Jeremiah Reinkoester

10/30/2013

Golf Scoring Averages

Introduction

Prior to this project, I had no overt fondness for golf statistics. I enjoy golf but have neither the time nor money to excel beyond my current, mediocre ability. However, there is a wealth of statistics about golf online. It was truly fascinating to see all the work that has been done to determine the greatest predictors of golfing success. It is a sport that can be broken down into the most minute statistical detail. For example, the PGA calculates and ranks players based on their *left rough tendency,* which they define as “the average score relative to par score when the tee shot comes to rest in the left rough (regardless of club) and the distance of the drive is determined with a laser.” With so many explanatory variable options, I could not resist.

Background Information and Data

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 am assuming the reader has an elementary understanding of the game of golf. I found all of my data and information at http://www.pgatour.com/stats.html. For all of the data I used and the calculations, see the attached excel sheet titled “Golf Stats Regression”. The first major difficulty I encountered was choosing my explanatory variables. The PGA Tour lists around one hundred various statistics for determining a player’s ability. I try to choose a mix of variables, a few seemingly obvious variables, like driving distance and driving accuracy, as well as a few unexpected variables. I was hoping an odd variable would strongly correlate scoring average. The following is a list of my variables, along with definitions and acronyms:

- 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.
- Driving Distance (DD)—the average number of yards per measured drive.

- Driving Accuracy (DA) — the percentage of time a tee shot comes to rest in the fairway.
- 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))
- 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.
- Scrambling (S)—the percent of time a player misses the green in regulation, but still makes par or better.
- Bounce Back (BB)—the percent of time a player is over par on a hole and then under par on the following hole.
- Proximity to Hole (PH)—the average distance the ball comes to rest from the hole (in feet) after the player's approach shot.
- 3-Putt Average (3-P)—the percent of time 3 or more putts were taken for a hole.

Note: I will use the acronyms quite frequently throughout the paper.

There were statistics on roughly 140 players. I chose 20 players for my analysis. It would be better to use more players, but manually logging the statistics was quite time consuming. The following are the statistics for the players with the top 20 scoring averages in 2013:

2013	SA	DD	DA	GR	SGP	S	BB	PH	3-P
Steve Stricker	68.95	283.60	70.65	71.16	0.73	65.57	22.22	33.83	3.07
Tiger Woods	68.99	293.20	62.50	67.59	0.42	60.00	19.46	33.92	2.87
Justin Rose	69.27	296.60	63.57	68.89	-0.19	60.71	22.82	32.17	3.70
Henrik Stenson	69.29	290.90	70.09	71.96	0.01	57.28	22.08	35.75	3.51
Adam Scott	69.34	297.80	61.84	68.80	0.00	56.38	15.00	34.08	2.59
Sergio Garcia	69.58	291.00	61.28	67.46	0.61	57.45	26.38	35.92	2.38
Matt Kuchar	69.59	284.90	58.93	65.84	0.40	63.55	17.09	33.92	1.85
Charl Schwartzel	69.69	296.10	59.87	65.85	0.30	55.02	22.92	34.08	2.29
Jordan Spieth	69.70	289.40	67.80	66.94	0.18	61.07	20.09	32.92	3.18
Keegan Bradley	69.75	300.60	62.82	66.54	0.25	60.88	22.27	35.25	2.36
Jason Day	69.76	299.30	58.03	64.93	0.37	61.39	15.12	36.33	2.71
Phil Mickelson	69.77	287.90	57.30	66.67	0.66	58.55	22.34	35.17	2.92
Webb Simpson	69.81	285.40	63.30	66.67	0.31	57.95	24.17	34.00	2.71
Brandt Snedeker	69.82	281.30	62.57	65.68	0.69	60.86	22.27	33.17	2.74
Luke Donald	69.84	278.10	62.87	62.16	0.53	61.52	22.58	34.08	2.39
Jim Furyk	69.86	275.30	70.47	68.30	0.09	59.55	18.32	31.25	2.78
Jason Dufner	69.94	285.90	64.81	67.53	-0.23	59.11	20.93	34.17	4.26
Bill Haas	70.05	288.20	62.31	67.79	0.26	62.01	21.00	35.42	2.45
Zach Johnson	70.10	278.80	69.68	68.14	0.37	59.66	23.22	33.50	3.01
Freddie Jacobson	70.11	287.30	56.26	61.67	0.45	59.66	20.90	34.50	2.78

Correlation

I first study the correlation between the explanatory variables. As the exhibit below indicates, there are a few variables with strong correlation. Greens in Regulation and Driving Accuracy have a 0.68 correlation. Also, 3-Putt Avoidance and Strokes Gained-Putting have a negative correlation of -0.55. This is to be expected as putting ones drive on the fairway likely improves the ability to put shots on the green, and 3-Putt Avoidance and Strokes Gained-Putting are both measures of ones putting ability. As one will see, the strongest model does not include these pairs.

	DD	DA	GR	SGP	S	BB	PH	3-P
DD	1.00							
DA	-0.41	1.00						
GR	0.10	0.68	1.00					
SGP	-0.28	-0.26	-0.33	1.00				
S	-0.28	0.16	-0.03	0.31	1.00			
BB	-0.20	0.16	0.05	0.26	-0.17	1.00		
PH	0.37	-0.36	-0.01	0.25	-0.21	0.32	1.00	
3-P	-0.03	0.44	0.42	-0.55	-0.06	0.12	-0.20	1.00

Model I

Model I is the constrained full model—it includes all eight explanatory variables. Using the Regression Add-In in Excel, I have the following regression analysis:

Model I

Regression Statistics	
Mult. R	0.79624
R^2	0.633997
Adj. R^2	0.367814
S.E.	0.268014
Obs.	20

ANOVA

	df	*SS*	*MS*	*F*	*Sign. F*
Regression	8	1.368709	0.171089	2.381804	0.091626
Residual	11	0.790147	0.071832		
Total	19	2.158855			

	Coeff.	*S.E.*	*t Stat*	*P-value*
Intercept	80.95816	4.288114	18.87967	9.91E-10
DD	-0.02847	0.01331	-2.13872	0.055729
DA	0.00166	0.028251	0.058754	0.954202
GR	-0.08723	0.042835	-2.03649	0.066505
SGP	-0.66805	0.370932	-1.80101	0.099149
SGP	-0.01217	0.030131	-0.40392	0.694015
BB	0.014901	0.025881	0.575749	0.576374
PH	0.10271	0.06389	1.607593	0.136225
3-P	-0.08079	0.155133	-0.52078	0.612844

We can see our Adjusted R^2 is not very high at 0.367814. Also, several of the coefficients' P-values are high. It would seem that eliminating DA might improve the model. This leads us to Model II.

Model II

Model II

Regression Statistics	
Mult. R	0.796167
R^2	0.633883
Adj. R^2	0.420314
S.E.	0.256644
Obs.	20

ANOVA

	df	*SS*	*MS*	*F*	*Sign. F*
Regression	7	1.368461	0.195494	2.968052	0.047219
Residual	12	0.790395	0.065866		
Total	19	2.158855			

	Coeff.	*S.E.*	*t Stat*	*P-value*
Intercept	81.05073	3.819001	21.22302	6.97E-11
DD	-0.02888	0.010843	-2.66319	0.020669
GR	-0.08534	0.026994	-3.16137	0.0082
SGP	-0.67348	0.344011	-1.95772	0.073926
S	-0.01177	0.028107	-0.41877	0.682779
BB	0.015219	0.024233	0.628032	0.541751
PH	0.101971	0.059984	1.699973	0.114885
3-P	-0.08063	0.148531	-0.54288	0.597149

We can see that R^2 has not decreased much, which is good, and Adjusted R^2 has increased quite a bit. Although our P-values are improved it might be beneficial to remove the highest one, which is S. We will do this in our next model. Also, note that our F-stat has increased—all of these things are signs that this model is better than the original.

Model III

Model III

Regression Statistics	
Mult. R	0.7928
R^2	0.628532
Adj. R^2	0.457085
S.E.	0.248371
Obs.	20

ANOVA

	df	*SS*	*MS*	*F*	*Sign. F*
Regression	6	1.35691	0.226152	3.666047	0.023564
Residual	13	0.801946	0.061688		
Total	19	2.158855			

	Coeff.	*S.E.*	*t Stat*	*P-value*
Intercept	80.10787	2.985293	26.83417	9.06E-13
DD	-0.02828	0.010403	-2.71863	0.017559
GR	-0.08623	0.026042	-3.31127	0.005624
SGP	-0.72988	0.306345	-2.38254	0.03315
BB	0.018852	0.021898	0.860877	0.404904
PH	0.104861	0.057665	1.81846	0.092099
3-P	-0.09192	0.141358	-0.65024	0.526861

Again, we see improvement in our Adjusted R^2 and F-stat, both signs that our model is improving. We continue to see if we can improve upon this model by removing the coefficient, 3-P, with the highest P-value—an indication that this coefficient might be 0.

Model IV

Model IV

Regression Statistics	
Mult. R	0.785143
R^2	0.61645
Adj. R^2	0.479468
S.E.	0.243197
Obs.	20

ANOVA

	df	*SS*	*MS*	*F*	*Sign. F*
Regression	5	1.330827	0.266165	4.500227	0.011792
Residual	14	0.828028	0.059145		
Total	19	2.158855			

	Coeff.	*S.E.*	*t Stat*	*P-value*
Intercept	79.89577	2.905606	27.49711	1.39E-13
DD	-0.02697	0.009992	-2.69883	0.017296
GR	-0.09113	0.024411	-3.73311	0.002227
SGP	-0.62238	0.252533	-2.46455	0.027269
BB	0.015031	0.020656	0.727709	0.478786
PH	0.103326	0.056416	1.831489	0.088387

Again, our Adjusted R^2 and F-stat are greater without significantly lowering R^2. It would seem that BB is now the obvious choice to be removed. We will do this in Model V. At this point, we have removed Driving Accuracy, Scrambling, 3-Putt Avoidance, and now Bounce Back. Scrambling and Bounce Back were rather cool statistics that I hoped would be strong indicators of a player's Scoring Average. However, I suspect the top players in the world are able to play fairly consistent golf, thereby avoiding the necessity for too much Scrambling and Bounce Back. 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. There are quite a few articles online regarding the subject. Here is a link to Professor Mark Broadie's paper on the subject: http://www.columbia.edu/~mnb2/broadie/Assets/putting_strokes_gained_20110113.pdf. Here is a concise explanation on the PGA Tour's website: http://www.pgatour.com/stats/academicdata/shotlink.html.

Model V

Model V

Regression Statistics	
Mult. R	0.775849
R^2	0.601942
Adj. R^2	0.495794
S.E.	0.239353
Obs.	20

ANOVA

	df	*SS*	*MS*	*F*	*Sign. F*
Regression	4	1.299506	0.324877	5.670743	0.005509
Residual	15	0.859349	0.05729		
Total	19	2.158855			

	Coeff.	*S.E.*	*t Stat*	*P-value*
Intercept	80.23604	2.822406	28.42825	1.84E-14
DD	-0.02821	0.009689	-2.91175	0.010736
GR	-0.08831	0.02372	-3.7229	0.002041
SGP	-0.58634	0.243714	-2.40584	0.029487
PH	0.107284	0.055266	1.941242	0.071257

This model is very hopeful. Again we have seen improvement in Adjusted R^2 and the F-stat. Also, our P-values are all below 10%. We will check one more model, but it will be clear that this is our best model. We eliminate PH in our next model.

Model VI

Model VI

Regression Statistics	
Mult. R	0.708476
R^2	0.501939
Adj. R^2	0.408552
S.E.	0.259235
Obs.	20

ANOVA

	df	*SS*	*MS*	*F*	*Sign. F*
Regression	3	1.083613	0.361204	5.374858	0.009428
Residual	16	1.075242	0.067203		
Total	19	2.158855			

	Coeff.	*S.E.*	*t Stat*	*P-value*
Intercept	80.75609	3.043043	26.53794	1.18E-14
DD	-0.0174	0.008588	-2.02645	0.059724
GR	-0.08885	0.025688	-3.45859	0.003234
SGP	-0.38327	0.238406	-1.60762	0.127471

We see a clear decrease in all of our indicators, R^2, Adjusted R^2, and F. Also, our P-Value for SGP has jumped up above 10%. It is interesting to note how the removal of Proximity to Hole has caused the Strokes Gained-Putting variable to dramatically reduce the likelihood of its necessity in the model.

At this point it seems evident that Model V—which includes Driving Distance, Greens in Regulation, Strokes Gained-Putting, and Proximity to Hole—is the best of the presented models. If one is interested, I continued eliminating the explanatory variable with the highest P-value. This led to Greens in Regulation as the only explanatory variable with a P-Value of 0.007374. The F-stat is 9.114441, but the Adjusted R^2 is only 0.299266, indicating that this is not as good a fit as Model V. Not enough of the variability in the data set is accounted for by this model. With only one variable, seemingly correlated to the response variable, I would expect the large F-Stat.

F-Test

It is of interest to determine if our constrained model is adequate for different groups of golfers. Using an F-Test, we will determine if the simple model is a better model than the unconstrained model for different sets of golfers.

In our data set of the golfers with the top 20 scores, we will compare the even ranked golfers with the odd ranked golfers. I would not suspect that we would need the unconstrained model in this situation.

To be explicit, let

$$Y = \beta+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4$$

be the model describing the Scoring Average of the even ranked golfers and let

$$Y = \gamma+\gamma_1X_1+\gamma_2X_2+\gamma_3X_3+\gamma_4X_4$$

be the model for the odd ranked golfers.

Also, let

$$Y = \beta+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+(\gamma-\beta)D + (\gamma_1-\beta_1)X_1D+(\gamma_2-\beta_2)X_2D+(\gamma_3-\beta_3)X_3D+(\gamma_4-\beta_4)X_4D \quad *$$

be the unconstrained model, where D is a dummy variable. Notice that when D=1, we get the odd model and when D=0 we get the even model. The following table better explains our variables:

2013	SA	DD	GR	SGP	PH	D	D*X1	D*X2	D*X3	D*X4
Steve Stricker	68.95	283.60	71.16	0.73	33.83	1	283.60	71.16	0.73	33.83
Tiger Woods	68.99	293.20	67.59	0.42	33.92	0	0.00	0.00	0.00	0.00
Justin Rose	69.27	296.60	68.89	-0.19	32.17	1	296.60	68.89	-0.19	32.17
Henrik Stenson	69.29	290.90	71.96	0.01	35.75	0	0.00	0.00	0.00	0.00
Adam Scott	69.34	297.80	68.80	0.00	34.08	1	297.80	68.80	0.00	34.08
Sergio Garcia	69.58	291.00	67.46	0.61	35.92	0	0.00	0.00	0.00	0.00
Matt Kuchar	69.59	284.90	65.84	0.40	33.92	1	284.90	65.84	0.40	33.92
Charl Schwartzel	69.69	296.10	65.85	0.30	34.08	0	0.00	0.00	0.00	0.00
Jordan Spieth	69.70	289.40	66.94	0.18	32.92	1	289.40	66.94	0.18	32.92
Keegan Bradley	69.75	300.60	66.54	0.25	35.25	0	0.00	0.00	0.00	0.00
Jason Day	69.76	299.30	64.93	0.37	36.33	1	299.30	64.93	0.37	36.33
Phil Mickelson	69.77	287.90	66.67	0.66	35.17	0	0.00	0.00	0.00	0.00
Webb Simpson	69.81	285.40	66.67	0.31	34.00	1	285.40	66.67	0.31	34.00
Brandt Snedeker	69.82	281.30	65.68	0.69	33.17	0	0.00	0.00	0.00	0.00
Luke Donald	69.84	278.10	62.16	0.53	34.08	1	278.10	62.16	0.53	34.08
Jim Furyk	69.86	275.30	68.30	0.09	31.25	0	0.00	0.00	0.00	0.00
Jason Dufner	69.94	285.90	67.53	-0.23	34.17	1	285.90	67.53	-0.23	34.17
Bill Haas	70.05	288.20	67.79	0.26	35.42	0	0.00	0.00	0.00	0.00
Zach Johnson	70.10	278.80	68.14	0.37	33.50	1	278.80	68.14	0.37	33.50
Freddie Jacobson	70.11	287.30	61.67	0.45	34.50	0	0.00	0.00	0.00	0.00

Using the Regression Add-In in excel, I calculated the following:

	RSS	RegSS	TSS
Odd Model	0.342	0.7648	1.1068
Even Model	0.365	0.6684	1.0334
Unconstrained	0.7069	1.4519	2.1589
Constrained	0.8593	1.2995	2.1589

Notice that we already have listed these statistics for the constrained model in the Model V analysis. To make sure I was calculating the F-Test correctly, I did it two ways. The first way utilized the even, odd, and constrained model, as in the illustrative worksheets. The calculation was

$$F = \frac{RSS_C - RSS_{Odd} - RSS_{Even}}{RSS_{Odd} + RSS_{Even}} \times \frac{20-10}{5} = 0.431202$$

The second way utilized the constrained and unconstrained models as follows:

$$F = \frac{Reg_U - Reg_C}{Res_U} \times \frac{20-10}{5} = 0.431202$$

Notice that n=20 (number of players), k+1 = 10 (number of variables in the unconstrained equation—see *), and q = 5 (number of restrictions in the null hypothesis or number of coefficients in the unconstrained model). To see the full analysis see the “F-Test” tabs in the attached spreadsheet. Finally, the P-value was FDIST(0.431202, 5, 10) = 0.817136. This clearly indicates that the null hypothesis should be accepted. In other words, the coefficients for the variables involving D should be 0. So the simpler, constrained model is the better choice. To put it another way, both the odd and even players can be modeled with the same regression model.

I also ran an F-Test comparing the top ten and bottom ten players. The F-Stat was not as favorable in this analysis. I calculated an F-Stat of 2.999628 and a P-Value of 0.065578. The null hypothesis would still be accepted at the 5% level. I was not sure if I should normalize the explanatory variables. I attempted to do this on the “F-Test Normal” tab. This resulted in an even higher F-Stat. I may have incorrectly normalized, or it could just be that the simpler model is not as good when breaking the groups from top to bottom instead of even and odd.

Conclusion

Based on my analysis, the constrained Model V is the best model for predicting Scoring Average. The explanatory variables are Driving Distance, Greens in Regulation, Strokes Gained-Putting, and Proximity to Hole. The F-Test further indicates that Model V is a decent model for predicting Scoring Average, as breaking the data into different groups still results in a preference for the simpler model. There are, of course, many limitations and drawbacks to my model. For one, it would be of interest to explore lots of combinations of explanatory variables. As I mentioned, there are around 100 explanatory variable options on the PGA website. It would be cool to find some obscure variable that is strongly correlated to a player’s scores. 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. 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.