Lucky or Good

By: Nick Vogl

Sometimes it's better to be lucky than to be good. That's how the old saying goes anyways. However, when evaluating players it is important to know how much of their performance is due to luck and how much is skill. In baseball, that is why we have expected statistics, to get a better idea of what should have happened instead of what actually happened. This allows us to make better predictions of the future.

To understand expected statistics, we must first understand the underlying statistic. Let's use batting average for example. Batting average is used by taking the number of hits and dividing by at bats. This is a way to measure how frequently a batter gets a hit successfully. However, some hits are bound to be 'lucky.' In other words, sometimes a player might hit the ball poorly, but it happens to just barely get by a defensive player for a successful hit. At the same time a different player can hit the ball the same way, but this time it may be caught for an out. Similarly, a player can be 'unlucky' and hit a ball very hard, but a defensive player makes a great play getting the batter out. Again, a different player can hit the ball the exact same way, but have different results due to factors solely outside of the batters control.

How do expected stats work and how do they differ from actual stats? Expected stats try to make up for this lucky factor that is inherently included in the underlying statistics. The way they do this is whenever a batter hits a ball, instead of only using the results of just that specific ball, the expected stat will use the result of every ball hit at the same speed and angle. This gives a better idea of what should happen as opposed to what did happen.

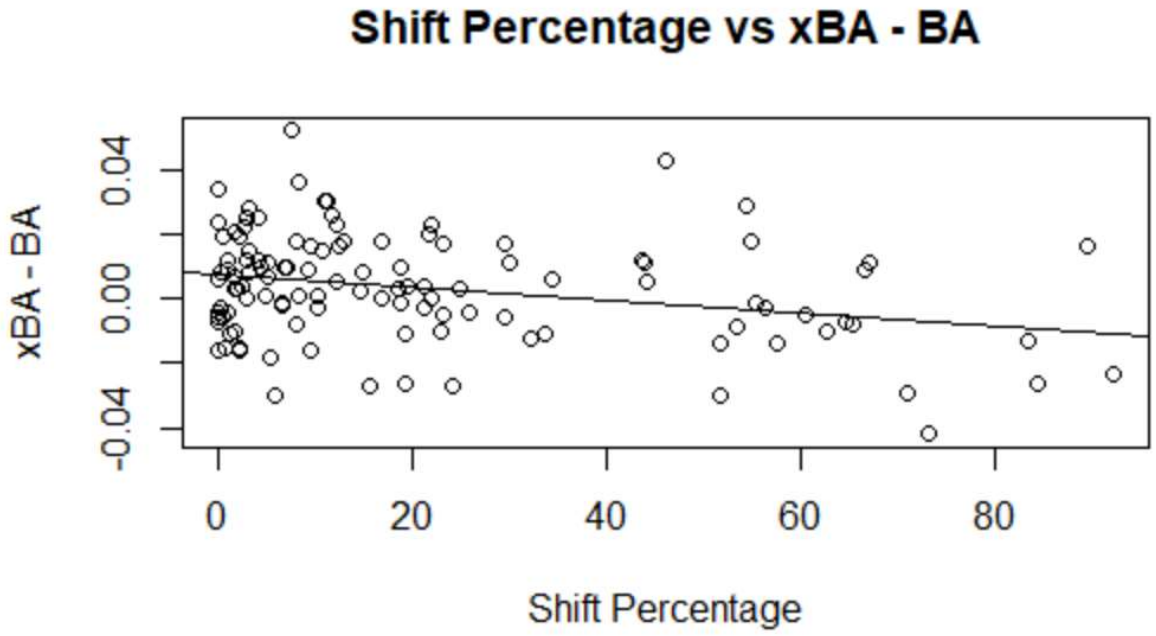
Let's continue seeing how this calculation is different using batter average. In both cases the denominator is at-bats, and goes up by one every time a batter goes to the plate. However, the numerator of the fraction is different. For batting average the numerator is number of hits. In other words, every time a batter goes up, the top part of the faction goes up by 1 if the batter gets a hit, and 0 if he does not get a hit. When calculating the expected batting average, every time a batter goes to bat, the numerator goes up by a decimal ranging from 0 to 1. The decimal represents the percent of the time that all balls hit at the same speed and angle successfully land as a hit. This way if two batters hit the ball the exact same way, but one has a defender make a remarkable play and gets out while the other gets a hit, their expected batting averages will be the exact same even though their actual batting averages differ.

Using expected statistics seems like it should be a more fair and accurate way to evaluate talent and predict future performance of players, but is it? I would argue largely yes, except for it is missing one key component. Expected statistics measure how hard a ball is hit and the vertical angle it is hit at to understand its trajectory. The problem with this is it only provides a 2-dimensionl projection, but baseball is played in a 3-dimensionl space. This in particular is becoming more and more important as teams continue to employ more and more defensive shifts. This means if a batter consistently hits the ball to the same side of the field, the opposing team can place their defenders there to give a greater chance of getting the batter out. In this case a batter can hit the ball well, but hit it right to where the defense is playing. On the other hand, a batter that routinely hits the ball to different parts of the field might get more hits because the defense can't predict where the ball will go before it is hit. This means the difference

between expected statistics and their respective underlying statistics is not only a 'luck' factor, but also a predictability factor.

How can we improve expected statistics in order to compensate for a hitter's predictability? My proposal is to create a new statistic that includes a component for how often a player is shifted on. The way of doing this is to first try to predict the residual, or difference between, a player's actual statistics and expected statistics. Once a predicted residual exists, adding this to the expected statistic will give us our new statistics, namely our expected shift statistics.

Let's continue looking at how this calculation can take place using batting average. To do the calculation, I will use 2018 data from the website Baseball Savant, in efforts of trying to predict the 2019 data. To start, let's look at a scatter plot that shows the relationship between the percent of time a player was shifted on and the residual between expected batting average (xBA) and batting average (BA).



In examining this graph, we can clearly see a negative relationship. This means as a player is shifted on more frequently, the difference between his expected batting average and actual batting average goes down. This makes intuitive sense as the more a player is shifted on the more often he may hit a ball well as predicted by the speed and angle of the ball, but also right at the defense.

Since this appears to be a linear relationship, the next step is to try to predict the residual by using a linear model. This is the output from R programing.

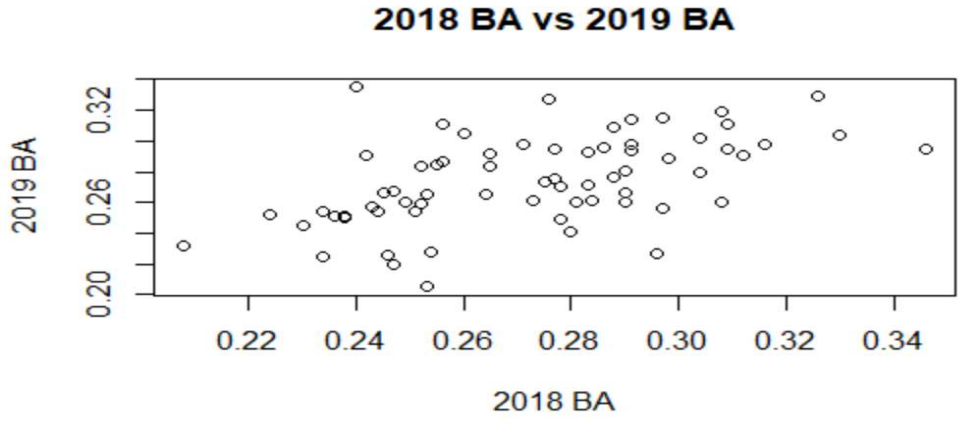
```
Call:
lm(formula = xbadiff ~ shift_percent, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.036283 -0.011220 -0.000075  0.011688  0.046053

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.411e-03  1.965e-03   3.772 0.000258 ***
shift_percent -1.979e-04  6.246e-05  -3.168 0.001964 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpreting the output, it means our model is:

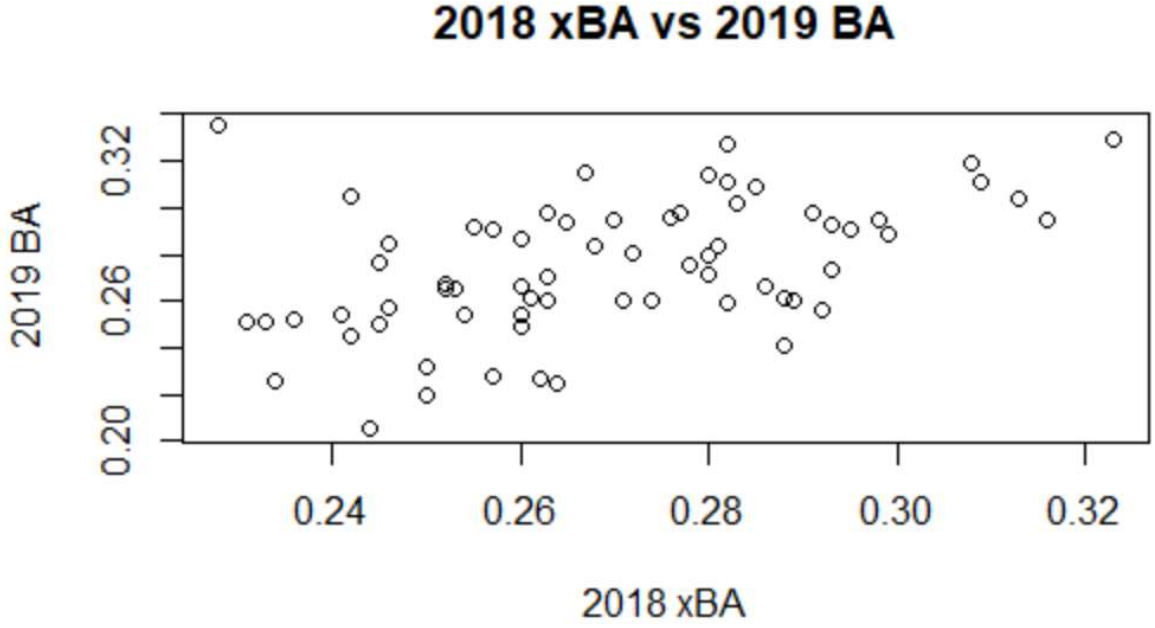
$$\text{Prediction} = 0.007411 - 0.0001979 * \text{Shift \%}$$

The column Pr(>|t|) represents the p-values. Typically a p-value less than .01 is significant, which is true for both the intercept and slope in this models. That means this model is good, and no variables need removed. If we add the prediction of this model to a players expected batting average, we will get what I will call expected shift batting average, or xsBA.

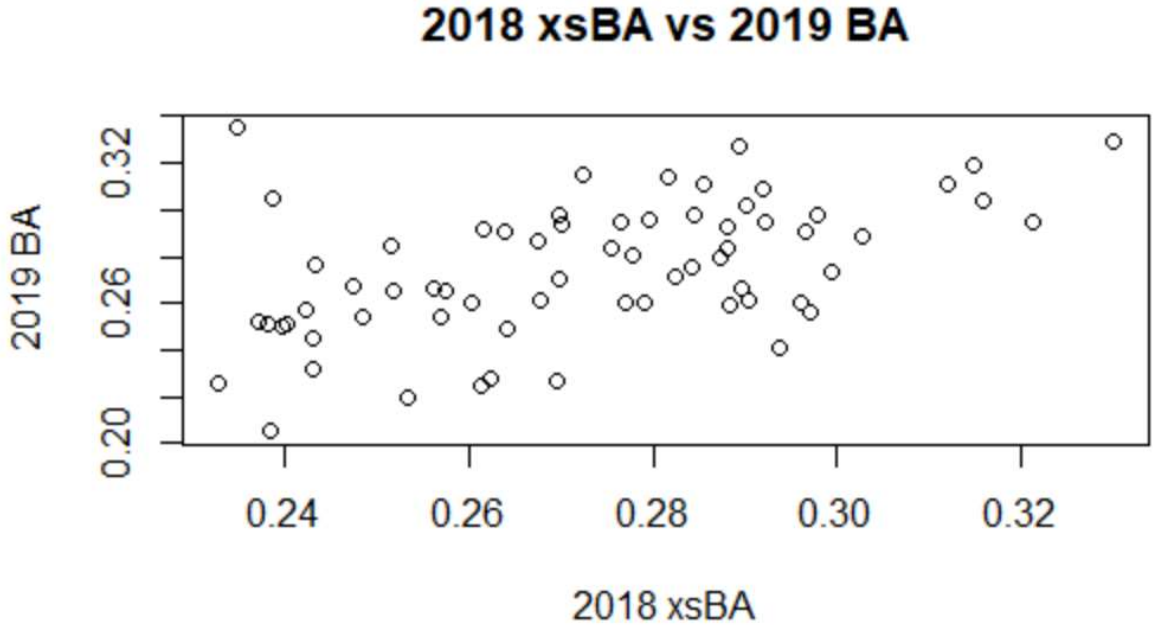
To summarize, as of now we have three statistics, BA, xBA, and xsBA. BA is a measure of how frequently a player gets hits. xBA tries to predict a players future BA. xsBA tries to improve the prediction from xBA. By verifying that the model used to produce xsBA is significant, we know the xsBA reduces the residual of BA and xBA. In other words, we know xsBA is closer to a player's current batting average relative to xBA's distance to batting average. However, that isn't really what we are interested in. After all if we really wanted to try to predict a player's current batting average we would just use their batting average, since we already know it. The purpose of xBA is to try to predict a player's future batting average. So now let's see how well a players BA, xBA, and xsBA from 2018 predict that same players BA from 2019.



The first plot is a graph of 2018 batting average vs 2019 batting average. This graph give a mean square error, or MSE, of 0.000791. MSE is a measure of how well a plot can be fit with a straight line. In particular, the smaller the MSE is the closer the dots are to a straight line, and the better the predictions are.



This second graph is for 2018 xBA vs 2019 BA. The MSE of this graph is 0.00723. Seeing how the MSE shrunk from the first graph to the second shows that in 2018, a player's xBA showed a better prediction from that players 2019 BA than their 2018 BA. This is a good thing seeing how that is the main purpose of xBA is to try to improve the prediction of a players future BA.



Our final plot is for 2018 xsBA vs 2019 BA. The MSE of this plot is 0.000681, which is the lowest and best yet. This suggest that our new statistic, xsBA, is actually the best of the three predictors.

What does all this mean? Expected statistics are great tools and can certainly be used to help make predictions on a player's future performance. However, just like any statistic, expected statistics certainly have their flaws that need to be understood. In particular, expected statistics do a poor job accounting for players that consistently hit the ball in the same direction and get 'beaten' by the shift. To compensate for this flaw, my recommendation is to create new statistics including a shift adjustment similar to xsBA. Even if someone decides to not use these proposed statistics and instead uses expected statistics to try evaluate players, I would highly recommend looking at a hitter's spray chart and shift percentage before drawing any conclusions. After all, the name of the game is to hit it where they ain't.