Stanford Men's Varsity Golf Team Performance Analysis: Informing Putting and Approach Shot Practice and Strategy on a Risk-Reward Basis

Lindsey Kostas, Luke Lefebure, Sam Sarpong, Chris Sebastian

INTRODUCTION

Two of the most important shots in golf are the approach shot and the putt. These shots can separate the good from the great players. Approach shots are shots made towards the green after a tee-shot and putts are any stroke made on the putting green.

The following analysis examines the approach shot and putting performance of members of the Stanford Men's Varsity Golf Team with the goal of helping each player achieve his best possible score by improving shot selection (e.g. risk vs. reward) and practice regimens. To that end, this analysis identifies the players who may improve their scores most by focusing on putting and those who should focus on approach shot accuracy.

DATA

This analysis is based on a Shots to Hole dataset provided by the Stanford Men's Varsity Golf Team. Shots to Hole is a diagnostic golf software that allows golfers to input information about each hole they play, such as the distance to the hole and club selection for each shot.

The dataset contains 2646 entries, with each entry corresponding to a hole that was played. This includes data for the Stanford Golf Team's nine players for both practice and competition rounds played between September 2014 and early November 2014, with the exception of one player who also recorded data from the 2013 spring and summer seasons. See **Figure 1** for a sample row in the dataset.

The dataset is missing many entries because players often failed to record complete information for a given hole. This problem is particularly pervasive in the shot difficulty, shot result, putt break, and putt slope data. Incomplete data is somewhat less pervasive, but equally problematic, for the club used data. The only complete data is the number of strokes, the distance of each stroke, and the lie of each stroke. For this reason, the analysis focused on this subset of the data. However, the small size of this dataset, the disparity in the number of courses played by each player, and the type of round, practice or competition, played by each player, still limited the reliability of the results obtained from this data.

METHODS

The raw data set was first parsed to retrieve the initial putt distance and number of putts for each hole played. Using the multinom function in R's nnet library, separate multinomial logistic regression models were then trained on each player's data to examine how putting performance varies with distance from player to player. These models used the number of putts as a categorical response variable and recorded distance to hole on the initial putt as a quantitative predictor variable. The player's probability of taking each number of putts (i.e. 1, 2, 3, or 4) from a given initial putt distance ranging from one foot to 115 feet was computed with the coefficients estimated from his training model. A multinomial logistic regression model was also trained on the players' combined data to obtain "team average" probabilities of taking a certain number of putts from each distance. The expected number of putts from each distance for each player and the "team average" were then calculated from these regression models with the equation: 1*P(1 putt) + 2*P(2 putt) + 3*P(3 putt) + 4*P(4 putt). These probabilities necessarily sum to 1 because no player took five or more putts.

The model's fit was assessed with a 10-fold cross-validation misclassification error rate using the erroforest function in the ipred package in R. This assessment was used to compare the predictive accuracy of the fitted multinomial logistic regression model against other classification models that were investigated for each player before selecting the final multinomial logistic regression predictive model. These other classification methods that were trained on the data included linear discriminant analysis, quadratic discriminant analysis, naïve bayes, and random forest.

Proceeding with the results generated by the multinomial regression models, a "strokes gained putting" metric was then calculated by subtracting a player's actual number of putts from a given distance from the "team average" expected number of putts from that distance. For example, if a player 2 putts from an initial distance of 10 feet and the "team average" expected number of putts from 10 feet is 1.7, this player's "strokes gained putting" on this hole is -0.3. This metric was then averaged for each player and multiplied by 18 to get an 18-hole average. Thus if a player's 18-hole average is 1.2 "strokes gained putting," then he is expected to take 1.2 fewer putts per round than the average player if both players faced the same initial putt distances on every hole. For comparison purposes, a more traditional putting statistic which does not account for distance, average number of putts per hole, was calculated.

Two visualizations were then developed for each player's expected number of putts function. One graph depicts the marginal decrease in expected number of shots by getting one foot closer to the hole from each putt distance. This marginal decrease is calculated by subtracting the expected number of strokes from x feet from the expected number of strokes from x-1 feet. The other graph plots the difference between each player's expected number of putts and the "team average" expected number of putts to identify the relative putting strengths and weaknesses of each player.

A similar analysis was performed on each player's short game and approach shots data. The player's short game was defined as any shot 50 yards or less from the green (excluding putting) that landed on the green, and approach shots were defined as any shot between 50 and 200 yards from the green that landed on the green. The short game and approach shot distances to the hole were plotted against the resulting putt distance for each player. Finally, linear regression models were fitted to this data, but these models produced no discernable relationship between initial distance and resulting distance for any player.

RESULTS

The multinomial logistic regression model provided the optimal model fit to the data for each player, yielding the lowest 10-fold cross-validation misclassification error rate. The next-best performing model was a random forest model which produced higher error rates for all players except for Patrick Grimes and Viraat Badhwar. Despite the random forest model's better performance for Grimes and Badhwar, the multinomial logistic regression models were adopted because the respective visualizations of the predicted putting probability data from the random forest model for each player shows that this model generally over-fits the relatively small training sample. **Table 1** summarizes the misclassification error rates of the multinomial logistic regression and the random forest models for each player. **Figure 2** illustrates the random forest model where the top panel represents one putts, the middle panel represents two putts, and the bottom panel represents three puts. **Figure 2** only shows the analysis for one of the nine players, but the data for each of the remaining eight players produced comparable results. The irregular curves with multiple peaks and troughs in **Figure 2** show that the random forest model over fits the data.

A sample of the fitted multinomial logistic regression estimates of putting probability is presented in **Table 2**. This sample shows the probability estimates for one player for initial putts between 1 and 5 feet from the hole. Similar estimates were made for all players from all distances between 1 and 115 feet. **Figure 3** shows each player's probability of one putting as a function of initial putt distance. As expected, the function for one putt is monotonically decreasing in distance and this observation was consistent across all players (i.e. the further away a player is from the hole, the less likely he is to reach the hole in one putt). The function for two putts, shown in **Figure 4**, also behaves as expected by increasing until it peaks in the range from 20 to 35 feet from the hole and then decreasing. The only exception to this relationship is Franklin Huang who exhibits a strictly increasing function that peaks and then flat-lines at approximately 30 feet from the hole. Finally, **Figure 5** shows that the function for three putts also behaves as expected by monotonically increasing in distance. With the exception of two players, Huang and Grimes, the function for three putts initially increases at an increasing rate and then reaches a point of inflection in the range between 30 and 60 feet from the hole at which point it begins to increase at a decreasing rate. The two exceptions have functions that increase at an increasing rate (Grimes) and increase at a decreasing rate (Huang) for the entire range of distances. The analysis performed for four putts is omitted from the appendix because only one of the nine players, Badhwar, had data for four putts. Badhwar's four putts function has a similar shape as the three putts function with the main difference being it is shifted to the right.

Table 3 shows a summary of players' putting metrics, including the 18 hole average of "strokes gained putting" and average number of putts for each hole. **Table 4** illustrates that the rankings of the players' putting ability based on these metrics differ significantly.

Figure 6 shows the difference between a player's expected number of putts and the "team average" expected number of putts as a function of distance. A player whose function is above zero for a given distance is doing worse than the average member of the team from that distance and a player whose function is below zero is doing better than the average member of the team. This graph illustrates which players are the best and worst at long, short, and mid-range putts.

Figure 7 depicts the marginal decrease in expected strokes gained by getting one foot closer to the hole for each player. A one foot improvement between about 10 and 12.5 feet improves the score the most for all players, illustrated by the locations of the local maximums of each player's curve. Bradley Knox, David Boote, Grimes, and Jeffrey Swegle have steeper curves, while Jim Liu and Dominick Francks have much flatter curves. In general, a steeper curve represents a greater marginal benefit from getting one foot closer to the hole, whereas the flatter curves mean that getting one foot closer isn't expected to change a player's score significantly.

A similar analysis was performed for approach shots by analyzing both lie to distance and distance to distance relationships. **Figure 8** depicts a scatterplot of all shots that landed on the green. The distance to the hole on the initial shot appears on the horizontal axis, and the resulting putt distance is on the vertical axis. **Figure 9** depicts the best fit regression output that corresponds to **Figure 8**. **Figure 9** illustrates that while the coefficients from the regression are statistically significant, indicated by near zero p-values, there is a general absence of correlation among the data, indicated by the R-squared value of .29. This low R-squared value implies that due to excessive noise in the data there is no discernable relationship between the distances to the hole from which an approach shot is hit and at which the approach shot lands. A similar analysis was conducted for each player and produced similar results. These results establish that distance from which an approach is hit cannot be used to predict where the approach will land.

DISCUSSION

Table 3 and **Table 4** illustrate the difference between using "strokes gained putting" and average number of putts per hole to evaluate putting ability. Ranking the players on these two metrics produces very different results. These differences exist because the strokes gained statistic takes into account the distance putted whereas the average number of putts does not. In general, players that rank the best for the "strokes gained putting" metric have the best putting ability. As a result, a player who ranks better on the "strokes gained putting" metric than on the average number of putts metric can attribute his high number of average putts to the greater distance to the hole and not because of inferior putting skills. For example, Knox ranks second in strokes gained but seventh in average putts. This suggests that, after accounting for distance putted, Knox has the second best putting ability on the team despite taking, on average, more putts than his teammates. Thus, Knox may improve his score the most by placing his approach shots closer to the hole as opposed to trying to improve his putting.

Figure 6 provides insight into a player's putting strengths and weaknesses. Francks is the only player who is worse than average for the entire range of distances, and Huang is the only player who is better than average for the entire range of distances. The remaining players' performances fluctuate based on distance. Maverick McNealy and Jim Liu perform worse than average for short putts but better than average for middle and long range putts. Badhwar, Boote, Knox, and Swegle perform better than average on short putts but worse than average on middle and long range putts. Grimes performs worse than average on middle range putts, but better than average on short and long range putts. Based on the analysis of each player's putting ability, a player stands to gain most if he improves from the distance where his curve reaches its maximum because this is where he is performing worst relative to his teammates. For example, a player who performs as Grimes should work on putts from around 12 feet from the hole. Insofar as all players perform relatively similar and close to average for mid-range putts between 15 and 40 feet, this convergence around the team average suggests that practicing mid-range putts may confer the least marginal benefit. Instead, players are better served practicing short and/or long putts.

Figure 7 divides the players into two groups: those who should focus on improving putting and those who should focus on improving approach shots. **Figure 7** also establishes a risk-reward relationship between a player's expected score and his approach shot. In general, a steeper curve represents a greater marginal benefit from getting one foot closer to the hole, whereas the flatter curves indicate that getting one foot closer doesn't change a player's score significantly. Based on a risk-reward analysis, only the players with steeper curves have an incentive to go for a riskier approach shot that places the ball closer to the hole because the marginal gains may outweigh the potential costs. Players like Knox, Boote, Grimes, and Swegle, who all have steep curves, can benefit most by simply getting a little closer to the hole on their approach shot. Because these players may want to play more aggressively, they should also work on improving approach shot accuracy. Conversely, players like Liu and Francks, who have flatter curves, do not have as great a marginal benefit from getting closer to the hole, so the increased risk of an aggressive approach shot may outweigh the benefit of positioning the putt at a safer, but farther, distance from the hole. These flat curve players, given their current putting ability, are better served by focusing on improving their putting rather than trying to get closer to the hole on their approach shot. Further, all players' curves peak within the same ~3 foot range and behave similarly at the tails, suggesting that this three foot region is the range of distances in which players like Liu and Francks, with the lowest maximum gains and flattest curves, have the most opportunity for improvement. For such players, practicing putting from these distances will raise the local maximums for their respective curves and provide them with the same opportunity as their teammates to improve their score by getting closer to the hole.

Figure 8 and the corresponding regression in **Figure 9** show a slight positive trend in the data, but this line is skewed by a large cluster of points in the bottom left. This cluster of points represents chip shots and other shots that likely resulted from missed approach shots. Discarding this cluster (shots with an initial distance of less than about 50 yards) and re-running the regression results in a much flatter line. The re-examined results suggest that there is no relationship between initial shot distance and resulting putt distance. Unlike putts, which are performed on a relatively homogeneous surface every time, shots from off of the green are influenced by a multitude of factors, including lie, course conditions, and pin placement. This dichotomy did not distort the results obtained from the analysis of the putt data, but rendered the analysis of the approach and short game shots data useless. If the dataset had been sufficiently detailed and large enough to filter short and approach shots by their relative difficulty, the results obtained by this analysis may be very different and more conclusive.

LIMITATIONS AND FUTURE ANALYSIS

The initial goal was to apply David Romer's "Go-For-It" analysis of the fourth down in football to golf. That is, improve match strategy with a single function that provides the Stanford Golf Team with sufficient information to determine how aggressively a player should hit an approach shot (i.e. when should a player "go-for-it"). The crux of the golf "go-for-it" model depends on the outcomes of different strategies; a balance between the score a player attains with a riskier shot by attacking a difficult pin with the score produced by a much safer shot such as when the player lays a shot up.

This analysis could not be performed due to the absence of any data about the player's level of aggression on any given shot. It was also not possible to infer such level with any reasonable certainty solely from the raw data. For example, the data did not consistently specify metrics, such as club selection, from which the players' level of aggression could be inferred. As previously noted, the only metrics that were consistently reported by players were the distance to the hole and number of shots taken. Further, even if the players' levels of aggression could be reasonably and reliably inferred, a single function golf go-for-it model would also require knowledge of course topography, weather conditions, and pin locations - data that was not recorded and, in any event, impossible to obtain.

Even the relatively complete data, distance, lie, and number of shots, was not sufficient to create a single golf go-for-it model. As a result, the analysis was modified to create a model for each member of the golf team. This model attempted to calculate a player's expected score on a given hole based on the probabilities a player assigns to reaching various positions on the course. To that end, the model was a weighted linear combination of functions that attempted to describe each player's short game and putting ability as a function of distance to the hole. Specifically, the player would assign a probability to his next shot for each of the scenarios described by one of the model's functions, based on how aggressive he hit his current approach shot. For example, $E(score) = w1*p(x)+w2*b(x)+w3*c(x)$, where $p(x)$ is putting as a function of distance, $b(x)$ is bunker shots as a function of distance, $c(x)$ is chip shots as a function of distance, and the wi's are the player selected probability weights. Armed with this function, a player, ideally, could determine the optimal strategy by comparing the expected score from multiple approach strategies and their associated probability weights.

But even this analysis proved to have limited utility. Again, the analysis was hampered by the small size of the dataset. With very few courses played by each player, there was not sufficient data to produce any significant or reliable relationship between the distance from the hole or the lie of the approach shot and the distance hit on the approach shot. Specifically, the bunker and chip shot data was not sufficient to permit the use of this equation. As a result, the analysis was limited to putting performance as it relates to distance from the hole.

In an attempt to create a limited putting centric go-for-it model, the available data was used to determine if a player is maximizing his putting ability. This analysis required data about holes for which all nine players had played, as well as, knowledge of course topography and pin location. As a result, the analysis was necessarily limited to one of the Stanford Golf Course holes. However, even this scaled back approach was hindered by insufficient data; there were only 30 entries for rounds played on the Stanford Golf Course, with multiple players only playing this course once. Since the analysis of putting performance consisted of on average calculations, the small dataset precluded any reliable simulation of a given hole which could be used to analyze whether a player was playing optimally.

Ultimately, the analysis was limited to putting metrics as a function of distance from the hole. While this analysis produced interesting results about which aspects of a player's game he should practice and which players stand to gain the most with riskier shots, the analysis is essentially limited to one aspect of the game – putting. However, if sufficient data can be compiled, this analysis can form the basis for the construct of models that may analyze a player's entire game. That said, the results achieved by this analysis can still inform players about their putting strengths and weaknesses, the importance of this aspect of their game to their final score, and, to an extent, how much risk a player should take on his approach shot.

APPENDIX

Table 1.

10-Fold Cross-Validation Misclassification Error Rates

Player Name	**Multinomial Logistic Regression**	**Random Forest**
Bradley Knox	16.20%	16.76%
David Boote	21.07%	23.06%
Dominick Francks	27.27%	30.42%
Franklin Huang	23.94%	28.19%
Jeff Swegle	22.22%	22.78%
Jim Liu	20.26%	26.29%
Maverick McNealy	22.59%	23.59%
Patrick Grimes	**19.21%**	**17.22%**
Viraat Badhwar	**21.73%**	**20.61%**

Table 2.

Multinomial Logistic Regression Putting Probability Estimates for the first 5 feet of first putt distance to the hole for player Bradley Knox

Distance (feet)	1 Putt	2 Putts	3 Putts	4 Putts	Probability Sum
1	0.9315	0.0684	0.0001	0.0000	1
2	0.9129	0.0869	0.0002	0.0000	1
3	0.8898	0.1010	0.0002	0.0000	1
4	0.8615	0.1381	0.0003	0.0000	1
5	0.8274	0.1722	0.0005	0.0000	1

Table 3.

Putting Ability Metrics

Player Name	**Strokes Gained**	**Average First Putt Distance**	**Average Number of Putts per Hole**
Franklin Huang	1.020109413	18.05792	1.656370656
Bradley Knox	0.547945187	19.21229	1.709497207
Viraat Badhwar	0.510772366	17.62953	1.699164345
Maverick McNealy	0.371290794	15.88704	1.637873754

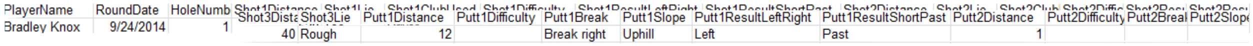
Jeff Swegle	0.248766885	17.44111	1.688888889
Patrick Grimes	0.002671094	16.00662	1.665562914
Jim Liu	-0.14262183	18.65517	1.711206897
David Boote	-0.327432108	15.94632	1.701789264
Dominick Francks	-1.766484383	18.70979	1.807692308

Table 4.

Player Putting Ability Rankings

Rank	Player's Strokes Gained	Players Average Putts
1	Franklin Huang	Maverick McNealy
2	Bradley Knox	Franklin Huang
3	Viraat Badhwar	Patrick Grimes
4	Maverick McNealy	Jeff Swegle
5	Jeff Swegle	Viraat Badhwar
6	Patrick Grimes	David Boote
7	Jim Liu	Bradley Knox
8	David Boote	Jim Liu
9	Dominick Francks	Dominick Francks

Figure 1.



PlayerName	RoundDate	HoleNumb
Bradley Knox	9/24/2014	1

Shot3Dista	Shot3Lie	Putt1Distance	Putt1Difficulty	Putt1Break	Putt1Slope	Putt1ResultLeftRight	Putt1ResultShortPast	Putt2Distance	Putt2Difficulty	Putt2Breal	Putt2Slope
40	Rough	12		Break right	Uphill	Left	Past	1			

Figure 2.
Random Forest Probabilistic Classification Predicted Putting Probability Function for Bradley Knox *Observed over-fitting

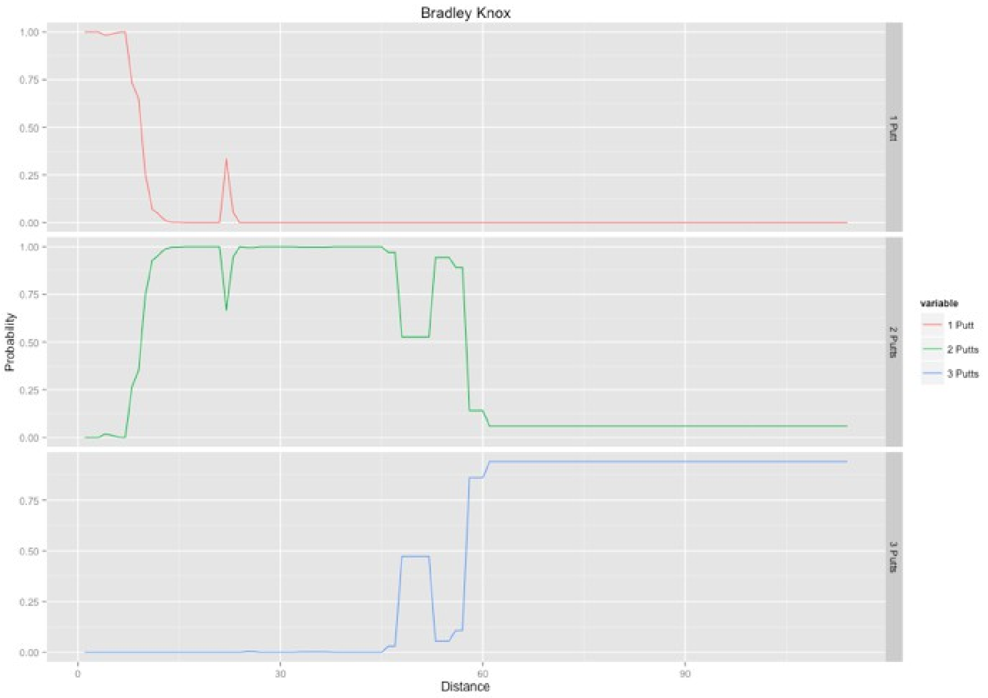


Figure 3.

Multinomial Logistic Regression Putting Probability Functions for One Putt

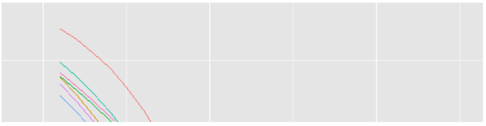


Figure 4.

Multinomial Logistic Regression Putting Probability Functions for Two Putts

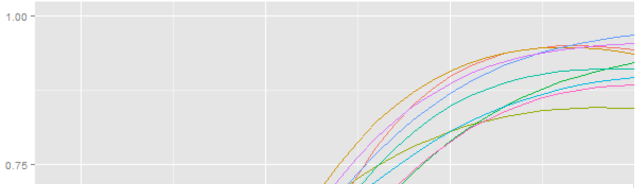


Figure 5.

Multinomial Logistic Regression Putting Probability Functions for Three Putts

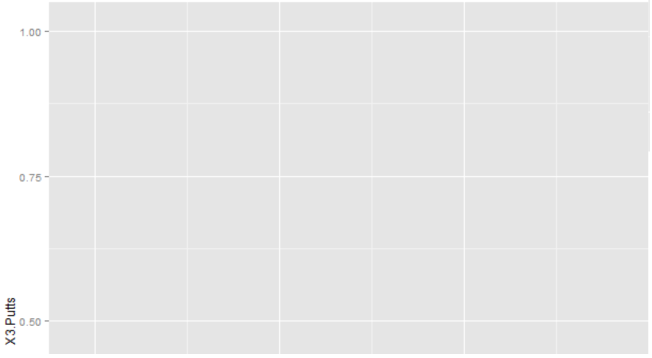


Figure 6.

Expected Difference Putting

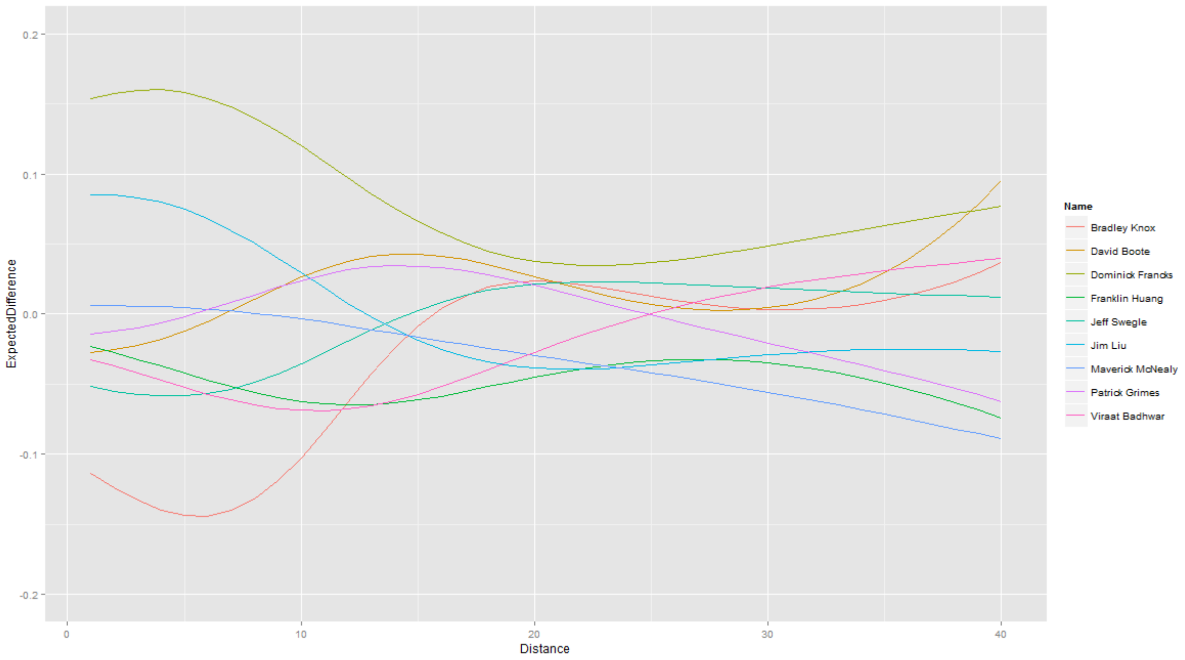


Figure 7.

Strokes Gained for Putts 1 Foot Closer to the Hole

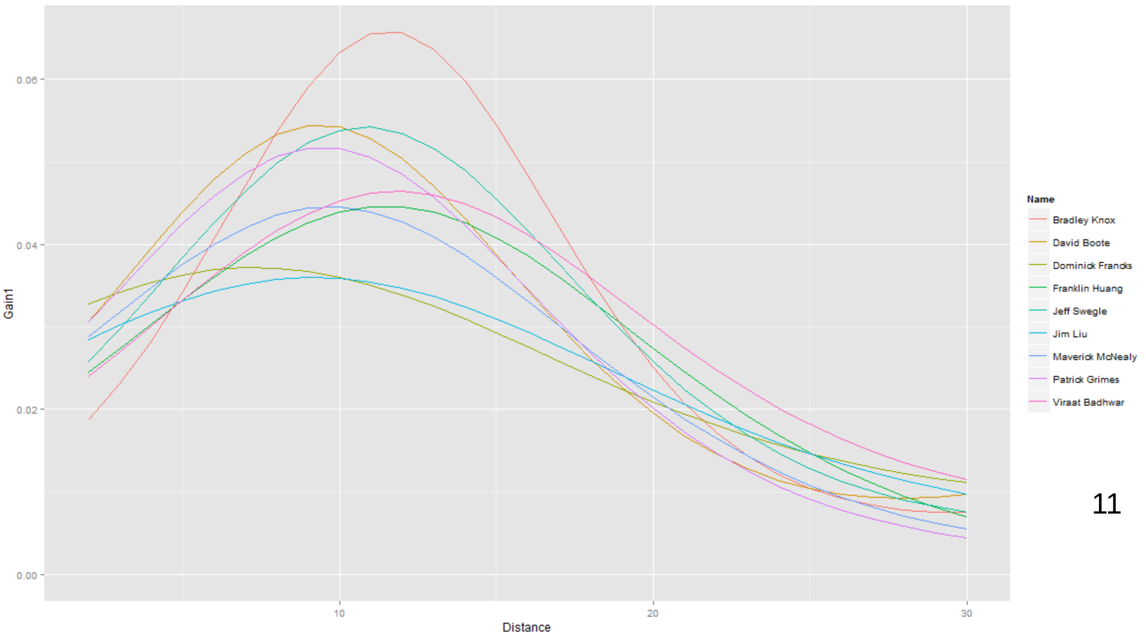


Figure 8.

Distance Hit on Approach Shots

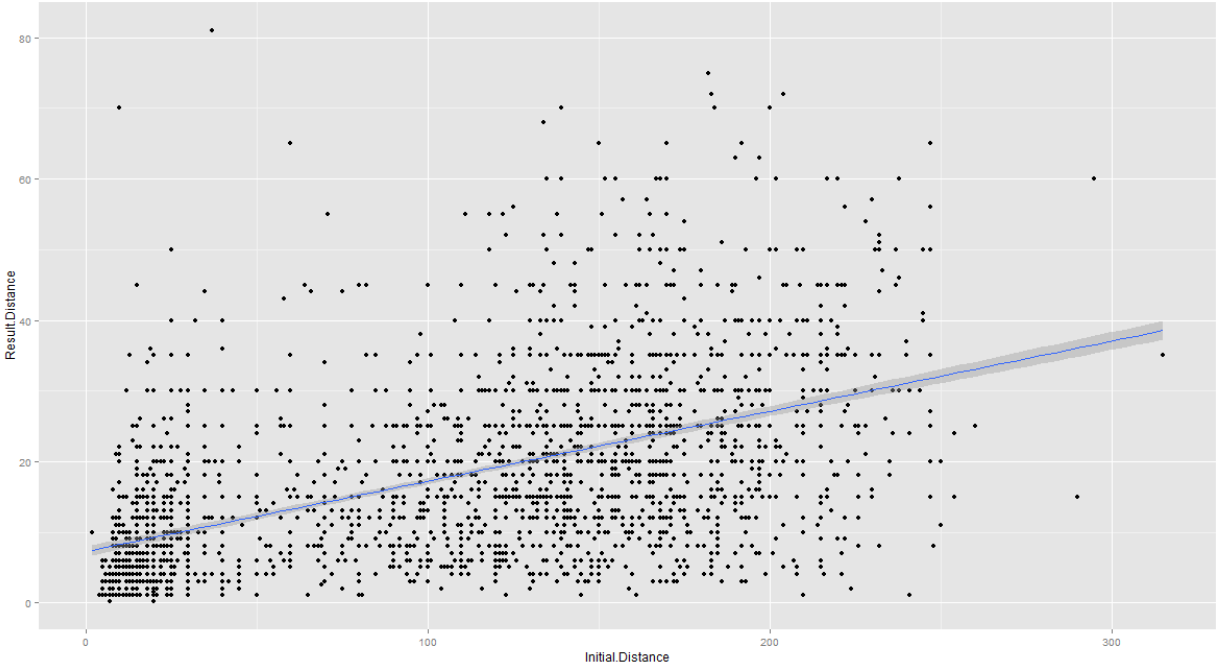
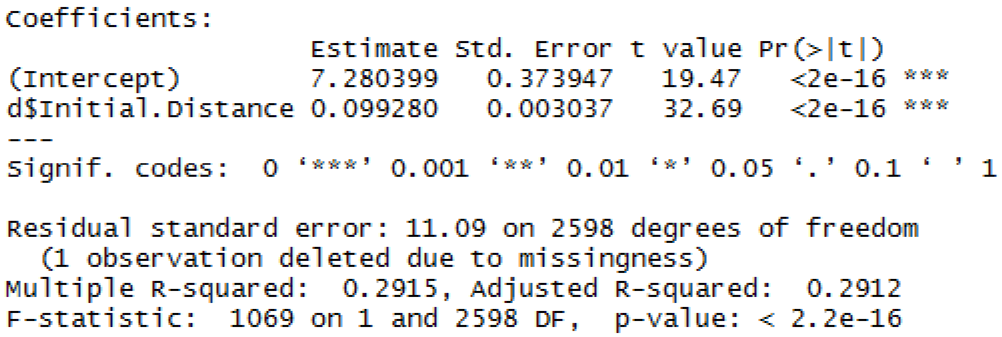


Figure 9.

Linear Regression Output for Distance Hit on Approach Shots Data



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        7.280399   0.373947   19.47   <2e-16 ***
d$Initial.Distance 0.099280   0.003037   32.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.09 on 2598 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2915, Adjusted R-squared:  0.2912 
F-statistic:  1069 on 1 and 2598 DF,  p-value: < 2.2e-16
```