Estimating Winning Percentage in Softball

Using Stein Estimators

Bailey Thompson

A Senior Project paper presented to
The Mathematics Department
Bachelor of Science
Mathematics
Seattle University
June 1, 2020

1 Abstract

In this paper, I show a strong association between runs scored and runs allowed to winning percentage of a softball team in the most recent NCAA seasons. To do so, I use a Stein-estimator based data from the 2017, 2018, and 2019 NCAA seasons to develop a model for calculating what a team's runs and runs allowed would be given their scoring tendencies in the beginning of the season. Using that estimation, I then create a linear regression to determine what a team's winning percentage would be given their scoring data, before using the Bill James Pythagorean Winning Percentage for an even more accurate estimate. Since the Pythagorean Winning Percentage formula was originally used for baseball, I optimized the formula exponent to fit softball specifically. With the cancellation of the 2020 NCAA seasons due to COVID-19, I thought it would be fitting to determine who may have been on top in this 2020 season using data from the first 3 weeks of play before campuses were shut down.

Table of Contents

2 Introduction

There are two integral parts to the game of softball. Offense and Defense. One of the age-old debates among fans of many sports is relative merits of a prolific offense vs. a stingy defense. While obviously a team would like to excel in both areas, does one aspect of the game have a stronger influence on winning than the other? If a team excels in both areas, we consider them elite, but, historically, how important are runs to the game of softball? Offensive strategy is mostly just to hit the ball skillfully to let the batter reach base and advance other runners around the bases to score runs. The count of balls and strikes is an indicator of how aggressive the batter should be. The offense may try to sacrifice, with the batter deliberately making an out in order to advance runners. Defensive strategy is more complex, as particular situations (number of outs and positions of base-runners) and particular batters call for different positioning of fielders and different tactical decisions. The defense may decide to allow a run if it can achieve one or multiple outs.

I want to explore the relationship between the amount of runs a team scores and their winning percentage. Of course a team needs to score runs to win, but at some point is is all relative to the number of runs you allow to be scored against you. We can start by taking numbers from the most recent season of NCAA Softball, focusing on win percentage, average runs scored per game, and average runs allowed per game. I decided to use runs scored and runs allowed per game instead of earned runs in hopes that this will somewhat absorb fielders' choices and errors made by both teams.

Throughout the course of this project, the 2020 NCAA Women's College World Series was cancelled, and with that all games after March 8th due to COVID-19. With this information, I thought it would be appropriate to create some kind of estimate for which teams were on track for success before the abrupt cutoff in games.

3 What does it take to win?

In the following section, I describe two specific research topics I explored during this project. The first is the Bill James Pythagorean Winning Percentage. I went through the process of understanding the history behind the derivation of this equation as well as computed my own exponent for softball since I couldn't find any previous work that had been done. Next, for the bulk of my research, I studied Stein Estimators, and more specifically the James-Stein estimator for prediction. This is what I used to create my approximations based on the first ten games of each of the past three seasons, and tested it's accuracy against what really happened.

3.1 Bill James and the Pythagorean Winning Percentage

In order to create these biases, I want to assume that runs scored is a good predictor of winning percentage. In Figures 1 and 2, we see that runs per game and runs allowed per game trended with overall winning percentage for the 2019 season. It is the same case with the past 3 seasons as well.

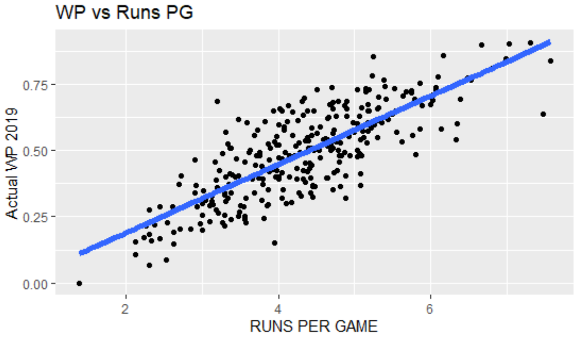

Figure 1: runs scored per game vs overall winning percentage in 2019

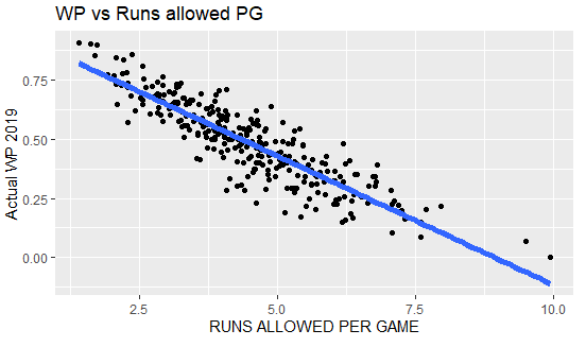

Figure 2: runs allowed per game vs overall winning percentage in 2019

Although, at some point, it doesn't matter how many runs you score so long as you hold your opponents to as little runs as possible. That being said, I created a variable called run differential, which I defined to be the difference between runs scored and runs allowed. In Figure 3, we see that there is a strong, positive, linear correlation between run differential and overall win percentage. So, you win more games if you outscore your opponents and limit the amount of runs scored against you. It makes sense. I then performed a linear regression on the variable run differential, and we see this in the green line in Figure 4. The association is obviously there, but it wasn't exactly how I wanted it to be when calculating win percentage. Since linear regressions go on infinitely, it doesn't behave accurately at the extremes.

Upon further research, I realized something like this had been perfected in the past. In 1985, sports analyst Bill James created a formula to estimate the amount of games a baseball team "should" win in a season given the amount of runs they've scored (RS) and allowed (RA). This formula, dubbed the "Pythagorean Winning Percentage" for its use of squares, has been used in many sports to determine an estimate for winning percentage based on scoring, however the exponent tends to vary. The formula is as

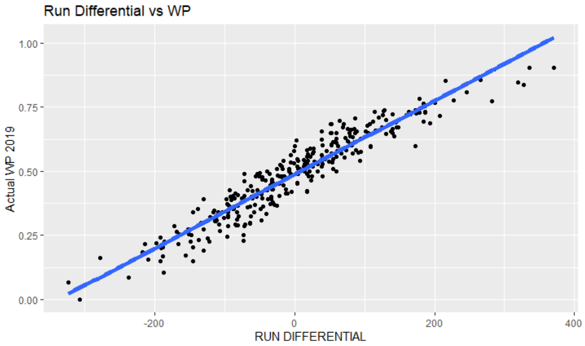

Figure 3: run differential vs overall winning percentage in 2019

follows:

$$WINNINGPCT = \frac{RS^2}{RS^2 + RA^2}$$

Figure 4 shows us that my linear regression (green line) is slightly less accurate than the Pythagorean Winning Percentage Formula (orange line). The mean of the residuals is near zero like with the linear regression. The Pythagorean Formula is very similar to my linear regression model. It also has the added benefit of reacting better at the extremes since the linear regression is linear it goes on infinitely and the Pythagorean does not.

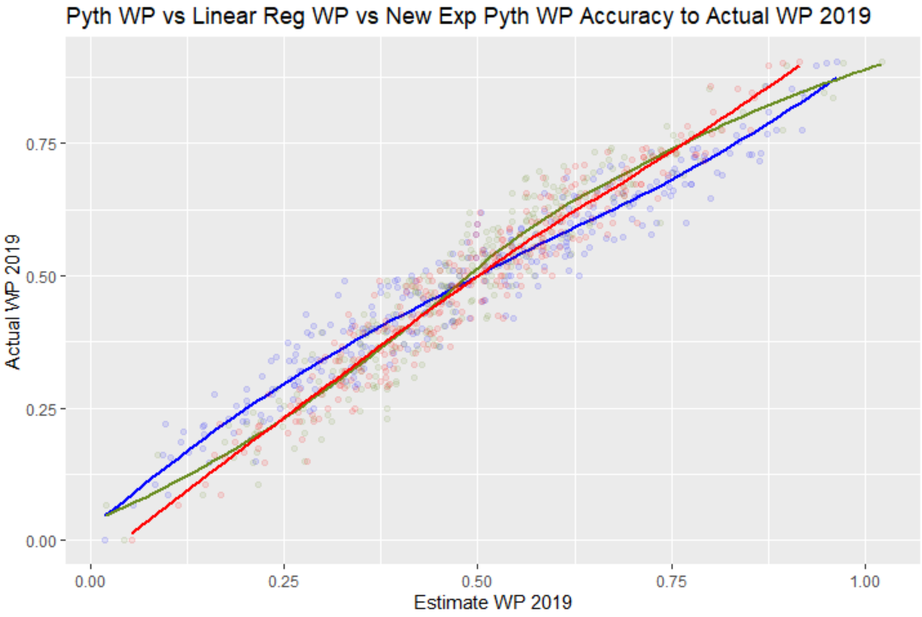

Figure 4: my created linear regression vs Pythagorean Winning Percentage model

Later analysis showed that the a better exponent for the Pythagorean Winning Percentage in baseball was around 1.88, and this exponent may be way different for other sports. So I decided to take a look at the last 3 seasons in NCAA softball and determine what a teams' winning percentage should look like depending on their run differentiation. Expanding the winning percentage formula and substituting 2 with our unknown exponent β, we see that:

$$\frac{W}{W+L} = \frac{RS^{\beta}}{RS^{\beta}+RA^{\beta}}$$

$$W = RS^{\beta}(W+L)/(RS^{\beta}+RA^{\beta})$$

$$WRS^{\beta} + WRA^{\beta} = WRS^{\beta} + LRS^{\beta}$$

$$WRA^{\beta} = LR^{\beta}$$

$$\frac{WRA^{\beta}}{L} = \frac{RS^{\beta}}{RA^{\beta}}$$

$$\frac{W}{L} = \frac{RS^{\beta}}{RA^{\beta}}$$

$$log(\frac{W}{L}) = \beta * log(\frac{RS}{RA})$$

$$\beta = \frac{log(\frac{RS}{RA})}{log(\frac{W}{L})}$$

Using this formula for the last three NCAA seasons, making sure to get rid of outliers (teams who won or lost no games that would disrupt the logs), I was able to come up with a constant approximately equal to 1.45. All constants ranged from 1.4 to 1.5 within that three year span. We see the accuracy of the Pythagorean Win Percentage with an exponent of 1.45 in the orange line in Figure 4.

For simplicity and accuracy purposes at the extremes, I will be using the Pythagorean Winning Percentage formula with an exponent of 1.45 in the final computation. In my final calculation, using this exponent instead of 2 decreased my mean squared error by on average about 45% for estimates over the past three seasons.

3.2 Empirical Bayes - Stein Estimator

Stein's paradox is the phenomenon that when three or more parameters are estimated simultaneously, there exist combined estimators more accurate on average (that is, having lower expected mean squared error) than any method that handles the parameters separately.

James-Stein estimators are a popular example of this, being a biased estimator of the mean of Gaussian random vectors. It can be shown that the James–Stein estimator dominates the "ordinary" least squares approach, i.e., it has lower mean squared error.

I am going to use the James-Stein estimators in an empirical Bayesian approach to parameter estimation. Using the first 10 games of a season, I would like to predict how many runs per game will a team score and allow, using how they scored at the beginning of the season to do so.

The James-Stein Estimator, as used by Efron and Morris when calculating batting average over a given season, is simplified to:

$$\hat{\theta}_i = \hat{\beta}\overline{y} + (1 - \hat{\beta}) - y_i,\ i = 1, ...k$$

where $\hat{\theta}_i$ is the biased estimator for the corresponding θ_i, $1-\hat{\beta}_i$ is the shrinking factor, y_i is the average of dataset i, and $\overline{y}$ is the grand average of averages in our approximation. This model resembles a linear regression $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Efron and Morris simplified a way for finding $1 - \beta$ even further, defining the shrinking factor to be

$$c = 1 - \frac{(k-3)\sigma^2}{\sum(y-\overline{y})^2}$$

where k is the number of unknown variables, σ^2 is the variance of the individual observations, and $\sum(y - \overline{y})^2$ reflects the variance from mean to mean.

4 Using the Stein Estimator

In 1977, Phil Everson used statistics from the 2006-2007 National Basketball Association (NBA) to estimate the average points per game scored by each team in a season using the first 10 games. Expanding off of Efron and Morris' idea of using a Stein Estimator to introduce deliberate biases, Everson used the average points per game for each NBA team in the first 10 games of the season to predict how they would score in the remaining 72 games. Although his results weren't exactly as anticipated with regards to the actual season, I figured that softball seasons were much shorter and runs tended to vary by less, so this may show more accuracy for me.

First I needed to establish what type of distribution scoring models in softball. Overall scoring of runs and runs allowed per game are negative binomial distributions and average runs scored/allowed per game can be modeled with a normal distribution as seen in Figures 5 and 6. Since all the teams I am measuring are playing in the same season, it's reasonable that team means (θ_is) would be similar with few outliers. This normal distribution also seems like a good description for how θ_is would vary.

With this information about scoring in mind, the Stein-Estimator is described as a two-level normal hierarchical model, Everson created a very comprehensible representation of what Efron and Morris described as " The Big Picture" in his publication "A Statistician Reads the Sports Pages" (2007) shown in Table 1.

The distributions on the left side of the table show Y_is vary about their respective θ_is. This is where we get the two-level structure. On the right side of the table are the marginal distributions of Y_is and the conditional distributions of each θ_i given $Y_i = y_i$. σ and therefore V are considered known in the Stein approximation. μ and A represent the mean and variance of the k unknown θ_is.

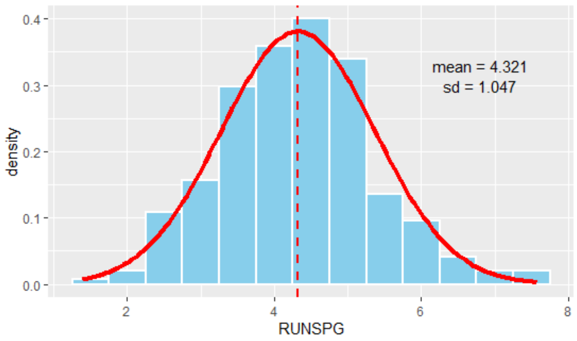

Figure 5: showing the normal distribution of runs scored per game

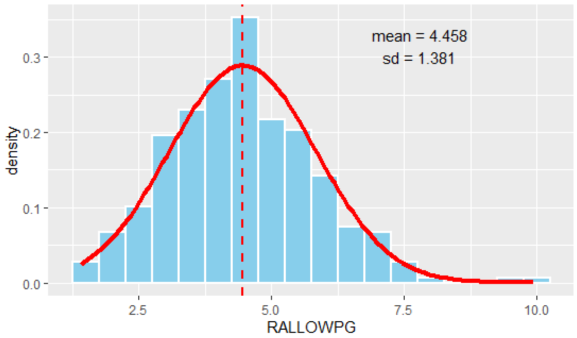

Figure 6: showing the normal distribution of runs allowed per game

4.1 Predicting Runs Scored and Runs Allowed

Using the same two-level normal hierarchical model Efron and Morris described, I computed these approximations for runs per game and runs allowed per game for every team in NCAA softball. Ten games were used because the James-Stein estimator requires equal variance, which is dependent upon the number of games you look at, but it was still early enough in the season to where this would actually be an early season estimate. Taking these considerations, I created my Stein Estimate for both runs scored per game and runs allowed per game in the 2017, 2018, and 2019 seasons. My results for each variable in 2019 are shown in Table 2.

4.1.1 Analysis

Table 2 shows my results modeling runs scored and runs allowed per games using James-Stein Estimators. My estimate for the standard deviation $\hat{\sigma} = 1.784$ for runs scored per game. Since our values ranged from 0.55 to 10.5 runs per game, this range isn't that large for $k = 295$ estimates. Same goes for our mean, with this range of values, our μ doesn't seem unusual. In both runs scored and allowed, $\hat{\beta}$ is small, and therefore $\hat{A}$ is large (> 0), which means most of the weight is put on the observed average, $\overline{y}$, rather than the overall mean μ.

$Y_i \mid \theta_i, \mu, A \overset{\text{indep}}{\sim} N(\theta_i, V)$	$Y_i \mid \mu, A \overset{\text{indep}}{\sim} N(\mu, V + A), i = 1, ..., k$
$\theta_i \mid \mu, A \overset{\text{indep}}{\sim} N(\mu, A)$	$\theta_i \mid \mu, A, y_i \overset{\text{indep}}{\sim} N(\beta\mu + (1-\beta)y_i, (1-\beta)y_i, (1-\beta)V),$ $B = V(V + A), V = \sigma^2/n$

Table 1: "The Big Picture" visualized by Everson

	Runs Scored	Runs Allowed
k	294	294
n	10	10
$\hat{\sigma}$	1.784	2.009
V	0.318	0.404
μ	4.418	4.816
$\sum(y - \overline{y})^2$	935.767	1186.812
$\hat{\beta}$	0.0993	0.0994
$\hat{A}$	2.886	3.661
$\sqrt{\hat{A}}$	1.699	1.913

Table 2: Calculations for Stein Approximations for Runs Scored and Allowed Per Game in first 10 games of 2019

Since Stein estimates assume a constant standard deviation, our variances for the average runs/runs allowed per game of $n = 10$ games is $V = 0.318$ and $V = 0.404$ respectively. To account for the uncertainty in the remaining 50 games, I will compute the standard deviation for runs scored per game to be $\sqrt{1.784^2/10 + 1.784^2/50} = 0.618$ and the standard deviation for runs allowed per game to be $\sqrt{2.009^2/10 + 2.009^2/50} = 0.696$. In most situations, the variances for k estimates will not all be the same as Stein estimations infer. Even if all the standard deviations were around the same point, a different value for n would result in a different variance value. For example, before March 10th, 2019, all 295 teams had played between 10-18 games. For my data, I took the first 8-10 games from each team depending on what was available for me in order to get an average value of n at 10. However, for a true season estimate based on games before March 10th, teams in the Ivy League who had only played 8 games each would have much more variance in my end answer than teams in the SEC who had already played 18 games. A larger n means a smaller V and therefore a smaller β, which puts more weight on the observed value y_i rather than the overall estimated mean μ.

Figure 7 shows how far my estimates strayed from the true average, as well as how this compared to if we were to just take the average runs/runs allowed per game from the first 10 games.

Figure 7: This graph shows the comparison between of runs/runs allowed per game from the first ten games, and after a Stein approximation to estimate the remainder of the season.

4.2 Winning Percentage

After finding what our teams scoring tendencies are given the first ten games of the season, I then used this along with the Bill James Pythagorean Winning Percentage Formula with my exponent of 1.45 to calculate the win percentage for every team based on the first 10 games of the 2017, 2018, and 2019 seasons. I then compared this to the actual win percentage for 2019 as well as the Pythagorean Winning Percentage Formula if we were to have just used the average runs/runs allowed per game from the first ten games. My results are shown in Figure 8.

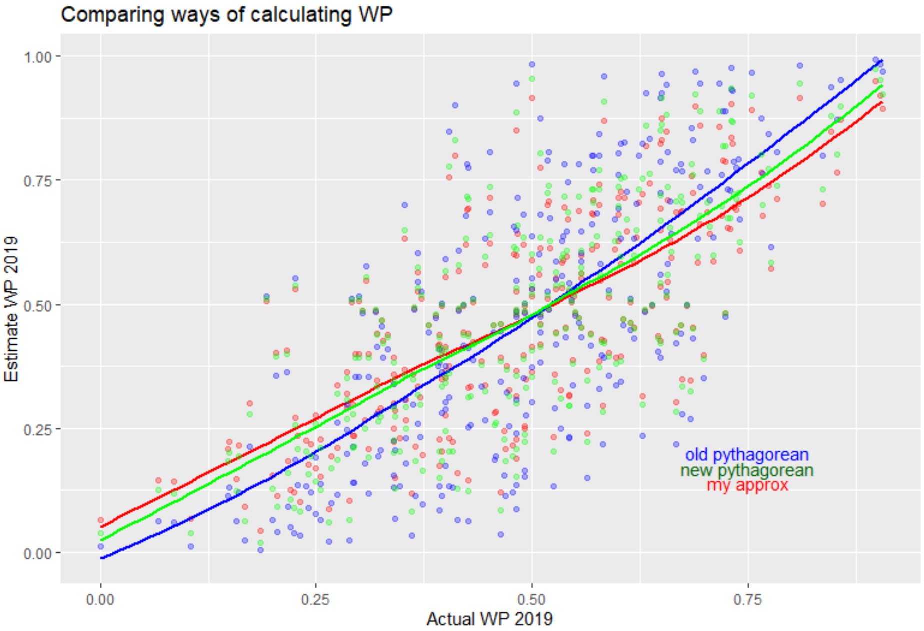

Figure 8: This graph shows the different ways of calculating winning percentage, all based on the Pythagorean Winning Percentage formula. The points are representative of each team in the approximations.

The blue line, labeled "Old Pythagorean" indicates our predicted winning percentage for 2019 from the first 10 games using the average runs per game from the first 10 games, without estimating what those approximations would look like at the end of the season. The green line, labeled "New Pythagorean" uses the same data, but with the new exponent of 1.45 we derived. Finally, the red line is representative of my approximation, in which we created Stein approximations for runs scored and runs allowed and then generated our predicted winning percentage.

We can see that the red dots are brought closer to the line, as the Stein approximations anticipate regression to the mean. I thought this would be at the expense of good teams that continue to do well, but it turns out this approximation was most successful at the higher extreme. Both of the other approximations estimated teams to have much higher win percentages than they did, while mine dragged them down closer to the mean. I think this is because with softball seasons being only about 60 games long, one game loss results in about 0.0167 point deduction from your overall season winning percentage. Considering even the best teams in softball lose 5-6 games per year, the high approximation for Louisiana to have a 0.993 winning percentage from the old Pythagorean and 0.974 for the new Pythagorean is nearly impossible. My approximation gave Louisiana a predicted winning percentage of .949, meaning they would have gone about 57-3 with a 60 game season. In reality, they were 52-6, playing only 58 games, and their win percentage was 0.897. My approximation created an overall mean square error of 0.0133, while the old Pythagorean gave 0.0255 and the new Pythagorean gave 0.0158.

All three lines perform well at the center of the graph, with the new Pythagorean estimate seeming to be the most accurate for middle-grade teams. However, at the extreme points, my approximation takes over. By anticipating some regression to the mean, we bring the win percentage of good teams down and of poor teams up, giving an overall better season estimate for better teams since, as I said before, even the best teams lose 5-6 games a season.

What I found most interesting about this computation was the fact that I got different top teams for each of my estimates.

5 2020 Predictions

As mentioned earlier, the cancellation of all World Series competitions left sports lover shook, so with my approximations I decided to predict the 2020 season using Stein Estimators and our new formula for calculating win percentage. Although this doesn't tell exact head-to-head game outcomes, it could still be useful to see which team will win the most games. Besides, in the past 10 years, the winner of the WCWS was in the top 5 of winning percentage 80% of the time, and they have always been in the top 10. That being said, Table 4 shows the top 10 teams as calculated by beginning of the season Stein approximations using our new formula for the Pythagorean Winning Percentage.

	Old Pythagorean	New Pythagorean	My Approx
1	Louisiana	Louisiana	Louisiana
2	Loyola Chicago	Loyola Chicago	UCLA
3	UCLA	UCLA	Virginia Tech
4	Virginia Tech	Virginia Tech	Loyola Chicago
5	Florida	Florida	Florida
6	Oklahoma	Oklahoma	Georgia
7	Georgia	Georgia	Oklahoma
8	Wisconsin	Wisconsin	Wisconsin
9	Coastal Caro.	Coastal Caro.	Coastal Caro.
10	Missouri	Missouri	Alabama

Table 3: Calculations for highest winning percentage calculated in three different ways.

6 Limitations

Stein-Estimators do have some limitations, however. With the above results it's easy to think the Stein estimate improves on every y_i as an estimate of its mean θ_i, but it doesn't always. The problem arises when estimating $k > 2$ means simultaneously, using the sum of squared errors from the true means for the k estimates as a measure of overall error. With a simple regression model, any variable may appear to be associated with others. Efron & Morris say the Stein-Estimate "anticipates regression to the mean" and improves the estimates by "borrowing strength from the ensemble." For a single average, there is an equally likely chance of our approximation being an overestimate or an underestimate. If we take a look at the first 10 games from the 2019 season, we would see that the average run differential for Mississippi Valley is -7.75. A run-rule in softball applies when a the winning team has scored 8 more runs than their opponents in the fifth inning or after. Without any outside knowledge that 7 runs is a very large run differential, just knowing that this was the largest (negatively speaking) run differential in our sample set, one could still judge that this was likely an overestimate of how Mississippi Valley's true run differential.

It's also important to think about a teams' strength of schedule prior to the beginning of their first 10 games. With 295 teams, someone is bound to play much worse pitching or hitting teams at some point. For example, after season was over, Seattle University softball was posted to have had the 7th most difficult schedule in the league prior to the cancellation of the 2020 season. This means that, overall, we'd expect

	Massey Rankings Week 5 - 2019:
1	Oklahoma
2	UCLA
3	Florida St
4	Alabama
5	Washington
6	Florida
7	James Madison
8	Wisconsin
9	Louisiana
10	Virginia Tech

Table 4: Massey Rankings during Week 5 poll of 2019.

them to lose more games and still rank higher than a team with say the 70th hardest schedule who played worse teams but came out with a higher record. When trying to implement something to take this into account within my model, I used a preseason poll of strength of schedule as voted on by coaches throughout the league. With the limitation I had in getting any most data on softball seasons prior to 2016, and any data prior to 2009, it made developing a system to track the true strength of teams nearly impossible. After using this preseason poll, along with my linear regression, it actually gave me a worse error than I originally had just using runs. After that, it was clear that runs were determined by many factors, including strength of schedule (if you're playing harder teams, you'll score less) and it was a more accurate depiction if I left other variables out. This, as well as the fact that

	Actual Top 10 2019:
1	Oklahoma
2	UCLA
3	Louisiana
4	Alabama
5	Washington
6	Florida St
7	James Madison
8	Virginia Tech
9	Northwestern
10	Michigan

Table 5: Calculations for highest winning percentage calculated in three different ways.

	Winning Percentage
1	UCLA
2	LSU
3	Oregon
4	Duke
5	Mississippi St.
6	Texas
7	Oklahoma St.
8	Georgia
9	Arizona St.
10	Arizona

Table 6: Calculations for highest winning percentage using beginning of season statistics for 2020.

References