**Relationship Between**

Kelly Murphy

University of Montana MSBA

Spring 2018



**Shot Location and PGA TOUR Player Performance**

**Around the Green Using ShotLink Data**

TABLE OF CONTENTS

EXECUTIVE SUMMARY ii

INTRODUCTION 1

OVERVIEW OF TOPIC 3

RESEARCH QUESTION & HYPOTHESES 7

DATA 8

ANALYSIS 13

RESULTS 17

LIMITATIONS 24

Recommendation 25

CONCLUSION 27

WORKS CITED 28

APPENDICES 29

# EXECUTIVE SUMMARY

In this report, I analyzed the effects that different “lies” (fairway, primary rough, green side bunker) have on player performance around the green on the PGA TOUR. Specifically, I was curious about green side bunkers (sand traps), and whether or not they negatively impact players the way they are designed to compared to shots from the fairway and primary rough. Bunkers are meant to be a hazard that penalizes players who hit into them. Courses are strategically designed with this in mind. If bunkers are not impacting players in the way they are intended to, then players are able to take advantage of course designs, and actually play holes differently than the course architect intended. For example, bunkers are often used to create risk reward scenarios. A player can be aggressive with his shot, but he risks hitting the ball into a strategically placed bunker. If bunkers are not properly punishing players, then there is no risk involved with these shots.

Using ShotLink data, I ran several different analyses to determine to what extent bunkers are significantly impacting players, and how much bunkers are impacting players. These analyses included ANCOVA, ANOVA, and Multiple Linear Regression. These analyses controlled for many environmental factors to ensure that the shots included in the data were comparable, and also explored how distance affects the impact of each different shot location.

The results of the analyses suggest that green side bunkers are not having the intended effect on PGA TOUR players. Players are not penalized proportionately in the bunker, as they are from the primary rough. In fact, from certain distances, players are better off in the bunkers than they are in the primary rough.

Based on the results of these analyses, I recommend that the PGA TOUR changes their rules, with regards to bunker maintenance, for the 2019 season. I recommend that, for the 2019 season, the PGA TOUR stops raking bunkers at Web.com Tour events. By applying this rule change to the Web.com Tour, it allows the PGA TOUR to have a low-pressure testing environment, where they can determine whether or not the rule change would be viable on the PGA TOUR.

# INTRODUCTION

Since its inception in 2003, ShotLink has changed the way that golf is played on the PGA TOUR. ShotLink is a dataset comprised of every shot taken on the PGA TOUR since 2003. ShotLink uses lasers to map out each course, and measures each shot made by each player during every tournament. Traditionally, ShotLink data have been used by players and coaches to improve performance, and by television networks to enhance their tournament broadcasts. The PGA TOUR utilizes the data to enhance the fan experience, but they do not fully utilize the data when it comes to enhancing their own product, the golf game itself. The PGA TOUR has access to one of the most comprehensive datasets in professional sports, yet they still rely heavily on. As a result, the rules of the PGA TOUR are often outdated, and unable to keep up with how the sport is constantly changing at the professional level.

Recently, the PGA TOUR has begun to utilize their data to adjust the regulations of the game, but all of the research done focuses on distance. Specifically, the research focuses on how far players are hitting the ball off the tee (the first shot of each hole). The concern is that players today are able to hit the ball farther than ever before, and it’s giving them an unfair advantage over the courses that they are playing on.

While distance is an important part of the game today, it is definitely not the only area of the game where players are gaining an unfair advantage over the course. Focusing specifically on tee shots tells only a very small part of the story, because that only accounts for 18 shots per round.

This paper uses ShotLink data to help paint a more complete picture of where, specifically, PGA TOUR players are taking advantage of courses. The purpose of this research is to determine the effect that different locations have on shots around the green. These different locations include the fairway, primary rough, and green side bunker.

The fairway is made up of grass that is cut short. This should be the preferred location, because the short grass allows for players to make clean contact with the ball on their shots (**Figure 1**). The primary rough is longer grass than the fairway (usually two to four inches). Golf balls usually some to rest lower within the rough, as opposed to sitting on top of the grass (**Figure 2**). This makes it much more difficult for players to make clean contact, because at impact there is going to be grass between the ball and club. Shots from the bunker are hit off of sand (**Figure 3**). This is a hazard that is meant to be the most penalizing of the three locations, because of the way that the golf club interacts with the sand. As the club makes contact with the sand, the sand slows down the club dramatically.

**Figure 1: Ball Lying in the Fairway**



**Figure 2: Ball Lying in the Primary Rough**



**Figure 3: Ball Lying in the Bunker**



Specifically, the research compares shots around the green from the fairway, primary rough, and green side bunkers in order to determine whether or not each respective location has the intended effect on PGA TOUR players.

The remainder of this report will describe the analyses conducted to determine the effect of lies on shots from around the green. This includes an overview of the topic, the research question and hypotheses, a description of the data, and the analysis itself, followed by the limitations of the research and my recommendation.

# OVERVIEW OF TOPIC

Since its implementation in 2003, ShotLink has been used in many different ways, in several different domains.

Mark Broadie who is considered the “Godfather of Golf Analytics” used ShotLink data to actually improve golf analytics in his paper titled “Assessing Golfer Performance on the PGA TOUR”. In his paper, Mark Broadie used ShotLink data to introduce a new statistic in golf called “strokes gained”. The strokes gained statistic does just what the name implies. It categorizes each shot by type (i.e. drive, approach, around the green, and putts,), and then compares each player to the average, in order to determine where players are gaining or losing strokes during a round of golf. Today, strokes gained is one of the most widely used metrics in golf analytics due to its reliability in identifying a player’s specific strengths and weaknesses during a given round.

ShotLink data is also an integral part of PGA TOUR tournament broadcasts each week. J.P. Newport’s Wall Street Journal Article, “Stats, Chaos in the TV Trailer”, outlines exactly how crucial is to each golf broadcast. ShotLink is such a crucial part to golf broadcasts, that the PGA TOUR has hired a full time ShotLink expert, who’s job it is to work with the production team each week. His job is to use the data to enhance the broadcasts and help tell a story to the viewer. It is very interesting. There is so much going on around the course all at the same time, and he has to identify the most compelling statistics that he can “sell” to the producer. Turnbull claims that only 5% to 10% of stats end up making it on-air. This article shows how ShotLink is used to effectively tell stories, and enhance broadcasts for the networks.

Players are also utilizing ShotLink data in order to improve their performance. Take, for example, the Golf Digest article “Info Seekers”. This article by Guy Yocom goes into detail about the different ways that PGA Tour player Zach Johnson utilizes ShotLink data. Johnson has hired statistician Peter Sanders to help coach him on how to approach tournaments. With the help of Sanders, and ShotLink data, Johnson has made a lot of meaningful changes to his game. For example, he has learned that for his game it is important that he stays in the fairway, even if that means that he is trading accuracy for distance. He is very good with his approach shots, so he can afford to be farther back as long as he is in the fairway. He also learned that on par fives he should not be laying up to a distance he is confident with. He is better off going for the green and just getting as close as possible on his second shot. Johnson was skeptical at first because he is one of the best players on Tour from 70-100 yards out. However, since changing his strategy, Johnson has been performing better on par fives. The stats have also affected the tournaments that Johnson plays in. ShotLink data showed that Johnson does not perform as well at courses with Poa annua greens. Johnson will avoid tournaments at these courses when possible while planning his schedule.

Other tour pros, such as Brandt Snedeker, have also discovered the power of ShotLink data. Josh Sens’s Golf Magazine article, “It All Adds Up” describes the positive impact ShotLink data has had on Snedeker’s performance. Entering his fourth British Open, Snedeker had never made the cut. Before his fourth British Open, Snedeker started working with an analyst. The analyst told Snedeker that he is one of the best putters on Tour, so he doesn’t need to be aggressive. Instead of trying to knock the ball close to the pin, Snedeker should only focus on getting the ball on the green. With this strategy, Snedeker finished third in the British Open that year. Later that same year Snedeker went on to win the Tour Championship, earning him an $11.4 million paycheck. This article shows how effective the use of data is for some players on the PGA Tour, and gives a sense of the monetary value that this information can have.

Finally, there is the paper by Hickman and Metz, which uses ShotLink data to show how working with peers will affect performance in a business setting. This paper is an example of using ShotLink data to help answer a question completely unrelated to golf. The authors felt that by using ShotLink data they would be able to eliminate confounding variables that could not be accounted for in previous studies that tested for peer effects on performance in the workplace. They did this by looking at putting performance in individual tournaments and tournaments that were played as a team, then comparing how much better or worse players performed when they were able to observe and learn from their teammate’s shot, rather than playing as an individual. The results showed that being able to observe and learn from a peer does have an effect on performance, both positive and negative. Therefore if one’s peer performs well, a person is more likely to perform well too; however, if one’s peer performs poorly, then a person is more likely to also perform poorly.

The purpose of these examples is to show how effectively ShotLink data are being used in several different ways. The PGA TOUR, however, is not fully utilizing these data for their own organization. This lack of utilization would be similar to a retail business allowing customers and wholesalers to utilize the business’s data, but not using it for themselves. The rest of the golf industry is becoming more data-driven every day, while the PGA TOUR is still driven largely by tradition.

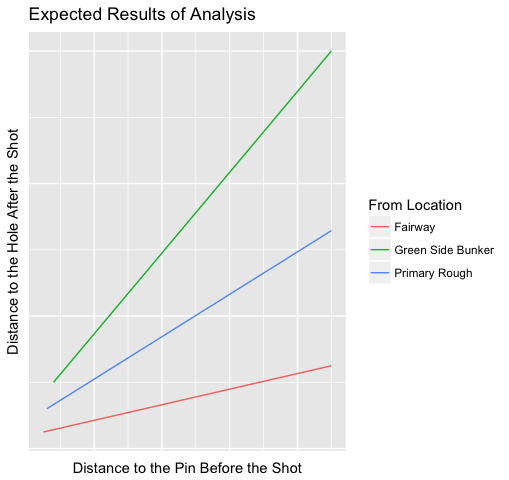
On March 5, 2018 the United States Golf Association (USGA) released their 2017 distance report, which relies heavily on ShotLink data. The report states than any further increase in how far players are hitting the ball at the professional level would be undesirable for the sport. As mentioned earlier, distance is a very small part of the game. My report looks beyond distances in order to determine whether or not additional changes, not focused on distance, should be considered. It remains to be seen whether or not the PGA TOUR will make any changes, but a shift to more data-driven decision-making offers the potential to enhance the product on the PGA TOUR.

# RESEARCH QUESTION & HYPOTHESES

## RESEARCH QUESTION

As mentioned earlier, this paper explores the effects of different lies on shots from around the green, and how distance impacts performance from each lie. Specifically, I analyze shots from the fairway, primary rough, and green side bunkers in order to determine how different shot locations affect a player’s performance for that shot. I am using distance to the hole following the shot as my dependent variable, in order to measure how a player performed on a shot. I am separating these data by the location that each shot was taken from in order to determine the effect that each different shot location has on performance. I also use distance to the pin before the shot as another independent variable, in order to determine the impact that distance has on performance from each different shot location. **Figure 4** visualizes what the results of the analysis should look like, if each shot location has the intended effect on a player’s performance for that shot. Courses are designed to penalize players most for hitting shots into bunkers, followed by hitting shots into the primary rough; hitting the ball from the fairway is the preferred lie. If each lie impacts players the way that it is intended, then I expect the best performance to come from shots from the fairway, followed by shots from the bunker, and shots from green side bunkers leading to the worst performance. I also expect that the impact of each lie will be compounded as the distance to the pin before the shot increases.

**Figure 4**

****

## HYPOTHESES

My hypotheses for the analyses are as follows:

**H1**:Shots from the fairway are associated with the best shot performance of the three shot locations analyzed.

**H2**: Shots from green side bunkers are associated with the worst shot performance of the three shot locations analyzed.

**H3**: When analyzing the interaction of shot location and distance, the best shot performance will be associated with shots from the fairway.

**H4**: When analyzing the interaction of shot location and distance, the worst shot performance will be associated with shots from green side bunkers.

# DATA

As mentioned above, for this analysis I am using ShotLink data from the PGA TOUR. ShotLink is a state-of-the-art platform used for collecting data in real time during a PGA TOUR tournament. Before the start of a tournament the entire course is mapped out using lasers. During a tournament, volunteers use similar laser devices to collect data on each individual shot taken by each player. I was granted access to the ShotLink data through the ShotLink Intelligence program. ShotLink Intelligence is a not-for-profit opportunity that allows graduate level professors and students, such as myself, to utilize the ShotLink data in their studies.

ShotLink data have several different layers of granularity. From the highest level of granularity to the lowest: there are data on each individual shot taken by each player, data on each round played by each player, and data on how each player performed at the event as a whole. In order to perform my analysis I focused on data at the individual shot level.

In my analysis I used data from only the 2017 season. The data go back to the 2003 season; however, this is a very large sample size (1,545,874 observations), and such a large sample size affects the results of the analysis by artificially inflating the significance of the results. By looking at only 2017 data, the sample size is reduced to 78,174 observations. Also, recent data are more relevant to my analysis than data from 15 years ago. I am curious about how the lies are affecting players on the PGA TOUR *today*.

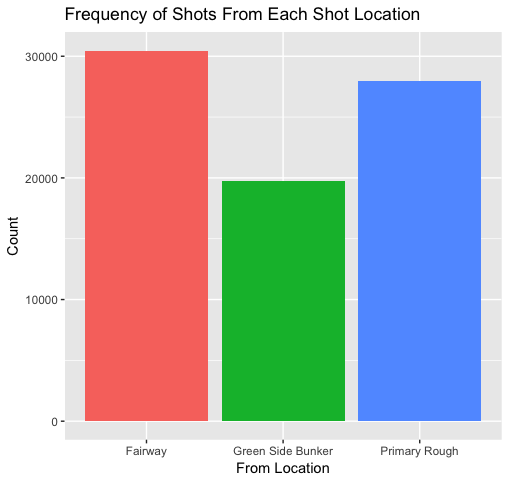
In order to ensure that all the data are consistent and comparable, I included many environmental factors, as covariates. These covariates include lie, elevation, and slope. I was able to control for these factors by ensuring that the data only included shots that were made under normal circumstances: a standard lie for each respective shot location, and a shot from level ground. Obviously, in golf no two shots are ever the same, so it is important to control for as many covariates as possible to ensure that the shots are comparable.

Variables:

The following descriptions represent the raw data used in the analysisbefore any data transformations have occurred. These data have been reduced from the full dataset of 16,133,849 observations, down to the 78,174 observations from the 2017 season, which are relevant for the analysis.

|  |  |  |
| --- | --- | --- |
| **Variable Type** | **Name** | **Description** |
| Dependent | Distance to the Hole After the Shot | Distance, in inches, from the position of the ball after it comes to rest at the end of a shot and the pin, as measured by a laser device recording the coordinates of the ball position when it comes to rest, and the coordinates of the cup on the green, calculating the distance between them. |
| Independent | Distance to the Pin Before the Shot | Distance, in inches, from the position of the ball before the shot was taken and the pin, as measured by a laser device recording the coordinates of the ball position before the shot, and the coordinates of the cup on the green, calculating the distance between them. |
| From Location – Fairway | Shots that were taken from the fairway; 30,466 (39%) shots were taken from the fairway. |
| From Location – Green Side Bunker | Shots that were taken from green side bunkers; 19,780 (25%) shots were taken from green side bunkers. |
| From Location – Primary Rough | Shots that were taken from the primary rough; 27,928 (35%) shots were taken from the primary rough. |
| Covariates | Year | Denotes the four-digit year of the event. The data have been stripped down to only include shots from only the year 2017. |
| Strokes Gained Category – Around the Green | Describes the statistical category that the stroke is assigned to. This analysis is only concerned only with shots from around the green, so I filter the data based on shots that fall into the ‘Around the Green’ category. |
| Elevation | Elevation of the player’s feet with respect to the ball. For example, a player’s feet can be above the ball, below the ball, or with the ball. For this analysis the data are limited to shots on an elevation with the ball. |
| Slope | The slope of the ground that the shot was taken from. This describes the slope on the target line of the shot. For example, the slope for a shot can be uphill, downhill, or level. The dataset used for the analysis only includes shots on a level slope. |
| Lie | Description of the lie of the ground that the ball is sitting on. For example the ball can have a good lie, buried lie, N/A, or be in a divot from a previous shot. In this analysis, the data are limited to the standard lies for each respective shot location. |
|  | In the Hole Flag | Dichotomous variable that indicates whether or not the shot finished in the hole. This analysis only includes shots that did not go in the hole. |

**Figure 5: Frequency of Shots From Each Shot Location**



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 1: Descriptive Statistics – Distance to the Pin** | | | | | | |
|  | **n** | **Mean** | **Median** | **Std. Dev.** | **Min.** | **Max.** |
| **Before Shot** | 78,174 | 705.64 | 623 | 363.36 | 122 | 2756 |
| **After Shot** | 78,174 | 124.84 | 82 | 136.94 | 1 | 2143 |

Because most of the variables used for the analysis are categorical, a traditional correlation matrix that runs between each of the variables is not possible. I can, however, determine the correlation between distance to the hole before the shot and distance to the hole after the shot. These two variables are continuous, and therefore, a correlation between the two variables can be determined. There is a weak positive relationship (0.3056) between distance to the hole before the shot and distance to the hole after the shot. This means that as distance to the hole before the shot increases, distance to the hole after the shot also increases overall, and vice versa.

# ANALYSIS

The goal of this analysis is to determine the effect that different locations around the green have on a player’s performance for that shot, and determine if the different lies have any significant impact on PGA TOUR players’ performance. Specifically, I analyzed shots around the green from the fairway, primary rough, and bunker to determine how each different lie affects how close to the hole players are able to get the ball for that shot.

In order to assess difference in shot performances between the different lies, I used an ANCOVA model. ANCOVA is used for a continuous dependent variable, a categorical independent variable, and a control variable.

In order to determine the effects that the shot locations have on shot performance, I used distance to the hole in inches following the shot as my dependent variable. This is the best continuous variable for the situation because it is the best measure of performance for a single shot. Other measures, such as a score relative to par or strokes made following the shot, rely on additional shots. Too many confounding variables are introduced in this situation, because each following shot has its own variables that affect it.

My categorical independent variable for the model was the location from which the shot was taken from (fairway, primary rough, or bunker). Finally, the variable that I controlled for was the distance away from the pin that the shot was taken from. So in plain English, I assessed whether different shot locations have a significant effect on the ball’s final proximity to the hole, controlling for distance.

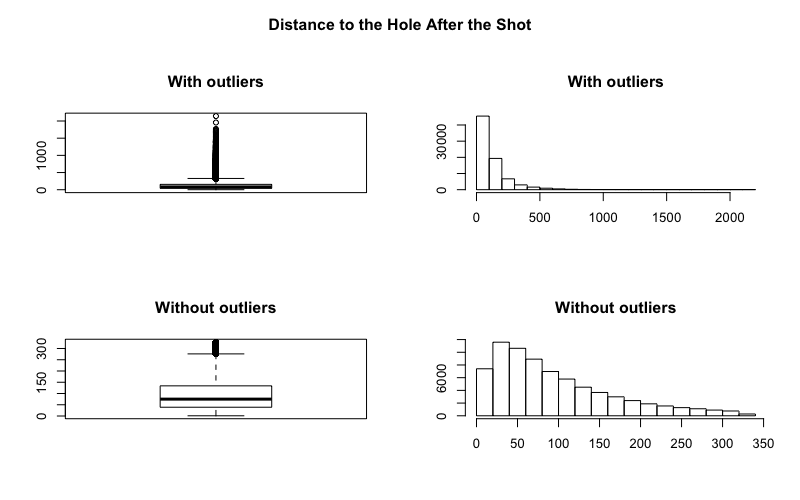
I also ran a multiple linear regression because running a regression gives information that an ANCOVA does not supply. Where an ANCOVA tells whether or not there is a significant difference between the fairway, green side bunker, and primary rough on performance, a multiple linear regression gives details about how strong the effect on the dependent variable is for each independent variable, and in which direction that effect is. A regression also tells how much variance in performance is explained in the model by each shot location. More specifically, the R-squared value tells how much variance in proximity to the hole is explained by the shot location and distance.

Before beginning my analysis, I cleaned the data in order to meet assumptions needed to run the models. One assumption that needs to be met is normality, or that all of the data are normally distributed. The first step I took to make the data normally distributed is removing all shots that went in the hole, because the amount of zeros in the dataset has an effect on the analysis that I’m trying to run. As an exploratory analysis, I ran proportion tests of shots in the hole between the different shot locations. **Table 2** shows the proportions of shots in the hole for each shot location. The difference between each shot location is significant (p < 0.01). Isolating these shots specifically could be an interesting future analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2: Proportion of Shots in the Hole From Each Shot Location** | | | |
| **Shot Location** | **Shots in the Hole** | **Total Shots** | **Proportion** |
| Fairway | 811 | 31,277 | 2.59% |
| Primary Rough | 444 | 28,372 | 1.56% |
| Green Side Bunker | 202 | 19,982 | 1.01% |

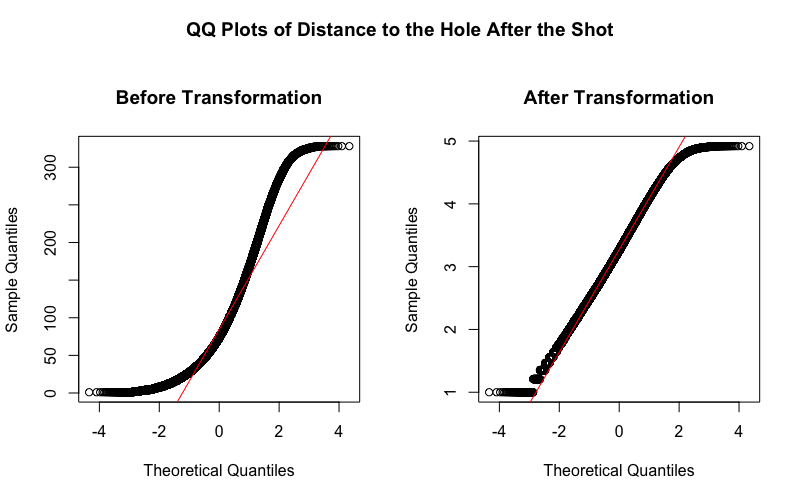
The next step was removing all outliers from the distance to the pin before the shot variable. Outliers heavily skewed the data. For my analysis I’m trying to keep shots as consistent as possible, and because each shot is different, it is very possible that outliers were caused by a confounding variable that is not included in my model. Even after removing outliers from the data, there is still enough data to conduct an effective analysis. The effect that the outliers had on the distribution of the data can be seen in **Figure 6**.

**Figure 6: Impact of Removing Outliers**

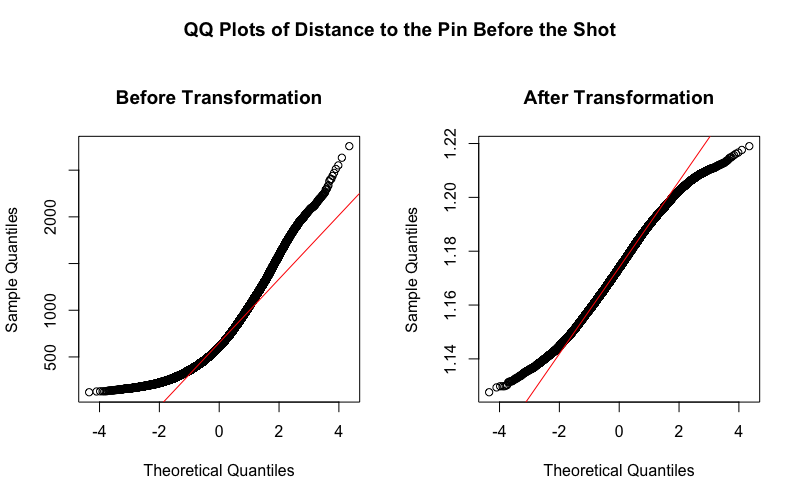


Once the outliers are removed, the data has a more normal distribution, but there is still some skew. To address the skew, I transformed the data, using Tukey’s Ladder of Powers, in order to determine the most effective way to transform each variable. **Figure 7** and **Figure 8** show the results of the transformation of each variable on QQ Plots, which visualize the normality of data before and after the transformation.

**Figure 8: Data Transformation of Distance to the Hole After the Shot**



**Figure 7: Data Transformation of Distance to Pin Before the Shot**



# RESULTS

Once the data were transformed, I conducted the analysis. For the ANCOVA I ran a model that explored whether there was an effect on distance to the hole after the shot, at different locations and different distances. I also explored the interaction between location and distance in order to determine whether distance to the hole before the shot magnified or diminished the effect of the shot location on performance.

The results of this model can be seen in **Table 3** (see appendix). The interaction between location and distance before the shot each exhibit a significant relationship with distance to the pin after the shot, I could not continue with an ANCOVA because the model does not pass the assumption that there’s no interaction between the independent variable and covariate.

I could, however, still run a normal ANOVA in order to determine if there were a difference between shot locations without controlling for distance. **Table 4** (see appendix) shows the output of the ANOVA, while **Table 5** shows descriptive statistics for the transformed distance to the hole following the shot**.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 5: Descriptive Statistics for the ANOVA model**  *( Transformed Distance to Hole Following Shot ~ From Location)* | | | |
| **From Location** | **Mean** | **Median** | **Std. Deviation** |
| Fairway | 3.12 | 3.10 | 0.76 |
| Green Side Bunker | 3.47 | 3.48 | 0.73 |
| Primary Rough | 3.35 | 3.35 | 0.77 |

The results show that there is a significant (p<0.01) difference between the different shot locations, and shots from the fairway are associated with the best performance, followed by shots from rough, with shots from the green side bunker just behind shots from the rough. Based on these results, H1 and H2 are not supported.

In order to paint a better picture of how each shot location affects shot performance, I also ran a multiple linear regression model. The regression model explored the effect that shot location have on the distance to the hole following the shot, along with how distance away from the pin before the shot affects the distance away from the hole after the shot. The model also explains the interaction between shot location and distance away from the pin further. **Table 6** shows the results of the regression.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 6: Regression Analysis to Determine Impact of Location and Distance**  *(Distance to Hole Following Shot ~ Transformed Distance to Pin + From Location + Transformed Distance to Pin: From Location)* | | | | | |
| Coefficients | Estimate | Std. Error | | T-Value | P-Value |
| (Intercept) | -22.6455 | 0.3072 | | -73.721 | <2e-16 \*\*\* |
| Transformed Distance to the Pin | 21.9342 | 0.2615 | | 83.871 | <2e-16 \*\*\* |
| Location – Green Side Bunker | 16.6440 | 0.6038 | | 27.565 | <2e-16 \*\*\* |
| Location – Primary Rough | 4.5600 | 0.4487 | | 10.162 | <2e-16 \*\*\* |
| Location – Green Side Bunker: Transformed Distance to the Pin | -13.8753 | 0.5140 | | -26.996 | <2e-16 \*\*\* |
| Location – Primary Rough: Transformed Distance to the Pin | -3.6471 | 0.3824 | | -9.537 | <2e-16 \*\*\* |
| --- |  |  | |  |  |
| Significance | \*\*\* 0.001 | \*\* 0.01 | | \* 0.05 | . 0.1 |
| -- |  |  | |  |  |
| Residual Std. Error: 0.7015 on 72579degrees of freedom | | | | | |
| Multiple R-Squared: 0.1694 | | | Adjusted R-Squared: 0.1693 | | |
| F-Statistic: 2960 on 5 and 72579 DF | | | p-value: <2e-16\*\*\* | | |

The R-squared for this model is 0.1694 (p<0.01), meaning that 16.94% of the variance of distance to the pin following the shot is explained using these variables. Given how many factors might affect each shot, and how no two shots are exactly the same, this is a respectable adjusted R-squared value. The results of the regression show that all factors are significantly related to the distance to the hole following the shot (p < 0.01). Looking specifically at the results for shot location, shots from the fairway lead to the closest proximity to the hole, followed by the primary rough, followed by shots from green side bunkers. These findings support the traditional assumptions of how shot locations are designed to punish players; however, by looking at the interaction between shot location and distance to the pin before the shot, and the corresponding plot (**Figure 9**), it is evident that shots from the bunker do not have the same effect as shots from the fairway and rough.

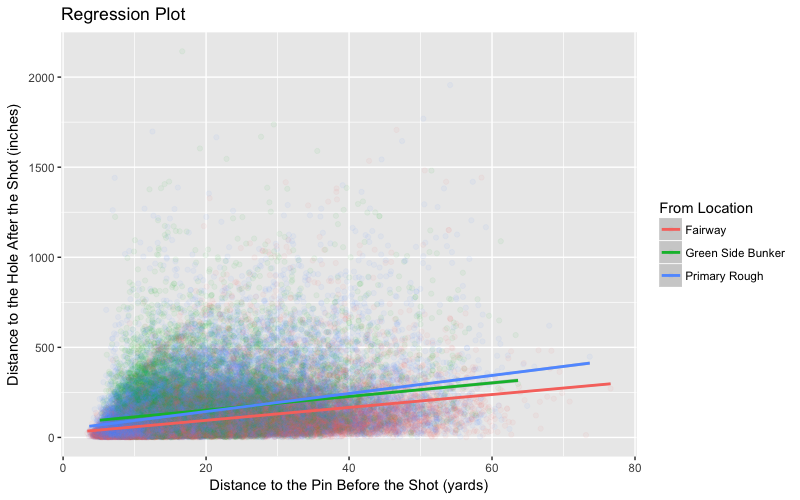
If they had the same effects, then the slopes of all three lines would be, more or less, parallel. If anything, I’d expect the slopes to get incrementally steeper from fairway to rough to bunker, as illustrated in Figure 4 because distance to the hole will compound the effects of each shot location. Figure 9 shows that this is not the case. The slopes of fairway (red) and rough (blue) are roughly equal, but the slope of the green side bunker line (green) is drastically shallower, to the point where shots from the bunker have an equal effect as shots from the fairway.

**Figure 9: Regression Plot Using Transformed Data**



I suspected that this was a result of most of the outliers that were removed being from shots from the bunker. To test this suspicion, I ran the regression again without removing any outliers. This model isn’t as effective because the data don’t meet the necessary assumptions as well as the transformed data, but it can still be supplemental to the original model. The plot (**Figure 10**) of this model illustrates a more predictable relationship between the shot location and distance to the pin prior to the shot. Each shot location has a similar starting position on the graph for close shots. As the distance before the shot increases, the distance of the lines between fairway and rough increase. This is what I’d expect: the farther away a player is, the more penalized he is by unfavorable shot locations. Looking at the relationship between shots from the green side bunker and distance to the hole on performance, however, is once again different. The slope of this line is almost parallel to the fairway line, and intercepts the line for primary rough. These results do not support H3, so I reject this hypothesis. The results of these analyses suggest that, on the PGA TOUR, shots from the bunker are not proportionately penalizing players compared to the rough. Players may, in fact, be better off inside the bunker than in other locations.

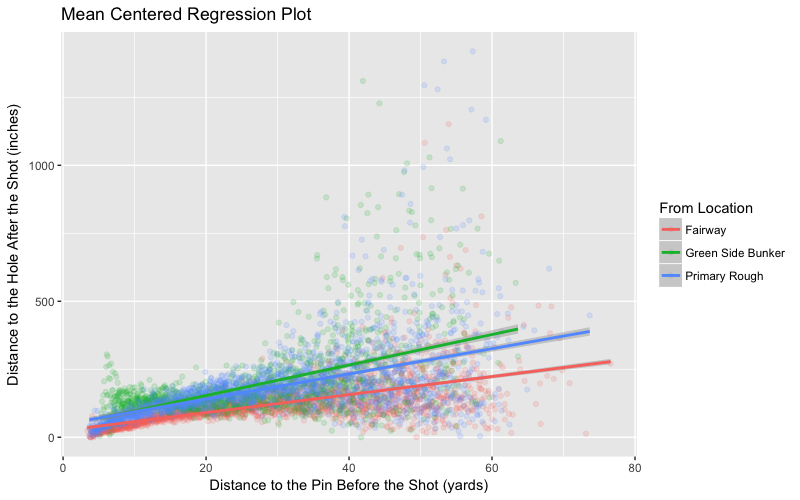
**Figure 10: Regression Plot Using Original Data**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 7: Mean Centered Regression Analysis to Determine Impact of Location and Distance**  *(Avg. Distance to Hole Following Shot ~ Distance to Pin + From Location + Distance to Pin: From Location)* | | | | |
| Coefficients | Estimate | Std. Error | T-Value | P-Value |
| (Intercept) | 24.25 | 5.105 | 4.750 | 2.09e-06 \*\*\* |
| Distance to the Pin | 0.092 | 0.004 | 22.320 | <2e-16 \*\*\* |
| Location – Green Side Bunker | 16.994 | 7.94 | 2.14 | 0.0324 \* |
| Location – Primary Rough | 23.01 | 7.93 | 2.141 | 0.00164 \*\* |
| Location – Green Side Bunker: Distance to the Pin | 0.0637 | 0.00690 | 9.230 | <2e-16 \*\*\* |
| Location – Primary Rough: Distance to the Pin | 0.0368 | 0.00595 | 6.189 | 6.53e-10 \*\*\* |
| --- |  |  |  |  |
| Significance | \*\*\* 0.001 | \*\* 0.01 | \* 0.05 | . 0.1 |
| -- |  |  |  |  |
| Residual Std. Error: 102.6 on 5270 degrees of freedom | | | | |
| Multiple R-Squared: 0.3378 | | Adjusted R-Squared: 0.3372 | | |
| F-Statistic: 537.8 on 5 and 5270 DF | | p-value: <2.2e-16\*\*\* | | |

In addition to running a regression with transformed data, I also used mean centering to run an additional regression. This means that I grouped all shots together that had the same distance to the hole before the shot, and found the average distance to the hole after the shot for each distance. By running the regression using means of distance to the pin after the shot, I was able to keep shots that went in the hole in the analysis, because there were no longer any zeros within the data for distance to the hole following the shot after calculating the means. **Table 7** shows the results of the analysis, while **Figure 11** plots the regression on a graph. When using mean centered data, the analysis is able to explain 33.78% of the variance compared to 16.94% when using transformed data. The results of the regression tell a much different story when using mean centered data, compared to using transformed data. When controlling for distance, shots from the bunker actually lead to better performance than shots from the rough. Compared to shots from the fairway, being in the rough alone increases the distance to the hole after the shot by 23.01 inches, while being in the bunker only increases the distance to the hole after the shot by 16.994 inches. These results support H1, but fail to support H2. Figure 11 shows a different interaction than the previous regression plot. In this case, the interaction between distance and shot location is closer to the expected results. From closer distances, the performance from each shot location is similar; however, as distance increases, the penalty for hitting into an unfavorable location increases. In this case, this is also true for shots from the green side bunker. The penalty for hitting into the bunker compared to the rough isn’t as great as I would expect, however. These results support both H3 and H4.

**Figure 11: Mean Centered Regression Plot**

****

# LIMITATIONS

The largest limitation with this research is the nature of the data. No two shots in golf will ever be truly identical, so this makes it difficult to compare between shots. ShotLink data are rich enough that controlling for covariates is possible to an extent; however, there are a nearly unlimited amount of factors that can affect a single golf shot. There are atmospheric factors such as temperature, humidity, wind, and altitude that play a large role in every golf shot. ShotLink does not currently record atmospheric data for each shot, but it is only a matter of time before ShotLink is setting up mobile weather stations throughout the course to record the atmospheric data at the time of each shot. Until then, these variables will not be able to be controlled for.

There are also a number of confounding factors that ShotLink will never be able to include in their data. These factors include the pressure that a player is facing, crowd distractions, or even a player’s health on a particular day. The possibilities of “intangible” confounding factors are endless, and illustrate one of the difficulties faced with analyzing sports data.

While there are many challenges faced with this research, there are steps that could be done in future research to help mitigate the effects of some of these covariates. Future research could limit data to single players in order to control for different skill levels between players. Future analyses could also limit data to a single course in order to ensure that the conditions as consistent as possible between each shot.

While there are many limitation faced with this research, some of which will never be able to be addressed, there are many steps that can be taken to mitigate the limitations. And limitations should not be a barrier to performing analyses on the data. The data are still extremely valuable, and worth analyzing.

# Recommendation

Based on the results of the analyses, I recommend that the PGA TOUR take a different approach to how they prepare courses for tournaments. The data suggests that bunkers are not impacting players as much as intended. Course architects strategically place bunkers around the green to challenge players, and to entice risk-reward situations on approach shots. Because bunkers don’t penalize PGA TOUR players, they are able to take advantage of the course design, since they don’t have to be worried about missing their shot into the bunker. In fact, in many cases hitting a shot into the bunker is a *good* miss. This is counter-intuitive to the way that the game is meant to be played.

The biggest reason that players are able to take advantage of green side bunkers is because of the conditions within them. Under the current rules, bunkers are raked after each shot. This creates a very favorable and consistent lie within the bunkers. Players at the professional level would prefer to be in the bunker because they know exactly what to expect when they hit it in the bunker. Nearly every time they can expect smooth sand that doesn’t pose a problem, and allows them to hit the ball close to the pin.

**I suggest that beginning in the 2019 season, the PGA TOUR stops the practice of raking bunkers at Web.com Tour events.** By not raking the bunker, the conditions will be inconsistent, and much less favorable. This will increase the penalty assessed on a player for hitting a shot into the bunker. The Web.com Tour is essentially the minor leagues to the PGA TOUR. By using the Web.com Tour as a trial for this rule change, it provides the PGA TOUR with a lower pressure environment to test out this new rule. At the conclusion of the 2019 season, the PGA TOUR can analyze the data to determine the effectiveness of this new rule, and determine whether or not the rule should be applied on the PGA TOUR.

This is also a cost effective solution. There would be no costs incurred to the PGA TOUR or to the courses that host the tournaments. During the week of the tournament, the host course would simply stop raking their bunkers. At the conclusion of the tournament, the course would rake the bunkers, and there would be no impact on recreational golfers for the rest of the year.

# CONCLUSION

The purpose of this report was to determine the effect that different “lies” around the green have on a player’s performance for that shot, and determine if the different lies have any significant impact on PGA TOUR players’ performance. Using ShotLink data, I ran analyses that suggested that bunkers are not currently penalizing players effectively. Based on the results I recommend that the PGA TOUR stop raking bunkers for Web.com Tour events in 2019. This will act as a testing environment for the new rule, and based on the results, the rule can be implemented to the PGA TOUR for the 2020 season.

# WORKS CITED

Broadie, Mark. "Assessing Golfer Performance Using Golfmetrics." In *Science and Golf V: Proceedings of the 2008 World Scientific Congress of Golf*, 253-262. Ed. D. Crews and R. Lutz. Mesa, AZ: Energy in Motion Inc., 2008.

Broadie, M. (2012). Assessing Golfer Performance on the PGA TOUR. *Interfaces*, *42*(2), 146-165.

Burke, M. (2012). ShotLink Is Making Golf Easier For Hacks And Harder For Pros. *Forbes.Com*, 30.

Hickman, D. C., & Metz, N. E. (2016). Peer Effects in a Competitive Environment: Evidence from the PGA TOUR. .

Leahy, B. (2014) *Predicting Professional Golfer Performance Using Proprietary PGA Tour “Shotlink” Data*. A dissertation submitted in partial fulfillment of the requirements of Dublin Institute of Technology for the degree of M.Sc. in Computing (Data Analytics) July 2014.

Martin, L. (2016). *Sports Performance Measurement and Analytics: The Science of Assessing Performance, Predicting Future Outcomes, Interpreting Statistical Models, and Evaluating the Market Value of Athletes*. Old Tappan, NJ: Pearson Education, Inc.

Newport, J. P. (2014). Stats, Chaos in the TV Trailer. *Wall Street Journal - Eastern Edition*. p. A14.

Sens, J. (2013). It All Adds Up. *Golf Magazine*, *55*(8), 84.

Yocom, G. (2017, August). Info Seekers. *Golf Digest*, 94-96.

# APPENDICES

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 3: Testing ANCOVA Interaction Assumption**  *(Distance to Hole Following Shot ~ Transformed Distance to Pin + From Location + Transformed Distance to Pin: From Location)* | | | | | |
|  | Degrees of Freedom | Sum of Squares | Mean Squares | F-Value | P-Value |
| Transformed Distance to the Pin | 1 | 5253 | 5252.8 | 10673.67 | <2e-16 \*\*\* |
| From Location | 2 | 1671 | 835.6 | 1697.98 | <2e-16 \*\*\* |
| Transformed Distance to the Pin: From Location | 2 | 359 | 179.3 | 364.42 | <2e-16 \*\*\* |
| Residuals | 72579 | 35718 | 0.5 |  |  |
| --- |  |  |  |  |  |
| Significance | \*\*\* 0.001 | \*\* 0.01 | \* 0.05 | . 0.1 |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 4: ANOVA**  *(Distance to Hole Following Shot ~ From Location)* | | | | | |
|  | Degrees of Freedom | Sum of Squares | Mean Squares | F-Value | P-Value |
| From Location | 2 | 1544 | 771.9 | 1351 | <2e-16 \*\*\* |
| Residuals | 72582 | 41457 | 0.6 |  |  |
| --- |  |  |  |  |  |
| Significance | \*\*\* 0.001 | \*\* 0.01 | \* 0.05 | . 0.1 |  |