Adam Wells, Zach Dammel, Bradley Knox
Stats 50 Final Project
December 10, 2014

Value of *Course* Position

In class, we learned about the value of field position in football - about how the expected outcome varies depending on the yardage. For our final project, we decided to apply that same concept to golf, a favorite sport of two-thirds of our group. We decided to look at how the expected score on a golf hole varies not only for different yardages, but for different conditions: rough and fairway. We weren't sure what we expected to find; we only knew that there would, in fact, be interesting findings. In this abstract, we will elucidate those findings and explain the methodology we used to uncover them.

Methodology:

Due to the limited availability of golf statistics, we decided to work under a few assumptions. These are as follows: (1) the hole is a par 4, (2) the flag is centered on the "green", which is a circle around the flag with a radius of 20 yards, and (3) there are no bunkers or hazards.

The first step in calculating the expected shots to hole was to find the probabilities of landing in a certain yardage interval given a starting point on the course. We called this our "transition matrix". Since no data is readily available on this, we took the 200 players at each yardage on the fairway that got closest to the flag on their next shot. We narrowed down the sample to players with a certain number of shots, as to eliminate outliers, and took every third of these players. Then we found the mean and standard deviation of yards to hole. We then normalized this distribution and found the probability of landing in a certain range (x1,x2) by the formula $\varphi(x2)-\varphi(x1)$.

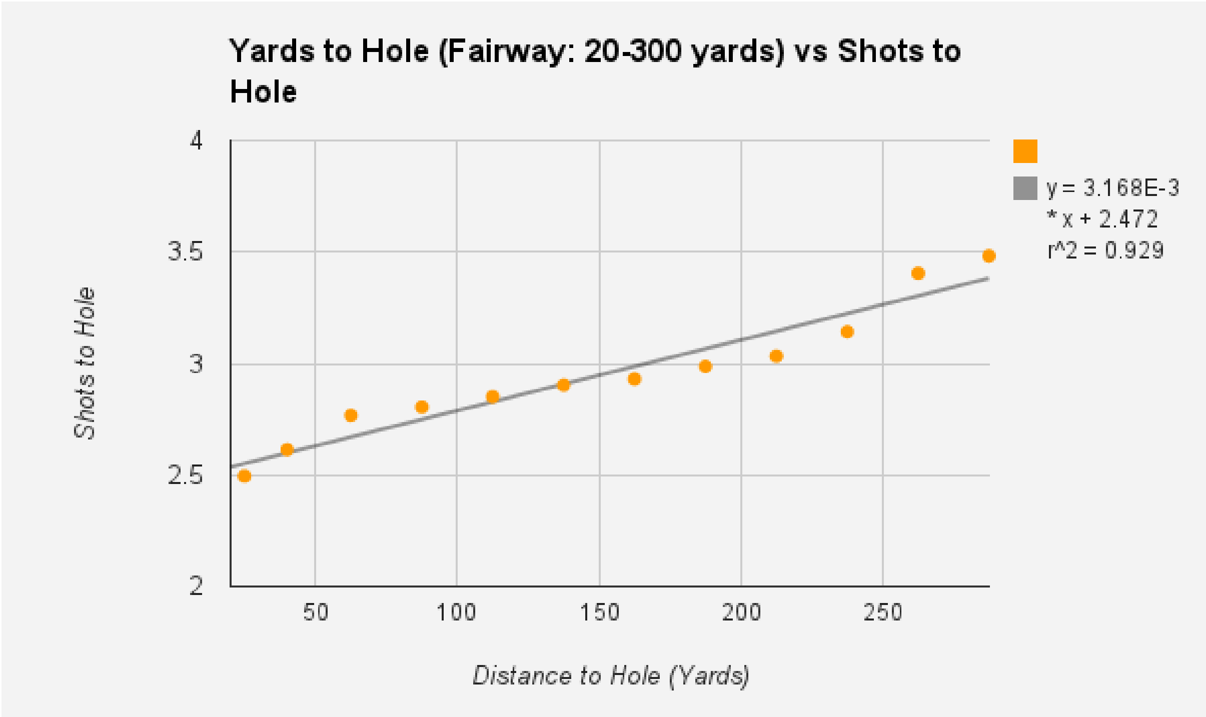
We then repeated this process for players who begin on the rough.

After compiling this data, we calculated the expected number of shots to hole given a distance to the flag. On the PGA database, we found the expected number of putts to hole given a range of starting distances on the green. Denote this column vector as “G”. Additionally, denote each row in our transition matrix $T_{(x1,x2)}$, where “x1” and “x2” represent yardage range. First, we noted that all shots when starting closer than 225 yards ended up within 20 yards from the hole (keep in mind we are talking about the top 200 PGA players at each yardage range). So for all ranges smaller than 225 yards we found our expected number of shots to hole by: mmult($T_{(x1,x2)}$,G). We then added +1 to account for the shot needed to get to the green. For yardage ranges 225-300, we followed a similar process. This time, we saw that all top 200 players starting in this range hit the ball to within 75 yards of the hole. We extended vector “G” to include yardage ranges up to 75 yards (which we found above), call this “Y”. Multiplying vectors “$T_{(x1,x2)}$” and “Y”, after accounting for +1 strokes when necessary, result in an expected value for each range.

This entire process was repeated for yardages starting in the rough. And then we had everything we needed to make some pretty cool discoveries.

Score vs. Yardage: Direct Relationship?

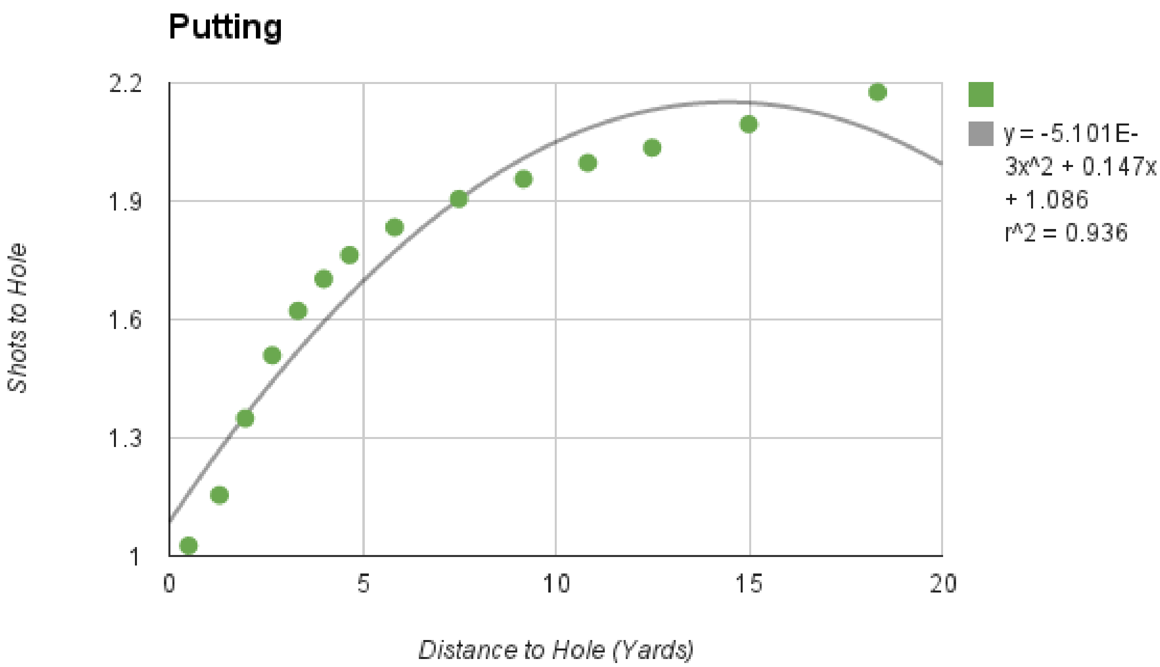
At a glance, you might think that for every additional yard gained you could expect a lower score. The goal, after all, is to ultimately get the ball so close to the hole that it falls in. But is it so straightforward?



As you can see, there is a peak between 50 and 75 yards and a dip between 175 and 250. Why is this the case? Let's analyze this from an intuitive perspective.

As a golfer, the peak makes sense. I hate being left with 50-75 yards. It's my worst yardage range and I actively avoid it. This is because the full, practiced golf swing can only take you so close. At some point, you have to make the gradual transition to chipping or putting. It is this in-between shot - lighter swings, half swings - that plagues golfers at these distances. These shots are less natural and more can go wrong. When you try to "ease up" on a swing, it is not unreasonable to "chunk it" and come up way short or, conversely, "blade it" and fly the green. For this reason, a golfer might want to either go for the green or lay up to about 100 yards, but advisably not in between.

The dip around 200 yards can be explained by another phenomenon. It has to do with putting. The farther you are from the green, the farther you expect to be from the hole after your next shot. And the difference in expected number of putts between two yardages decreases substantially as you move away from the hole. For example, the expected difference in number of putts from 1 to 3 yards is almost a half a stroke. The expected difference in number of putts from 10 to 12 yards, on the other hand, is less than a tenth of a stroke.



Thus, the relationship between yardage and expected score is *not* quite linear.

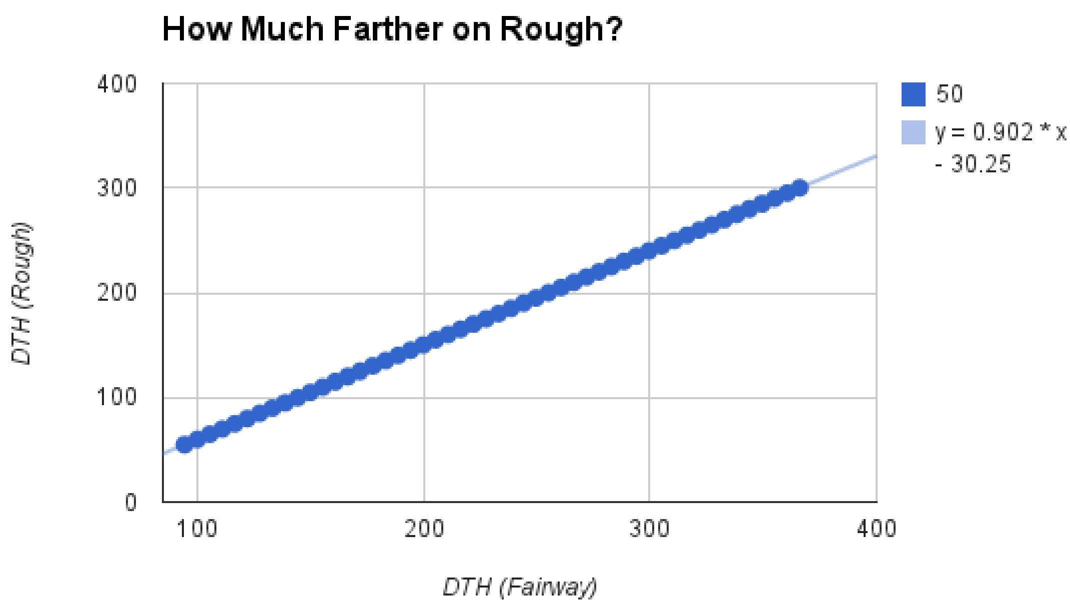
Long vs. Short: One of our main goals was to describe the advantage long hitters have over short hitters (or whether or not they actually even have an advantage!). For example, Zach Johnson is a top professional golfer who drives the ball an average of only 273.8 yards, but with a fairway accuracy percentage of 85%. John Daly, on the other hand, hits the ball a whopping 297 yards on average, but with only a 46% fairway accuracy percentage. Although it seems like Johnson might have an advantage because he hits almost twice as many fairways, when we plugged the numbers into a simulated 450 yard golf hole, Zach Johnson actually had a .02 stroke disadvantage due to his reduced driving distance.

We got this value by running linear regressions on expected number of shots to hole when starting on the fairway and rough. After getting these equations, we found the expected number of shots to hole Johnson and Daly would have if they landed on either the fairway or rough. Lastly, we weighted these expected values by each players accuracy statistics.

There are a few reasons why this value is not representative of the truth (as in reality Zach Johnson would probably have the advantage), which mostly has to do with our simplified model. When you miss the fairway, there are many possible states for the ball to being in besides the rough, including in the trees, a bunker, a hazard, or out of bounds. These other states are often much more penalizing than simply being in the rough. So, in reality, Zach Johnson has the upper hand off the tee. In other words, our model assumes the best worst-case scenario.

Fairway vs. Rough:

After running linear regressions on expected shots to hole when on the rough and fairway, we thought it would be interesting to find out how much closer the flag one would need to be on the rough to make a golfer indifferent between rough and fairway. To do this, we found a relationship between proximity to hole when on the fairway and rough when expected shots to hole were equal. Since we already had two equations for expected shots to hole for both fairway and rough, all we had to do was equate these equations to solve “distance to hole on the rough” in terms of “distance to hole on the fairway”. Our results are as follows (note: due to data constraints, our range on the fairway was ~85+ yards):



Value of Driving Distance and Accuracy:

Next, we decided to find an equation for expected shots to hole after a player's drive. Since we have expected shots to hole on the rough and fairway given a player's starting position on the course, we can combine those two equations and weight them on driving accuracy (the % of the time the drive ends on the fairway). Doing this we get:

$$E(S)=\alpha[(L-D)(.00372)+2.6)]+(1-\alpha)[(L-D)(.003042)+2.498)]$$

Where L is the length of the course, D is the driving distance, and α is the fairway percentage off the drive.

Where do we go from here: If we had more sophisticated data, we think an interesting application might be evaluating the decision of whether to "go for the green" or "lay up" on the second shot on a par 5. Our current analysis - which incorporates distance aspect of that consideration - gets us halfway there. The most important factor, however, is what surrounds the green. Are there bunkers, water hazards, or other dangers? This would be challenging because, since every hole is different - with varying degrees of danger - we would have to look at it on a case-by-case basis. It might be fun to ascribe some mathematics to the concept of risk versus reward.