ORDINAL LOGISTIC REGRESSION MODEL FOR THE MASTERS GOLF TOURNAMENT

ERIK L. HEINY

# Utah Valley University

Cody Frisby

# Utah Valley University

Address correspondence to: Erik L. Heiny, Department of Mathematics, Utah Valley University, Orem, UT 84058, or email [erik.heiny@uvu.edu](mailto:erik.heiny@uvu.edu)

Introduction

It is April of 2016 and Jordan Spieth, the defending champion and second ranked player in the world, is walking towards the 10th tee in the final round of the Masters. Spieth has moved to seven under par coming off four straight birdies to close out his front nine, and has a five shot lead. His pursuers include Jason Day, the number one ranked player in the world, but Day is seven shots back at even par and beginning play on number 11. Paired with Day is Dustin Johnson who is six shots back at one under par. On the 12th tee are two Europeans, Lee Westwood and Danny Willett. Westwood has been a European Ryder Cup stalwart throughout his career playing on nine teams. He is largely considered one of the best players to never win a major, but at even par and seven shots back, his chances seem remote. Willett, relatively unknown in America despite being ranked number 12 in the world, is Spieth’s closest pursuer at two under par and five behind.

An old saying at the Masters is “The tournament doesn’t start until the 10th tee on Sunday”. There are eagle and birdie opportunities on the back nine, but also water hazards and potential big numbers. In fact, the cut at the Masters is to the low 50 and ties, but due to how quickly players can make up or lose ground at Augusta National, anyone within 10 shots of the lead after two rounds also makes the cut (even if outside the top 50). But we are talking about Jordan Spieth here. He knows how to win at Augusta, he won the Masters and U.S. Open the year before, and was in contention at the British Open, the third leg of a potential grand slam, all the way to the 72nd hole. With a five shot lead and only nine holes to go, what is the probability that Spieth doesn’t win? If Spieth doesn’t win, who would most likely catch him? Would it be Willett, his closest pursuer, or maybe Day, the number one player in the world?

To simulate the Masters, it might be tempting to predict round-by-round scores for each player. If the object is simply to predict who will win prior to the start of the tournament, this approach may indeed work quite well. However, it cannot answer the question posed above. In a 2010 article, McHale assessed the fairness of the golf handicapping system in the UK by using ordered logistic regression to model hole-by-hole scores as a function of covariates. He then used this model to simulate scores for players of different handicaps and estimate probabilities of winning hole-by-hole matches for each player. This study will follow McHale’s lead and use ordinal logistic regression to produce probabilities of eagle, birdie, par, bogey, double bogey, and triple bogey, for each player, on each hole, in each round, for every year. Simulations of the Masters tournament can then be conducted to answer several interesting questions. How well does this model forecast the Masters? What are the best predictors of Masters success? Is it driving or putting, or just world ranking? How has Tiger’s probability of winning changed over the years? In his prime, a common question was should you take Tiger or the rest of the field? Was Tiger’s probability of winning really high enough to make this a reasonable question? Leading into the 2017 Masters, Dustin Johnson was playing the best golf of his life. He had won his last four starts and was the number one ranked player in the world. Then he had an accident the day before the Masters, injured his back and had to withdraw. What were his chances of winning? And finally, what were Spieth’s chances of winning back in 2016? Ordinal logistic regression and simulation will help us answer all these questions.

Methods

Hole-by-hole scores were available for each Masters participant from 1983 to 2017 using the PGA Tour’s ShotLinkTM database. The key fields in this data set were player, year, round, hole, par, and score relative to par or RTP score. The other covariates used in the model were strokes gained (SG), which are multiple measures of player skill developed by the PGA Tour, the official world golf rankings (OWGR), and the number of Masters appearances.

The object of our model is to develop probabilities of different scores for each player on each hole they play. The possible scores on any hole are eagle or better (2 shots less than par or RTP = -2), birdie (1 shot better than par or RTP = -1), par (RTP = 0), bogey (1 shot more than par or RTP = 1), double bogey (2 shots more than par or RTP = 2), and triple bogey or worse (3 shots more than par or RTP = 3). Since we want to estimate probabilities for an ordinal response, score, an ordered logistic regression model is appropriate. Ordinal logistic regression takes advantage of the cumulative nature of the response by using cumulative logits. A common type of cumulative logit, referred to as a proportional odds model (POM), is given below as defined in *Categorical Data Analysis* by Agresti:

(1)

Notice that *cumulative* probabilities are being modeled here. *Y* represents the player’s score on a hole, *j* goes from 1 to *J*-1 (*J* = 6 in this study, the no. of possible scores on a hole), is the intercept for the *j*th cumulative logit, is a vector of coefficients for the covariates or fixed effects, and is the vector containing values for the covariates which are year, round, hole, no. of Masters, strokes gained, and official world golf ranking. These cumulative logits can be easily converted into cumulative probabilities, and then into probabilities for each possible score as demonstrated below.

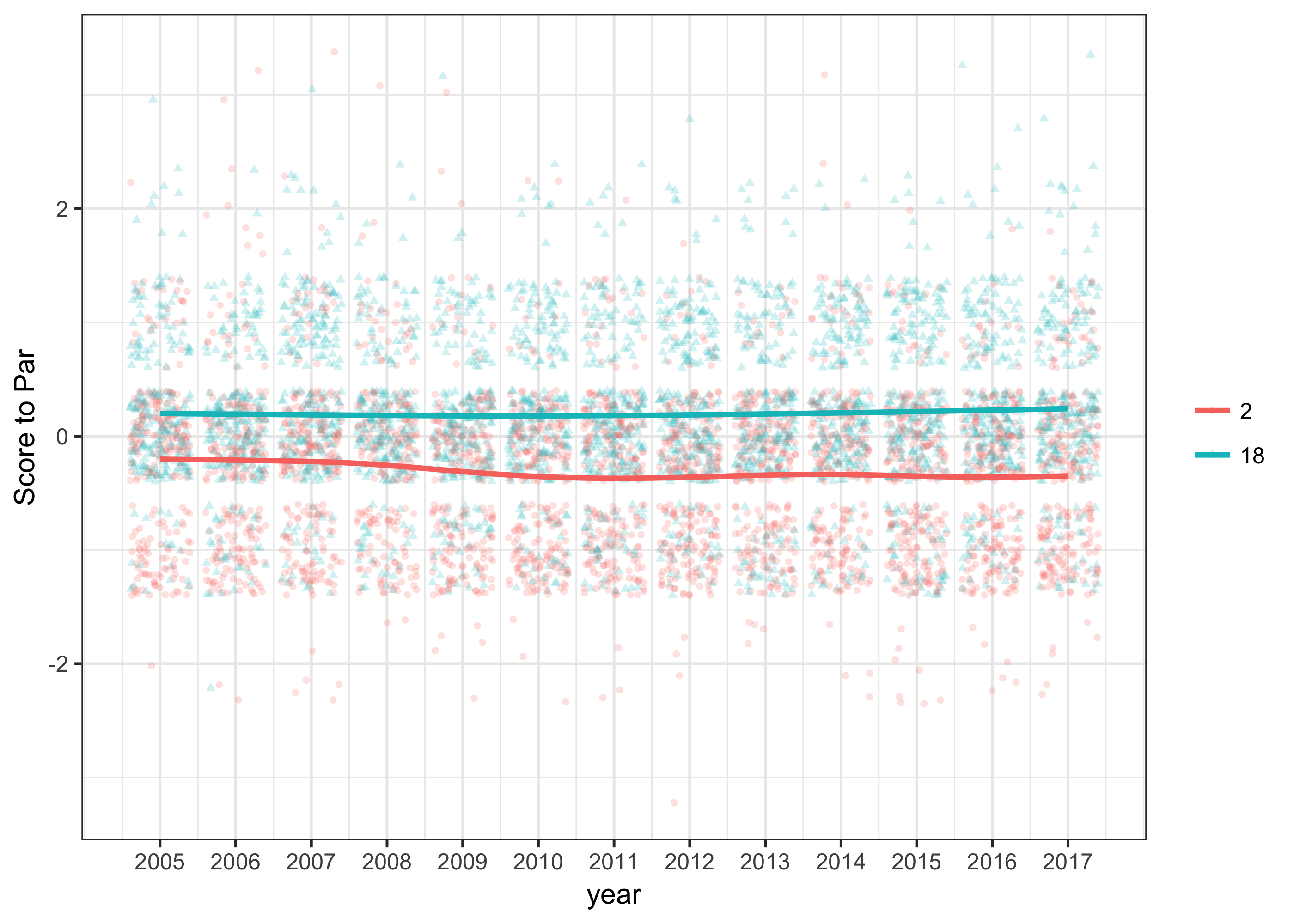
(2)

(3)

(4)

(5)

For this study, it is important that the covariates in are good predictors of score. Let’s take a closer look at the covariates being used in this model. The actual hole being played has a big impact on score. All golfers know that different holes on a course will vary in difficulty. Figure 1 below is a plot of scores by year for hole number 2, a par five where the green is reachable in two shots, and hole number 18, a long uphill par four. It is clearly seen that hole 2 will tend to produce lower scores relative to par than hole 18. To account for the varying difficulty of each hole at August National, hole is included as a categorical variable in the model.

Figure 1. Scoring on Holes 2 and 18 by Year. 

Year and round were also included as categorical variables. The course will vary in difficulty from one year to the next, and from one round to the next. These differences are most likely due to weather conditions, firmness of the greens, pin locations, etc. In the 2017 Masters, the first two rounds had very high winds, and then the weekend had almost perfect weather. A 2013 article by Balsdon gives some evidence that PGA Tour players will change risk strategies and change playing styles when near the cut line, which would affect round two scores. Regardless of the reasons, including year and round helps account for a significant portion of the variability in players’ scores, and gives us a more precise estimate of how players’ specific skills such as strokes gained, world ranking, and number of Masters appearances, affect score.

The strokes gained concept, initially developed by Professor Mark Broadie of Columbia University, measures PGA Tour players’ skills in different categories relative to the average player on Tour. The different skill categories, as defined on the PGA Tour website (<http://www.pgatour.com/news/2016/05/31/strokes-gained-defined.html>), are listed below:

* Strokes Gained Off-the-Tee (SG Tee): measures player performance off the tee on all par fours and par fives.
* Strokes Gained Approach-the-Green (SG Approach): measures player performance on approach shots. Approach shots include all shots that are **not** from the tee on par four and par five holes and are **not** included in strokes gained around-the-green and strokes gained putting. Approach shots include tee shots on par threes.
* Strokes Gained Around-the-Green (SG Around): measures player performance on any shotwithin 30 yards of the edge of the green. This statistic does **not** include any shots taken on the putting green.
* Strokes Gained Putting (SG Putt): measures how many strokes a player gains (or loses) on the greens.

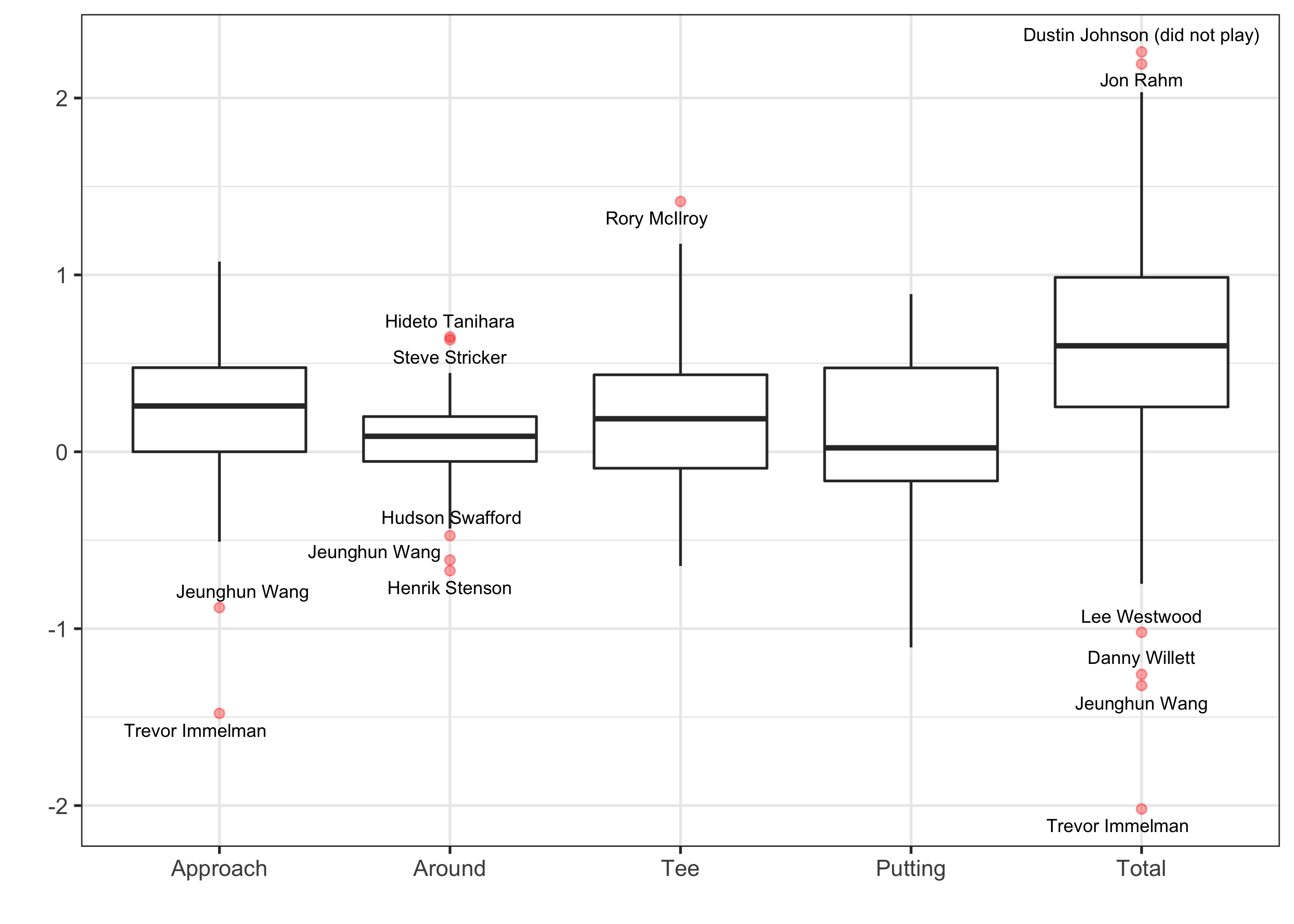
Note: When fitting the model, SG Tee was set equal to zero on all par threes because it only measures skill off the tee on par four and par five holes.

To illustrate how these statistics work, consider Sergio Garcia, the 2017 Masters champion. He had a SG Tee value of 0.737 for the 2016 PGA Tour season. That means that over the course of one round, Sergio was 0.737 shots better than the average PGA Tour player with his tee shots on par four and par five holes. For a full tournament with four rounds, Sergio would pick up about three shots off the tee, relative to an average player. In the other strokes gained categories, Sergio’s values were SG Approach = 0.482, SG Around = -0.031, and SG Putting = -0.388. The strokes gained statistics are additive, so Sergio’s SG total = 0.737 + 0.482 – 0.031 – 0.388 = 0.801. That means that overall Sergio was 0.801 strokes better per round than the average player on the PGA Tour.

It should be noted that strokes gained are only computed for selected rounds in PGA Tour events. For Sergio in 2016, his statistics were based on 31 measured rounds. Additionally, the strokes gained statistics in this study were adjusted so that they represented the prior 12 months of play leading into the Masters. Since strokes gained data is only available going back to 2004, the Masters tournaments from 2005 to 2017 were included in this study. For Sergio leading into the 2017 Masters, his strokes gained values over the prior 12 months were adjusted to: SG Tee = 0.957, SG Approach = 0.085, SG Around = 0.148, SG Putt = -0.373, and SG total = 0.817. These values were based on 30 measured rounds.

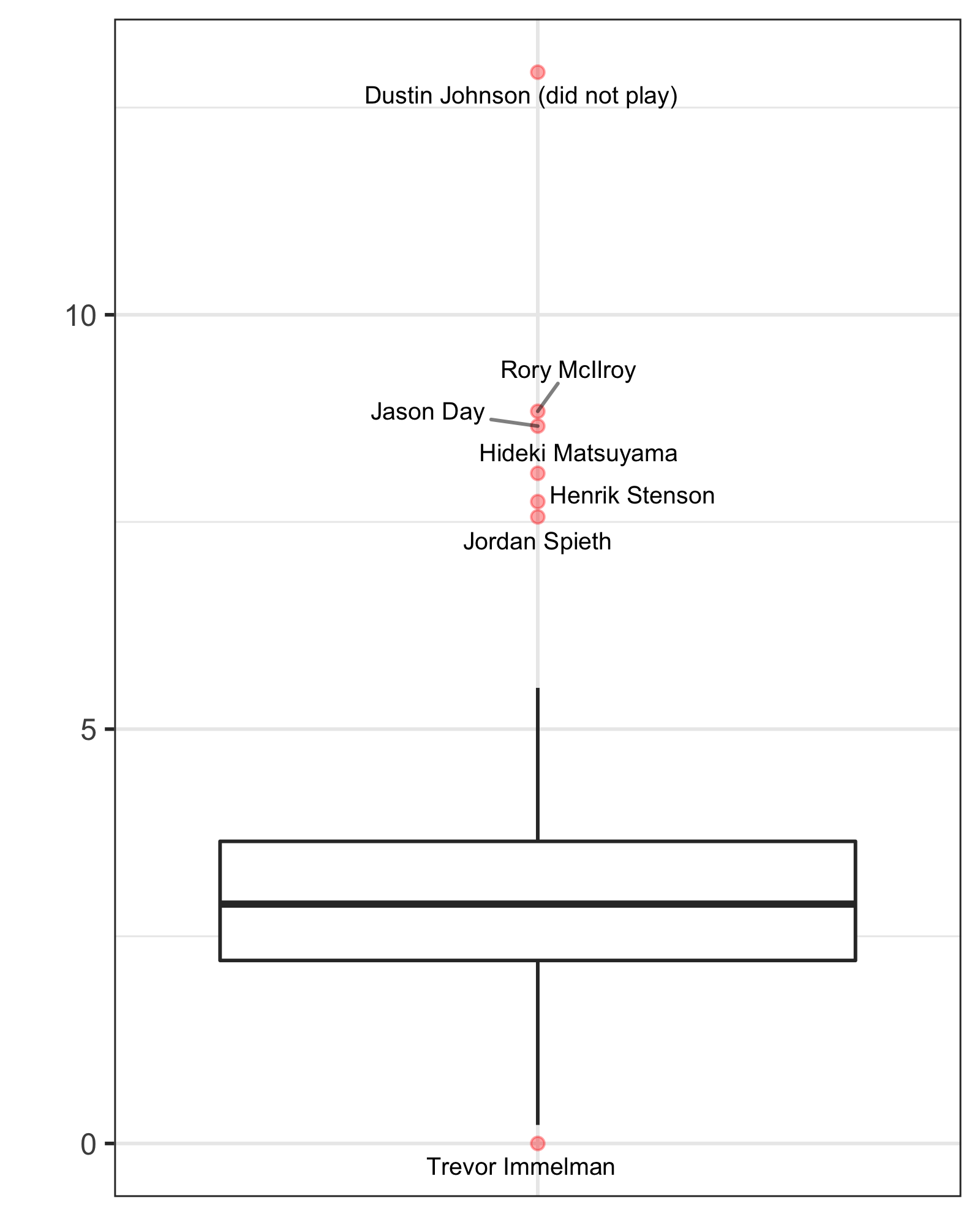
Some players were removed from the data set due to missing or limited strokes gained data. The Masters field always includes several older former champions as well as younger amateurs that do not play on the PGA Tour at all. To be included in the data set, a player needed to have at least 10 measured rounds in the 12 months prior to the Masters. Some of the former champions removed from the 2017 data set were Fred Couples, Bernhard Langer, Larry Mize, Mark O’Meara, Sandy Lyle, and Ian Woosnam. Couples had a high finish in 2017, and Langer was in the second-to-last group for the final round of 2016 before fading, but generally these players do not have a big impact on the outcome of the tournament.

Figure 2 below contains boxplots for all four strokes gained categories, as well as strokes gained total for 2017. The PGA Tour adjusts strokes gained so that the average value for each category is zero, but the center tends to be above zero for the Masters participants, most notably strokes gained total. This is not surprising considering the Masters has a limited field and not all PGA Tour players qualify for the tournament. Something else to note from Figure 2 is how close the skill levels are between these players. Just looking at SG total, the middle 50% are within 0.75 shots of each other per round. Finally, Figure 2 contains the SG values for Dustin Johnson who did not play in the 2017 Masters due to a back injury suffered the day before the tournament began. Nevertheless, being the number one player in the world and winner of his four prior starts leading into the Masters, he is included as a player of interest.

Figure 2. Strokes Gained Data for 2017 Masters. 

The Masters field has always had a significant international presence over the years. Some of the foreign-born champions include Gary Player, Seve Ballesteros, Nick Faldo, Bernhard Langer, Sandy Lyle, Jose Maria Olazabal, and Angel Cabrera, just to name a few. Many of these players do not play full-time on the PGA Tour, and it is important to have a covariate that accounts for play around the globe. The official world golf rankings are computed each week using a point system that takes into account tournament results from the PGA Tour, European Tour, Asian Tour, Japan Tour, and others. The ranking system is a two year rolling average that accounts for field strength, extra points for major championships, and more emphasis on recent tournament results. Interested readers can go to <http://www.owgr.com/about> for more details.

This study used the official world golf ranking points(OWGR) of each player the week of the Masters. Figure 3 below is a boxplot showing the distribution for OWGR leading into the 2017 Masters. It is approximately normally distributed, but with the top six players as outliers. As mentioned earlier, Dustin Johnson did not play. Similar to strokes gained, the middle 50% are tightly bunched together and within 1.5 points of each other.

Figure 3. Official World Golf Ranking of 2017 Masters Participants 

The last covariate included in the model was the number of Masters played. Rarely do first-time players win the Masters. In fact, it has happened only three times. Horton Smith won the first Masters played in 1934, Gene Sarazen in 1935 with his “shot heard round the world” making a double eagle on the par five 15th in the final round, and Fuzzy Zoeller in 1979. It usually takes some time to learn how to play Augusta National, to get familiar with the large undulating greens, when to be aggressive, and when to play more conservatively. Including no. of Masters as a categorical variable gave the best results. If the player was making their 1st or 2nd Masters appearance, then no. of Masters was set equal to 1, otherwise it was 0.

For this data set, there is a clustering of observations, or scores, within each player. To account for this we will use a class of models, generalized linear mixed models (GLMM), which focuses on inferences about the individuals (or golfers) in the population. These are conditional models, (conditional on each player), that allow us to estimate probabilities of score on each hole that are player specific. To do this we modify equations (1) and (2) above by adding in a random player effect for the *ith* player, , where .

(6)

(7)

It should be noted that observations within each cluster are assumed to be independent. In other words, a golfer’s score from one hole to the next is assumed to be independent. Is this reasonable? To answer this gets into the “hot-hand” controversy. Interested readers can refer to papers by Arkes and Clark, listed in the Additional Readings section, which address this issue on the PGA Tour. We will simply state that Clark found no evidence of correlation in PGA Tour players’ scores from one hole to the next, and Arkes found a significant “cold hand effect” in three, six, nine, and 18 hole stretches. However, this effect was very minimal. For example, Arkes found that for every shot over par in a three hole stretch, the PGA Tour player would average approximately 0.01 shots over par for the next three holes. Furthermore, empirical evidence for this data set supports the assumption. Autocorrelation at lag one, correlation of score from hole to hole, was equal to 0.0139. In the McHale study regarding the fairness of handicaps in the UK, it was 0.016.

The model defined in equations (6) and (7) above is still a proportional odds model. This model enforces the “parallel lines assumptions”, which means that the effects of the covariates do not change across the *J* -1 cumulative logits. If these effects are allowed to change, it is a non-proportional odds model, or a partial proportional odds model (PPOM) if the effects for some, but not all the covariates, change across the *J* – 1 cumulative logits. See equations (8) and (9) below. After testing the “parallel lines assumption” for each effect, the effects for hole, round, year, no. of Masters, SG Tee, and SG around were allowed to vary across the *J* – 1 cumulative logits, and the effects for SG Approach, SG Putt, and OWGR continued to use the “parallel lines assumption”.

(8)

(9)

Both models, POM and PPOM, were fit to the data in this study. The SAS procedures PROC GLIMMIX and PROC NLMIXED were used to fit the POM and PPOM respectively. As shown in Table 1 below, the PPOM provided a better fit based on the log-likelihood, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). However, due to ease of interpretation regarding covariates using the POM, results from both models are discussed in the following sections.

Table 1. Fit Statistics for Models POM and PPOM

|  |  |  |
| --- | --- | --- |
|  | POM | PPOM |
| -2lnL | 118,500 | 117,094 |
| AIC | 118,588 | 117,454 |
| BIC | 118,747 | 118,107 |
| n | 58,986 | 58,986 |

Cross-Validation Analysis:

To determine the predictive ability of the model, the Masters was forecasted using the PPOM for years 2014, 2015, 2016, and 2017. For 2014, the model coefficients were estimated using data only through 2013. Probabilities of score on each hole were then computed for the 2014 players, and the Masters tournament was simulated 10,000 times using R. Each simulation had a 36 hole cut to the low 50 and ties, or anyone withing 10 shots of the lead, and playoffs were used in case of ties at the end of 72 holes. The players were ranked according to their estimated probability of winning, and spearman’s rho was computed between the model rankings and the actual finish. For comparison, spearman’s rho was also computed between the Vegas rankings (http://www.vegasinsider.com/) and actual finish, as well as between the model rankings and Vegas rankings. To make these comparisons the same, the older former champions and the amateurs that do not play on the PGA Tour were also removed from the Vegas rankings and actual finish. This process was repeated for years 2015, 2016, and 2017, and the results are included in the Table 2 below:

Table 2. Spearman’s Rho between Model Rankings, Vegas Rankings, and Actual Finish

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 2017 | 2016 | 2015 | 2014 |
| Model vs. Actual | 0.4810 | 0.3787 | 0.5290 | 0.2695 |
| Vegas vs. Actual | 0.5250 | 0.3550 | 0.5706 | 0.2267 |
| Model vs. Vegas | 0.8887 | 0.8873 | 0.8442 | 0.8928 |

The model has similar predictive power to the posted Vegas odds prior to the tournament. In some years the correlation between the model and actual results is low and in others it is high, essentially depending on how well the favorites played. The strokes gained and world rankings data both show that these players are very close in ability and it is difficult to separate them. The reader should keep in mind that each Masters played is just one realization of what could have happened that year. Considering that the model predicts as well as the Vegas posted odds, it indicates that the model gives reasonable probabilities of what could have happened for each Masters, and these probabilities are the real interest. Table 3 below lists the top 14 players for 2017 so as to include the actual champion, Sergio Garcia. The table includes rank, and estimated probability of winning for both the PPOM and Vegas, as well as the actual finish and score for each player. Not included in the table below, is the probability of winning for world number one, Dustin Johnson. He had a 0.2028 chance of winning when included in the model, and Vegas had him listed at 5 to 1 or a 0.1667 chance of winning.

Table 3. 2017 Masters Predictions for PPOM and Vegas Odds

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| name | PPOM Rank | PPOM Prob. | Vegas Prob. | Vegas  Rank | Actual Finish | Score |
| McIlroy, Rory | 1 | 0.1191 | 0.1111 | 2 | 7 | 285 |
| Day, Jason | 2 | 0.0864 | 0.0606 | 3 | 22 | 290 |
| Spieth, Jordan | 3 | 0.0863 | 0.1333 | 1 | 11 | 287 |
| Matsuyama, Hideki | 4 | 0.0653 | 0.0435 | 7 | 11 | 287 |
| Rose, Justin | 5 | 0.0483 | 0.0323 | 8 | 2 | 279 |
| Fowler, Rickie | 6 | 0.0409 | 0.0526 | 5 | 11 | 287 |
| Scott, Adam | 7 | 0.0336 | 0.0278 | 11 | 9 | 286 |
| Rahm, Jon | 8 | 0.0331 | 0.0526 | 5 | 27 | 291 |
| Stenson, Henrik | 9 | 0.0323 | 0.0303 | 10 | 58 | 150 |
| Kuchar, Matt | 10 | 0.0259 | 0.0179 | 16 | 4 | 283 |
| Hatton, Tyrrell | 11 | 0.0251 | 0.0179 | 16 | 76 | 156 |
| Casey, Paul | 12 | 0.0240 | 0.0244 | 14 | 6 | 284 |
| Mickelson, Phil | 13 | 0.0237 | 0.0556 | 4 | 22 | 290 |
| Garcia, Sergio | 14 | 0.0235 | 0.0323 | 8 | 1 | 279 |

Fixed Effects

The cumulative logits in (6) and (8) were designed to find the probability of a triple bogey or worse (RTP = 3 or j = 1), then the probability of a double bogey or worse (RTP = 2 or j = 2) , etc., up to the probability of a birdie or worse (RTP = -1 or j = 5). Of course the probability of an eagle (RTP = -2 or j = 6) is one minus the probability of birdie or worse. The full model was developed using 58,696 observations from years 2005 through 2017. While the PPOM provided a better fit, the fitted POM coefficients are given in Table 4 below due to their ease of interpretation relative to the PPOM.

Table 4. Estimated Coefficients for, **,** the Fixed Effects in the POM

|  |  |  |
| --- | --- | --- |
|  | coefficient | p-value |
| α1 (triple bogey or worse) | -5.4122 | <.0001 |
| α2 (double bogey or worse) | -3.3611 | <.0001 |
| α3 (bogey or worse) | -0.9032 | <.0001 |
| α4 (par or worse) | 2.1643 | <.0001 |
| α5 (birdie or worse) | 5.927 | <.0001 |
| hole 1 | 0.2138 | <.0001 |
| hole 2 | -1.6314 | <.0001 |
| hole 3 | -0.6399 | <.0001 |
| hole 4 | 0.2145 | <.0001 |
| hole 5 | -0.05217 | 0.2888 |
| hole 6 | -0.2388 | <.0001 |
| hole 7 | -0.08299 | 0.096 |
| hole 8 | -1.4304 | <.0001 |
| hole 9 | -0.3127 | <.0001 |
| hole 10 | 0.04909 | 0.3188 |
| hole 11 | 0.4085 | <.0001 |
| hole 12 | -0.1878 | 0.0002 |
| hole 13 | -1.7337 | <.0001 |
| hole 14 | -0.3093 | <.0001 |
| hole 15 | -1.6608 | <.0001 |
| hole 16 | -0.3847 | <.0001 |
| hole 17 | -0.1031 | 0.0368 |
| hole 18 | 0 | . |
| sg\_approach | -0.0502 | 0.0437 |
| sg\_around | -0.03668 | 0.3287 |
| sg\_tee | -0.122 | <.0001 |
| sg\_putt | -0.02025 | 0.4 |
| OWGR\_april | -0.02652 | <.0001 |
| round 1 | 0.139 | <.0001 |
| round 2 | 0.1813 | <.0001 |
| round 3 | 0.1403 | <.0001 |
| round 4 | 0 | . |
| no\_masters\_category 1 | 0.04489 | 0.0245 |
| no\_masters\_category 0 | 0 | . |
| year 2005 | -0.04007 | 0.3488 |
| year 2006 | -0.0423 | 0.327 |
| year 2007 | 0.2997 | <.0001 |
| year 2008 | -0.02854 | 0.5091 |
| year 2009 | -0.2398 | <.0001 |
| year 2010 | -0.1284 | 0.0026 |
| year 2011 | -0.2497 | <.0001 |
| year 2012 | -0.08938 | 0.0329 |
| year 2013 | -0.09805 | 0.0205 |
| year 2014 | -0.00216 | 0.9594 |
| year 2015 | -0.2577 | <.0001 |
| year 2016 | 0.08403 | 0.0434 |
| year 2017 | 0 | . |

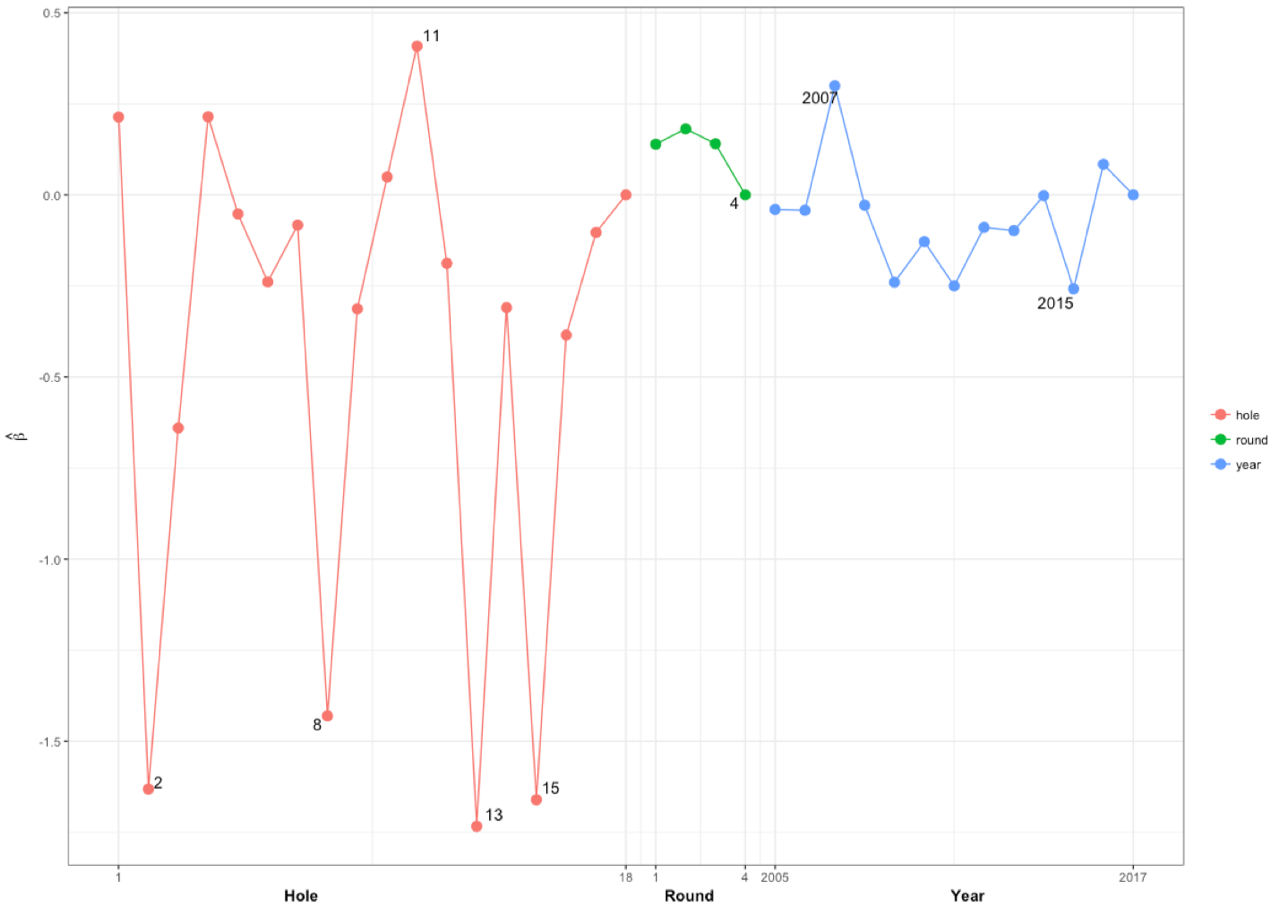
In general when looking at Table 4, a negative coefficient means that scores will tend to get better (or lower) as the covariate increases, and a positive coefficient means that scores will get worse (or higher). For the continuous covariates, strokes gained and world ranking, all coefficients are negative which means players of higher skill will tend to have lower scores. Based on the magnitude of the coefficients and p-values, SG Tee is the most important of the SG skills at the Masters. Interestingly SG Around, and SG Putt are not statistically significant. However, it would be a mistake to conclude that chipping and putting are not important at the Masters. The players that do well each year almost certainly performed well in these areas. It just means that using the chipping and putting performance of a PGA Tour player to predict scoring at the Masters, will not be as effective as using their driving performance.

For the categorical covariates, year, round, hole, and no. of masters, all coefficients are relative to a baseline category. The baselines are year equal to 2017, round equal to four, hole equal to 18, and no. of Masters more than two. The positive coefficient on no. of Masters in Table 4, shows that the inexperienced players were at a disadvantage, and tended to shoot higher scores as expected. The impacts for hole, round, and year can be more clearly seen in Figure 4 below. Based on the magnitude of the coefficients, hole had the largest impact on scoring. More specifically, the large negative coefficients for the par fives, holes 2, 8, 13, and 15 indicate they were the easiest holes, and that the players had to do their scoring on the par fives at Augusta. The large positive coefficient for number 11, the long par four over 500 yards with a pond front and left of the green, indicates it was the toughest hole.

Year, while not as important as hole, also had a significant impact on scoring. Figure 4 also shows that 2007 was the year the course played the most difficult, and year 2015 was the year it played the easiest. 2007 had cold, windy weather, and Zach Johnson won with a score of one over par, which was the only year in this study where the winning score was over par. 2015 was the year that Jordan Spieth won at 18 under par, tying Tiger’s record score from 1997.

For the round effect, model coefficients indicate that round four had lower scoring than the other three rounds, but this effect was small relative to the effects of year, and especially hole. The authors believe that any effect due to round is probably a result of course setup, most specifically, pin placements. Viewers of the Masters will know that announcers often mention “traditional Sunday pin placements” for certain holes, and perhaps tournament organizers want better scoring on Sunday, or at least better chances of eagles for the “Sunday roars”. Regardless, scoring was better in round four over the course of this study.

Figure 4. POM Coefficients for Hole, Round and Year



To help the reader see a little more clearly how these coefficients turn into probabilities, consider the difference in holes 2 and 18 for a player playing in round four of 2017 with SG equal to zero for all skills, no. of masters played is more than two, and OWGR is equal to three (approximately the 50th percentile). Equation 10 below shows the explicit calculation for the first cell in Table 5.

Table 5. POM Cumulative Probabilities for Hole 2 vs. Hole 18

|  |  |  |
| --- | --- | --- |
|  | Hole 2 POM | Hole 18 POM |
| P(triple or worse) |  |  |
| P(double or worse) |  |  |
| P(bogey or worse) |  |  |
| P(par or worse) |  |  |
| P(birdie or worse) |  |  |

(10)

These cumulative probabilities are then easily converted into probabilities for each score.

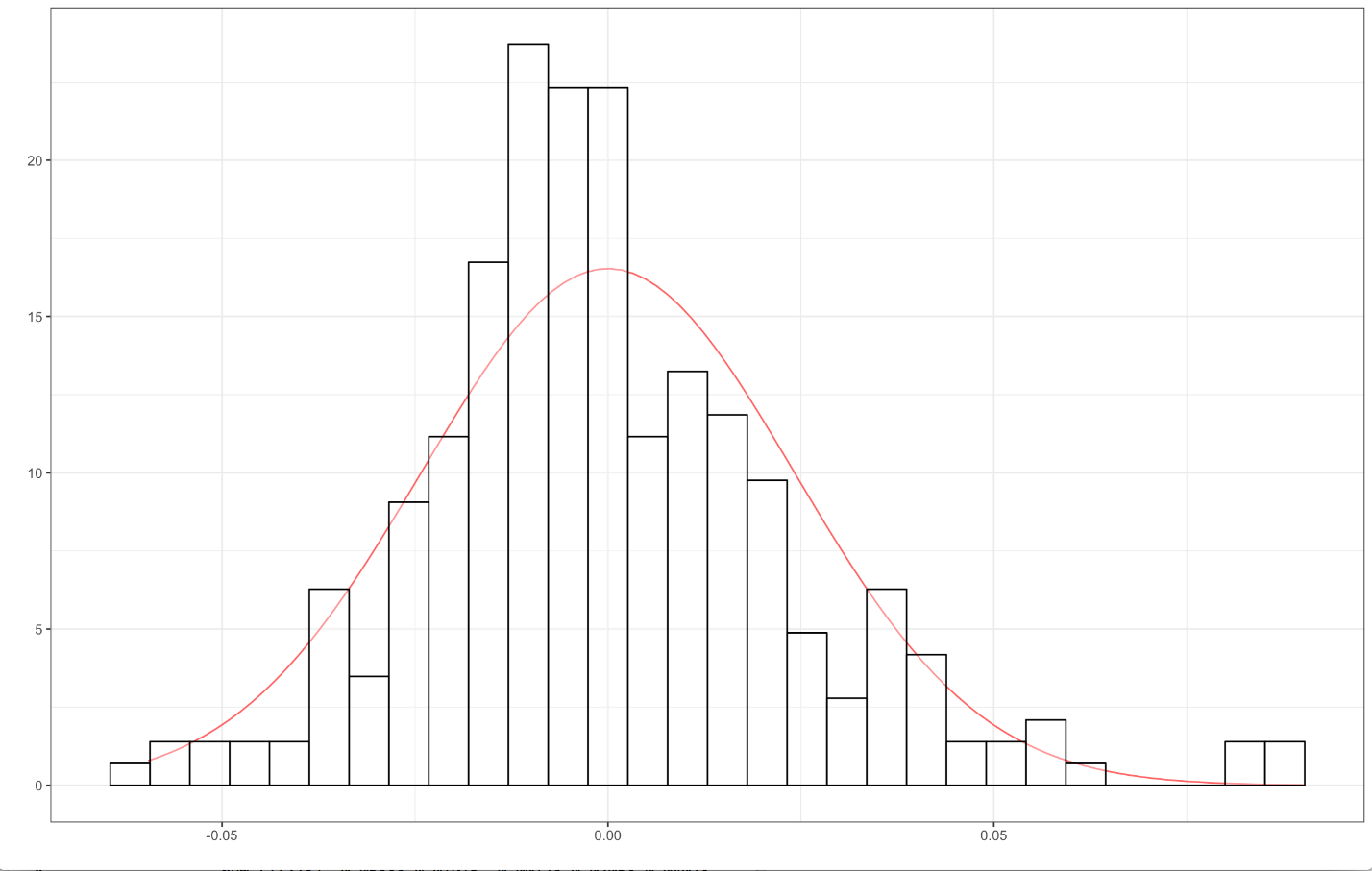
Table 6. POM Probabilities of RTP score for Hole 2 vs. Hole 18

|  |  |  |
| --- | --- | --- |
|  | Hole 2 POM | Hole 18 POM |
| P(triple or worse) |  |  |
| P(double) |  |  |
| P(bogey) |  |  |
| P(par) |  |  |
| P(birdie) |  |  |
| P(eagle) |  |  |

Table 6 clearly shows that hole 2 is going to yield lower scores relative to par. The reader can easily verify that this player’s average score would be 4.194 for hole 18 (0.194 over par), and 4.672 for hole 2 (0.328 under par). These results are consistent with Figure 1 from earlier.

Random Effects (Subject-Specific or Player Effects)

In addition to the fixed effects in the model discussed above, random player effects were also estimated. Each player has their own set of values for the player-specific covariates, strokes gained, world ranking, and no. of Masters. These alone give each player different score probabilities on each hole. Therefore in this context the random player effect, estimated from the clustering of observations within each player, can be interpreted as whether the player plays the Masters better or worse than expected, after adjusting for their skill (SG and OWGR) and experience (no. of Masters). Figure 5 below shows a histogram of the estimated random player effects. For this figure, the sign of the coefficient was changed so that positive coefficients indicate the player tends to have lower scores than expected at the Masters, and negative coefficients indicate the player tends to have higher scores than expected. The outliers in Figure 5, players who play the Masters particularly well, are Jordan Spieth, Justin Rose, Angel Cabrera, and Phil Mickelson. Spieth has the highest coefficient at 0.0901, and based on his brief record at the Masters to this point, it is no surprise. Spieth has one victory, two seconds, and an 11th place finish. Even after adjusting for his high skill level, he plays the Masters better than anyone. Rose is next at 0.0870. While he doesn’t have any Masters wins, he has finished second twice in the last three years, he has five top ten finishes, and has never missed the cut. Mickelson is fourth at 0.0830. Even though he has a better Masters record than Rose, he has also had better strokes gained statistics and usually a higher world ranking through the years. Players of note that played the Masters particularly poorly between 2005 and 2017, are Martin Kaymer and Ernie Els. Kaymer at -0.0496, is a two-time major champion who has missed the cut five times in 10 tries, and just recently had his best ever finish at 16th in 2017. Ernie Els at -0.0555, is a two-time winner of both the U.S. Open and the British Open. While he placed 2nd in the Masters in both 2000 and 2004, between 2005 and 2017 he has missed the cut five times and his best finish is 13th.

Figure 5. Random Player Coefficients (histogram) 

Simulations

For each year, 2005 to 2017, coefficients from the PPOM (for both fixed and random effects) were used to compute probabilities of different scores on each hole and each round for each player. The Masters tournament for each year was then simulated 10,000 times. Again all simulations involved a cut to the low 50 and ties, or anyone within 10 shots of the lead, and playoffs were run in case of a tie for 1st place. From these simulations, probabilities of winning for each player were estimated. These estimated probabilities of winning were used to further investigate some of the covariates included in the model. Figures 6 and 7 below show the probability of winning versus OWGR, and the probability of winning versus SG Total (the sum of all four strokes gained categories), respectively. These figures have also labeled some players of note, and the Masters champion each year is highlighted in blue. The curves were fit in R using the smoothing method “loess”, and show the overall trend of the data. World ranking seems to be the better predictor of Masters performance, due to the tighter and more discernable pattern. Both figures show that probability of winning begins to increase dramatically for the very top players.

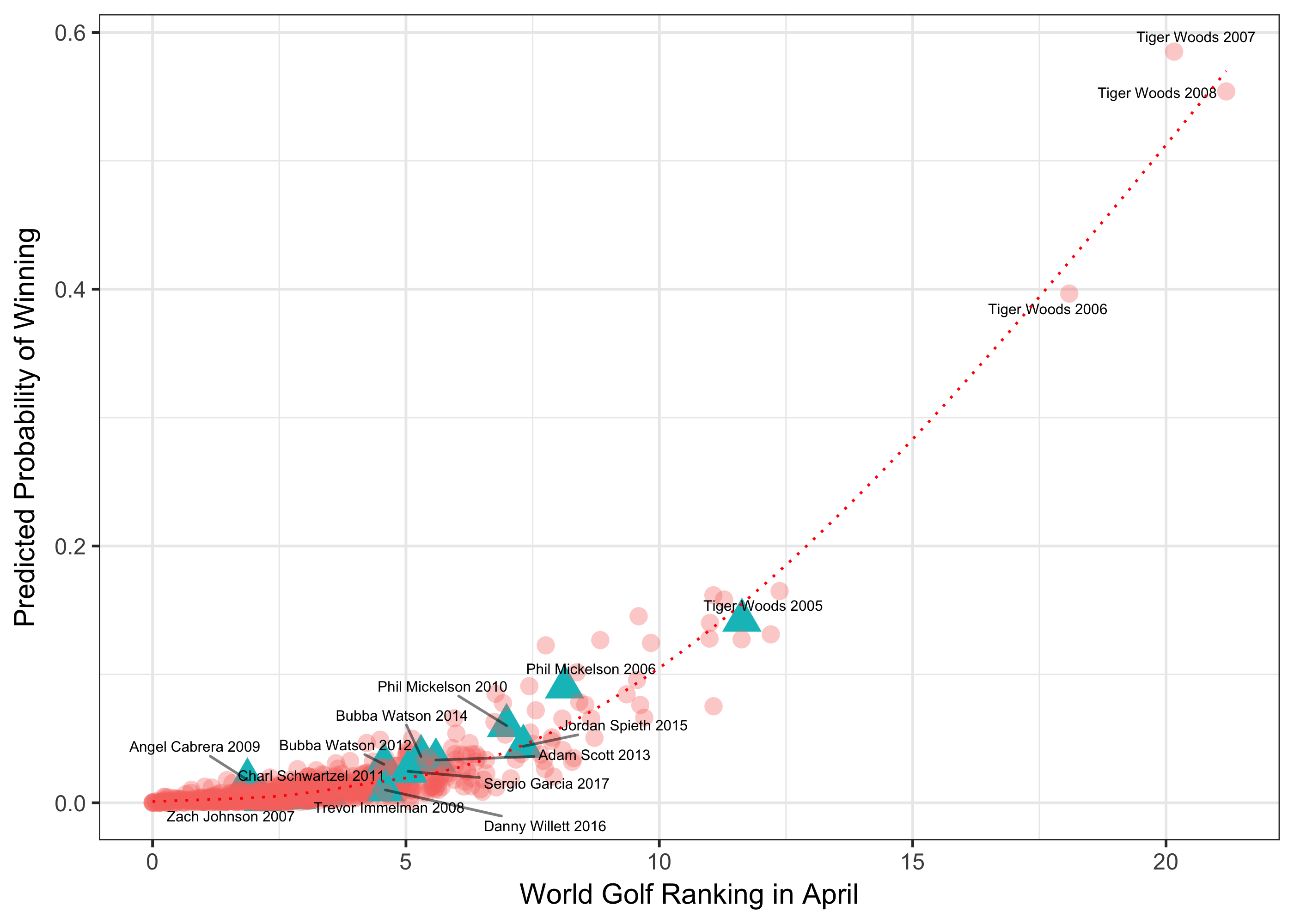
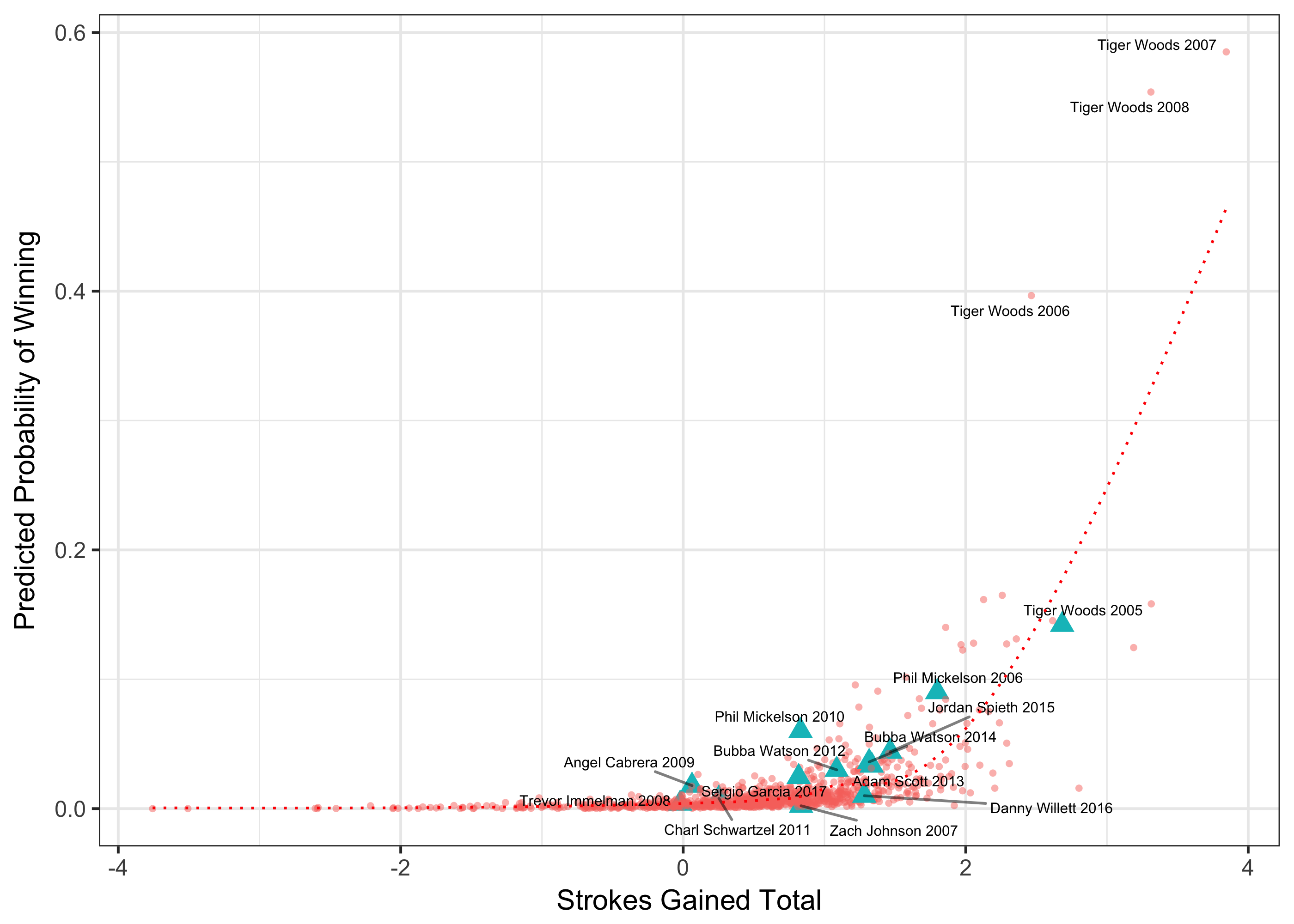
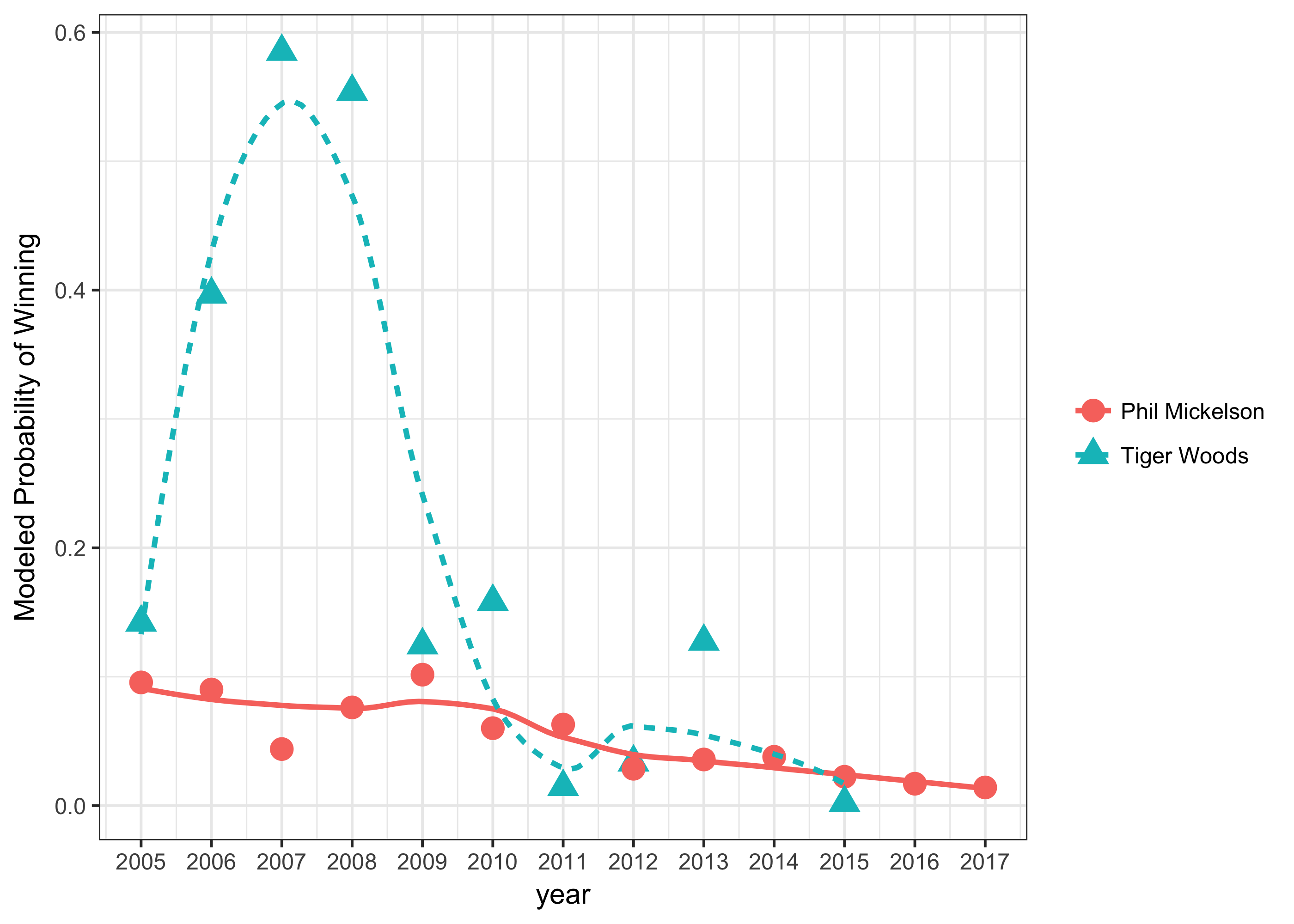
Figure 6. Probability of Winning vs OWGR 

Figure 7. Probability of Winning vs Strokes Gained Total 

Figures 6 and 7 also show that Tiger Woods’ probabilities of winning during 2006 (0.397), 2007 (0.585), and 2008 (0.554) were truly remarkable. His strokes gained values and world rankings were so high during these years that he actually had a better chance of winning than the rest of the field. In 2007 and 2008 his strokes gained total were 1.5 shots better than the second best player on the PGA Tour per round. His world ranking points were over 20, which was more than twice as much as the 2nd ranked player. When TV commentators used to ask each other if they would take Tiger or the field, it was a reasonable question. Despite Tiger’s extremely high probability of winning the Masters in that three year stretch, he wasn’t able to do it. He actually finished tied 3rd in 2006, tied 2nd in 2007, and 2nd alone in 2008. For comparison, Figure 8 below charts Tiger’s and Phil Mickelson’s probabilities of winning through the years. Note a steady but slight decline for Phil as he ages into his 40s, and the dominant years from 2005 to 2010 for Tiger.

Figure 8. Probability of Winning the Masters: Tiger vs Phil 

Finally we get back to Jordan Spieth in 2016. When Spieth walked to the 10th tee in the final round of 2016, what was his probability of winning? The simulation was run in “real-time” with each player starting on the hole they were playing when Spieth moved to number 10. Table 7 below lists the players’ scores and starting positions, strokes gained and world ranking, and their probability of winning based on 100,000 simulations.

Table 7. 2016 Spieth Simulation Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Player | Score | Tee Box | SG - Total | OWGR Pts | P(winning) |
| **Jordan Spieth** | **-7 or 245** | **10** | **1.86** | **11.000** | **0.9458** |
| Smylie Kauffman | +1 or 253 | 10 | 0.876 | 2.395 | 0.0004 |
| Hideki Matsuyama | +2 or 254 | 10 | 1.378 | 4.518 | 0.0004 |
| Dustin Johnson | -1 or 255 | 11 | 1.44 | 6.379 | 0.0135 |
| Jason Day | E or 256 | 11 | 2.259 | 12.374 | 0.0062 |
| **Danny Willett** | **-2 or 258** | **12** | **1.28** | **4.595** | **0.0247** |
| Lee Westwood | E or 260 | 12 | 0.052 | 1.977 | 0.0023 |
| Soren Kjeldsen | -1 or 262 | 13 | 0.124 | 2.446 | 0.0067 |

Note: Matsuyama was playing in the group ahead of Spieth and had not completed the 10th hole. Since simulations cannot start a player in the middle of a hole, he was started on the 10th tee.

According to the model, Spieth had a 95% chance to win. A playoff would have resulted only 4% of the time. As golf fans know, Spieth proceeded to bogey 10 and 11, and then hit the ball in the water twice on the par three 12th and take a quadruple bogey seven. This opened the door for Willett, who did not let his chance get away. The authors certainly do not know what was going through Willett’s mind when he suddenly found himself in the lead after Spieth’s meltdown on 12, but the pressure of holding the lead himself did not faze him. He played the last six holes in three under par, and won comfortably by three shots over Spieth and Westwood. He went from a two or three in a 100 chance with seven holes left, to the green jacket.

A similar simulation was performed for the 2017 Masters. Sergio Garcia finally broke through and won his first major in a playoff with Justin Rose. However, on the 13th hole in the final round, Garcia trailed Rose by two shots. Rose hit a perfect tee shot, setting himself up for an eagle or birdie opportunity at the reachable par five. Garcia meanwhile hit his tee shot left of Ray’s creek into a bush and had to take a penalty drop. It would have been fun to start a simulation from this point, but a limitation of this model is it has to start players at the beginning of a hole. Garcia was able to pitch out of the trees down the fairway, and then pitch onto the green and make his putt for a par. Meanwhile Rose hit his second shot to the back of the green, but it took him three shots to get down and so he also made a par. Despite Garcia managing to avoid losing any more ground, he was still two shots behind with five holes to play. What were his chances of winning from this point? The simulation results from 100,000 repetitions indicate that he still only had a 17% chance to win. A playoff between the two would have happened 11% of the time.

Table 8. 2017 Simulation: Garcia vs. Rose

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Player | Score | Tee Box | SG - Total | OWGR Pts | P(winning) |
| Justin Rose | -8 or 260 | 14 | 1.148 | 4.490 | 0.83245 |
| Sergio Garcia | -6 or 262 | 14 | 0.817 | 5.043 | 0.16755 |

Looking again at Sergio and Rose from their tee shots on 13, a reasonable result would be a bogey for Sergio, and a birdie for Rose. If Sergio went to the 14th tee four shots behind, the probability of him winning was only 0.033 in 100,000 simulations. This just emphasizes how important Sergio’s great par save was to keeping him in the tournament.

Conclusion

This study has shown that ordinal logistic regression can be used to effectively estimate probabilities of different scores on each hole for each player in the Masters golf tournament. Cross-validation analysis shows that the model has similar predictive abilities to the pre-tournament odds listed in Las Vegas. Predicting scores by hole allows for interesting “real-time”, in-round simulations that cannot be conducted with a round-by-round predictive model.

Amongst the covariates in the model, the official world golf rankings appear to be the best single predictor of Masters performance. The model also showed using strokes gained statistics, that long-game skills on the PGA Tour (driving and approach shots) are better predictors of Masters success than the short game skills of chipping and putting. In addition, including a random player effect in the model allows for estimation of which players tend to play the Masters better or worse than their skill level would indicate.

For future researchers interested solely in pre-tournament predictions, they might want to experiment with a round-by-round model and see if they get better results. Other options might be to make some adjustments to the covariates. For example, instead of using the prior 12 months of strokes gained data, the prior 24 months could be used. This might be helpful in getting more accurate predictions for the European players that don’t play as much on the PGA Tour. A weighting system could also be used, with strokes gained statistics in more recent tournaments weighted more heavily. The authors feel that the best improvement could be made by including pin positions, and wind conditions. While this information was not available for this study, a historical record of pin locations and wind conditions probably does exist, or at least could be easily recorded in future years.

Additional Readings:

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics.

Arkes, J. (2016). The Hot Hand vs. Cold Hand on the PGA Tour*. International Journal of Sports Finance*, 11, 99 – 113.

Baldson, E.M. (2013). Risk Management with Tournament Incentives. *Journal of Quantitative Analysis in Sports*, 9, 301-317.

Clark, R.D. (2005). Examination of Hole-to-Hole Streakiness on the PGA Tour. *Perceptual and Motor Skills*, 100, 806-814.

McHale, I.G. (2010). Assessing the Fairness of the Golf Handicapping System in the UK. *Journal of Sports Sciences*, 28(10), 1033-1041.