Abstract

The purpose of this research is to find the solution to one simple question in the realm of golf: what is necessary to win at the highest level of play? And if not win, per say, then what stats create the most consistent results when playing at the highest level?

In order to do this, we need to establish what our definition of 'winning', and/or 'being consistent'. For the sake of making our results more consistent to reproduce, we will expand the scope of 'winning' to be the same as 'being consistent'. This is because unlike the standard team sport, golf is individually based. This means that when competing in a tournament field, which can have up to 200 people, winning can be difficult. So, our parameter for 'being consistent' will be measured by, among other things, the players holding a substantial amount of 'Top 10 Finishes' at the end of each season; ranked by the Official World Golf Ranking list for each year's (sample year) top 20 ranked individuals. A more thorough explanation of our methods and procedures will be explained below accordingly.

The results of our estimated model show that the key to 'being consistent' on the PGA tour lies not in a player's ability to accumulate 'Top 10 Finishes' but rather to maintain high percentages in 'Scrambling', 'Putting Inside 5ft', 'Sand Saves', 'GIR', and a low percentage in '3 Putt Avoidance' relative to score.

Introduction

How do we accurately predict what golf stats are required to be a top tier PGA tour player? What sort of variables can we use to gage how good one must be to be a winner on the PGA tour? Regardless of outside variables such as weather conditions, course conditions and the like, is this estimator consistent?

The prediction for what it takes to be a winner on the PGA and a consistent player at the highest level will most likely have to do with how well this player plays compared to the rest of the field. This player must be better than the tour's average in the golf stats that matter to be not only a winner on tour, but overall, a top tiered player.

As it turns out from our findings, our original definition was not displayed in the interpretation of the data given.

Materials and Methods

All data information is provided free courtesy of the PGA Tour's official stats website. From here we were able to abstract potential variable statistics such as 'SG:Total', 'GIR', 'Driving Accuracy', etc. to add to our spreadsheet in Excel. All Data Analysis is conducted through Excel's "Data Analysis" pack extension.

We must first go ahead and define all the parameter(s) and statistics used in any of our tests for purposes of clarity for the results of our findings. Every percentage statistic is represented in our spreadsheet from a value between 0 and 1.

Top 10s: The total amount of times a player has finished a tournament at or within the top 10 on the leaderboard throughout one sample year (season). (PGA Tour)

OWGR: The golf ranking attributed to each player sampled by the International Federation of PGA Tours. "The Official World Golf Ranking, which is endorsed by the four major championships and the five professional tours which make up the International Federation of PGA Tours, is issued every Monday, following completion of the previous week's tournaments from around the world. This statistic is the average number of points earned per event in the last 104 weeks. These points are awarded based upon finish position as well as the strength of the field. The points are initially worth double their original value and decline gradually over this two-year period. There are 8 13-week periods and points decline by .25 times their value each period." (PGA Tour)

GIR – Greens in Regulation: A percentage attributed to the calculated number of greens that have been hit in regulation over all greens (or all holes played) for a particular season. Regulation meaning the expected number of strokes required to get the ball on the putting surface (or green). Expected Values are 1 (*or* less) stroke on a Par 3, 2 (*or* less) strokes on a Par 4, and 3 (*or* less) strokes on a Par 5. If the value for the hole's GIR stroke is higher than the expected, it counts as a missed GIR. If the value of our GIR Stroke is less than the expected, it is counted as a GIR and GUR percentage. (PGA Tour)

Scrambling: A binomial golf statistic that is defined as the percentage of times a player misses the green in regulation, but still scores a par or better. For example, if on the first shot of a par three, a player misses the green, and then proceeds to chip on and one putt, this counts as a successful scrambling attempt. These may also be referred to as 'up and downs.' (PGA Tour)

Putting Inside 5ft: A binomial statistic that is defined as the percentage of times a player makes a putt inside 5 feet. (PGA Tour)

Scoring: A statistic that counts (on average) how many strokes per round completed for a specific player. (PGA Tour)

SG: AVG: [Shots Gained: Average *or* Shots Gained: Total] "The per round average number of strokes the player was better or worse than the field of players on the same course and event (for the whole sample year)" for all the shots on the course. It is the sum of SG: OTT, SG: APR, and SG: ARG. (PGA Tour)

SG: OTT: [Shots Gained: Off The Tee] " " for all strokes off of the tee box (the tee shot). (PGA Tour)

SG: APR: [Shots Gained: Approach] " " for all strokes approaching the green. (PGA Tour)

SG: ARG: [Shots Gained: Around the Green] " " for all stokes around the green; including: greenside bunkers, the fringe, and the putting surface. (PGA Tour)

3-Putt Avoidance: The percent of time 3 or more putts were taken for a particular hole (total 3-putts, 4-putts, etc./ total holes played). (PGA Tour)

Driving Distance: The average number of yards per measured drive. These drives are measured on two holes per round. Care is taken to select two holes which face in opposite directions to counteract the effect of wind. Drives are measured to the point at which they come to rest regardless of whether they are in the fairway or not. (PGA Tour)

Driving Accuracy: The percentage of times the tee shot came to rest in the fairway regardless of the club chosen. (PGA Tour)

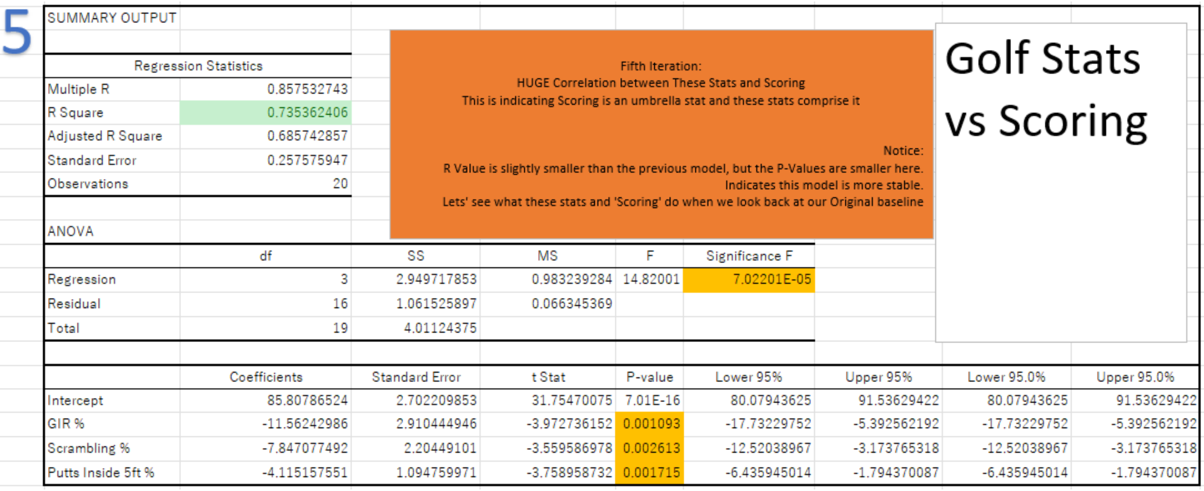
Sand Saves: The percentage given to a player who was able to successfully get the ball 'up and down' out of a greenside bunker within 2 strokes or less from that point, regardless of score. Example: The ball rests in a greenside bunker of a par 4 in 3 strokes and is hit in the hole after 2 more strokes. Even though this is a failed Scramble attempt, it still counts as a successful Sand Save. (PGA Tour)

Proximity to the Hole: The average distance a ball comes to rest from the hole (in feet) after the player's approach shot. The shot must not originate from on or around the green (within 30 yards or closer) and is measured by laser. (PGA Tour)

The procedures are described as follows. What we want to do is create a model from one sample year as accurately as possible and see if it can be applied to the other years with similar results.

First, we take the data information from a sample year and find the Top 20 individuals from that sample according to the OWGR. Then, lookup the amount of ‘Top 10 Finishes’ attributed to that sample player for all players, then the same step is repeated using the other data sheets for each golf statistic for the sample year for all players (GIR, Scrambling, etc.). Then, we input the Tour Average for that statistic below its respective statistic in the spreadsheet and calculate all our sample’s averages and standard deviations for each statistic. Although this is not necessary for a multi-regression analysis, it is useful for contrasting our sample to ‘the field’ (the average for the Tour Players for that statistic) for later use in interpretation. After running the ‘Regression’ function within the ‘Data Analysis’ tool pack, with the Dependent being ‘Top10s’ and the Independent Variables being ‘GIR’, ‘Scrambling’, ‘Putts Inside 5ft’, ‘Scoring’, ‘SG: AVG’, ‘SG: OTT’, ‘SG: APR’, and ‘SG: ARG’, we are given a table with the associated R-Squared value and the P-Values of our statistics used. We are also given a Significance F ‘p-value’ that indicates the significance of the model ran.

Using this information, we then start narrowing down through several attempts (Tables 1 – 9 in ‘Regression Models for 2018’) to find the proper set up for our parameter variable and independent variables. The best model that was found was the Fifth Iteration (Table 5) which showed very low p-values for our variables and model with the highest R-Squared model and can be found below. This was then applied when adding more variables to the model. In particular, the SG statistics in Table H, however we noticed that all ‘SG’ stats besides ‘SG: AVG’ gave errors in the p-value estimator. This makes sense and will be discussed in the discussion section. After this, using Table E as our model reference, we added some other statistics to test their validity in the model. The variables added include: ‘Driving Accuracy’, ‘Driving Distance’, ‘3-Putt Avoidance’, and ‘Proximity to the Hole’.



5

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.857532743
R Square	0.735362406
Adjusted R Square	0.685742857
Standard Error	0.257575947
Observations	20

Fifth Iteration:
HUGE Correlation between These Stats and Scoring
This is indicating Scoring is an umbrella stat and these stats comprise it

Notice:
R Value is slightly smaller than the previous model, but the P-Values are smaller here.
Indicates this model is more stable.
Lets' see what these stats and 'Scoring' do when we look back at our Original baseline

Golf Stats vs Scoring

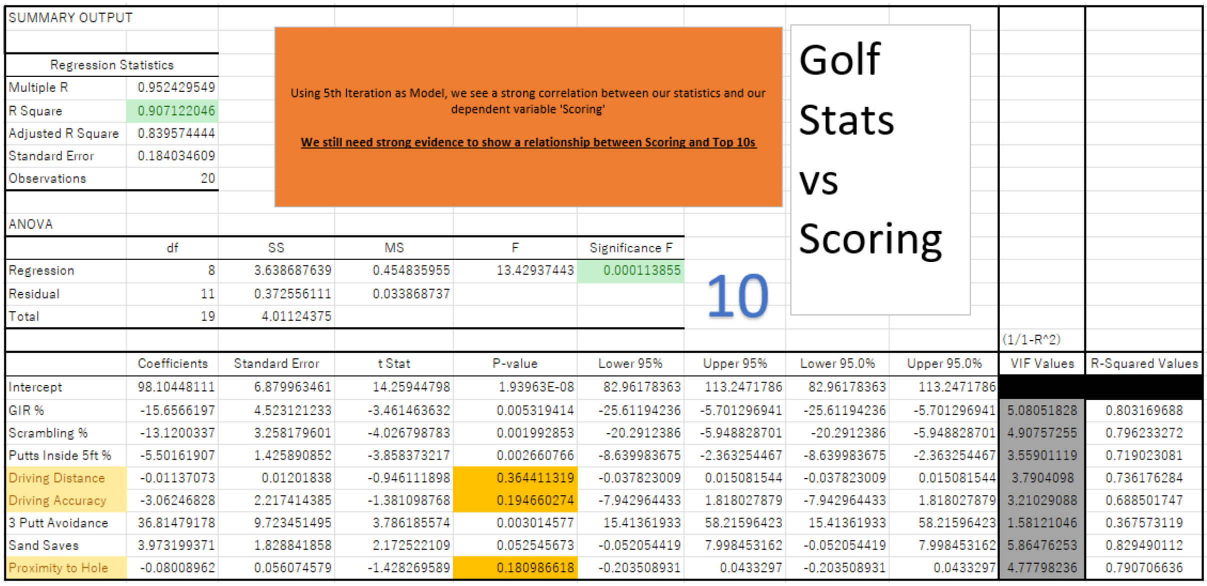
ANOVA

	df	SS	MS	F	Significance F
Regression	3	2.949717853	0.983239284	14.82001	7.02201E-05
Residual	16	1.061525897	0.066345369		
Total	19	4.01124375			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	85.80786524	2.702209853	31.75470075	7.01E-16	80.07943625	91.53629422	80.07943625	91.53629422
GIR %	-11.56242986	2.910444946	-3.972736152	0.001093	-17.73229752	-5.392562192	-17.73229752	-5.392562192
Scrambling %	-7.847077492	2.20449101	-3.559586978	0.002613	-12.52038967	-3.173765318	-12.52038967	-3.173765318
Putts Inside 5ft %	-4.115157551	1.094759971	-3.758958732	0.001715	-6.435945014	-1.794370087	-6.435945014	-1.794370087

After running the model again (Table 10, 2018 Regressions) we see high p-values in some of the new variables added while maintaining a good Significance F value and see a great increase in our R-Squared value (to about 90%) so we decide to run a ‘Correlation’ Table Test from the ‘Data Analysis’ Tool pack to see how much these variables are related to one another (Table 11, 2018 Regressions). In combination with a Variance Inflation Factor calculation

(Table Set 1, VIFs 2018) for our variables, we weed out variables that make the model unstable and arrive at our most stable model (Table 12, 2018 Regressions) with the highest R-Squared value and lowest VIF values (Table Set 2, VIFs 2018), relatively, thus far. We try to see if the 'SG: AVG' statistic is a stable statistic by adding it to this model (Table 13, 2018 Regressions) but from the results of our VIF test afterwards on Table M (Table Set 3, VIFs 2018) not having 'SG: AVG' in our model is much more stable; even if it is the case that with it we can see a slightly higher R-Squared value it is not really worth the trade-off of including it in the calculation.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.952429549
R Square	0.907122046
Adjusted R Square	0.839574444
Standard Error	0.184034609
Observations	20

Using 5th Iteration as Model, we see a strong correlation between our statistics and our dependent variable 'Scoring'

We still need strong evidence to show a relationship between Scoring and Top 10s

Golf Stats vs Scoring

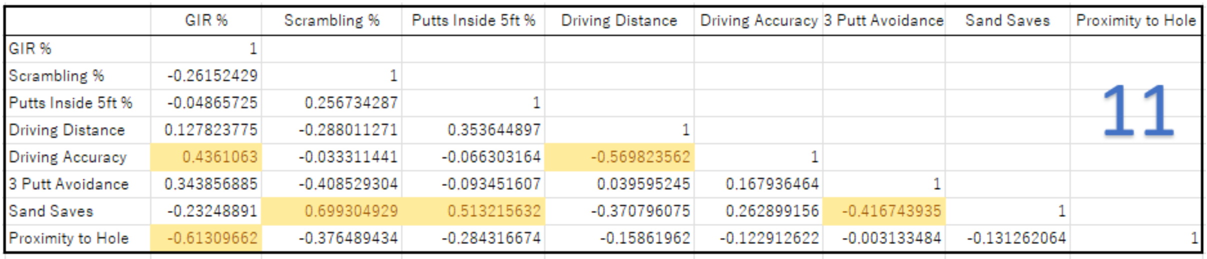
ANOVA

	df	SS	MS	F	Significance F
Regression	8	3.638687639	0.454835955	13.42937443	0.000113855
Residual	11	0.372556111	0.033868737		
Total	19	4.01124375			

10

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%	(1/1-R^2) VIF Values	R-Squared Values
Intercept	98.10448111	6.879963461	14.25944798	1.93963E-08	82.96178363	113.2471786	82.96178363	113.2471786		
GIR %	-15.6566197	4.523121233	-3.461463632	0.005319414	-25.61194236	-5.701296941	-25.61194236	-5.701296941	5.08051828	0.803169688
Scrambling %	-13.1200337	3.258179601	-4.026798783	0.001992853	-20.2912386	-5.948828701	-20.2912386	-5.948828701	4.90757255	0.796233272
Putts Inside 5ft %	-5.50161907	1.425890852	-3.858373217	0.002660766	-8.639983675	-2.363254467	-8.639983675	-2.363254467	3.55901119	0.719023081
Driving Distance	-0.01137073	0.01201838	-0.946111898	0.364411319	-0.037823009	0.015081544	-0.037823009	0.015081544	3.7904098	0.736176284
Driving Accuracy	-3.06246828	2.217414385	-1.381098768	0.194660274	-7.942964433	1.818027879	-7.942964433	1.818027879	3.21029088	0.688501747
3 Putt Avoidance	36.81479178	9.723451495	3.786185574	0.003014577	15.41361933	58.21596423	15.41361933	58.21596423	1.58121046	0.367573119
Sand Saves	3.973199371	1.828841858	2.172522109	0.052545673	-0.052054419	7.998453162	-0.052054419	7.998453162	5.86476253	0.829490112
Proximity to Hole	-0.08008962	0.056074579	-1.428269589	0.180986618	-0.203508931	0.0433297	-0.203508931	0.0433297	4.77798236	0.790706636

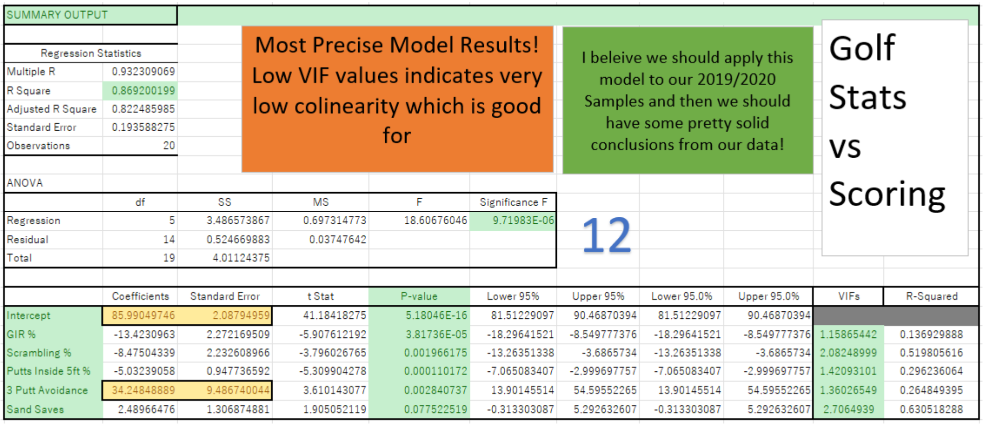
As you can see by adding some new variables above, 3 out of 5 of them had unstable P-values.



	GIR %	Scrambling %	Putts Inside 5ft %	Driving Distance	Driving Accuracy	3 Putt Avoidance	Sand Saves	Proximity to Hole
GIR %	1							
Scrambling %	-0.26152429	1						
Putts Inside 5ft %	-0.04865725	0.256734287	1					
Driving Distance	0.127823775	-0.288011271	0.353644897	1				
Driving Accuracy	0.4361063	-0.033311441	-0.066303164	-0.569823562	1			
3 Putt Avoidance	0.343856885	-0.408529304	-0.093451607	0.039595245	0.167936464	1		
Sand Saves	-0.23248891	0.699304929	0.513215632	-0.370796075	0.262899156	-0.416743935	1	
Proximity to Hole	-0.61309662	-0.376489434	-0.284316674	-0.15861962	-0.122912622	-0.003133484	-0.131262064	1

11

We will reference Table 11 in the discussion.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.932309069
R Square	0.869200199
Adjusted R Square	0.822485985
Standard Error	0.193588275
Observations	20

Most Precise Model Results! Low VIF values indicates very low colinearity which is good for

I beleive we should apply this model to our 2019/2020 Samples and then we should have some pretty solid conclusions from our data!

Golf Stats vs Scoring

ANOVA

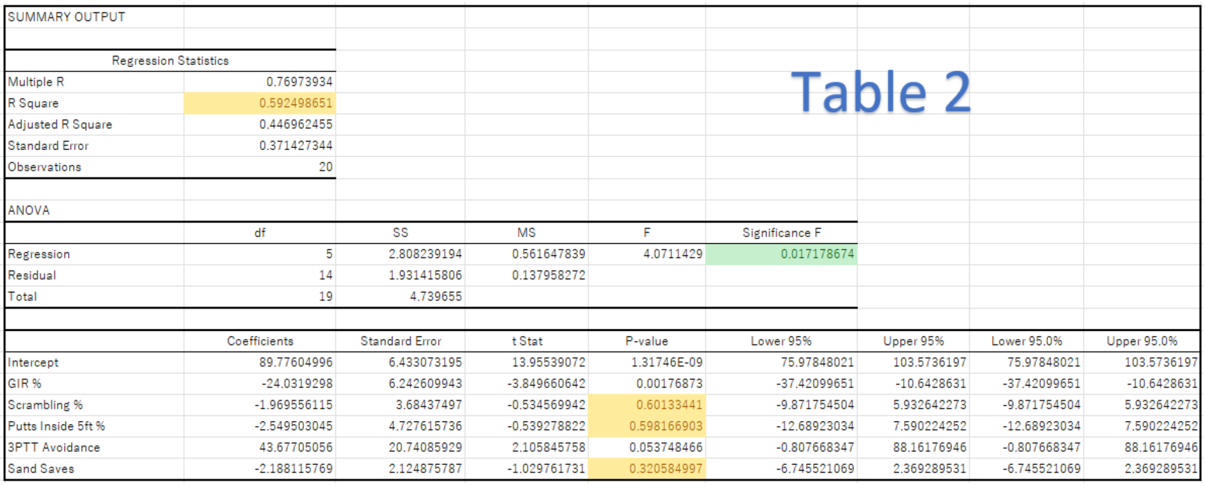
	df	SS	MS	F	Significance F
Regression	5	3.486573867	0.697314773	18.60676046	9.71983E-06
Residual	14	0.524669883	0.03747642		
Total	19	4.01124375			

12

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%	VIFs	R-Squared
Intercept	85.99049746	2.08794959	41.18418275	5.18046E-16	81.51229097	90.46870394	81.51229097	90.46870394		
GIR %	-13.4230963	2.272169509	-5.907612192	3.81736E-05	-18.29641521	-8.549777376	-18.29641521	-8.549777376	1.15865442	0.136929888
Scrambling %	-8.47504339	2.232608966	-3.796026765	0.001966175	-13.26351338	-3.6865734	-13.26351338	-3.6865734	2.08248999	0.519805616
Putts Inside 5ft %	-5.03239058	0.947736592	-5.309904278	0.000110172	-7.065083407	-2.999697757	-7.065083407	-2.999697757	1.42093101	0.296236064
3 Putt Avoidance	34.24848889	9.486740044	3.610143077	0.002840737	13.90145514	54.59552265	13.90145514	54.59552265	1.36026549	0.264849395
Sand Saves	2.48966476	1.306874881	1.905052119	0.077522519	-0.313303087	5.292632607	-0.313303087	5.292632607	2.7064939	0.630518288

After taking out the variables with high P-values from Table 10, the Table prior is our most precise table from our model, and it also officially shifts our focus from our original hypothesis of Top 10s being the dependent variable indicator to directly how these independent variables affect the Score (in strokes) of a golfer for any given round.

Now applying the model to our other two sample years, we get their respective results as seen below.



SUMMARY OUTPUT

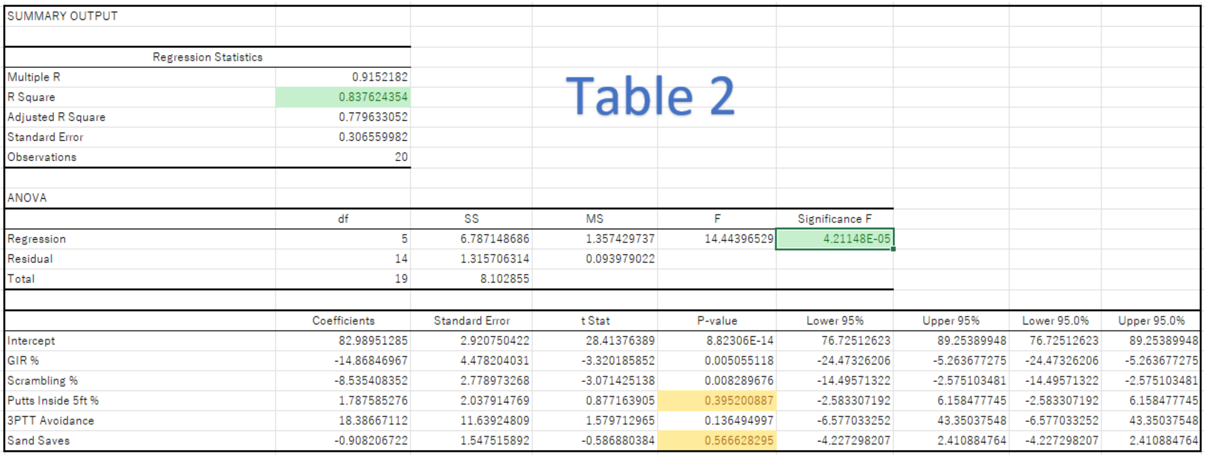
Regression Statistics	
Multiple R	0.76973934
R Square	0.592498651
Adjusted R Square	0.446962455
Standard Error	0.371427344
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	5	2.808239194	0.561647839	4.0711429	0.017178674
Residual	14	1.931415806	0.137958272		
Total	19	4.739655			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	89.77604996	6.433073195	13.95539072	1.31746E-09	75.97848021	103.5736197	75.97848021	103.5736197
GIR %	-24.0319298	6.242609943	-3.849660642	0.00176873	-37.42099651	-10.6428631	-37.42099651	-10.6428631
Scrambling %	-1.969556115	3.68437497	-0.534569942	0.60133441	-9.871754504	5.932642273	-9.871754504	5.932642273
Putts Inside 5ft %	-2.549503045	4.727615736	-0.539278822	0.598166903	-12.68923034	7.590224252	-12.68923034	7.590224252
3PTT Avoidance	43.67705056	20.74085929	2.105845758	0.053748466	-0.807668347	88.16176946	-0.807668347	88.16176946
Sand Saves	-2.188115769	2.124875787	-1.029761731	0.320584997	-6.745521069	2.369289531	-6.745521069	2.369289531

(Table 2, Regressions 2019)



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9152182
R Square	0.837624354
Adjusted R Square	0.779633052
Standard Error	0.306559982
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	5	6.787148686	1.357429737	14.44396529	4.21148E-05
Residual	14	1.315706314	0.093979022		
Total	19	8.102855			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	82.98951285	2.920750422	28.41376389	8.82306E-14	76.72512623	89.25389948	76.72512623	89.25389948
GIR %	-14.86846967	4.478204031	-3.320185852	0.005055118	-24.47326206	-5.263677275	-24.47326206	-5.263677275
Scrambling %	-8.535408352	2.778973268	-3.071425138	0.008289676	-14.49571322	-2.575103481	-14.49571322	-2.575103481
Putts Inside 5ft %	1.787585276	2.037914769	0.877163905	0.395200887	-2.583307192	6.158477745	-2.583307192	6.158477745
3PTT Avoidance	18.38667112	11.63924809	1.579712965	0.136494997	-6.577033252	43.35037548	-6.577033252	43.35037548
Sand Saves	-0.908206722	1.547515892	-0.586880384	0.566628295	-4.227298207	2.410884764	-4.227298207	2.410884764

(Table 2, Regressions 2020)

Results

Given the results of our testing, it is reasonable to conclude that our 2018 model is acceptable for estimating a 'good score' - in strokes – when given several percentage values for PGA Tour players. This means that to get a 'good score' we need to maintain high percentages in 'Scrambling', 'Putting Inside 5ft', 'Sand Saves', 'GIR', and a low percentage in '3 Putt Avoidance' to obtain a 'good score' and therefore be a 'good player'.

However, we need to thoroughly explain the high variance of our 2018 model's application to the other two sample years. Both models (Table 2, Regression 2019/2020) show sufficient p-values for the regression itself – 0.017178674 and 4.21148E-05 respectively - however, a few statistical variables show some abnormally high p-values. How can this be the case?

Well, in both sample years, there was a player that played very seldomly but still maintained a Top 20 OWGR (Tiger Woods, Table 1, 2019/ 2020). This can affect our results if said player has statistics that deviate significantly from the sample average – which in our case, does occur. Therefore, it would not be unreasonable to then see a high spike in p-values for these for these statistics when there is a sample player whose statistics deviate significantly form the sample average. So, the model can still be used, however, it should be prefaced that it is applied to random samples with a small enough standard deviation.

We also must explain the effectiveness of the intercept in our 2018 Model results (Table 12, Regression 2018). Why is it that the model predicts a starting score (total score in strokes) of 85 – with a standard error of 2.09 - and not the traditional stroke to par in golf – which is 72? This could have something to do with the relationship of our variables to scoring. A Tour player's round average in 2018 was measured to be 70.89 strokes (Table 3, 2018), with our sample's average being slightly below that at 69.619. This means that our model was predicting a score of at least +12 strokes relative to par and approximately +14 strokes relative to the Tour average. This can be adjusted manually and would produce a more accurate number as a result. It is worth mentioning that an average score for an *average golfer* is about 90 strokes; but could this be a coincidence. We would need more analysis outside of the scope of our research to find out if that were the case, but it is interesting to notice, nonetheless.

One other variable that needs a bit of explanation is the '3 Putt Avoidance' statistic. It is confusing with its wording, since it is not calculating the percentage of *avoided* 3 putts, it is calculating the percentage of 3-puuts – or more – taken on a hole over all holes played in a seasonal sample year (PGA Tour). This means it has a direct relationship with strokes added: the more 3-putts, the more strokes you take in a round's score, whereas all other statistics found have a reverse relationship - saving strokes in a round – reducing the total score, which is the ultimate objective in golf fundamentally: getting the ball in the hole in the least number of strokes possible over the course of 18 holes. This may also need an adjustment since our model calculated the average number of putts in a round to be the coefficient of 34 strokes, which depending on your source, is not the tour average in general (it much less than this value). Our model calculated the average number of putts an *average golfer* takes in a round (Beall, para 3). So, this can also be adjusted accordingly and does again beg the question if our model is calculating total score estimated for an average golfer. This, however, is still inconclusive since our sample data has no influence from an average golfer in the slightest; however, it brings some interesting insight.

With the model we did go with in the end – Table 12 mentioned prior – we were able to run VIF tests on all our variables to ensure each of them of enough non-collinearity. Depending on the research you reference, suggested VIF values vary from being less than 5, to 7, or even as high as 10 as being sufficient for proving non-collinearity. For my research I went with the threshold of 5 or less being my metric. This is because even though the model from Table 10 is in fact significant, some p-values were greater than 0.05, which lead me to question those variables and their stability in incorporating them into our final model. So, after removing the suspect variables, it is clear that in Table 12, both the VIF and p-values were low enough to be accepted – deeming the variables in the model chosen not too collinear.

Discussion

The interpretation of our new conclusion – that our model interprets variables that correlate to obtaining a good score in any given round of golf – suggests a few things about the sport of golf in general.

For one, it suggests that Driving Distance nor Accuracy account for any significance for being a good Tour player. This does not that Driving the Golf Ball does not matter; it simply means from our data that all tour players are consistent enough off the tee – relative to each other – that it is a statistic that bodes very little importance when trying to be better than the rest of the field. One obvious observation that would not be included here is if your tee shot results in you losing strokes such as hitting a bad shot. But as stated above, tour players are consistent enough that this does not happen often so when comparing to other tour players, you will see very little difference in score. This also does not mean that distance is irrelevant too, however from our sample, there was no clear indication that hitting the golf ball further off the tee provided a significant advantage compared to the rest of the field. This could have to do with the increase in dispersion of the tee shot the farther ball goes, since naturally there is more room for error the longer the shot – or stroke - is.

For two, a lot of our model's focus is on the shots *saved* from everywhere besides the tee box. Specifically, in 'Greens in Regulation' which has the highest absolute coefficient value besides putting – '3 Putt Avoidance'. This means that it is paramount to hit the green – no matter how far from the flag it may be – as a tour player, in regulation. All tour players are good enough – in a sense – to put the ball in the hole on the green within 2 strokes or less. This can be seen across all sample year tables as the '3 Putt Avoidance' statistic is averaging about 3 percent across the board.

Lastly for three, our model suggests that having a good scramble percentages correlates to having a lower overall score. This can be interpreted as players who are in bad positions are making the best possible decisions and executions out of a terrible situation and not giving strokes back to the field. In other words, tour players, when put in a bad position are still making par when they miss the green in regulation to not lose a shot to the field. Whether that be from a poor tee shot or approach shot.

So, our model essentially is stating, if you have good GIR percentages as well as avoiding 3-putting or worse 97 percent of the time, alongside having a strong scrambling percentage compared to the field (a significant deviation from the Tour Average), you will – as a tour player – be very well off compared to the field of competition. This can also apply to amateur golfers who want to improve their game.

Another remark on the final sentence in the results section. '*Not too collinear*' is referring to the fact that these variables do in fact share a relationship and that is: our model predicts strokes, so it is natural to assume each stroke taken impacts one of the variables in our model in a certain way since our total score is changing. In other words, since you count all strokes taken at the end of the round, it makes sense that some of these strokes can be attributed to an attempt at a 'Green-in-Regulation', 'Scrambling', or etc. The point of the VIF test was to ensure that our data collected was independent enough to be used as its own variable in the model and to not include two variables that related to well with one another. An example of these kind of variables would have been 'Driving Distance' and 'Proximity-to-Hole'.

To continue the remark, the two variable statistics mentioned before share to closely of a relationship. If a player drives the ball very far on average, it implies their approach shots are shorter than the rest of the field on average. This then trickles down to being able to hit the green in regulation, since tour pros excel at hitting greens the closer, they are to the flag – therefore affecting the 'GIR' and 'Proximity-to-Hole' statistics. And we see this being represented in our 'Collinearity Test' chart (Table 11, Regressions 2018). Values are attributed to this chart from 0 to 1, being how much the data from each variable relates to one another.

This does not mean that data can or cannot be used – per say – but rather how much that data correlates to other variables set of data. Which as discussed before, in golf: all strokes count, so to some degree you can argue certain statistics count a certain percentage to a golfer's total score. Conversely, you can also say some statistics represent the same exact thing – such as 'Proximity-to-Hole' and 'GIR'; with over a 60% correlation according to the test from Table 11 – and should be omitted because of it. Sand saves was not omitted since in the model it held its own by giving off small VIF and p-values; even with the 'Correlation Test' saying it relates heavily to 'Scrambling' and 'Putt Inside 5ft'. But that makes sense, since if a tour player makes a bad shot from out of the bunker, they may have to attempt another or end up putting more than twice on a hole or not Parring the hole all together, leading to a failed Scramble statistic.

Acknowledgements

This project could not have come to full fruition if it were not given the wonderful help and guidance of Dr. Matthew Hersh. It was Dr. Hersh's expertise that advised we can triple check certain variables for consistency by finding their associated VIF values in the respective multi-regression function. This is what ensured that the variables I found were in fact

significant to the specific model that was created.

Computer Code

Although computer code could have been used to organize the initial data input, it was not necessary. Also, the data analysis for this research was conducted solely with the 'Data Analysis' tools in the extension pack tool in Excel.

Reflection

This project has taught me to become more comfortable with Data Analysis on Excel. Prior to this research, I have only used the Data Analysis tool seldomly in a Physics Lab course here and there. Needless to say, this research has strengthened my ability interpret relationships between different sets of data, which is very useful for the application of it in the work force.

This project has also been a very personal one to me, since it involves a sport that I love dearly. I have given me a better picture of what it takes to be one of the best golfers in the world quantifiably. Due to conflicts with the variable coefficients units not being shared, it makes finding an equation for this quantification of the original thesis difficult. We have found, however, an accurate relationship between the aforementioned golf statistics relative to scoring so it does beg the question: is it not reasonable to gage the metric of success and consistency on the PGA Tour with being able to produce a low score? I believe our regression model data proves just that, since the series of the trails with the regression model tool in Excel showed us, which variables were inconsistent with our model and one of them was the 'Top 10s' Parameter itself. Shifting our focus to show a parameter for lowest golf score gives us a better picture as to what a good player on the PGA Tour looks like. It also is de facto a proof of consistency (which was needed in the original thesis), since maintaining good percentages in the stats that were found to be significant gives us a good idea of what 'playing good' on tour is about.

Literature Credit

All work cited is courtesy of the PGA Tours Player Stats Website which is powered by Shot Link Technology. A revolutionary system that combines the power of laser radar shot tracing devices such as Trackman launch monitors and multiple topographic nodes on the golf course to show precise distances hit (alongside attributes such as spin rates, launch angles, apex height, etc.) as well as giving information for the remaining distance to the hole ("Shotlink Data | PGA TOUR Stats"). All these tools combined can then be used to create data comparing players against one another and gives the audience an idea of how accurate a professional Tour player is, the percentage of conversions on a hole of a certain distance, or even what a normal score for them to shoot is.

Works Cited

Beall, Joel. “How Many Putts Does the Average Golfer Make? New Data Shows You Need More Time on the Practice Green...AND the Range.” Golf Digest, Golf Digest, 17 Mar. 2017, www.golfdigest.com/story/how-many-putts-does-the-average-golfer-make-new-data-shows-you-need-more-time-on-the-practice-greenand-the-range. Accessed 23 Nov. 2021.

PGA Tour. “PGA TOUR: Stats Leaders.” PGATour, PGA Tour, www.pgatour.com/stats.html.

“Shotlink Data | PGA TOUR Stats.” PGATour, PGA Tour, www.pgatour.com/stats/academicdata/shotlink.html. Accessed 30 Nov. 2021.