Generalized Linear Mixed Models for Putting Performance on the PGA TOUR

Scott Manski

Advisor
Dr. Eric Nordmoe
Mathematics Department Kalamazoo College
Kalamazoo, MI

"A paper submitted in partial fulfillment of the requirements of the degree of Bachelor of Arts at Kalamazoo College"

January 23, 2015

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Dr. Eric Nordmoe for his assistance in this project. Dr. Nordmoe spent many hours explaining the statistical concepts used in this analysis in addition to working with me through the R packages and commands. This project could not have been completed without his assistance. I would also like to thank the Heyl Foundation for funding this project. Finally, I would like to thank Dr. Paul Roback for providing the R scripts he used in his research as a reference for commands.

Generalized Linear Mixed Models for Putting Performance on the PGA TOUR

Scott Manski

Abstract

In golf, putting is considered to be the single most important aspect of the game. With approximately 40 percent of total strokes having been putts in 2013 on the PGA TOUR, it is obvious that putting performance is a significant aspect of a player's overall performance. Analysts are always trying to find new ways to measure a player's putting performance on the golf course. For many years, measuring putting performance has been restricted to simply measuring putts per round or putts per green. In this paper, we concentrate on quantifying the various factors that contribute to putting performance.

While the probability of making a putt as a function of distance has been modeled in past research, a mixed effects model has not yet been used to incorporate additional variables representing random effects such as individual or season-to-season variability. A series of generalized linear models and generalized linear mixed models using the logit link function were used in this analysis. The response variable used was whether or not the putt finished in the hole, making the logit link function ideal for this analysis. The BIC and McFadden's R^2 model selection criteria were used for model comparison. The best model consisted of fixed effects; *Distance* and *Putt.For*, and random effects; *Year* within *Player* and *Hole* within the *Course*. An analysis of the best linear unbiased predictors (BLUPs) also provides insight into the conditions for which putting performance is at its best.

Additionally, a generalized linear mixed model was fit for the 2014 season and the BLUPs were used as a ranking system for putting performance. The results were compared to the rankings provided by strokes gained putting and the ranking systems showed moderate consistency.

KEYWORDS: PGA TOUR, putting performance, lme4, generalized linear mixed models

1. INTRODUCTION

"Half of golf is fun; the other half is putting."-Peter Dobereiner. In golf, putting is considered to be the single most important aspect of the game. A putt is a shot that originates on the green, the closely mown area around the hole. With approximately 40 percent of total strokes having been putts in 2013 on the PGA TOUR, it is obvious that putting performance is a significant aspect of a player's overall performance. Analysts are always trying to find new ways to measure a player's putting performance on the golf course. For many years, measuring putting performance has been restricted to simply measuring putts per round or putts per green. In this paper, we concentrate on the probability of making a putt as a measure of a player's putting performance.

Traditionally, a player's putting performance has been measured by the number of putts per round or by the number of putts per green. However, these statistics fail to take into account the distance of the putts. Theoretically, a player that misses more greens is more likely to have shorter putts, which will reduce this player's putts per green. Originally introduced in 2008 and later revised, Professor Mark Broadie of Columbia University developed a measure of putting performance called "strokes gained" (Broadie 2012). Broadie's idea transformed into a very highly remarked golf statistic, strokes gained putting. In short, strokes gained putting is a measure of putting performance that not only accounts for the distance of putts, but also for putting performance of other competitors. Strokes gained putting has truly revolutionized the way golf analysts and players think about putting performance on the PGA TOUR.

Broadie was able to develop strokes gained putting using individual shot statistics from the ShotLINK data set. Begun in 2003, ShotLINK is a program sponsored by CDW to capture a wide variety of statistics for each event, round, hole and stroke taken by every player in the world's most prestigious professional golf tours, including the PGA TOUR, the Champions TOUR and the WEB.com TOUR. For the PGA TOUR, ShotLINK uses approximately 10,000 volunteers annually to record such data with the use of laser technology, digital imaging and on-course scorers. ShotLINK has revolutionized data collection among

the world's top tours. The vision of the ShotLINK system is to "Turn data into information, information into knowledge, and knowledge into entertainment" (PGA TOUR, Inc 2014).

Putting performance at the individual putt level can be thought of as a binary or success-failure response variable. Intuitively, we consider a putt that finishes in the hole as a success, and any putt that does not finish in the hole as a failure. To model the probability of a success, a generalized linear model is used with the primary independent or predictor variable being the distance to the hole. A generalized linear model is an extension of the linear model, and it allows for more general forms of the distribution of the response. An extension of the generalized linear model is the generalized linear mixed model, a model that combines both fixed and random effects. This allows us to explicitly incorporate systematic variability in putting performance across golfers, courses, etc. Detailed descriptions of the models considered are provided in Section 3.

Recently, generalized linear models and generalized linear mixed models have been widely used in sports statistics. In baseball, Albert (2006) used a binomial random-effects model as a measure of a pitcher's performance. In soccer, McHale and Szczepański (2014) used two mixed effects models for identifying goal scoring ability of players in the European Football League. Also in soccer, Groll and Abedieh (2013) used a generalized linear mixed model to incorporate team-specific random effects for predicting the outcome of the European football championship 2012. In college basketball, Noecker and Roback (2012) used a generalized linear mixed model to calculate the probability of the next foul being on the home team to provide insight into how officials even out foul calls in NCAA basketball.

Generalized linear models have also been used in golf science. Pope and Schweitzer (2011) used a generalized linear model to predict the probability of making a putt as a function of distance and a series of dummy variables for eagle, birdie, par, bogey, and double bogey situations. Also in golf, Fearing, Acimovic and Graves (2011) were able to fit the probability of making a putt as a function of distance using a generalized linear model. They also modeled the expected remaining distance to the hole after a putt as a function of distance using a gamma regression. From these models, they were able to use a Markov model to

model the expected number of putts remaining to hole out as a function of distance. While the probability of making a putt as a function of distance has been modeled in past research, a mixed effects model has not yet been used to incorporate additional variables representing random effects such as individual or season-to-season variability.

While the length of a putt is a very important variable when it comes to the probability of successfully making the putt, there are many other variables that can affect the probability of success. Some of these other variables could include the player, course, year, time of year, weather and significance of the putt. It is clear that there are many other variables that can affect the outcome of a putt. In this paper, we attempt to model the probability of a putt being made as a function of distance, and we attempt to strengthen this model by incorporating several of these additional variables. We will refer to these variables as random effects variables. A variable is considered to be a random effect when the levels of the variables are a sample of some conceptual larger population of effects. For example, the player is a random effect because the players in our data set do not make up the entire population of PGA TOUR players for which the model is representing. In a simple model that only accounts for distance, we are ignoring any effect of these random effects variables. This type of model is a generalized linear model. By incorporating these other variables, we are able to determine any variation among players, courses, years, etc. Models that include random effects are mixed models, and in this analysis, these models are generalized linear mixed models.

In this project, R (R Core Team 2014), a programming language for statistical computing and graphics, in combination with RStudio was used for statistical analysis. RStudio is a widely used open source integrated development environment for R. Inside of R, the `glm2` (Marschner 2014) package and the `lme4` (Bates, Maechler, Bolker and Walker 2014) package were used for our statistical models. These packages contain functions for fitting generalized linear models and generalized linear mixed models.

2. DATA

For this project, we are interested in statistics at the individual stroke level. ShotLINK provides detailed data for every stroke taken on the PGA TOUR by every player since the 2003 season. Because we are interested in the putting performance, the data set was filtered by *From Location (Scorer)*. According to the ShotLINK documentation, *From Location (Scorer)* is defined as the "general location from which the shot began as recorded by the walking scorer." Therefore, the data set was filtered to include only presumed putts, strokes whose *From Location (Scorer)* was *Green*. This also means that any stroke taken with a putter that did not start from the green was not included in the analysis.

The other variables included in the data set were *Distance to Pin*, *In the Hole Flag*, *Distance to the Hole After the Shot*, *Player Number*, *Player First Name*, *Player Last Name*, *Tournament Year*, *Permanent Tournament #*, *Course Name*, *Round* and *Hole Number*. The definitions of these variables are provided by ShotLINK and are shown in Appendix A.

An additional variable, *Putt.For* was also added to the data set. *Putt.For* is the score relative to par for that particular putt, calculated by the difference between *Hole Score* and *Par Value*. For example, if a player is putting for a three on a par four, *Putt.For* is *Birdie*. If a player is putting for six on a par five, *Putt.For* is *Bogey*. *Putt.For* contains 5 levels; *Eagle or Better*, *Birdie*, *Par*, *Bogey* and *Double Bogey or Worse*.

In addition, several other filters were included to ensure that there would be enough information to assess the importance of individual variability in building the model of putting performance. Putts occurring in the 5th round (or playoff round) of a tournament were excluded from the data set due to a limited number of occurrences. Likewise, putts taken from over 125 feet were also removed. Every course included in the data set must have been played in more than one year. For a player to be included in the data set, the player must have played in the fourth round of at least one tournament, the player must have played in more than one year and the player must have hit at least 1000 putts, all among the 13 years of collected statistics. The final data set consists of a total of 4,837,287 putts divided among 545 players at 66 courses since 2003.

2.1 Exploratory Data Analysis

Table 1 shows the number of players, courses, tournaments, rounds and putts for each of the 13 years of collected data. The number of players from each year ranged from 293 to 337 players, for a total of 545 different players for the 13 years. Likewise, the number of courses played in each year ranged from 32 to 42, for a total of 66 different courses played for the 13 years.

Table 1: Summary table for the number of players, courses, tournaments, rounds and putts for each year in the data set.

Year	Players	Courses	Tournaments	Rounds	Putts
2003	301	39	39	14,054	386,928
2004	319	41	41	15,081	434,028
2005	337	42	42	15,378	443,887
2006	331	41	41	15,337	445,992
2007	325	39	39	14,095	411,013
2008	335	39	39	14,555	425,899
2009	328	38	38	13,788	400,201
2010	322	38	38	13,886	405,962
2011	333	37	37	13,540	394,440
2012	329	35	35	12,806	373,497
2013	312	32	32	11,841	344,254
2014	293	33	33	12,044	350,199
Overall	545	66	456	167,106	4,837,287

Figure 1 shows the fraction of putts made as a function of distance using the spline smoothing function for each of the levels of *Putt.For*. This plot shows the variability in the fraction of putts made for each of the levels of *Putt.For*.

Figure 2 shows the histogram of average putts per round for each of the 545 players in the data set. The mean of the average putts per round is 29.543 putts with a standard deviation of 0.412 putts.

Figure 3 shows the histogram of average putts per round for each of the 66 courses in the data set. The mean of the average putts per round is 29.484 putts with a standard deviation of 0.606 putts.

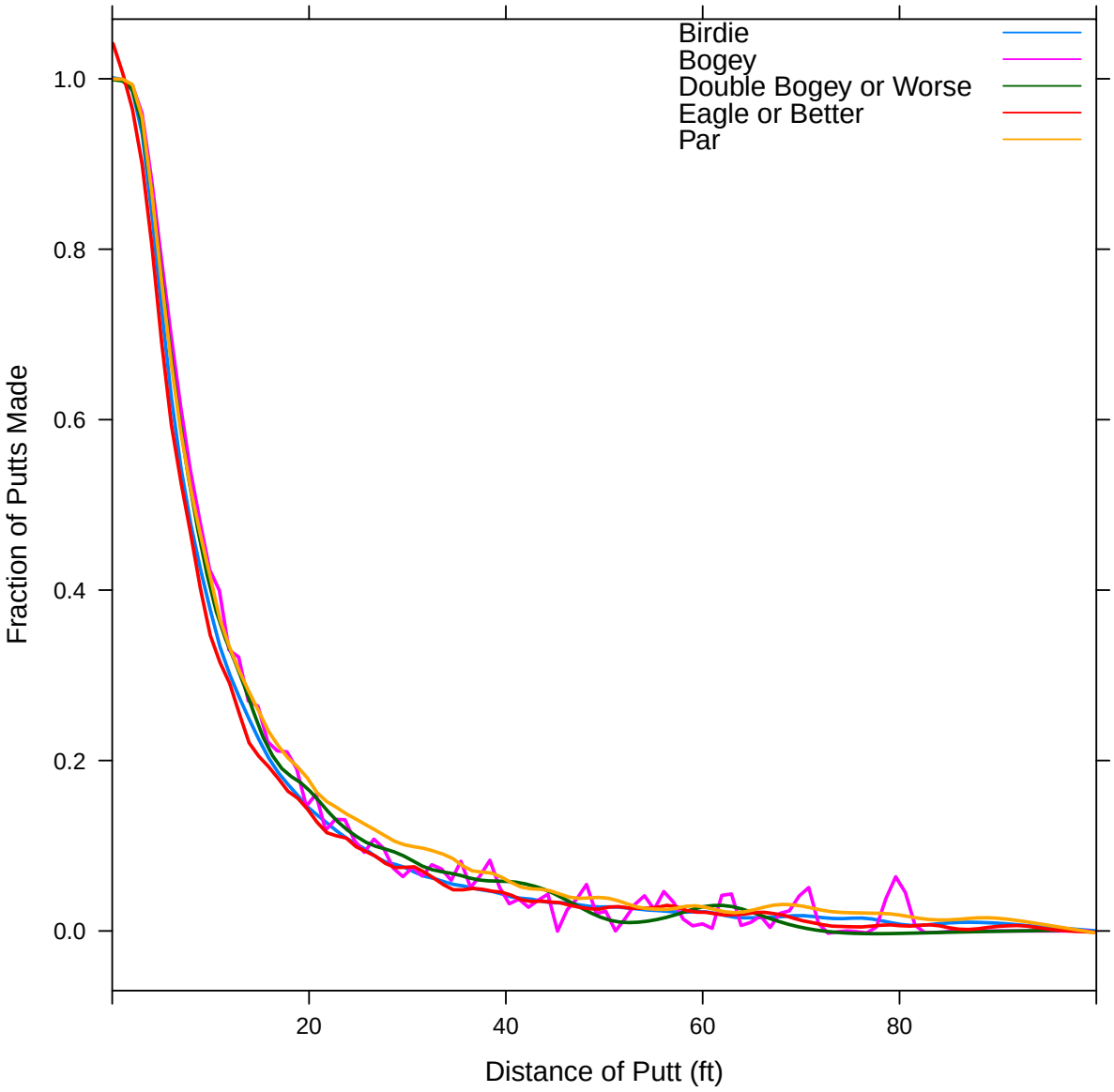

Figure 1: A plot of the fraction of putts made for each level of Putt.For as a function of distance using the spline smoothing function.

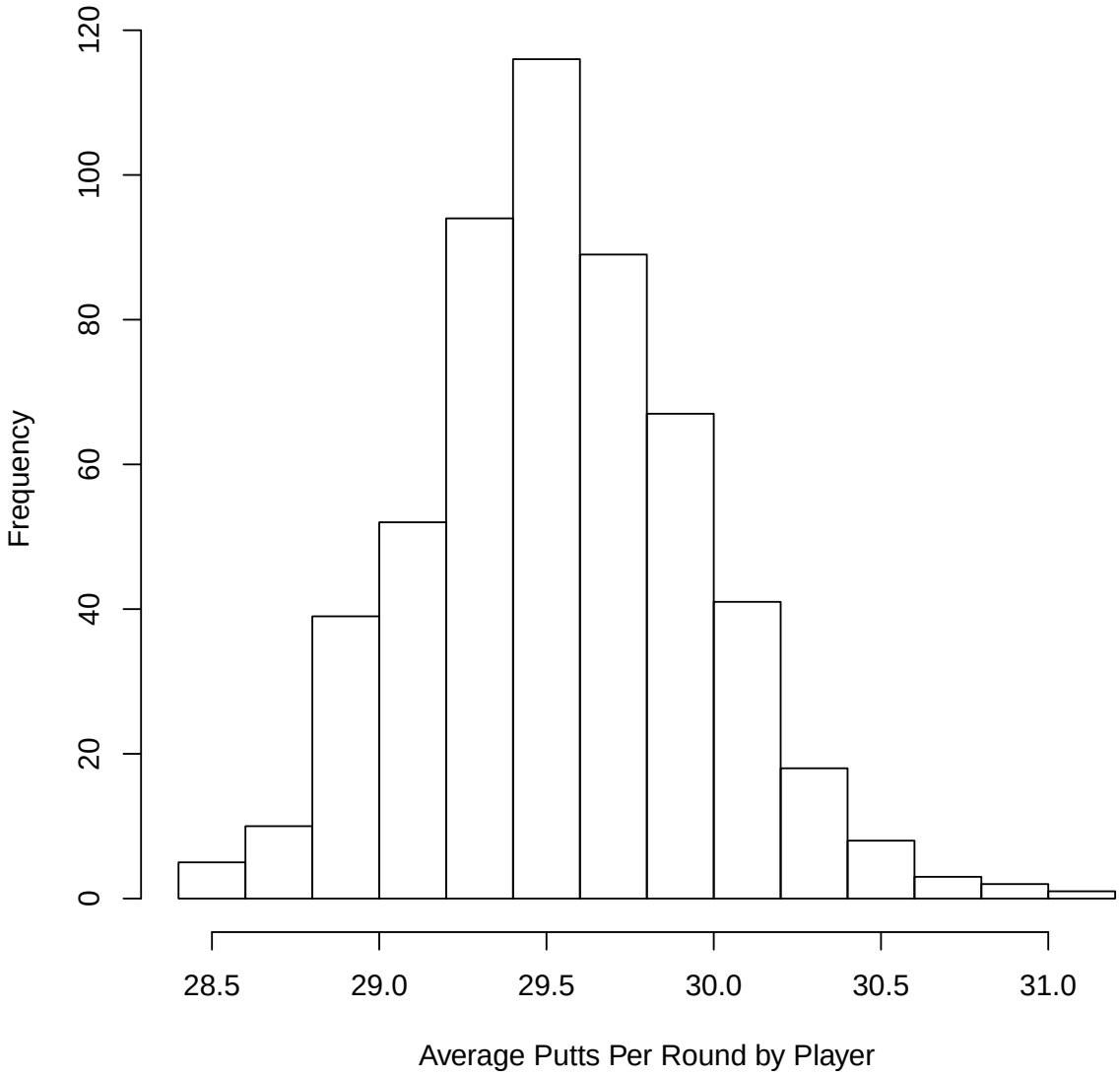

Figure 2: Histogram of the average putts per round for each of the 545 players in the data set.

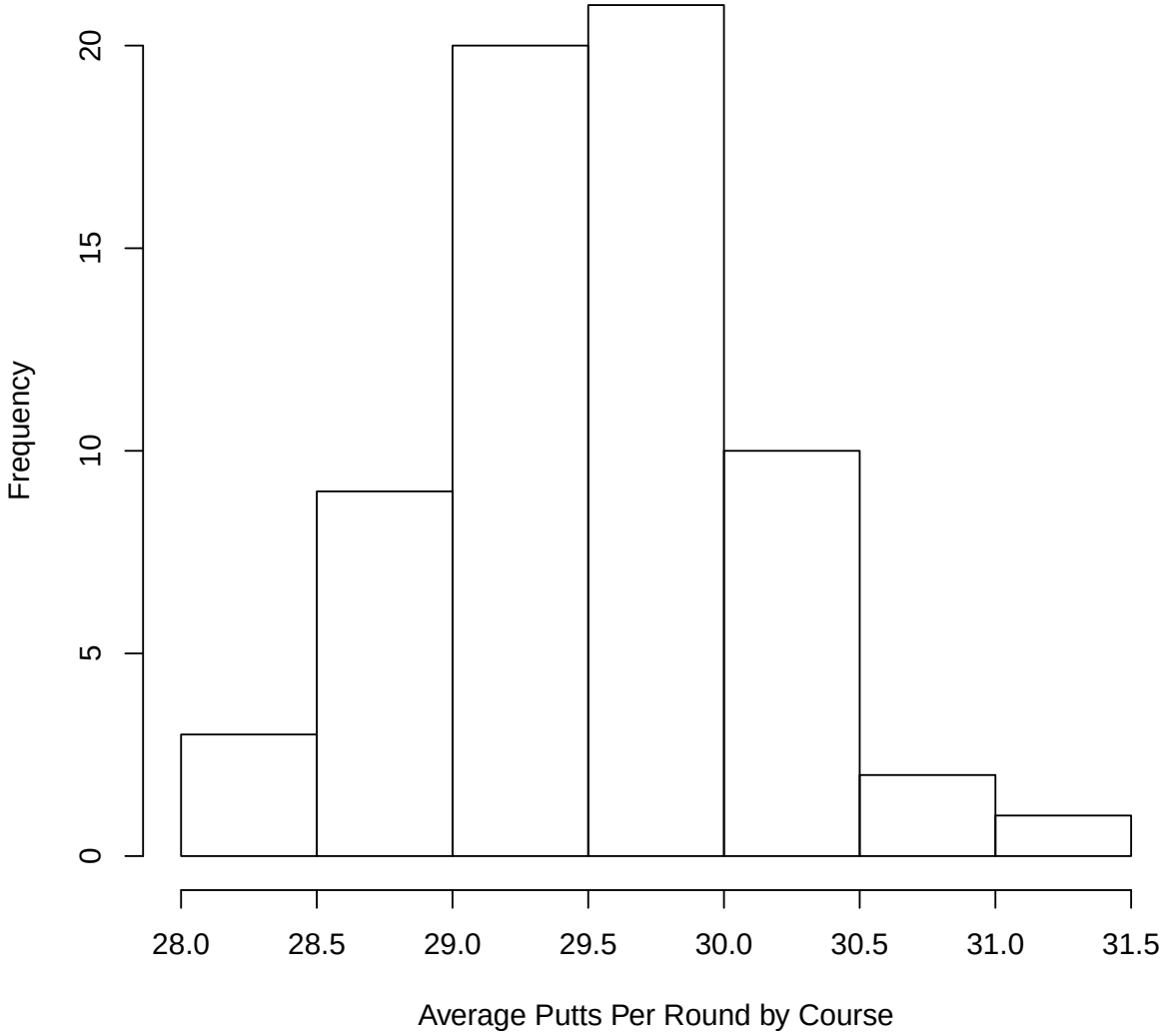

Figure 3: Histogram of the average putts per round for each of the 66 courses in the data set.

2.2 Possible Predictor Variables

In this section, we discuss the possible predictor variables that will be considered in the models to come. These variables contain information that ranges from the year the putt was taken, to the hole the putt was taken on, all the way down to the specific stroke.

Distance

Distance is the main fixed effects variable that is considered in this analysis. It is obvious that the probability of making a putt will be most influenced by the length of the putt.

Putt For

The score relative to par which the putt is for may also show variation in the probability of making a putt. While each putt carries the same weight, putts for par may have different psychological implications than putts for birdie or putts for bogey. As a result, *Putt.For* is considered as a fixed effects variable.

Round Number

A tournament in golf is played over a four day period. As each day progresses, the amount of pressure felt by the leading performers continues to rise. As a result, a variation in probability of making a putt may be experienced as the tournament progresses. This variability is considered as a fixed effects variable for *Round Number*.

Player

The player is the most obvious random effects variable to be considered in predicting the probability of making a putt. Like any other sport, some players are better in a certain aspect of the game than others. There are very good putters on the PGA TOUR and there are some players that struggle with putting. Estimating variability associated with the random effect *Player* allows us to account for the importance of individual differences in putting skill when predicting the probability of a successful putt.

Year Nested Within Player

While there are differences in putting performance from player to player, it is also likely that individual players may show variability in putting performance from year to year. A random effects term for year nested within player allow our models to account for variability

in individual players' putting performances from year to year.

Course

Unlike any other sport, the playing field is inconsistent for the game of golf. Each year, the PGA TOUR plays on roughly 50 different courses, with presumably wide variations in the difficulty of their greens, presenting varying levels of challenge to the putting abilities of the competitors. As a result, a random effects term for *Course* is also considered due to the random variability of the playing conditions.

Hole Nested Within Course

We can take the variability in the course one step further by considering the variability of each individual hole within the course. There are 18 different holes at each of the courses played by the PGA TOUR. A random effects term for hole nested within course allows our models to account for variability between the 18 different holes at each course.

3. METHODS

3.1 Generalized Linear Models

A linear model is commonly written as,

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

and has one random effect, the error term ϵ_i. The parameters of the model are the regression coefficients, $\beta_1, \beta_2, \ldots, \beta_p$, and the error variance, σ^2. In most cases, $x_{1i} = 1$, making β_1 the intercept (Fox 2002).

An extension of the linear model is the generalized linear model. *Generalized linear models* are models that incorporate coefficients in a linear predictor but allow for more general forms of the distribution of the response. The linear predictor, η, determines the conditional mean of the response, μ, according to a *link function*, g. The probability model for a binary response is the Bernoulli distribution, a distribution with only two possible

values: 0 and 1. For a Bernoulli distribution, it is easy to see that the expected value μ must satisfy $0 \leq \mu \leq 1$. We do not want to have restrictions on the values of the linear predictor so we equate the linear predictor to a function of μ that has an unrestricted range. In the case of the Bernoulli distribution, we equate the linear predictor to the expected response by the *logit* link function. That is,

$$\eta = \text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right).$$

3.2 Generalized Linear Mixed Models

When the linear predictor incorporates random effects in addition to fixed-effects parameters, we call them mixed effects models (Bates 2010). *Fixed effects* are defined as parameters associated with an entire population or with certain repeatable levels of experimental factors while *random effects* are associated with individual experimental units drawn at random from a population (Pinheiro and Bates 2002). In this analysis, the distance of the putt will be a fixed effect while the player will be a random effect.

When a generalized linear models contains both random effects and fixed effects, the resulting model is a generalized linear mixed model. For example, let us model the probability of making a putt using distance of the putt as a fixed effect and the player as a random effect. Equation 1 shows a generalized linear mixed model using the logit link function with *Distance* as a fixed effect and *Player* as a random effect.

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 d_i + \epsilon_{j_i} \tag{1}$$

In Equation 1, p_{ij} is the probability of putt i hit by player j going in, d_i is the distance of putt i, β_0 and β_1 are the regression coefficients, and ϵ_{j_i} is the random effects term for player j hitting putt i. All random effects are assumed to be independently and identically distributed as normal random variables with mean zero and separate variance terms. Therefore, $\epsilon_{j_i} \sim N(0, \sigma_j^2)$.

3.3 Model Selection Criteria

In this project, two model selection criteria are considered. The first model selection criterion used is the Bayesian Information Criterion (BIC). The BIC incorporates the likelihood in addition to a penalty term for the number of parameters in the model. As a result, we can use the BIC as a method to find to best model, the model with maximum likelihood and minimal terms. The BIC is evaluated as,

$$BIC = -2 \log \text{Lik} + n_{par} \log(N),$$

where n_{par} denotes the number of parameters in the model and N denotes the sample size used to fit the model. For the BIC, lower values indicate a better model (Schwarz 1978).

The second model selection criterion used in this paper is McFadden's R^2. McFadden's R^2 is a pseudo R^2 calculated from the log likelihood of the null and full models. In this case, the null model is the intercept model while the full model is the model of interest. McFadden's R^2 is evaluated as,

$$R^2 = 1 - \frac{\log \text{Lik}_{Full}}{\log \text{Lik}_{Null}}.$$

For McFadden's R^2, higher values indicate a better model (Faraway 2006).

3.4 Best Linear Unbiased Predictors

One advantage of using a generalized linear mixed model is the random effects associated with each of the groups, called best linear unbiased predictors (BLUP). While random effects are not technically parameters in mixed models, they do behave in a similar way. Often times, it is advantageous to "estimate" random effects terms. It is convention to *estimate* fixed effects and *predict* random effects. BLUPs are the conditional modes of the random effects evaluated at the conditional estimates of the regression coefficients. BLUPs provide insight into how the individual subjects are related to one another (Pinheiro and Bates 2002).

4. RESULTS

4.1 Simple Models—Generalized Linear Models

In this analysis, the generalized linear models will be our simple models. This is simple due to only fixed effects being considered. Three different generalized linear models were fit to the data. As in Fearing, Acimovic and Graves (2011), each model consisted of a 4th degree polynomial for distance and a log-distance term. GLM1 only takes into account *Distance* as a fixed effects variable while GLM2 incorporated *Putt.For* as a fixed effects variable and GLM3 incorporated *Round Number*. The purpose of this was to determine whether or not *Putt.For* and *Round Number* should be included in the extended models (the generalized linear mixed models). The coefficients, standard errors and model selection criteria are shown in Table 2. McFadden's R^2 and BIC values for the three models indicate that *Putt.For* improved the original model while *Round Number* did not. As a result, *Putt.For* is treated as a fixed effect in the generalized linear mixed models.

Table 2: Model coefficients, standard errors and model selection criteria for models using *Distance*, *Putt.For* and *Round Number* as fixed effects variables.

Variable	GLM1	GLM2	GLM3
(Intercept)	-5.6470(0.0495)	-5.8462(0.0497)	-5.6480(0.0495)
Distance	6.8614(0.0736)	7.0502(0.0740)	6.6813(0.0736)
ln(Distance)	-5.6045(0.0265)	-5.5902(0.0265)	-5.6045(0.0265)
Distance^2	-1.9961(0.0300)	-2.0724(0.0302)	-1.9961(0.0300)
Distance^3	0.2943(0.0059)	0.3073(0.059)	0.2943(0.0059)
Distance^4	-0.0161(0.0004)	-0.0169(0.0004)	-0.0161(0.0004)
Eagle or Better		-0.0667(0.0139)	
Par		0.1798(0.0035)	
Bogey		0.2555(0.0075)	
Dbl Bogey or Worse		0.1276(0.0110)	
Round 2			-0.0188 (0.0038)
Round 3			-0.0074 (0.0043)
Round 4			-0.0096 (0.0043)
Model Selection Criteria			
BIC	2880329	2877274	2880351
R^2	0.5547	0.5552	0.5547

By rearranging the *logit* function, the probability of making a putt as a function of distance can be determined. Equation 2 shows the probability of making a putt i as a

function of length of the putt, d_i.

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \ln(d_i) + \beta_2 d_i + \beta_3 d_i^2 + \beta_4 d_i^3 + \beta_5 d_i^4)}} \tag{2}$$

Using Equation 2, we are able to plot the probability of making a putt as a function of distance with the empirical probability of making a putt. The empirical probability is calculated by assigning each putt to a bucket, where buckets are in 1 inch increments. For example, there were 18,421 putts taken from 10 feet (120 inches) and 7237 of those putts were made. Therefore, the empirical probability of making a putt from 10 feet (120 inches) was calculated to be 0.393. Empirical vs model-based estimates from GLM1 are compared in Figure 4.

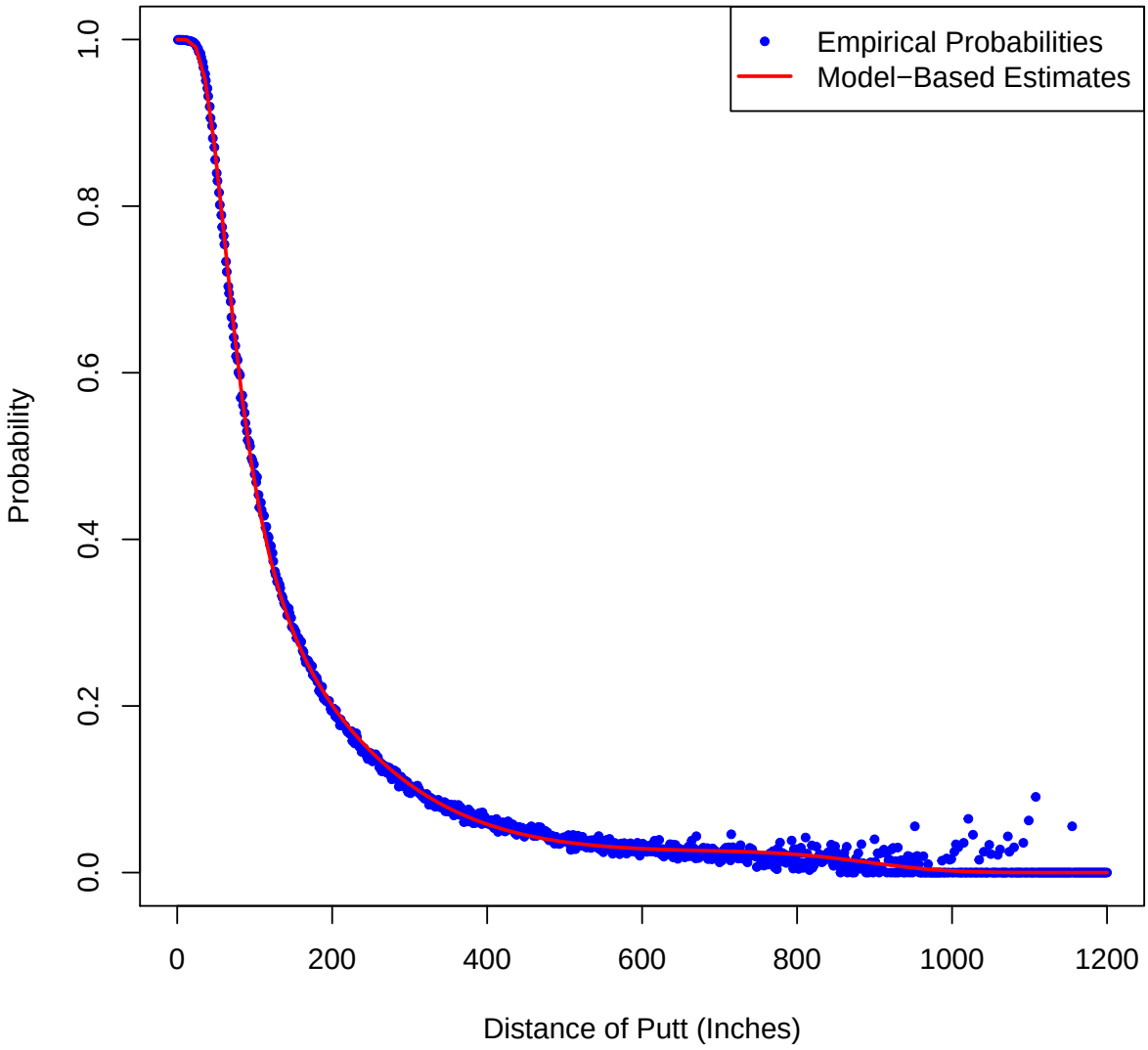

Figure 4: A histogram for the probability of making a putt given the distance from the hole. The blue line represents the empirical probability of making the putt while the red line represents model-based estimates obtained from the generalized linear model.

4.2 Extended Models—Generalized Linear Mixed Models

The extended models are generalized linear mixed models. Each model contains a different random effects term or nested random effects term. Five different generalized linear mixed models were fit, each with one of the following random effects terms: (1) *Player*, (2) *Year* nested in *Player*, (3) *Course*, (4) *Hole* nested in *Course*, and (5) *Year* nested in *Player* and *Hole* nested in *Course*. The model coefficients with standard errors, variance estimates associated with the random effects, and model selection criteria for the five generalized linear mixed models are shown in Table 3. Using BIC values, the best model for predicting the probability of making a putt is GLMM5, a generalized linear mixed model with *Distance* and *Putt.For* as fixed effects, and *Year* nested in *Player* and *Hole* nested in *Course* as random effects. From the R^2 perspective, little is gained from the more complex models.

4.3 BLUPs Analysis

One advantage of random effects modeling is the random effects associated with each of the groups, called best linear unbiased predictors (BLUP). BLUPs provide an insight into how the individual subjects are related to one another. All random effects estimates are from GLMM5. It is important to note, while the BLUPs provide insight into the relative putting performance for players, or the relative difficulty of the courses or holes, the margins of error for these estimates may be relatively large. Figure 5 shows the random effects estimates for each of the players in the data set. Several notable players are identified by name. Figure 6 shows the random effects estimates for the players with the highest estimates in the data set.

Additionally, Figure 7 shows the BLUPs for the 66 courses included in the data set. From the BLUPs, players have a higher probability of making a putt at Doral CC (Blue) than at any other course. On the other hand, players have a lower probability of making a putt at Pebble Beach Golf Links than at any other course.

Table 3: Model coefficients with standard errors, variance estimates associated with the random effects and model selection criteria for the five generalized linear mixed models.

Variable	GLMM1	GLMM2	GLMM3	GLMM4	GLMM5
(Intercept)	-5.8531(0.0499)	-5.8577(0.0450)	-5.8295(0.0507)	-5.8834(0.0508)	-5.8450(0.0509)
Distance	7.0439(0.0740)	7.0452(0.0740)	7.0273(0.0740)	7.0253(0.0740)	7.0183(0.0741)
ln(Distance)	-5.5919(0.0266)	-5.5945(0.0266)	-5.5825(0.0266)	-5.5813(0.0266)	-5.5843(0.0266)
Distance^2	-2.0696(0.0302)	-2.0698(0.0302)	-2.0653(0.0302)	-2.0635(0.0302)	-2.0604(0.0302)
Distance^3	0.3068(0.0059)	0.3068(0.0059)	0.3063(0.0059)	0.3059(0.0059)	0.3053(0.0059)
Distance^4	-0.0168(0.0004)	-0.0168(0.0004)	-0.0168(0.0004)	-0.0168(0.0004)	-0.0167(0.0004)
Eagle or Better	-0.0603(0.0139)	-0.0621(0.0139)	-0.0689(0.0139)	-0.0925(0.0143)	-0.0872(0.0143)
Par	0.1780(0.0035)	0.1781(0.0035)	0.1845(0.0035)	0.1942(0.0037)	0.1932(0.0037)
Bogey	0.2544(0.0075)	0.2547(0.0075)	0.2629(0.0075)	0.2729(0.0075)	0.2730(0.0076)
Dbl Bogey or Worse	0.1258(0.0110)	0.1275(0.0110)	0.1267(0.0110)	0.1329(0.0111)	0.1321(0.0112)
Variance					
Player	0.0071	0.0061			0.0062
Year:Player		0.0058			0.0051
Course			0.0068	0.0059	0.0054
Hole:Course				0.0049	0.0048
Model Selection Criteria					
BIC	2874999	2874437	2874759	2873911	2871289
R^2	0.5555	0.5556	0.5555	0.5557	0.5561

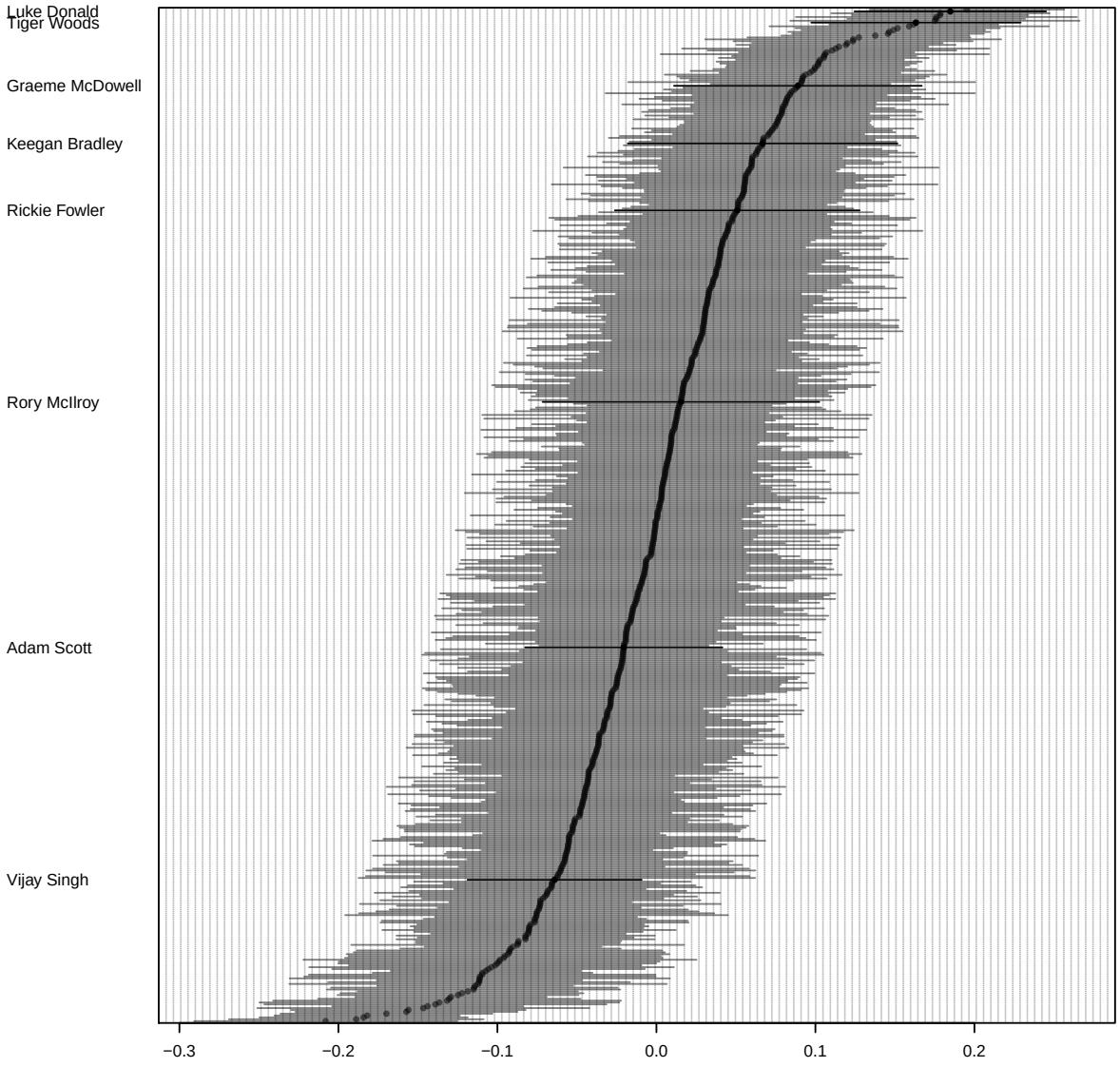

Figure 5: BLUPs for each of the 545 players in the data set. Several notable players are identified by name.

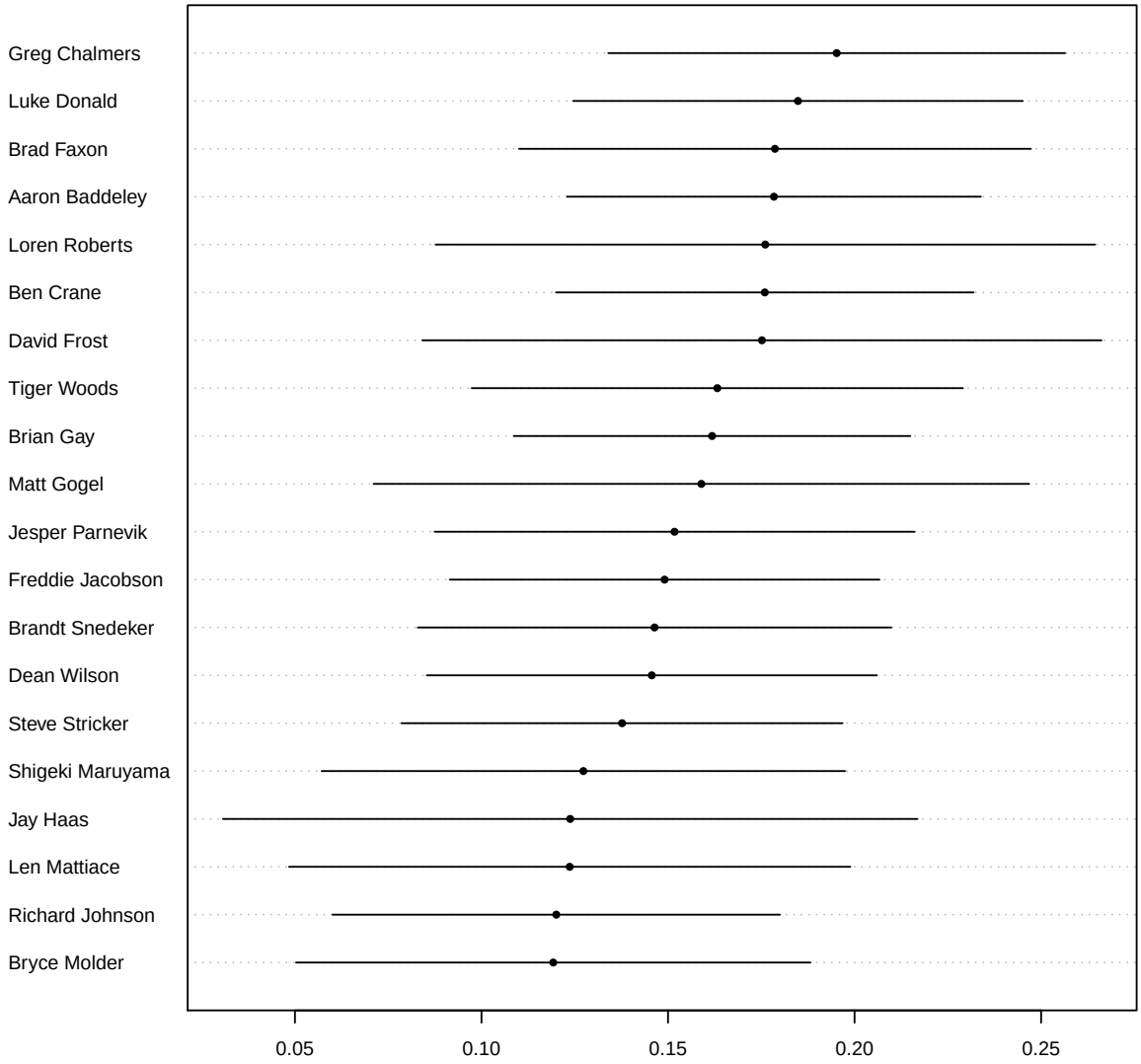

Figure 6: The top 20 players with the highest BLUPs.

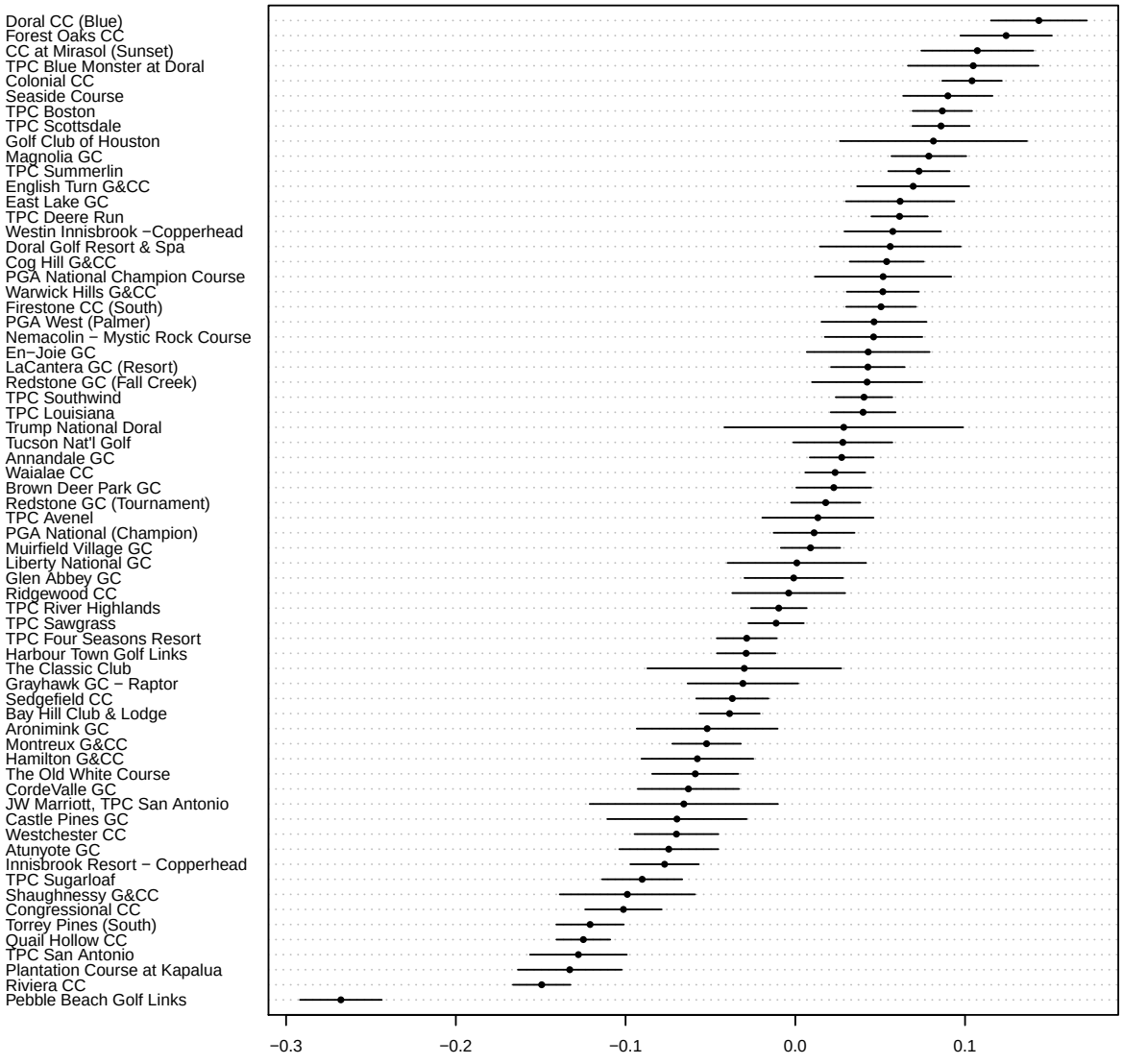

Figure 7: BLUPs for the 66 courses included in the data set.

4.4 The 2014 Season

A generalized linear mixed model containing fixed effects for *Distance* and *Putt.For*, and random effects for *Player* and *Hole* within *Course* was fit for the 2014 season. The BLUPs for *Player* were used as a ranking system for putting performance, and the results were compared to the rank given by average strokes gained putting. A comparison of the two ranking systems is shown in Figure 8. The Spearman rank correlation coefficient is 0.8273, indicating a strong, positive monotonic correlation between the ranking systems.

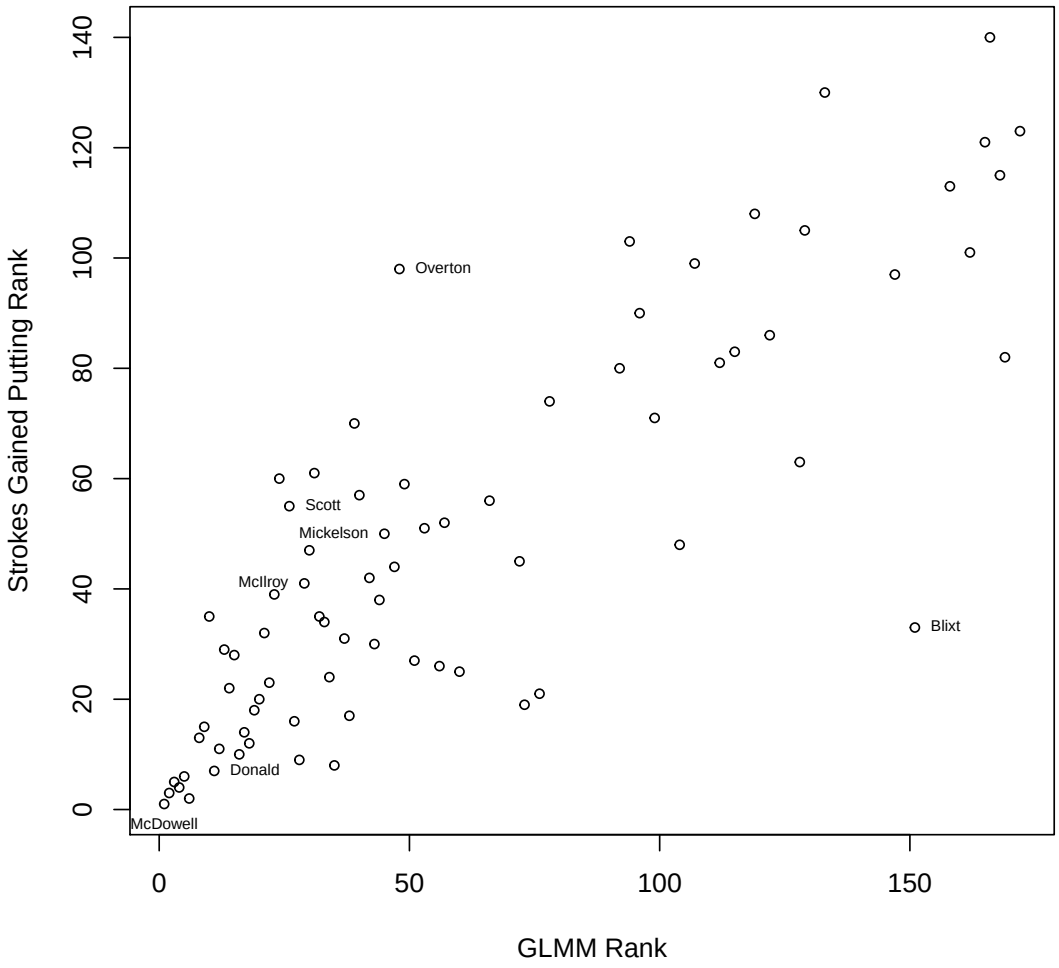

Figure 8: 2014 ranking comparison between the GLMM and strokes gained putting ranking systems. The Spearman rank correlation coefficient is 0.8273. Strokes gained putting provided from the PGA TOUR website (PGA TOUR, Inc 2015).

5. DISCUSSION

The probability of making a putt from a given distance was modeled using a series of generalized linear models and generalized linear mixed models. Each model consisted of a 4th degree polynomial for distance and a log-distance term, the same model used in Fearing, Acimovic and Graves (2011). The results of the three generalized linear models indicate that *Putt.For* is strongly associated with the probability of making a putt while *Round Number* is weakly associated with the probability of making a putt. *Putt.For* being strongly associated with the probability of making a putt is consistent with Pope and Schweitzer (2011). Additionally, a variety of random effects variables were considered, including player, year, course and hole. Each model was compared to the simple generalized linear model using the BIC and McFadden's R^2 model selection criteria. The best model consisted of fixed effects; *Distance* and *Putt.For*, and random effects; *Year* within *Player* and *Hole* within the *Course*.

Additionally, an analysis of the BLUPs also provides insight into the conditions for which putting performance is at its best. The BLUPs for *Player* and *Course* provide some insight into the best putters on tour in addition to the relative difficulty of the 66 courses included in the data set.

Finally, a generalized linear mixed model was used on the 2014 season and the BLUPs were used as a ranking system for putting performance. The results were compared to the rankings provided by strokes gained putting and the results showed moderate consistency. There was no indication from the PGA TOUR website as to the criteria needed for the players to be considered for the strokes gained putting ranking system. Fitting a generalized linear mixed model on the same data set used by the PGA TOUR for their strokes gained putting ranking system may show more consistency in the ranking systems.

Additional analysis may include incorporating additional variables such as weather conditions. Temperature and precipitation can be very influential to a golfer's performance and putting performance. This type of analysis is currently limited by the lack of data on weather conditions of golf tournaments. Generalized linear mixed models could also be utilized in other aspects of golf, including but not limited to, driving performance or short

game performance. A combination of these generalized linear mixed models could then be used to assess a player's overall performance on the golf course.

6. REFLECTION

This project started in the Spring of 2014. At that time, Mitch Wilson, the coach of the Men's Golf team, had informed me of the ShotLINK Prize Competition. I saw the competition as an opportunity for me to learn about some statistical methods in the context of the sport I love. I wrote a research proposal for Dr. Eric Nordmoe, my Senior Individualized Project (SIP) advisor, and for the Heyl Foundation, the foundation that would be sponsoring my summer research. Dr. Nordmoe and I met at the beginning of the summer to brainstorm ideas for the direction of the project.

The initial weeks were spent reading other scholarly articles written for the ShotLINK Prize Competition. Dr. Nordmoe had suggested learning more about generalized linear mixed models, so the literature search expanded to include research utilizing generalized linear mixed models in the context of sports. We thought that it would be interesting to use generalized linear mixed models to model the probability of making a putt in golf. This would mean that we would be using the ShotLINK data set at the individual stroke level.

The comma separated values file obtained from ShotLINK was easily imported into R Studio. While the data set contained many fields about each individual stroke, the data set needed to be filtered and additional variables needed to be added. These filters and additional variables are described in Section 2. While R contains many commands for easily filtering a data set, more complicated scripts were also needed for some of the filters, such as every course included in the data set must have been played in more than one year.

After we obtained the final data set, the generalized linear models and generalized linear mixed models were fit. For our models, we needed to determine which variables would be fixed effects and which variables would be random effects. Additionally, we needed to consider whether or not some of these variables needed to be nested. When we considered all of these variables and combinations of variables in our models, we needed a way to compare the models. We were able to settle on using the BIC and McFadden's R^2 model selection criteria. The final leg of the project was spent comparing the rankings given by the BLUPs for the generalized linear mixed model for the 2014 season with the strokes gained putting

rankings for the 2014 season.

While I was able to learn a lot about some widely used statistical techniques, I feel that I learned more about R and its packages. The majority of my time was spent learning about how to manipulate the data set in R in addition to summarizing the results in a well written paper. This paper was written as an R Sweave, a file written in LaTeX that also incorporates R code. I am thankful for the opportunity to learn about these statistical techniques and R. After this paper has been turned in for the SIP requirement, I plan to continue the project and ultimately submit the paper for publication.

APPENDIX A. SHOTLINK DETAILED DEFINITIONS

The PGA TOUR Shot Detail Export is designed to produce a semi-colon delimited text file suitable for use with most common spreadsheets and databases. The information is presented at the shot level for each tournament and player selected.

Tournament Year (4 digit numeric)
Four digit year of the event

Player Number (5 digit numeric)
A unique 4 or 5 digit number assigned to each player

Permanent Tournament # (3 digit numeric)
The unique 3 digit number assigned to each tournament. This number remains consistent with a tournament from year to year, whereas the tournament schedule number will vary based on the sequence of tournaments played.

Player First Name (text)
The players full first name

Player Last Name (text)
The players full last name

Round (1 digit numeric)
The Round number 1–6. Most PGA TOUR and Nationwide Tour events are 4 round events. Most Champions Tour events are 3 rounds.

Course Name (text)
The full name of the Course on which the shot was played

Hole Number (2 digit numeric)
Hole number, 1–18, on which the shot occurred

Par Value (1 digit numeric)
The Par Value for the hole being played

From Location (Scorer) (text)
General location from which the shot began as recorded by the walking scorer

Distance to Pin (5 digit numeric)
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.

In the Hole Flag (character Y/N)
Value indicating whether or not the shot finished in the hole. Y = yes. N = no.

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.

REFERENCES

Albert, J. (2006), "Pitching statistics, talent and luck, and the best strikeout seasons of all-time," *The Journal of Quantitative Analysis in Sports*, 2(1).

Bates, D. M. (2010), *lme4: Mixed-effects modeling with R*, New York: Springer.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014), *lme4: Linear Mixed-Effects Models Using Eigen and S4*. R package version 1.1-7.
URL: *http://CRAN.R-project.org/package=lme4*

Broadie, M. (2012), "Assessing Golfer Performance on the PGA TOUR," *Interfaces*, 42(2), 146–116.

Faraway, J. J. (2006), *Extending the Linear Model with R*, New York: Taylor & Francis Group.

Fearing, D., Acimovic, J., and Graves, S. (2011), "How to catch a Tiger: Understanding putting performance on the PGA Tour," *Journal of Quantitative Analysis in Sports*, 7(1).

Fox, J. (2002), *An R and S-Plus Companion to Applied Regression*, New York: Sage Publications.

Groll, A., and Abedieh, J. (2013), "Spain retains its title and sets a new record - generalized linear mixed models on European football championships," *Journal of Quantitative Analysis in Sports*, 9(1), 51–66.

Marschner, I. (2014), *glm2: Fitting Generalized Linear Models*. R package version 1.1.2.
URL: *http://CRAN.R-project.org/package=glm2*

McHale, I. G., and Szczepański, Ł. (2014), "A mixed effects model for identifying goal scoring ability of footballers," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(2), 397–417.

Noecker, C. A., and Roback, P. (2012), "New Insights on the Tendency of NCAA Basketball Officials to Even Out Foul Calls," *Journal of Quantitative Analysis in Sports*, 8(3).

PGA TOUR, Inc (2014), *What is ShotLink?*
URL: *http://www.shotlink.com*

PGA TOUR, Inc (2015), *Strokes Gained: Putting (2014).*
URL: *http://www.pgatour.com/stats/stat.02564.2014.html*

Pinheiro, J. C., and Bates, D. M. (2002), *Mixed-Effects Models in S and S-PLUS*, New York: Springer.

Pope, D. G., and Schweitzer, M. E. (2011), "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes," *American Economic Review*, 101(1), 129–157.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. R version 3.1.2.
URL: *http://www.R-project.org/*

Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6(2), 461–464.