Bayesian Analysis of PGA Tour Hole Difficulty

Stephen Merrill and William White

December 12, 2016

Motivation

Golf is a sport that requires both skill and strategy in order to have success. We both enjoy the sport and were interested in assessing the difficulty of individual golf holes and the factors that influence that difficulty. We had PGA Tour data from research with a professor and felt we could use it to evaluate golf hole difficulty and the variables that determine difficulty. We wanted to develop a flexible model that could be used on any golf hole on the PGA Tour. Our model could be used in many applications related to the PGA Tour and golf in general. For example, golf course design, which is a lucrative business that requires creativity to come up with unique and challenging golf holes. We feel that golf course designers could use the results of our model to find holes that are difficult in a certain aspect and draw inspiration for their design. Another natural application would be for PGA Tour players to use the model when planning their strategy on a hole. They could find valuable insights on how to lower their scores by applying these results to individual holes on different courses.

Data & Model

This analysis was done using data from the 2012 Northern Trust Open held at Riviera Country Club in Pacific Palisades, California. Ideally we would have liked to use one of the four major tournaments, which most of the top ranked professional players participate in, but none of that data were available. Instead, we chose this tournament because there is still a relatively high level of professional competition and the course has no water hazards, which we thought would be difficult to properly account for. The data comes from ShotLink, a data collection service used by the PGA Tour to gather data on every shot. Specifically, there is data on over 31,000 shots at the Northern Trust Open.

Our model uses the response of score relative to par on a hole, which usually will take on one of 5 values seen below in Table 1. This number is the difference between the expected number of strokes to complete a hole, called par, and the actual number taken by the golfer. We combined all scores less than -1 into the Birdie category and all scores greater than +2 into the Double Bogey category due to sparsity of data for these extreme scores. There is a natural ordering to golf scoring, and as a result we chose to treat our response as ordinal data and model it accordingly.

Name	Eagle	Birdie	Par	Bogey	Double Bogey
Score relative to par	-2	-1	0	+1	+2

Table 1: There is a natural ordinal structure to golf scores

In order to answer the research question of which variables affect course difficulty, we selected four covariates that cover a golfer's entire skill set and fit β coefficients for each one. There is a natural correlation between the covariates since one shot affects the next, and this is accounted for in the model, which is formally defined below. These variables of interest are given as:

- β_1: Driving distance $>$ 300 yards. This is a binary variable for whether the drive is hit over 300 yards or not. 300 yards is considered an average driving distance for a PGA player.

- β_2: Distance off from center in 5 yard increments. Off from center refers to the distance away from the center of the fairway, measured from the location where the drive lands.
- β_3: Distance to scramble in 25 yard increments. Scramble distance is the distance from the hole, accumulated over shots where the golfer is not yet on the green, but should be. For example, on a par four hole any shots after the second that do not originate on the green will contribute to scramble distance for that hole.
- β_4: Distance of first putt in 5 foot increments. This is the distance from the location of the first putt attempt to the hole.

We determined that effects were best seen by incrementing the distance measures, which makes sense because there is often a threshold of how far off a golf shot is from ideal placement that needs to be passed before there would be an impact on the score. For example, hitting a drive a few yards off from center will not matter until it is far enough off that the threshold of the fairway is passed and the ball ends up in the rough. Figure 1 shows relationships among these variables from the data and leads us to believe that a regression framework is appropriate for this analysis.

Latent Variable Bayesian Hierarchical Probit Model

The ordinal data is modeled with a latent variable Z, which uses a probit link to model the underlying distribution of the response on a standardized scale. We decided to use a latent variable bayesian probit model because it allowed us to fit regression coefficients for our covariates and we could account for covariance, knowing our covariates were not independent of one another. The model is used for ordinal response variables so it worked well with score relative to par, which is ordinal. We extended the model to be hierarchical so we could evaluate multiple holes at once while getting separate coefficient estimates for each hole. Another nice feature of a bayesian model is that we could get posterior predictive distributions for each hole, allowing us to rank the holes on predicted mean score relative to par on the hole.

See equations (1) - (5) below for the formal definition of the model. There are three γ cutpoints, which are fixed at 0, 1 and 2 as those are natural points to cut. Values for the prior on $\boldsymbol{\mu}$ are set as $\mathbf{m} = \mathbf{0}$ and $\mathbf{V} = \mathbf{I}$, a 4x4 identity matrix. Values for the prior on $\boldsymbol{\Sigma}$ are set as w = 1 + n_{holes} and $\mathbf{I}$ = a 4x4 identity matrix. The model is implemented by deriving complete conditionals (see Appendix) and using a Gibbs Sampling technique to obtain draws from the posterior distributions. The results here are from a run of 5,000 draws with a burn of 1,000. Figure 2 addresses the diagnostics of this model by showing that the trace plots from the Gibbs Sampler converge well.

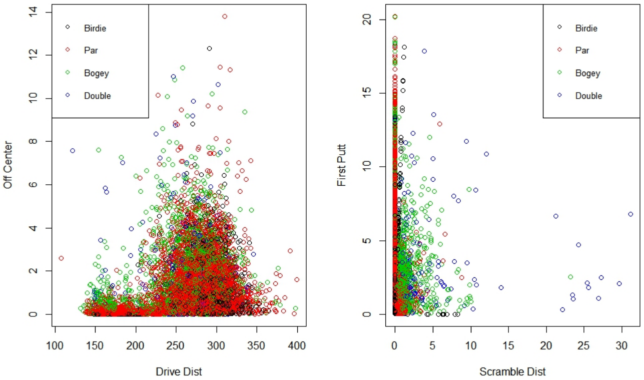

Figure 1: An exploratory glance shows that the data has linear relationships that are intuitively expected.

$$\mathbf{Z}_{ik} \sim \mathcal{N}(\mathbf{X}_k'\boldsymbol{\beta}_k, 1) \tag{1}$$

$$\mathbf{Y}_{ik} = j, \quad \text{if} \quad \gamma_{j-1} \leq \mathbf{Z}_{ik} \leq \gamma_j \tag{2}$$

$$j = \{-1, 0, 1, 2\}$$

$$k = 1, \ldots, n_{holes}$$

$$i = 1, \ldots, n_{obs}$$

$$\boldsymbol{\beta}_k \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3}$$

$$\boldsymbol{\mu} \sim MVN(\mathbf{m}, \mathbf{V}) \tag{4}$$

$$\boldsymbol{\Sigma} \sim InverseWishart(w, \mathbf{I}) \tag{5}$$

Results

Examining the β posterior distributions answers the research question of the effect of different factors on individual hole difficulty. As this is a probit model, the effect sizes are interpreted as the change in Z-score across the scale of the ordinal response. A few interesting observations from these results are highlighted below.

The first plot in Figure 3 shows that driving distance over 300 yards is much more beneficial on par 5 holes than par 4 holes, with almost no effect on par 3 holes. This result can be used to recommend that golfers drive as far as possible on par 5 holes, but use discretion on par 4 holes because a long drive could lead to trouble. Par 3 holes rarely if ever have drives over 300 yards because of how short they are, so precision placement of shorter drives is much more important. This can be seen especially in the effect of the off center distance. As par decreases, it is more and more vital to stay in the center of the hole. This makes sense because being off center on a par 4 or 5 hole allows for more opportunities to make corrections than on an unforgiving par 3. Scramble distance indicated a similar trend with par 3 holes being affected more. This result indicates that running into trouble and scrambling to make it to the green is worse on a par 3 hole. This was interesting because we believed scramble distance would be independent of par since scrambling is dependent on missing the green, irrespective of par. Finally, there is no clustering by par for first putt distance, nor is there a clear effect at all. Clearly putting is different for each hole, which is understandable because every hole has a different green with different challenges.

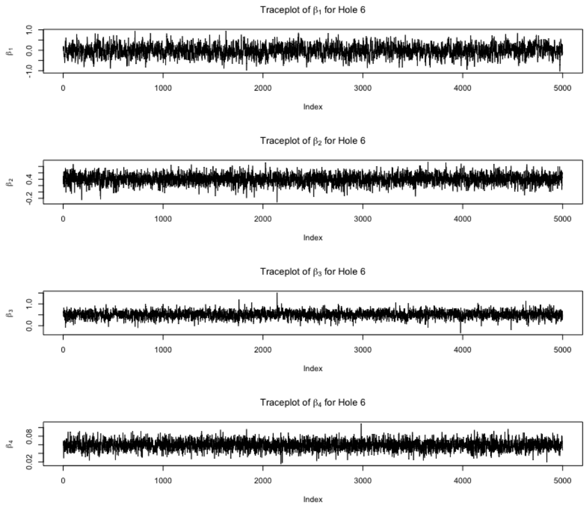

Figure 2: Trace plots for the $\boldsymbol{\beta}$ on Hole 6

Figure 4 shows a cumulative distribution of difficulty for three holes, obtained from the density of the means of the draws of each hole's Z-distributions. This distribution shows the difficulty of the hole in that it will produce scores with the same relative distribution. To answer the research question of direct hole difficulty evaluation, the cursory glance at hole difficulty shown by this plot can be formalized by extension to a posterior predictive distribution. This distribution reflects the probability of a hole's next score being a birdie, par, bogey, or worse and is seen in Table 2.

We found a cumulative distribution of difficulty of each hole and then found the proportion of the distribution between each of the fixed cut points to calculate our posterior predictive distribution on each hole. Table 2 shows the this calculated posterior predictive distribution This distribution reflects the discrete multinomial data with π_j probabilities of each hole's next score of being in category j. As would be expected, most of the holes have par as the score with the greatest probability of occurring. Some of the easier holes such as holes 1 and 17 have birdie as more probable than a par and some of the more difficulty holes such as 2 and 12 have a higher probability of bogey or worse.

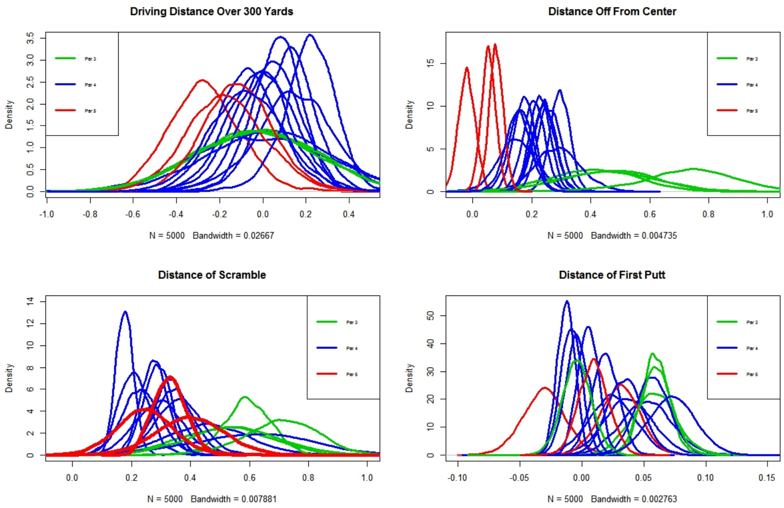

Figure 3: Posterior distributions for the $\boldsymbol{\beta}$ effects, with holes distinguished by par amounts

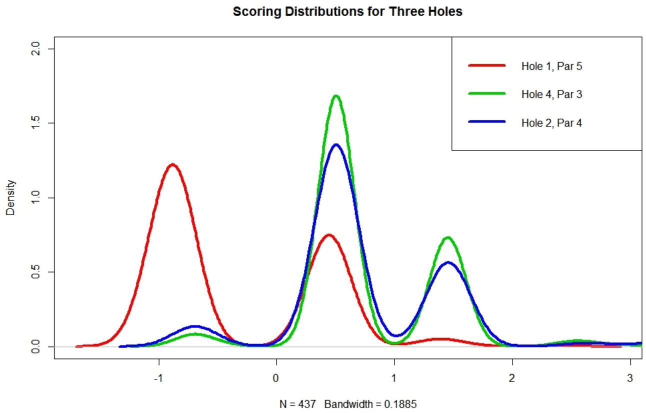

Figure 4: Hole 1 scores much lower than Holes 2 and 4. This scoring comparison between Par 3, 4 and 5 holes is fairly common.

We then took our posterior predictive distribution and calculated the mean score relative to par. We then ranked the holes from most difficult to least difficult based on this predicted mean score relative to par. To compare with what was observed in the data we also looked at the observed mean score to par for each hole and ranked them accordingly so we could see whether our results made sense. These predicted and observed ranks and scores can be see in Table 3. In Table 3 we can see that our ranks aren't exactly the same but the ranks of each hole are in the same general vicinity for the predicted and observed. Our predicted mean scores were less extreme, positive or negative, than the observed mean scores. This makes sense because we simulated 10,000 scores for our posterior predictive on each hole and there were only 437 observations per hole so we would expect more extreme values in the observations.

Figures 5 and 6 present visualizations of holes 1 and 12, which were ranked the least and most difficult, respectively. Hole 1 is a short par 5 that does not have any imposing hazards to worry about on the drive. Players do not have to be as cautious on the drive so they can go for length rather than accuracy. This then allows them to attempt to hit the green in 2 shots rather than the standard 3 shots for a par 5, which leads to more eagles and birdies. Hole 12 is a long par 4 with a thick tree line on the right side of the hole and a narrow green surrounded by thick rough (long grass) and a deep sand bunker. The β for driving accuracy on hole 12 is high suggesting that hitting far off from the center of the fairway raises your score significantly. This could be explained by the trees; if you get behind them it is next to impossible to hit the green on your second shot. Another factor is the narrowness of the green, which makes the angle from which you approach determine your margin for error.

Hole	Birdie	Par	Bogey	Double
1	0.585	0.289	0.108	0.018
2	0.271	0.360	0.249	0.120
3	0.338	0.364	0.230	0.068
4	0.272	0.356	0.267	0.106
5	0.300	0.357	0.235	0.107
6	0.372	0.358	0.208	0.062
7	0.356	0.360	0.214	0.070
8	0.301	0.356	0.237	0.106
9	0.282	0.351	0.247	0.120
10	0.329	0.366	0.227	0.078
11	0.403	0.356	0.186	0.056
12	0.264	0.348	0.272	0.116
13	0.322	0.351	0.239	0.088
14	0.280	0.353	0.271	0.097
15	0.286	0.352	0.254	0.109
16	0.398	0.367	0.182	0.052
17	0.400	0.371	0.183	0.045
18	0.275	0.359	0.256	0.110

Table 2: Posterior predictive distributions for each hole. Gives the probability of birdie, par, bogey and double bogey or worse for each hole.

Rank	Pred. Hole	Pred. Mean Score to Par	Obs. Hole	Obs. Mean Score to Par
1	12	0.240	15	0.311
2	2	0.218	4	0.293
3	4	0.207	12	0.293
4	9	0.205	2	0.259
5	18	0.200	18	0.256
6	15	0.186	9	0.222
7	14	0.184	14	0.215
8	5	0.149	8	0.211
9	8	0.148	13	0.188
10	13	0.093	5	0.162
11	10	0.053	7	0.076
12	3	0.028	16	0.073
13	7	-0.002	3	0.030
14	6	-0.040	10	-0.016
15	11	-0.105	6	-0.048
16	16	-0.111	17	-0.128
17	17	-0.126	11	-0.160
18	1	-0.442	1	-0.588

Table 3: Predicted mean score to par and rank by difficulty of each hole alongside the observed mean score and rank by difficulty of each hole.

Figure 5: Hole 1, (Par 5) the least difficult

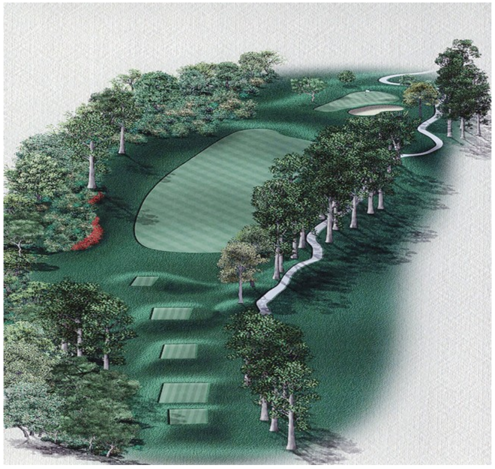

Figure 6: Hole 12, (Par 4) the most difficult

Conclusion

We feel that our model allows us to rank PGA Tour holes based on difficulty and evaluate how the covariates influence that difficulty. From our analysis we see that on average driving accuracy, measured by distance off from center, was much more influential on score than was hitting a drive over 300 yards. This is dependent, however, on the individual characteristics of the hole, such as its par. We also see the importance of scrambling distance on every hole that we analyzed, with the effect appearing to be influenced by par value for some unknown reason. The effect of first putt length is dependent on the hole and the characteristics of the green such as slope, speed and size. In the future we would like to extend this model to only look at putting, which is a unique part of golf. We could use this model to evaluate the effect of covariates such as slope, speed and distance on the number of putts taken on a green.

Appendix

Complete Conditionals

$$\boldsymbol{\beta}_k|* \sim MVN(\Sigma^*(\Sigma^{-1}\mu + X_k' Z_k), (\Sigma^{-1} + X_k' X_k)^{-1})$$

$$\boldsymbol{\mu}|* \sim MVN(V^*(V^{-1}\mu + \Sigma^{-1}\bar{\boldsymbol{\beta}}_k), (V^{-1} + \Sigma^{-1})^{-1})$$

$$\boldsymbol{\Sigma}|* \sim InverseWishart(w + n_{scores}, (I + \sum_{k=1}^{n_{holes}} (\beta_k - \mu)(\beta_k - \mu)')^{-1})$$

$$\mathbf{Z}_{ik}|*, Y_{ik} = j \sim \mathcal{TN}(\mathbf{X}_k'\boldsymbol{\beta}_k, 1, \gamma_{j-1}, \gamma_j)$$

Code

```
load(file = "~/STAT637/Stat637FinalGolf2012")
length<-5000
burn<-1000
h<-18
obs<-nrow(clean.dat)/h
clean.dat$score.to.par[clean.dat$score.to.par>2]<-2
clean.dat$score.to.par[clean.dat$score.to.par<(-1)]<-(-1)
y<-matrix(nrow = obs, ncol = h)
for(k in 1:h){
  y[,k]<-clean.dat$score.to.par[clean.dat$hole==k]
}

X<-array(0,dim = c(obs,4,h))
for (k in 1:h){
  X[,,k]<-as.matrix(clean.dat[clean.dat$hole==k,c("long.drive","drive.off.full",
  "dist.scramble","first.putt2")])
}

beta<-matrix(1,ncol=ncol(X),nrow = h)
beta.save<-array(dim = c(h,ncol(X),(burn+length)))
beta.save[,,1]<-beta

mu<-matrix(0,ncol=4,nrow=(burn+length))
mu[1,]<-rep(1,4)

m<-rep(0,4)
V<-10*diag(4)

w<-h+1
I<-diag(4)
Sigma<-array(dim = c(ncol(X),ncol(X),(burn+length)))
Sigma[,,1]<-I

z.o<-matrix(0,nrow=obs,ncol = h)
Z<-array(dim = c(c(obs,h,(burn+length))))
Z[,,1]<-z.o

gamma<-c(0,1,2)
```

```
birdie<-list()
even<-list()
bogey<-list()
dubbogey<-list()

for(k in 1:h){
  birdie[[k]]<-which(y[,k]==-1)
  even[[k]]<-which(y[,k]==0)
  bogey[[k]]<-which(y[,k]==1)
  dubbogey[[k]]<-which(y[,k]==2)
}

for(d in 2:(length+burn)){
  # update beta (k = 1,...,n_h)
  for (k in 1:h){
    sigstar <- solve( solve(Sigma[,,(d-1)]) + t(X[,,k])%*%X[,,k] )
    mustar <- sigstar %*% (solve(Sigma[,,(d-1)])%*%mu[(d-1),] + t(X[,,k])%*%Z[,k,(d-1)])
    beta.save[k,,d] <- mvrnorm(1,mustar,sigstar)
  }

  # update mu
  Vstar <- solve( solve(V) +  solve(Sigma[,,(d-1)]) )
  mstar <- Vstar %*% (solve(V)%*%m + solve(Sigma[,,(d-1)])%*%colMeans(beta.save[,,d]))
  mu[d,] <- mvrnorm(1,mstar,Vstar)

  # update Sigma
  wstar<-w+h
  S.mu<-matrix(0,nrow = 4,ncol = 4)
  for(k in 1:h){
    S.mu<-S.mu+(beta.save[k,,d]-mu[d,])%*%t((beta.save[k,,d]-mu[d,]))
  }

  Istar<-solve(I+S.mu)
  Sigma[,,d] <- riwish(wstar,Istar)

  # update Z
  for(k in 1:h){
    for (i in 1:obs){
      Z[i,k,d]<-rtnorm(1,t(X[i,,k])%*%beta.save[k,,d],1,gamma[3],Inf)
      if(y[i,k]==1)
      {
        Z[i,k,d]<-rtnorm(1,t(X[i,,k])%*%beta.save[k,,d],1,gamma[2],gamma[3])
      }
      if(y[i,k]==0)
      {
        Z[i,k,d]<-rtnorm(1,t(X[i,,k])%*%beta.save[k,,d],1,gamma[1],gamma[2])
      }
      if(y[i,k]==-1)
      {
        Z[i,k,d]<-rtnorm(1,t(X[i,,k])%*%beta.save[k,,d],1,-Inf,gamma[1])
      }
    }
  }
}
```