Rishika Malhotra-Dhir
Regression Analysis
Student Project
Jan – Mar 2014

Predicting Golf Scores

1 INTRODUCTION

My student project is based on golfing. I have always wanted to learn more about the sport and with the wealth of statistics available for golf online I decided to pursue the opportunity. It is truly fascinating to see all the work that has been done to determine the greatest predictors of golfing success.

I have chosen the top 20 players from the PGA tour who were ranked the highest in March 2014. I will determine what golfing related statistics are the best predictor of a player's scoring average. In this document, I will summarize my findings and what I've learned on this subject matter, as well as summarize my findings found in the accompanying excel file.

2 DATA AND VARIABLES

The PGA defines a players *scoring* average as "The weighted scoring average which takes the stroke average of the field into account. It is computed by adding a player's total strokes to an adjustment and dividing by the total rounds played. The adjustment is computed by determining the stroke average of the field for each round played. This average is subtracted from par to create an adjustment for each round. A player accumulates these adjustments for each round played." In this paper I attempt to model a player's scoring average (response variable) starting with eight explanatory variables.

I found all of my data and information at http://www.pgatour.com/stats.html

The following is a list of definition and acronyms that are used in this study:

Definitions

1. Scoring Average (SA) – the weighted scoring average which takes the stroke average of the field into account. It is computed by adding a player's total strokes to an adjustment and dividing by the total rounds played. The adjustment is computed by determining the stroke average of the field for each round played. This average is subtracted from par to create an adjustment for each round. A player accumulates these adjustments for each round played.
2. Driving Distance (DD)—the average number of yards per measured drive.
3. Driving Accuracy (DA) — the percentage of time a tee shot comes to rest in the fairway.
4. Greens in Regulation (GR)—the percent of time a player was able to hit the green in regulation. Note: A green is considered hit in regulation if any portion of the ball is

touching the putting surface after the GR stroke has been taken. (The GR stroke is determined by subtracting 2 from par (1st stroke on a par 3, 2nd on a par 4, 3rd on a par 5))

5. Strokes Gained-Putting (SGP)—the number of putts a player takes from a specific distance is measured against a statistical baseline to determine the player's strokes gained or lost on a hole.
6. Scrambling (S)—the percent of time a player misses the green in regulation, but still makes par or better.
7. Bounce Back (BB)—the percent of time a player is over par on a hole and then under par on the following hole.
8. Proximity to Hole (3-P)—the average distance the ball comes to rest from the hole (in feet) after the player's approach shot.
9. 3-Putt Average (PH)—the percent of time 3 or more putts were taken for a hole.

I have simply selected the top 20 players with the highest scoring average in March 2014.The following table summarizes the data used for this study:

2014 RANK	PLAYER NAME	Scoring Avg	Driving Distance	Driving Avg.	Greens in Regulation	Strokes Gained-Putting	Scrambling	Bounce Back	3-Putt Average	Proximity to Hole
1	Dustin Johnson	68.766	303.3	60.35%	76.74%	0.607	62.69%	37.04%	3.03	0.338
2	Graeme McDowell	68.987	276.0	69.46%	72.22%	0.646	63.33%	21.74%	3.10	0.445
3	Bubba Watson	69.216	320.8	61.79%	73.61%	0.275	65.26%	20.59%	2.96	0.313
4	Zach Johnson	69.233	281.3	77.51%	75.46%	0.393	63.21%	35.71%	2.89	0.327
5	Harris English	69.278	300.8	61.66%	74.07%	-0.161	65.48%	28.79%	2.91	0.351
6	Webb Simpson	69.311	289.6	65.68%	72.22%	1.401	65.83%	26.09%	2.99	0.352
7	Hideki Matsuyama	69.448	295.9	57.47%	67.93%	0.396	62.99%	29.09%	3.05	0.316
8	Charles Howell III	69.547	299.3	56.31%	73.41%	0.153	68.16%	19.75%	3.02	0.344
9	Will MacKenzie	69.624	294.0	62.48%	69.75%	0.599	59.18%	23.26%	2.95	0.363
10	Jimmy Walker	69.649	301.1	49.06%	70.10%	1.195	62.84%	20.00%	3.01	0.374
11	Graham DeLaet	69.687	303.9	59.62%	73.08%	0.059	56.35%	26.23%	2.98	0.351
12	Matt Every	69.731	283.9	59.34%	69.68%	0.665	60.21%	19.77%	3.00	0.334
13	Chris Stroud	69.768	284.6	65.13%	68.10%	-0.058	66.29%	18.46%	2.96	0.371
14	Hunter Mahan	69.785	293.5	61.65%	68.06%	1.033	57.39%	23.08%	3.07	0.373
15	Kevin Stadler	69.79	296.7	63.67%	69.91%	-0.228	55.90%	25.84%	2.93	0.354
16	Brendon Todd	69.799	281.3	64.84%	70.03%	0.698	62.93%	19.32%	2.96	0.343
17	Russell Knox	69.817	283.9	69.46%	72.38%	-0.014	60.34%	23.08%	2.90	0.313
18	Jason Kokrak	69.831	308.7	54.86%	67.20%	0.131	60.11%	16.46%	3.06	0.358
19	Brian Stuard	69.847	285.4	64.52%	67.78%	0.242	65.52%	25.81%	2.99	0.359
20	K.J. Choi	69.856	284.6	57.63%	69.57%	0.273	59.52%	28.85%	3.00	0.337

3 MODEL

The constrained full model includes all eight explanatory variables. It is defined as:
$Y = \alpha + ß_1X_1 + ß_2X_2 + ß_3X_3 + ß_4X_4 + ß_5X_5 + ß_6X_6 + ß_7X_7 + ß_8X_8$

Y = Scoring Average

α = Intercept

$ß_i$ = Least Squares Coefficients

X_1 = DD

X_2 = DA (%)

X_3 = GR (%)

X_4 = SGP

X_5 = S (%)

X_6 = BB (%)

X_7 = PH

X_8 = 3-P

Rishika Malhotra-Dhir
Regression Analysis
Student Project
Jan – Mar 2014

3.1 MODEL 1

Model I is the constrained full model—it includes all eight explanatory variables. The Microsoft Excel Regression tool provides the following results for this model:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9338
R Square	0.8721
Adjusted R Square	0.7790
Standard Error	0.1489
Observations	20.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8.0000	1.6625	0.2078	9.3722	0.0006
Residual	11.0000	0.2439	0.0222		
Total	19.0000	1.9064			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	85.9325	3.1638	27.1608	0.0000	78.9690
DD	- 0.0093	0.0046	- 2.0360	0.0666	- 0.0194
DA	- 1.5747	0.8950	- 1.7594	0.1063	- 3.5447
GR	- 5.8062	1.8602	- 3.1212	0.0097	- 9.9005
SGP	- 0.0789	0.0886	- 0.8914	0.3918	- 0.2738
S	- 2.4455	1.0695	- 2.2867	0.0430	- 4.7995
BB	- 1.8554	0.8401	- 2.2085	0.0493	- 3.7045
PH	- 1.9324	0.8988	- 2.1501	0.0546	- 3.9106
3-P	- 2.1676	1.5938	- 1.3601	0.2010	- 5.6755

The R^2 for this full model is 0.87. This indicates that 87% of the variation of the golfer's score can be explained by these 8 explanatory variables. However, we would still like to improve on this model by making it simpler by removing unnecessary variables.

The results above show that GR % is a very important factor in predicting a golfer's score. The results also show that SGP is not a good explanatory as it has the highest P-value of .3918, therefore it will be eliminated from the future models. We now move on to Model II, which excludes SGP as an explanatory variable.

3.2 MODEL 2

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9289
R Square	0.8628
Adjusted R Square	0.7828
Standard Error	0.1476
Observations	20.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7.0000	1.6449	0.2350	10.7821	0.0002
Residual	12.0000	0.2615	0.0218		
Total	19.0000	1.9064			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	86.5556	3.0592	28.2940	0.0000	79.8903
DD	-0.0083	0.0044	-1.8899	0.0832	-0.0179
DA	-1.4804	0.8811	-1.6802	0.1187	-3.4002
GR	-6.0738	1.8201	-3.3371	0.0059	-10.0393
S	-2.4800	1.0596	-2.3405	0.0374	-4.7886
BB	-1.8430	0.8328	-2.2131	0.0470	-3.6575
PH	-2.2066	0.8373	-2.6355	0.0218	-4.0308
3-P	-2.1204	1.5792	-1.3427	0.2042	-5.5612

The R^2 for this model is still approximately 0.87 after we have removed the explanatory variable of SGP. The F-Statistic has increased from 9.3722 in the previous model to 10.7821 in this model, which implies that this 7 variable regression model is a better fit.

The results above still show that GR % is a very important factor in predicting a golfer's score. The results also show that 3-P is not a good explanatory as it has the highest P-value of 0.2042, therefore it will be eliminated from the future models.

3.3 MODEL 3

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9177
R Square	0.8422
Adjusted R Square	0.7694
Standard Error	0.1521
Observations	20.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6.0000	1.6056	0.2676	11.5643	0.0001
Residual	13.0000	0.3008	0.0231		
Total	19.0000	1.9064			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	87.4889	3.0698	28.5002	0.0000	80.8571
DD	- 0.0067	0.0044	- 1.5382	0.1480	- 0.0161
DA	- 1.6772	0.8953	- 1.8734	0.0837	- 3.6114
GR	- 6.5327	1.8421	- 3.5464	0.0036	- 10.5122
S	- 2.3110	1.0841	- 2.1317	0.0527	- 4.6530
BB	- 1.4500	0.8034	- 1.8049	0.0943	- 3.1856
PH	- 2.8426	0.7114	- 3.9961	0.0015	- 4.3794

Although the adjusted R square has decreased the F statistic has increased to 11.5643 indicating that the 6 variable model is a better fit.

The results above still show that GR % is a very important factor in predicting a golfer's score. The results also show that DD is not a good explanatory as it has the highest P-value of 0.148, therefore it will be eliminated from the future models.

Rishika Malhotra-Dhir
Regression Analysis
Student Project
Jan – Mar 2014

3.4 MODEL 4

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9019
R Square	0.8135
Adjusted R Square	0.7469
Standard Error	0.1594
Observations	20.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5.0000	1.5508	0.3102	12.2122	0.0001
Residual	14.0000	0.3556	0.0254		
Total	19.0000	1.9064			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	85.1202	2.7821	30.5953	0.0000	79.1531
DA	-0.7729	0.7074	-1.0927	0.2930	-2.2901
GR	-7.7563	1.7406	-4.4560	0.0005	-11.4896
S	-2.1374	1.1296	-1.8922	0.0793	-4.5602
BB	-1.3113	0.8363	-1.5679	0.1392	-3.1051
PH	-2.6539	0.7341	-3.6153	0.0028	-4.2284

Although the adjusted R square has decreased the F statistic has increased to 12.2122 indicating that the 5 variable model is a better fit.

The results above still show that GR % is a very important factor in predicting a golfer's score. It is interesting to note that excluding DD from study suddenly increases the P-value for DA, S, BB and PH but decreases the P-value for GR. This indicates a strong correlation between DD, DA, S, BB and PH. The results also show that DD is not a good explanatory as it has the highest P-value of 0.293, therefore it will be eliminated from the future models.

3.5 MODEL 5

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8931
R Square	0.7976
Adjusted R Square	0.7436
Standard Error	0.1604
Observations	20.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4.0000	1.5205	0.3801	14.7758	0.0000
Residual	15.0000	0.3859	0.0257		
Total	19.0000	1.9064			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	83.9022	2.5655	32.7039	0.0000	78.4340
GR	-7.8798	1.7482	-4.5075	0.0004	-11.6059
S	-2.2366	1.1332	-1.9736	0.0671	-4.6519
BB	-1.4907	0.8254	-1.8061	0.0910	-3.2499
PH	-2.3423	0.6808	-3.4406	0.0036	-3.7934

Although the adjusted R square has decreased the F statistic has increased to 14.7758 indicating that the 4 variable model is a better fit.

The results above still show that GR % is a very important factor in predicting a golfer's score. It is interesting to note that excluding DA from study decreases the P-value for GR, S and BB. This indicates that eliminating DA was a wise choice.

The elimination of Driving Accuracy surprised me. But noting its strong correlation with Greens in Regulation, it seems that Greens in Regulation is more strongly correlated to a low Scoring Average. Similarly, 3-Putt Avoidance is strongly correlated to Strokes Gained-Putting. Strokes Gained-Putting is a fairly modern, academic development in golf. It is regarded as a better way to track putting ability.

This model is very hopeful. Again, although we have seen a decrease in R square, the F-stat has increased significantly. Also, our P-values are all below 10%. We will continue to check our model, but this is our best model so far. We eliminate BB in our next model.

3.6 MODEL 6

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8681
R Square	0.7536
Adjusted R Square	0.7074
Standard Error	0.1714
Observations	20.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3.0000	1.4366	0.4789	16.3081	0.0000
Residual	16.0000	0.4698	0.0294		
Total	19.0000	1.9064			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	84.2875	2.7314	30.8589	0.0000	78.4972
GR	- 9.6098	1.5623	- 6.1509	0.0000	- 12.9218
S	- 1.7654	1.1781	- 1.4984	0.1535	- 4.2629
PH	- 2.2798	0.7264	- 3.1386	0.0063	- 3.8197

Although the adjusted R square has decreased the F statistic has increased to 16.3081 indicating that the 3 variable model is a better fit.

The results above still show that GR % is a very important factor in predicting a golfer's score. It is interesting to note that excluding BB from study decreases the P-value for GR but increases the P-value for S, bringing it to more than 10%.

Given that F statistic has increased we should try a model by eliminating S.

3.7 MODEL 7

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8479
R Square	0.7190
Adjusted R Square	0.6859
Standard Error	0.1775
Observations	20.0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2.0000	1.3707	0.6853	21.7465	0.0000
Residual	17.0000	0.5357	0.0315		
Total	19.0000	1.9064			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	83.8204	2.8112	29.8169	0.0000	77.8894
GR	-10.1964	1.5669	-6.5074	0.0000	-13.5023
PH	-2.3514	0.7509	-3.1314	0.0061	-3.9356

Although the adjusted R square has decreased the F statistic has increased to 21.7465 from 9.3722 from the 8 variable model, indicating that the 2 variable model is a better fit.

The results above indicate that the greens in regulation and 3-putt average are the most important factors in predicting average scores for golf players. This is the best model so far. Again, although we have seen a decrease in R square, the F-stat has increased significantly. Also, our P-values are all below 10%.

Rishika Malhotra-Dhir
Regression Analysis
Student Project
Jan – Mar 2014

4 CONCLUSION

Based on my analysis, the constrained Model 7 is the best model for predicting Scoring Average. The explanatory variables are Greens in Regulation and 3-putt Average. There are, of course, many limitations and drawbacks to the proposed model. For one, it would be of interest to explore lots of combinations of explanatory variables. There are around 100 explanatory variable options on the PGA website.

I was hoping Scrambling and Bounce Back were such variables. They were not. Also, I am not sure that Scoring Average is the best indicator of a player's ability or the outcomes of tournaments. There are other potential options for the response variable.

In conclusion we found our best fit model to be:

$Y = 83.82 - 10.1964X_3 - 2.3514X_8$

Y = Scoring Average

X_3 = Greens in Regulation (%)

X_8 = 3-Putt Average

This shows that the key drivers behind scoring average for the top 20 golfers in 2014 are Green in Regulation % and 3-Putt Average.

Finally, the conclusion of my analysis is a bit of an obvious one: to improve one's score one needs to increase distance off the tee, hit approach shots on the green, and improve putting efficiency.