparing the James-Stein estimator to the maximum likelihood estimator to see which one performed better on our data set.

```
#Find errors for each method
errors <- estimators %>%
  mutate(
    js_pred_error_i = (JS - TRUE_ERA)^2, #Find each individual error for JS
    mle_pred_error_i = (MLE - TRUE_ERA)^2 #Find each individual error for MLE
  ) %>%
  summarise(
    JS_Prediction_Error = sum(js_pred_error_i)/N, #Find average error rate for JS
    MLE_Prediction_Error = sum(mle_pred_error_i)/N #Find average error rate for MLE
  )
```

Table 2 illustrates the average error for both the James-Stein estimator and the MLE. The James-Stein estimator outperforms the maximum likelihood estimator in this situation as the average error is a full run lower for the James-Stein estimator.

By calculating the error for each observation and finding the average error, we see that the James-Stein estimator did in fact outperform the MLE. The James-Stein estimator had an average error rate of 1.80 while the MLE had an average error rate of 2.92. Both of these

error rates are somewhat high, but that is in part due to the fact that ERA can fluctuate heavily from year to year. These error rates are significantly higher than the ones we saw in the papers by Brown (2008) and Jiang & Zhang (2010), but that is in part because they were predicting batter average which can only be between 0 and 1. As a result, it is not a surprise that the error rates for ERA were much higher than batting average. The individual error rates for each pitcher can be seen in the Shiny App that I created to help display my results. Now, I would like to look at some specific players to better illustrate how the James-Stein method works.

I am starting with the player who had the largest positive shrinkage, and that is Andrew Kittredge. Andrew Kittredge had such a large shrinkage because he had the worst ERA of any pitcher in the data set during the 2018 season with an ERA of 7.80 which is basically twice as high as the league average ERA for that season. Kittredge had a very high shrinkage because his ERA was very far from league average, so the estimator shrunk his prediction towards the league average much more than others. Kittredge had a shrinkage of 2.39 as his MLE prediction was 7.80 while his JS prediction was 5.41, but he ended up outperforming both predictions during the 2019 season. Kittredge ended up with an ERA of 4.21 in the 2019 season, so the James-Stein estimate was off by 1.20 while the MLE was off by 3.59. The James-Stein shrunk the predicted ERA for Kittredge back towards league average, and as a result, the James-Stein estimator outperformed the MLE in this specific instance.

Another extreme case comes from Blake Treinen, who had the largest negative shrinkage of any player in the data set. Blake Treinen, who had the lowest ERA in the 2018 season, had a large shrinkage as his predicted ERA was brought back towards league average. Treinen's shrinkage was -2.83 as his MLE prediction was 0.79 while our James-Stein estimator predicted an ERA of 3.62. The James-Stein estimator shrunk Treinen's prediction back towards the mean because it was unlikely that he could repeat such an incredible performance, and would instead regress. It was true that Treinen regressed in the 2019 season, but he regressed more than our estimator predicted. He ended up with an ERA of 4.95 which was a below-average ERA in the 2019 season, so both estimators predicted an ERA much lower than Treinen actually achieved. The James-Stein estimator still outper-

formed the MLE in this instance as it shrunk Treinen's prediction back towards the mean instead of assuming he would be able to repeat his performance like the MLE does.

Another example I want to look at is Brad Keller because his James-Stein prediction was the most accurate of the 302 pitchers in the data. Brad Keller had an ERA of 3.08 in 2018, but his James-Stein prediction was 4.20, so he had a shrinkage value of 1.12. In 2019, his actual ERA was 4.19, so our James-Stein estimator was only off by .01 while the MLE was off by 1.11. I wanted to highlight this example to show a very accurate prediction for the James-Stein estimator.

The final example I want to look at is Dan Straily because his James-Stein prediction had the largest error of any pitcher in the data. His predicted ERA for the 2019 season based on James-Stein was 4.47 while his actual ERA was 9.92. This resulted in an error of 5.45 which is very high, but this was due to Dan Straily's extremely poor performance during the 2019 season. In fact, Straily ended up leaving the MLB after the 2019 season to sign with a team in the Korean Baseball Organization. While the James-Stein estimator did a very poor job of predicting Straily's ERA, this is not surprising because nobody could have expected Straily to pitch so poorly after having an ERA of 4.13 in 2018. I wanted to highlight this example to show how much ERA can vary from year to year which is why some of the predictions are not very accurate. The purpose of the James-Stein estimator is to "improve overall performance, at a possible danger to individual estimates" (Efron & Hastie 2016). This is one case where the James-Stein prediction was not very close, but it did still outperform the MLE which predicted Straily would have an ERA of 4.13 in 2019.

To investigate the results further, the link to the Shiny App I created is `https://r.amherst.edu/apps/tmarshall21/james-stein/`. This allows for more exploration of the project for those that are interested.

From this, we can see that the James-Stein estimator brings extreme values back towards the mean while it does not influence values near the mean very much. In our case, this means it predicts improvement for pitchers with very high ERAs while it predicts regression for pitchers with very low ERAs.

Figure 4 and Figure 5 show how the predicted ERA compares to the actual 2019 ERA for each pitcher based on both the James-Stein and maximum likelihood estimator. All

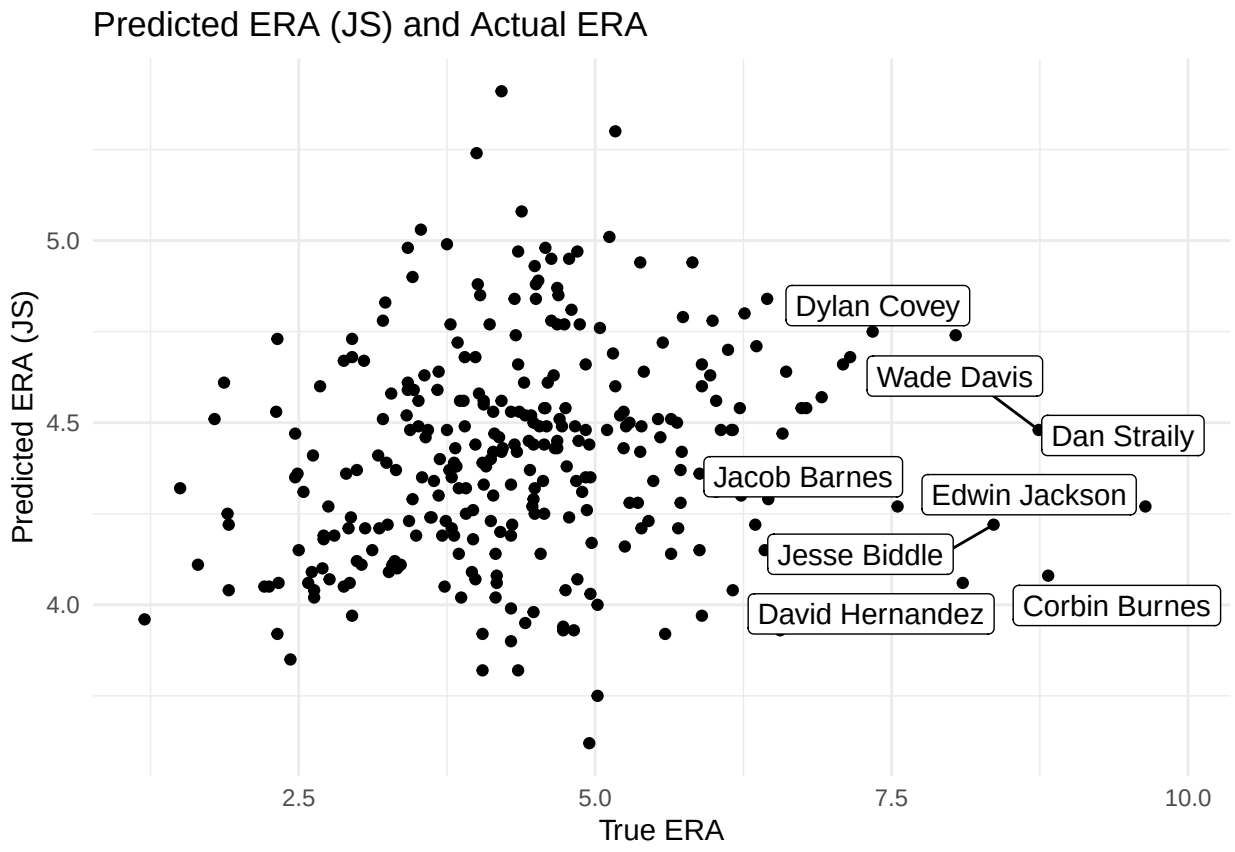

Figure 4: Comparing Predicted ERA(JS) with Actual ERA in 2019. The pitchers who had an actual ERA that was at least three runs away from their James-Stein prediction are labeled in the plot. The eight pitchers who met this criteria all had an actual ERA that was more than three runs higher than their prediction.

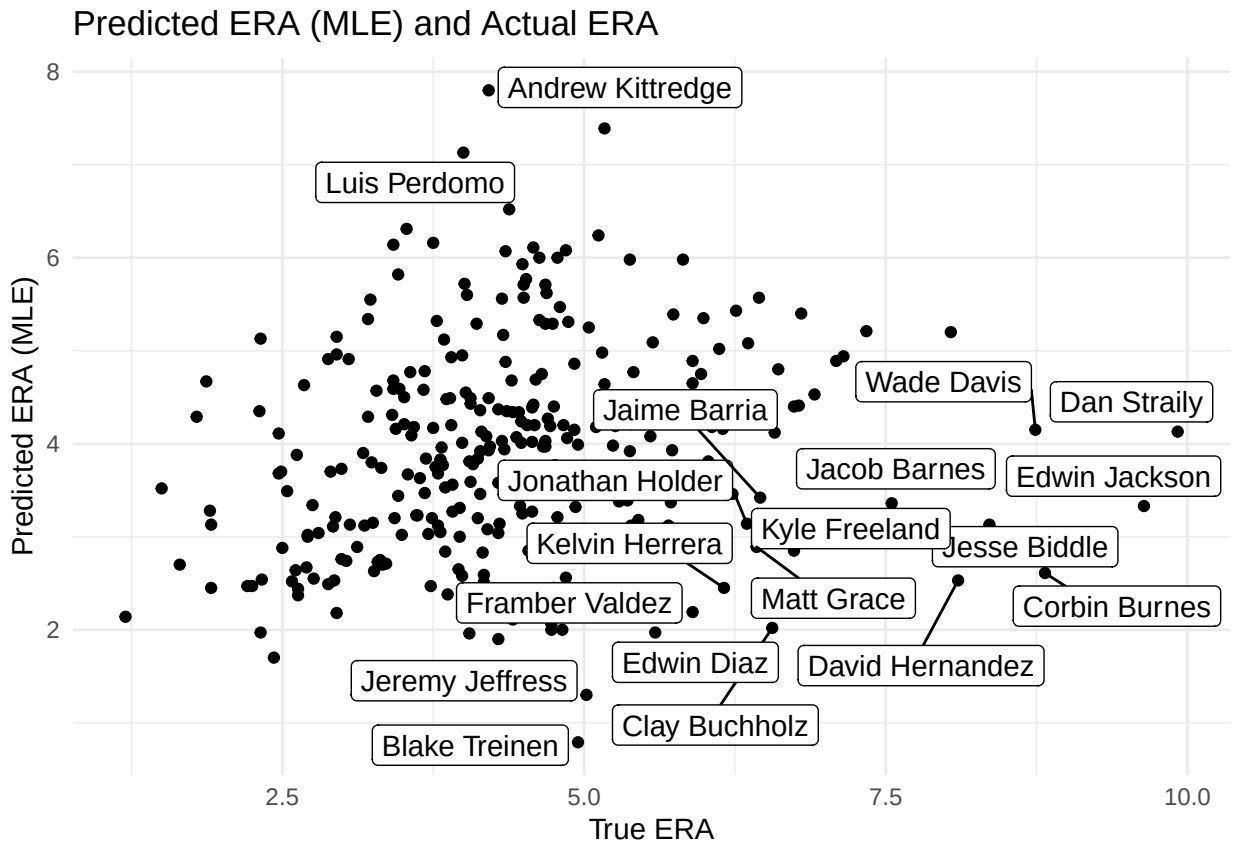

Figure 5: Comparing Predicted ERA(MLE) with Actual ERA in 2019. There were nineteen pitchers who had an actual ERA that differed from their prediction by at least three runs, and they are labeled in the plot. Seventeen of these pitchers had an ERA that was greater than their prediction by more than three runs while two pitchers had an actual ERA that was less than their predicted ERA by at least three runs.

pitchers who had actual ERA that differed by more than three runs from their predicted ERA are labeled in the plots. With the James-Stein estimator, only eight pitchers met this criteria, and all eight had an actual ERA that was at least three runs higher than their prediction. These pitchers all performed very poorly during the 2019 season as they all had ERAs over 7.5. For the maximum likelihood estimator, nineteen pitchers met this threshold. Seventeen of these pitchers had an ERA that was at least three runs higher than their prediction while two pitchers had an actual ERA that was at least three runs lower than their predicted ERA. From the analysis above, we can see that the James-Stein estimator does a better job of predicting ERA in the following season than the maximum likelihood estimator.

7 Conclusion

After analyzing both models, I found that the James-Stein estimator did in fact outperform the maximum likelihood estimator on my data. The James-Stein estimator had a smaller average error than the maximum likelihood estimator which is what I expected to happen. This is because the James-Stein estimator includes a shrinkage parameter that helps bring the estimates back towards the mean. The accuracy of the James-Stein estimator varied based on the player as it was very accurate for some players and not very accurate for others. The players that were consistent from 2018 to 2019 had much more accurate predictions than pitchers who had a large difference between their 2018 and 2019 performances. I think the James-Stein estimator could be more accurate in its prediction if I had used multiple seasons worth of ERA to predict the 2019 season because multiple seasons would be more reflective of a pitcher's true performance compared to a single season. However, this would have required more data and would have limited the number of pitchers that could be included since they would have needed to pitch in more than just the 2018 and 2019 seasons. I would be interested to try this in a future project to see if the James-Stein estimator would be more accurate as long as there is sufficient data available.

The James-Stein estimator has now been applied to earned run average as well as batting average (by Jiang & Zhang (2010) and Brown (2008) and a few others), but I am interested to see other applications of this method to baseball statistics. I think on-base percentage

is another offensive statistic that the James-Stein estimator could be applied to although it would likely be similar to batting average. Another pitching statistic that I would be interested to see the James-Stein estimator applied to is Fielding Independent Pitching (FIP). Fielding Independent Pitching is a similar statistic to ERA, but it only accounts for statistics that are within a pitcher's control (home runs, walks, hit by pitches, strikeouts). I would be interested to see if the James-Stein estimator performs better or worse on Fielding Independent Pitching compared to ERA. I think the James-Stein estimator could be applied to other baseball statistics as well, but I think these two applications in particular would be interesting.

Since the James-Stein estimator can be used in any situation where a large number of means are being estimated, it is not limited to baseball data. I am also interested to see other examples of the James-Stein estimator on different types of data. Other potential applications could include making predictions for countries and states or making predictions in education for schools. The James-Stein estimator is a very useful technique, so I look forward to exploring it even more in the future.

8 Appendix

The code below is the wrangling that reads in the data and gets us set up for the analysis.

```
#Read in data for 2018 and 2019
pitching2018 <- readr::read_csv("pitching2018.csv") %>%
  janitor::clean_names() %>%
  mutate(ERA = (p_earned_run/p_formatted_ip) *9) %>% #Calculate ERA
  select(last_name, first_name, year, player_age, p_game, p_formatted_ip, ERA) %>%
  mutate(player_name = paste0(first_name, " ", last_name))
pitching2019 <-readr::read_csv("pitching2019.csv") %>%
  janitor::clean_names() %>%
  mutate(TRUE_ERA = (p_earned_run/p_formatted_ip) *9) %>% #Calculate ERA
  select(last_name, first_name, year, player_age, p_game, p_formatted_ip, TRUE_ERA) %>%
  mutate(player_name = paste0(first_name, " ", last_name)) #Create name variable
```

```
#Join the two datasets
pitchingtotal <- inner_join(pitching2018, pitching2019, by = "player_name") %>%
  mutate(MLE = ERA, #MLE is ERA from 2018
         year.x = as.factor(year.x),
         year.y = as.factor(year.y)) %>% #Change year to factor variables
  rename(innings_pitched = p_formatted_ip.x,
         currentyear = year.x,
         nextyear = year.y) #Rename variables
```

References

Brown, L. D. (2008), 'In-season prediction of batting averages: A field test of empirical bayes and bayes methodologies', *The Annals of Applied Statistics* **2**(1), 113–152.

Efron, B. & Hastie, T. (2016), *Computer Age Statistical Inference*, Cambridge University Press.

Jiang, W. & Zhang, C.-H. (2010), 'Empirical bayes in-season prediction of baseball batting averages', *Institute of Mathematical Statistics* **6**, 263–273.

Willman, D. (2018), 'Pitching statistics for 2018 mlb season'.
URL: *https://baseballsavant.mlb.com/leaderboard*

Willman, D. (2019), 'Pitching statistics for 2019 mlb season'.
URL: *https://baseballsavant.mlb.com/leaderboard*