Stochastic Processes Project: Markov Insights in Baseball At-Bats

Charles Rak

April 2023

1 Abstract

The initial motivation was to gain insight about how baseball hitters vary in value throughout the different counts of an at-bat. Major League Baseball (MLB) pitch-by-pitch data collected over the course of four seasons was used to construct a discrete-time Markov chain to model the problem. Hitters' performances were analyzed based on their ability both to extend the duration of an at-bat and generate value at different counts. Results show that hitters are indeed far more productive when ahead in a count compared to behind. Although the ability to extend an at-bat by not swinging at poor pitches has recently become a coveted trait for professionals, results show little to no correlation between the length of an at-bat and its value.

2 Executive Summary

The goal was to develop a more appropriate way of modeling a hitters' weighted on base average (wOBA) throughout different counts in an at bat. The current standard in baseball when it comes to modeling wOBA by count is to consider the results of at-bat ending plays that take place at that count, and use these to model wOBA. Counts in this model will be treated how they are in standard baseball context, where the number of balls precedes the number of strikes. For example, if a hitter saw 60 pitches at 0-2 counts, and the at bat ended on 30 pitches, the calculations would then be done on the results of those 30 pitches. However, as this example illustrates, the values derived from this methodology are an inaccurate representation of what actually happens in this scenario. This is because in the other 30 pitches, the at-bat continued to an outcome that could have a different value not reflected in the calculation. To solve this, we aimed to incorporate the potential for value later on in the at bat into what is currently being measured. This would allow us to see if hitters who reach 0-2 counts are able to fight off pitches and produce later in the at bat. Being able to extend at bats and provide value later on is an valuable trait that prospects look for in players. The goal of our model is to determine exactly how valuable a trait like this actually is.

To build our model, we used the observed outcomes of every pitch from the beginning of the 2015 season to the end of the 2018 season to develop an idea of how much value exists later on in these at-bats, both on a league level and a individual player level. We then observed how often players added value later on in these at bats compared to how many more pitches they would, on average, extend the at-bat for at each count, using different metrics to do so.

Results indicate that there is almost no correlation between a players' ability to extend the length of an at-bat and that players' ability to generate value later on in an at-bat. However, results do show that league average wOBA is far higher at counts in which a player has more balls than strikes, as can be seen in Appendix 1, showing that the ability to take balls when behind in counts is indeed a valuable trait.

This model also allows for examination on where hitters may potentially be weak in their at bats. For example, some hitters turn to an extremely aggressive approach as soon as they are ahead in the count, and in doing this they sometimes fail to produce any value. The ability to identify these weaknesses provides a blueprint for a pitcher's approach, and similarly, a blueprint for how a hitter can improve. If a player has a certain point in an at-bat where they are exceedingly dangerous, teams will do as much as they can to put the player in that situation. Conversely, opposing teams will look for weaknesses in a players game and ways to take advantage of them.

3 Technical Report

3.1 Introduction

For context, the sport of baseball consists of a pitcher throwing baseballs to a hitter who is trying to hit the baseballs. The goal of the pitcher is to throw the baseballs past the hitter and in the strike zone, or get the hitter to hit the baseball softly and not very far, preventing the opposing teams from scoring runs. The goal of the hitter is to hit the baseball hard and far, get on base, and score runs.

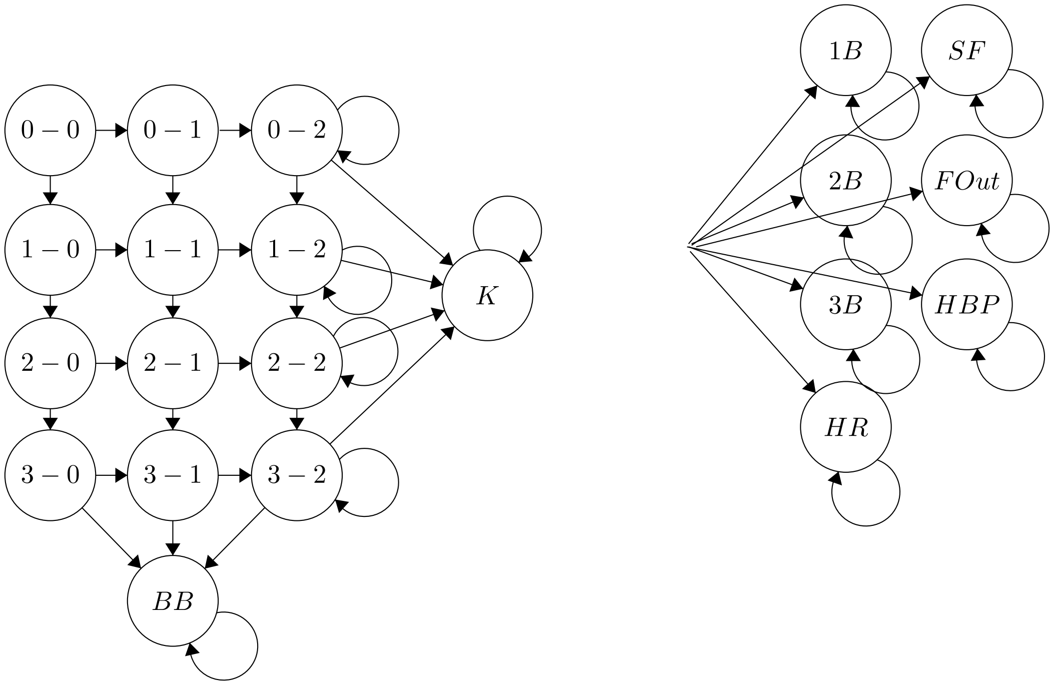
The rules work as follows: a strike occurs each time a pitcher throws a baseball that goes into the strike zone, or is swung at and missed or deflected out of play by the batter (foul). In all other scenarios, the pitch is said to be a ball. Once a hitter accumulates three strikes in an at-bat, they are said to be out. If a hitter accumulates four balls before three strikes, they get to take a base with what is called a walk. A **count** is used to represent the current number of balls and strikes in an at-bat, in the format [balls]-[strikes]. There can be [0,1,2,3] balls and [0,1,2] strikes before an at-bat is over, meaning there are 12 possible counts that can occur. After each pitch, the count either increases by a ball, a strike, or an at-bat ending event occurs. The only exception to this is when a ball is hit foul, in which case if the hitter already had two strikes, the count will remain unchanged. A common measurement for a hitters' value is weighted on-base average, or wOBA, which takes into account the value of each at-bat ending outcome and how frequently it occurs. Outcomes that are more likely to result in runs, like a triple or a home run, have higher values than outcomes that are less likely to result in runs, like a single or a double. The formula for wOBA is available in Appendix 2.

3.2 Methodology

The goal of the project to estimate a hitter's value throughout the count, with model weights consistent with the formula for wOBA. From a technical point of view, there was an understanding that the best way to model the process as a discrete-time Markov chain, as pitches in baseball happen in discrete time and the possible outcomes following each pitch can be modeled with a distribution. As the goal was to develop a Markov chain that transitioned after each pitch, it was essential to find pitch-level data sophisticated enough that it tracked all the inputs into the function we wanted to model, wOBA. From here, we used the observed outcomes of every pitch spanning four MLB seasons to generate a transition matrices for both the entire league and individual players. Following this, we were able to use absorption time and expected value formulas to calculate the expected time until the at-bat concluded, and the expected wOBA at each count. With these at our disposal, we can look to see if the ability to extend an at-bat (absorption time) correlates with the ability to generate value from an at-bat (wOBA), and look further into certain subsets of players to see if any trends arise.

3.3 Model Formulation

The states in the Markov chain would include the 12 possible counts, which is the set of numbers $(x, y) \in B \times S$ where $B = \{0, 1, 2, 3\}$ is the set of possible ball amounts and $S = \{0, 1, 2\}$ is the set of possible strike amounts, along with the nine possible at-bat ending outcomes: (strikeout, walk, field out, single, double, triple, home run, sacrifice fly, hit by pitch). We will treat these nine at-bat ending outcomes as absorbing states, as once the state is entered it is never left. At each count, there is a possibility of transitioning to seven of these absorbing states. The other two absorbing states, strikeout (K) and walk (BB) can only be reached from states with two strikes or three balls, respectively. With this defined relationship, we can draw our Markov chain, as can be seen in the diagram below. Note that the drawing was reduced for sake of cleanliness, and the 7 states on the right of the diagram can be directly accessed from any of the 12 count states in the three by four grid.



To generate the transition probabilities, we made use of data scraped from MLB.com publicly available on Kaggle [Appendix 3]. After downloading, the code in Appendix 4 can be used to load and restructure the data to retain only the necessary information about each pitch. Note that, as mentioned in the code, there are some redundancies within the data that must be removed. For example, a filter is applied to league data to get rid of plays that are classified as “In play, no out(s)”, denoted by code ‘D’, that are also labeled as an out of some form. As this is directly contradictory, it must be removed. Once the data has been cleaned, we can use it to generate our Markov chain. The function to create this is very extensive and would not fit in the report. It can be accessed on GitHub in Appendix 5.

3.4 Results

After we have generated our Markov chain can use R to calculate the expected time to absorption vector which will provide us insight on player’s ability to extend at bats. The function referenced in Appendix 6 can be used to do so. We then solve the following system of equations to get a sense of how much value players are able to add at each point in the count. Let P denote the transition matrix for the Markov chain, and $P_{i,j}$ denote the probability of transitioning from state i to state j. Using the formula for wOBA [Appendix 2], the weights(values) of each event are as follows. Events have value zero if not noted.

```
Walk - 0.69
Hit By Pitch - 0.72
Single - 0.89
Double - 1.27
Triple - 1.62
Home Run - 2.10
```

For cleanliness, the value for exiting into one of the absorbing states from state i is represented as follows:

$E[i, abs] = 0.89P_{i,16} + 1.27P_{i,17} + 1.62P_{i,18} + 2.10P_{i,19} + 0.72P_{i,21}$

Using this, the expected value at each count state can be represented by the following system of equations:

$$\begin{aligned}
\mathrm{E}[(0,0)] &= P_{(0,0),(0,1)}\mathrm{E}[(0,1)] + P_{(0,0),(1,0)}\mathrm{E}[(1,0)] + E[(0,0), abs] \\
\mathrm{E}[(0,1)] &= P_{(0,1),(0,2)}\mathrm{E}[(0,2)] + P_{(0,1),(1,1)}\mathrm{E}[(1,1)] + E[(0,1), abs] \\
\mathrm{E}[(0,2)] &= P_{(0,2),(0,2)}\mathrm{E}[(0,2)] + P_{(0,2),(1,2)}\mathrm{E}[(1,2)] + E[(0,2), abs] \\
\mathrm{E}[(1,0)] &= P_{(1,0),(1,1)}\mathrm{E}[(1,1)] + P_{(1,0),(2,0)}\mathrm{E}[(2,0)] + E[(1,0), abs] \\
\mathrm{E}[(1,1)] &= P_{(1,1),(1,2)}\mathrm{E}[(1,2)] + P_{(1,1),(2,1)}\mathrm{E}[(2,1)] + E[(1,1), abs] \\
\mathrm{E}[(1,2)] &= P_{(1,2),(1,2)}\mathrm{E}[(1,2)] + P_{(1,2),(2,2)}\mathrm{E}[(3,1)] + E[(1,2), abs] \\
\mathrm{E}[(2,0)] &= P_{(2,0),(2,1)}\mathrm{E}[(2,1)] + P_{(2,0),(3,0)}\mathrm{E}[(3,0)] + E[(2,0), abs] \\
\mathrm{E}[(2,1)] &= P_{(2,1),(2,2)}\mathrm{E}[(2,2)] + P_{(2,1),(3,1)}\mathrm{E}[(3,1)] + E[(2,1), abs] \\
\mathrm{E}[(2,2)] &= P_{(2,2),(2,2)}\mathrm{E}[(2,2)] + P_{(2,2),(3,2)}\mathrm{E}[(3,2)] + E[(2,2), abs] \\
\mathrm{E}[(3,0)] &= P_{(3,0),(3,1)}\mathrm{E}[(3,1)] + 0.69P_{(3,0),BB} + E[(3,0), abs] \\
\mathrm{E}[(3,1)] &= P_{(3,1),(3,2)}\mathrm{E}[(3,2)] + 0.69P_{(3,1),BB} + E[(3,1), abs] \\
\mathrm{E}[(3,2)] &= P_{(3,2),(3,2)}\mathrm{E}[(3,2)] + 0.69P_{(3,2),BB} + E[(3,2), abs]
\end{aligned}$$

Solving this system of equations results the in the expected wOBA for each count of the at-bat. While the math may not at first appear identical, the observations encompass all $AB+BB-IBB+SF+HBP$, and in calculating the cumulative probability that one of the value-generating states is reached, we are essentially performing the calculation $\frac{Event*EventValue}{AB+BB-IBB+SF+HBP}$ for each event, and summing them.

The function in Appendix 7 can be used to solve this system of equations.

Functionality also exists to generate player-specific data, in order to compare players to league average and others. The functions in Appendix 8 can be used to efficiently do so, with example included.

These values can then be further analyzed to compare players to one another, or to league averages. For a sanity test, we can compare the wOBA values of Mike Trout, one of the best hitters in this four-year span, to league average. As can be seen in Appendix 1, Trout's expected wOBA from each count is far better than that of league average, as we would expect to see for one of the leagues best hitters. There are two important takeaways from this graph. The first is that an outstanding hitter has a higher expected wOBA at every count compared to league average. This provides a confirmation that our calculations have been done properly. However, more importantly, we can see that for the league average hitter, counts in which a hitter has more balls than strikes (e.g. 2-0, 3-0) have far greater expected wOBA than counts in which they do not (e.g 0-2, 1-2). This confirms the idea that being able to extend an at-bat by not swinging at balls is generally a good approach.

We can now use the model to answer questions surrounding player traits and how they affect performances. To determine if there is any correlation between expected time to absorption and expected wOBA, we can randomly sample 50 hitters and see if there are any patterns.

As can be seen from the code and output in Appendix 9, none of the R correlation coefficients suggest a strong correlation between the two variables, though they are mostly consistent, especially in counts with 2 strikes. From this, we can conclude that the ability to extend an at-bat as measured by time to absorption is loosely correlated to adding value from a hitter's perspective.

However, this conclusion is very sensitive to the idea that time to absorption is something a hitter can control. While this does measure how long a hitters at-bat is on average, hitters cannot always control the pitches they are thrown, and if they are only thrown strikes, they have no choice but to swing. As such, a more accurate measurement of a players' ability to extend an at-bat is by observing chase rate, or the percent of pitches outside the strike zone that the player swings at. When we try to observe correlation between wOBA and chase rate, R correlation coefficients are not strong, but they are similarly consistent. As seen in code and output in Appendix 10, there is a loosely negative correlation between chase rate and wOBA. This makes sense intuitively, as pitches outside of the zone are nearly impossible to hit, so swinging at them is almost guaranteed to lose value.

Plots of both Absorption Time and Outside Zone Swing Rate vs wOBA for 0-2 counts are available in Appendices 11 and 12, respectively.

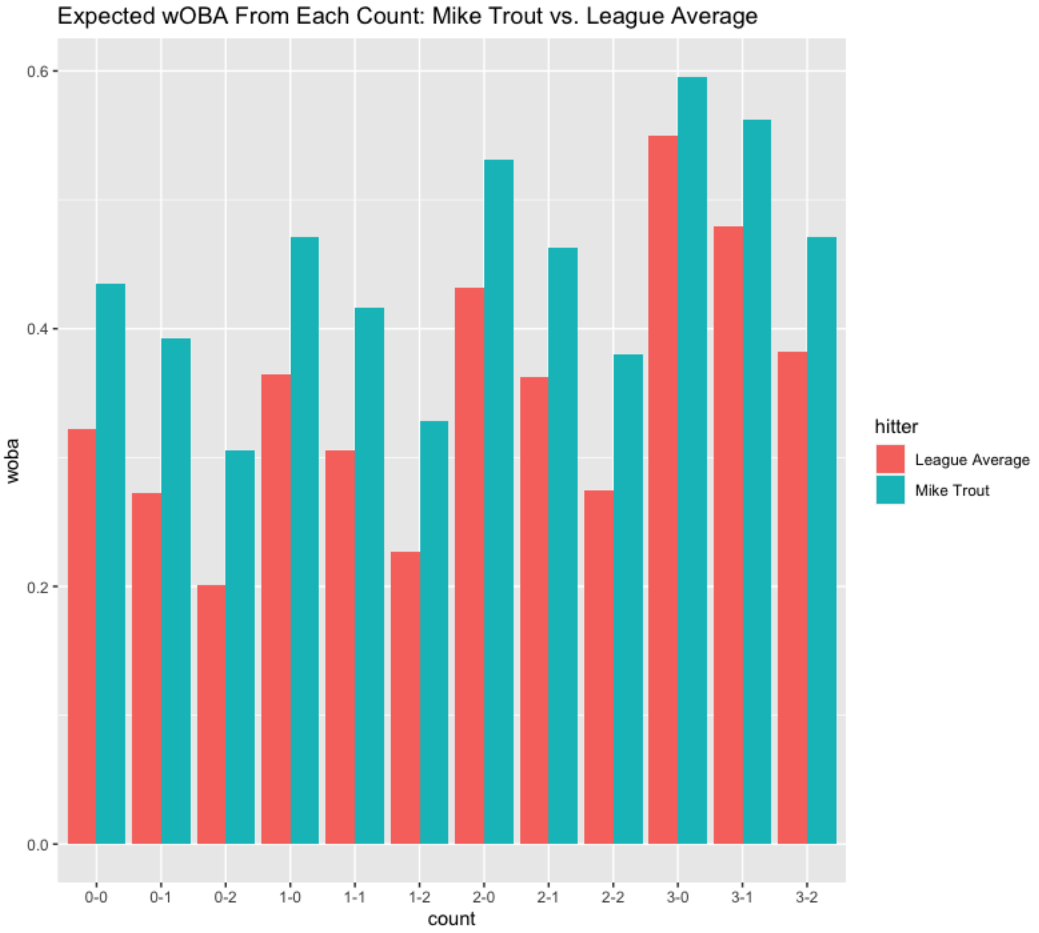
3.5 Conclusion

Our results show that there is loose correlation between a player's ability to extend an at-bat and their ability to generate value in an at-bat, as measured by wOBA. Even when separated into each count, R correlation coefficients did not exceed a magnitude of 0.5 when comparing expected absorption time to expected wOBA, and when comparing chase rate to expected wOBA. However, baseball is an extremely complicated sport, and there are a lot of factors that go into a hitters' success. While a low chase-rate or high expected absorption time may be a valuable trait, it certainly is not enough to generate value for a player on its own.

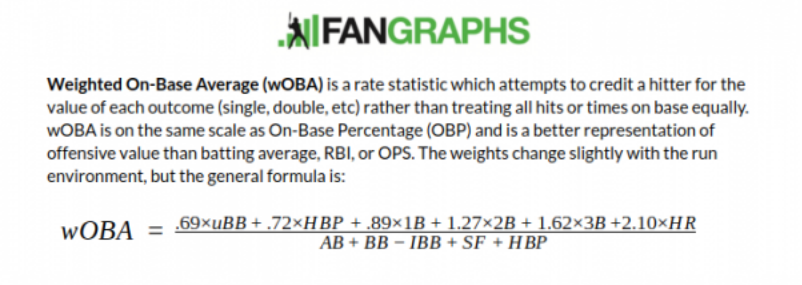
Potential expansions on the project can include the use of a Markov Chain to represent the outcomes of players at bats over the course of a season. In baseball, "cold streaks" are often chalked up to bad luck, however they can take a mental toll on players and potentially cause them to perform worse. Determining whether or not start state affects long-term distributions would be an insightful expansion possible with the data.

4 Appendix

1: Expected wOBA by Count



2: wOBA Formula



FANGRAPHS

Weighted On-Base Average (wOBA) is a rate statistic which attempts to credit a hitter for the value of each outcome (single, double, etc) rather than treating all hits or times on base equally. wOBA is on the same scale as On-Base Percentage (OBP) and is a better representation of offensive value than batting average, RBI, or OPS. The weights change slightly with the run environment, but the general formula is:

$$wOBA = \frac{.69\times uBB + .72\times HBP + .89\times 1B + 1.27\times 2B + 1.62\times 3B + 2.10\times HR}{AB + BB - IBB + SF + HBP}$$

3: MLB Pitch Level Datasets

4:

```
pitch_data <- read.csv(’pitches.csv’)
player_names <- read.csv(’player_names.csv’)
atbat_data <- read.csv(’atbats.csv’)
trimmed_pitch_data <- pitch_data %>% select(ab_id, b_count, s_count, outs,
                                            pitch_num, pitch_type, event_num,
                                            code, type)
trimmed_ab_data <- atbat_data %>% select(ab_id, batter_id, event, o)
#filter bad codes / errors in data
league_pitch_data <- left_join(trimmed_ab_data, trimmed_pitch_data, by = "ab_id") %>%
  filter(!((code == "D" &
              (event == "Grounded Into DP" |
                 event == "Groundout" |
                 event == "Forceout" |
                 event == "Sac Fly")) |
             code == "Z") )
```

5: Markov Chain Generation Function: Generates a Markov Chain transition matrix for a player given the player’s data.

6: Absorption Time Calculator: Calculates the expected absorption time from each state given the transition matrix.

7: wOBA Calculator: Utilizes bottom-up dynamic programming to solve the system of equations representing the expected wOBA from each count, given the transition matrix.

8: Get Player Values Function

Example:

```
trout_values <- get_player_values("Mike", "Trout")
> head(trout_values)
  count time_to_abs      woba     hitter
1   0-0    4.364128 0.4353267 Mike Trout
2   0-1    3.532367 0.3928437 Mike Trout
3   0-2    2.916253 0.3052465 Mike Trout
```

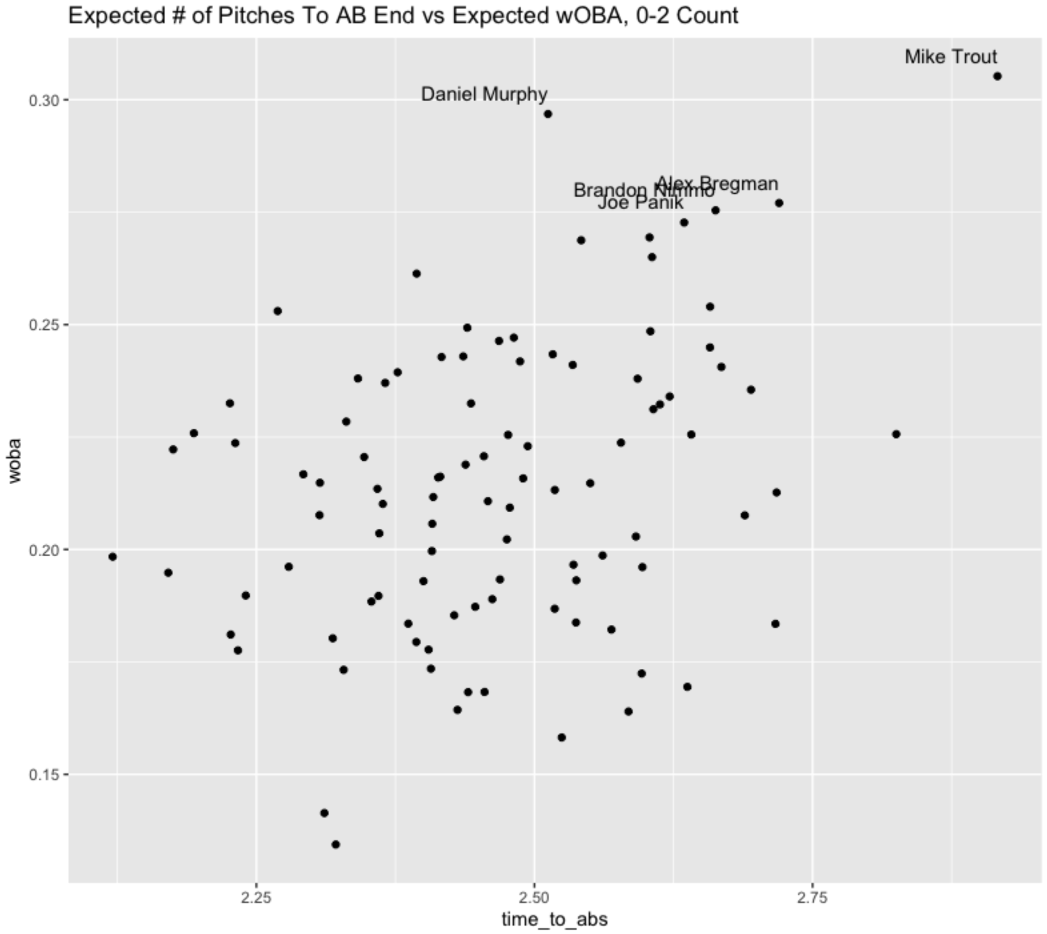
9:

```
player_sample <- woba_data$player_id %>% sample(100, replace=FALSE)
woba_abs_time <- data.frame()

for(player_id in player_sample){
  player_name <- get_player_name(player_id)
  first <- player_name[1]
  last <- player_name[2]
  vals <- get_player_values(first,last)
  woba_abs_time <- rbind(woba_abs_time, vals)
}
woba_abs_time %>%
  group_by(count) %>%
  summarize(r = cor(time_to_abs, woba))

OUTPUT
   count      r
   <chr>   <dbl>
 1 0-0    0.425
 2 0-1    0.237
 3 0-2    0.339
 4 1-0    0.386
 5 1-1    0.225
 6 1-2    0.342
 7 2-0    0.312
 8 2-1    0.271
 9 2-2    0.382
10 3-0   -0.387
11 3-1   -0.148
12 3-2    0.0362
```

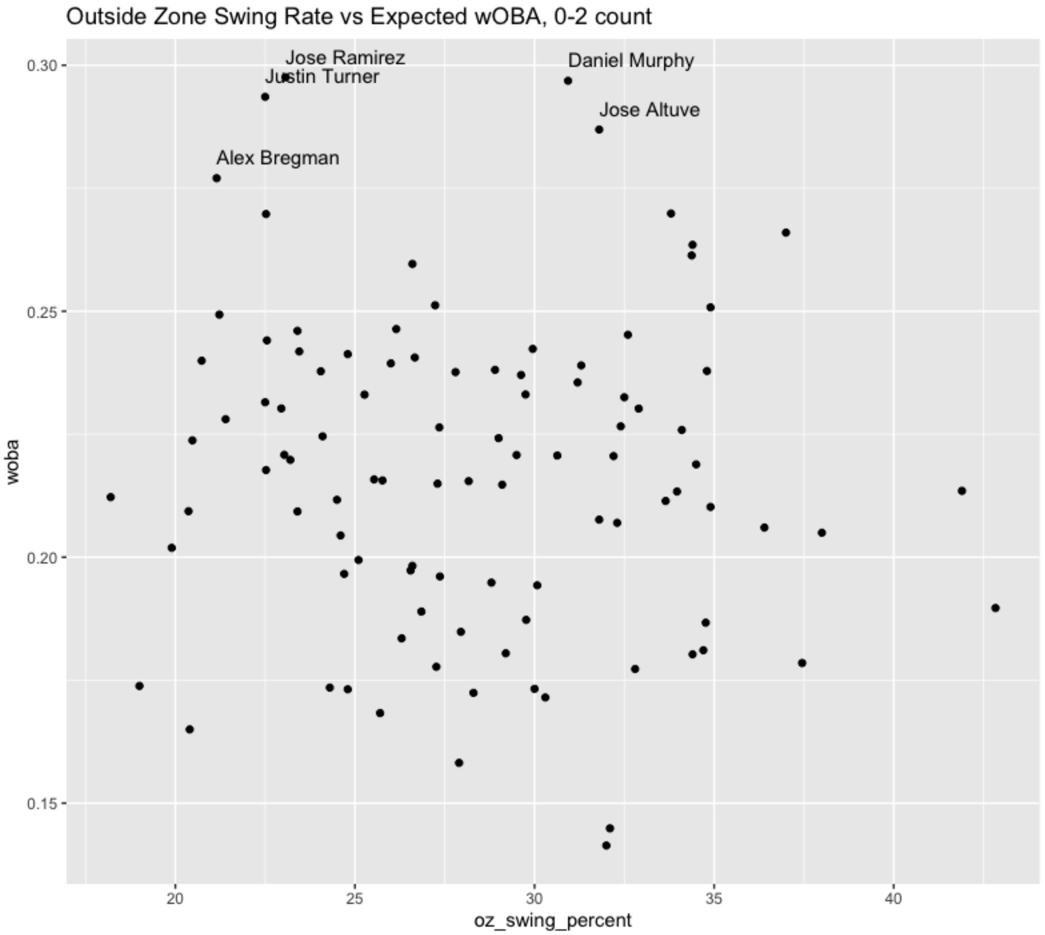
10:

```
badswing_sample <- badswing_data$player_id %>% sample(100, replace=FALSE)
woba_badswing <- data.frame()
for(pid in badswing_sample){
  player_name <- get_player_name(pid)
  first <- player_name[1]
  last <- player_name[2]
  vals <- get_player_values(first,last)
  player <- badswing_data %>%
    filter(player_id == pid)
  oz_swing <- player$oz_swing_percent
  vals <- cbind(vals, "oz_swing_percent" = rep(oz_swing, 12))
  woba_badswing <- rbind(woba_badswing, vals)
}
woba_badswing %>%
  group_by(count) %>%
  summarize(r = cor(oz_swing_percent, woba))
OUTPUT
   count       r
   <chr>   <dbl>
 1 0-0   -0.163
 2 0-1   -0.181
 3 0-2   -0.0858
 4 1-0   -0.191
 5 1-1   -0.199
 6 1-2   -0.160
 7 2-0   -0.223
 8 2-1   -0.249
 9 2-2   -0.227
10 3-0   -0.0589
11 3-1   -0.0537
12 3-2   -0.283
```

11: Expected wOBA as a Function of Absorption Time



12: Expected wOBA as a Function of Outside-Zone Swing Rate



Outside Zone Swing Rate vs Expected wOBA, 0-2 count