Investigating Significant Factors that Improve Golf Performance

by Comparing the Statistics of Stroke Gained by PGA Golfers

Jiahan Deng

2020/21/12

Abstract

This study analyzed the significant factors that improve the golf performance of the professional players. The linear regression, and the stepwise regression was performed for data analysis. The average score for the professional golf player is 72, and the average score could be improved by increasing the stroke gained during the competition.

Keywords

Performance analysis, Observational Study, Golf Score, PGA Tour, Stroke Gained

1. Introduction

A recent article analyzes the factors that affect the golf scores for amateur golfers (Suzuki, T., Sheahan, J.P., Okuda, I., & Ichikawa, D., 2020). While many people are watching the PGA Tour each year, and many people are trying to learn from the professional golfers, in this project, the goal is to find the significant factors that influence the average scores for the professional golfers.

PGA represents the Professional Golfers' Association of America, and the PGA Tour is the golf competition for those professional golfers every year. The game of golf is the process of using a club to strike a ball statically or several consecutive shots into the hole from the teeing ground. Golfers would be allowed to use up to 14 clubs in each game. There are a total of 18 holes in each game with different par, the predetermination number of strokes, for example if the hole is par 5, then the predetermination number of strokes would be 5 for this hole. The goal is to hit the golf ball from the tee to the hole with the fewest number of strokes.

The golf score could be affected by many factors, such as the fairway percentage, the average distance(driving distance), the average scrambling, and the average stroke gained(approach the green, around the green, the putting, and the off the tee gained). As a difficult sport, a small mistake might lead to severe punishment to the score.

The dataset used in this study contains 2312 observations from 2010 to 2018(Jong, 2019). Since golfers might play more than one game in the PGA Tour, the data shows their average performance of all the games he played each year. The response variable is "Average_Score" representing the performance of each golfer in the PGA tour each year. The predictor variable would be the factors above. Thus, this study aims to find the important factors which affect the golf score, and to see what changes could be made to improve the golf scores with modification of the practice focus.

One dataset would be used in this study. In the data section, both response variables and the predictors would be introduced, and the data cleaning process would be included in this section. For the model section, a linear regression model would be built. The analysis of p-value, R^2 and the model selection would also be described in this part. The final model diagnostic check, and its implication would be in the result section. The discussion part would include the summary, the explanation of the result, the weakness of the model, and the next step.

2. Data

The dataset was obtained from the Kaggle - PGA Tour Data (Jong,2019). The dataset originally contained 18 variables with 2312 observations. The dataset is based on the performance of the professional golfers during the PGA Tour. The target population is all the professional golf player's record. The frame population would be all units in the sampling frame (list of PGA golfers' record). The sample population is the player's record that has been recorded during the PGA Tour. The primary record (including rounds, distance of each stroke, scores, etc.) is done by the professional golf referees, and the statistical data are calculated based on the primary record. The dataset used in this report includes all the statistical data needed for the analysis.

2.1 Variable Selection:

There were some missing values in each year's PGA Tour. There were some observations containing only the golf player's name, the points they earned, number of wins, and the money they earned per calendar year. The missing values are omitted. After deleting the missing value, there are 1678 observations remaining. Moreover, 10 of 18 variables with 1 response variable (Average Score) and 9 predictors have been selected for this analysis. Histograms has been build to see each variable and their frequency in appendix 1.

2.1.1 Response variable:

Average Score: The average score through all rounds played in each year's PGA Tour (18 holes per round). The scores represent the performance of the golfers. The fewer the scores indicates the better performance.

2.1.2 Predictors:

Fairway percentage: this represents the percentage of the golf ball landed on the fairway, not out of bounds or into the bunker sand, after each stroke.

Average Distance: this represents the average distances of each stroke using a driver.

Average Scrambling(%): this represents the average percent of time where the golfer misses the green in regulation but still achieves the score on or lower the par score. The higher the scrambling rate, the better the performance of chipping and putting.

Average Total Stroke Gained (Average_SG_Total): this represents average of the sum of stroke gained of off the tee, around the green, approached the green, and putts in each round. (Pgatour.com, 2016) The higher the average total stroke gained represents the better performance of the golfers in general.

Since each player might have different strengths in each area (putts, approach the green, around the green, or off the tee), stroke gained in each area is used to analyzed separately.

Average Stroke Gained - Putts (Average_SG_Putts): this represents the average stroke gained with putting in each round.(Pgatour.com, 2016). The higher stroke gained in putts indicates better performance in putting.

Average Stroke Gained - approach the green (SG_APR): this represents the performance of golfers approaching shots. (Pgatour.com, 2016) The higher average stroke gained approach the green shows the better performance when hitting the ball approaching the green, which is usually the second and third stroke of each hole.

Average Stroke Gained - around the green(SG_ARG): this represents the performance of golfers shots within 30 yards from the green.(Pgatour.com, 2016). The higher the average stroke gained around the green, the better the performance of chipping.

Average Stroke Gained - off the tee (SG_OTT): this represents the performance of golfers off the tee of the par 4 and par 5 holes. (Pgatour.com, 2016) The higher the average stroke gained off the tee represents the better performance of driver and sometimes the initial shot for par 3 holes.

2.2 Description of Data:

Table 1: Basic information of the primary data

Variables	Min	Median	Max	Mean
Rounds	45	79.5	120	78.71
Fairway Percentage	43.02	61.43	76.88	61.44
Average Distance	266.4	290.6	319.7	290.8
Average Scrambling	44.01	58.27	69.33	58.12
Average SG Putts	-1.48	0.04	1.13	0.03
Average SG Total	-3.21	0.15	2.41	0.15
SG OTT	-1.72	0.06	1.49	0.04
SG APR	-1.68	0.08	1.53	0.07
SG ARG	-0.93	0.02	0.66	0.02
Average Score	68.7	70.9	74.4	70.9

3. Model

In the previous section, the variable selection and data cleaning was performed. 9 predictors were independent from each other, and the relationship between predictors and the response variable is linear. Linear regression models are chosen to estimate the relationship between the Average Score and the factors that influence the average score. In this section, the Akaike's Information Criterion(AIC) would be used to examine the model, and select the best model with the smallest AIC. Also, the Partial F Test would be performed to check whether the model selected by the AIC elimination is better than the original model. Last but not least, the p-values of all the predictors selected for the final model would be checked.

Original Model generated by RStudio:

$$AverageScore = 72.85 - 0.0003x_{Rounds} - 0.0040x_{FairwayPercentage} - 0.0031x_{AvgDistance} - 0.0103x_{AverageScrambling} - 0.8931x_{AverageSGPutts} - 0.0146x_{AverageSGTotal} - 0.9046x_{SGOTT} - 0.9292x_{SGAPR} - 0.8812x_{SGARG}$$

3.1 Akaike's Information Criterion (AIC):

AIC helps you to examine how well your model fits the data without overfitting the data. The smaller the AIC the better the model. A good model should have small AIC and high adjusted R^2 (Daignault, STA302W6B)

Starting with the original full model fitted and using AIC, the model contains 7 predictors, and has the smallest AIC of -5592.41. The 6 predictors are fairway percentage, average distance, average scrambling, average stroke gained Putts, stroke gained off the tee, stroke gained approached the green, and the stroke gained around the green.

The R^2 for this model (final model) is 0.9274, which indicates that 92.74% of the data fits the linear regression model. The adjusted R^2 for this model is 0.9271. Since both R^2 and adjusted R^2 is high (close to 1), the model did fit the data well.

3.2 Partial F Test:

The Partial F test would help to determine whether the model selected by the AIC (final model) is better than the original model. The null hypothesis for the partial F test is that the original model is not better. Since the p-value is 0.44 which is greater than 0.05, the null hypothesis would be accepted. So, the final model is better than the original model.

3.3 Coefficient and P-value Table (Table 2)

Variables	Coefficient	p-value
Intercept	72.844287	$<2e^-16$
Fairway Percentage	-0.004083	0.01214
Avg Distance	-0.003188	0.00513
Average Scrambling	-0.010443	$3.76e^-07$
Average SG Putts	-0.909596	$<2e^-16$
SG OTT	-0.920474	$<2e^-16$
SG APR	-0.945256	$<2e^-16$
SG ARG	-0.895954	$<2e^-16$

From the coefficient and p-value table (table 2) above, it is clear that all predictor variables in the final model have p-value smaller than 0.05, which means they are significant to the response variable.

3.4 Final Model:

The final model is generated by RStudio.

$$AverageScore = 72.844 - 0.004x_{FairwayPercentage} - 0.003x_{AvgDistance} - 0.010x_{AverageScrambling} - 0.910x_{AverageSGPutts} - 0.920x_{SGOTT} - 0.945x_{SGAPR} - 0.896x_{SGARG}$$

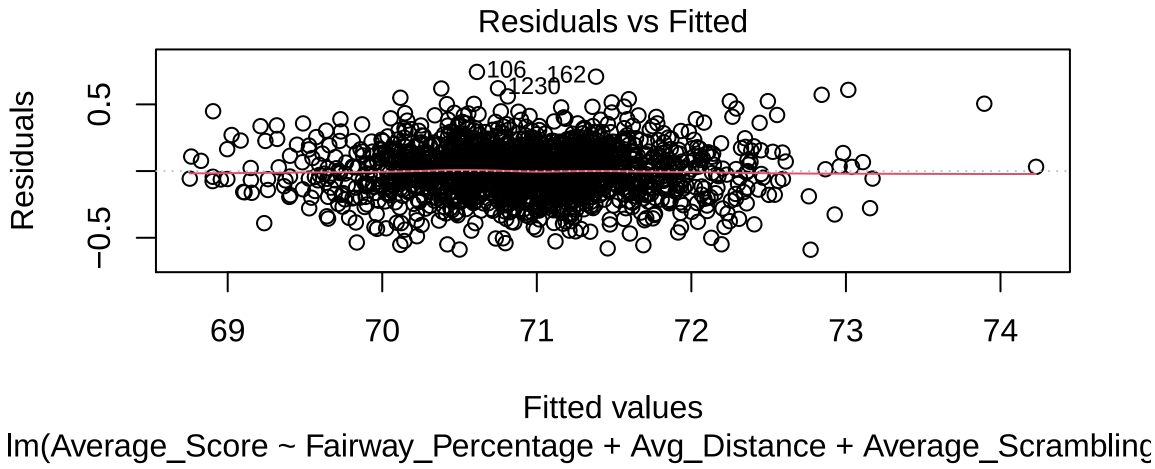
4. Validity and Result

In this section, the assumptions of the final model would be checked, and the result would be interpreted.

4.1 Assumption Check:

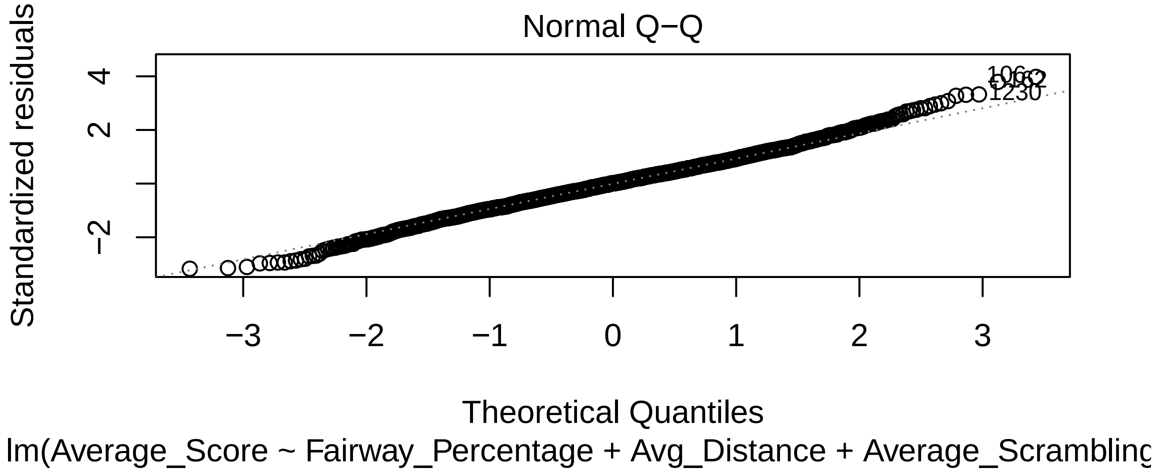
After the linear regression model has been built, the assumptions of the linear regression need to be checked to determine whether the final model works well for the data. In this section, the linearity, normality, homogeneity, outliers, and high leverage points of the final model(Daignault, STA302W3B) would be checked.

4.1.1 Linearity of the data:



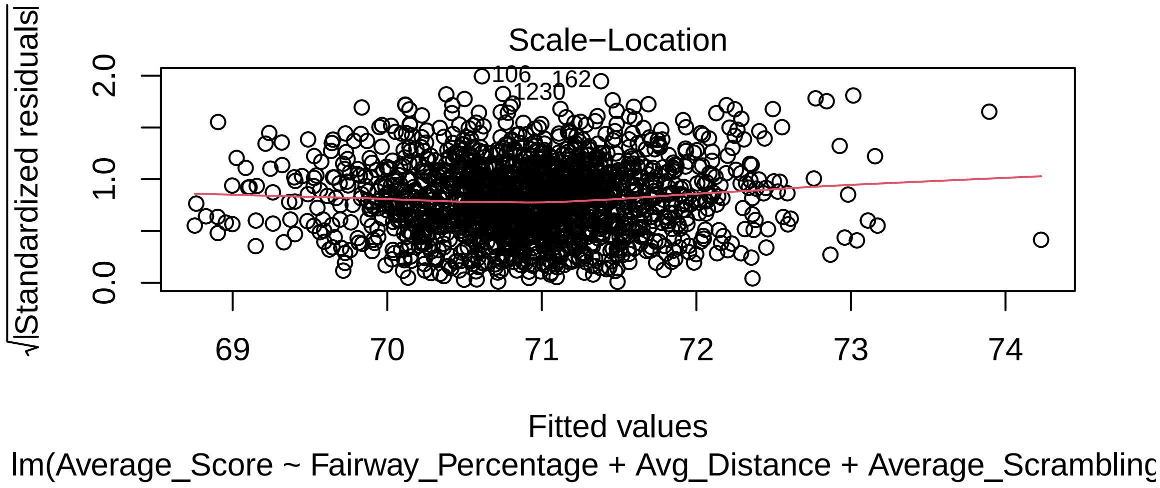
From the Residual vs Fitted Plot, the residuals are evenly spreaded around the horizon line without noticeable pattern. So, the relationship between the response variable - Average_Score and the predictors are linear.

4.1.2 Normality of residuals:



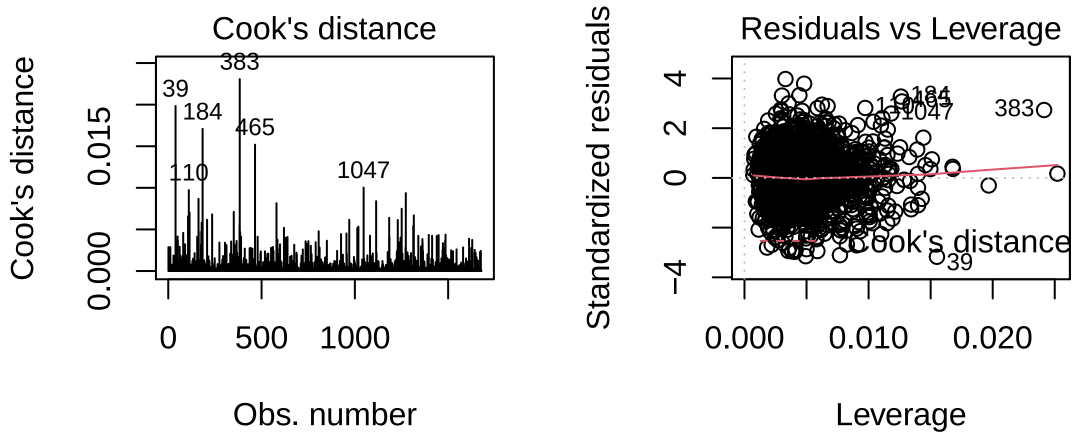
From the Normal Q-Q Plot, the Q-Q plot shows a straight-diagonal line, so the residuals of the linear regression model are normally distributed.

4.1.3 Homogeneity of variance:



From the Scale-Location Plot, since there is no distinct pattern, the residuals are spread equally along the ranges of predictors.

4.1.4 Extreme values:



From the cook's distance, and residuals vs leverage plots, 6 highest extreme points are labeled on the graphs.

4.2 Result:

$AverageScore = 72.844 - 0.004x_{FairwayPercentage} - 0.003x_{AvgDistance} - 0.010x_{AverageScrambling} - 0.910x_{AverageSGPutts} - 0.920x_{SGOTT} - 0.945x_{SGAPR} - 0.896x_{SGARG}$

The intercept of this model is 72.844, which indicates that the Average Score of the PGA golfers would be 72.844 when each of the predictors are at the average value. If the golfer could improve the fairway percentage for 1% while holding other factors unchanged, the average score would decrease by 0.004. If the golfer could hit longer drives(increase the Average Distance), 10 yard improvement in the driving distance while holding other factors constant, the average score would decrease by 0.03 stroke. If the golfer could increase the scrambling rate by 1% while holding other factors constant, the average score could be decreased by 0.01 strokes.

For the stroke gained, 1 point of Stroke Gained Putts could lead to 0.91 strokes decrease for the average score while holding other factors constant. By improving the Stroke Gained off the tee while holding other factors constant, the average score would decrease by 0.92 stroke. By improving the Stroke Gained Approached The Green while holding other factors constant, the Average Score would decrease by 0.95 stroke. By improving the stroke gained around the green while holding other factors constant, the Average Score would decrease by 0.9 stroke.

5. Discussion

5.1 Summary:

The goal is to find out the significant factors that would influence the golf average score. By performing the variable selection, regression selection, and the diagnostic check, the final model has been found and interpreted. As a result, the factors that affect the average score have been determined. This study demonstrates that the professional golfers could reduce the average score down by increasing fairway percentage, average distance, scrambling rate, stroke gained putts, stroke gained off the tee, stroke gained approached the green, and stroke gained around the green.

5.2 Conclusion:

Average Score:
Compared to the mean value of the Average Score of the dataset, the Average Score - 72.8 of the model is higher than the mean Average Score - 70.9 (refer to table 1) of the dataset. This might be due to the different strengths of golf players.

Fairway Percentage:
From the original dataset, the difference between the max and min of the player's fairway percentage is 33% (refer to table 1). From the model build in the model section, 33% diffence in fairway percentage would only lead to 0.1 strokes difference on the average score. The average score would not be affect siginifantly by the fairway percentage. However, For all the golf players, hitting the ball on the fairway would always be helpful for the following strokes. Golfers are always trying to avoid difficult situations as hitting the ball into the bunkers/bushes/water. Aiming the fairway, increasing the fairway percentage could avoid

penalty area efficiently.

Average Distance:
While all the amateur golfers are trying to optimize their driving distance in order to reduce their average score. For the professional golfer, compare the min Average Distance 266 yards to the max Average Distance- 319 yards (refer to table 1), this only leads to a difference of 0.16 Average Score. Although according to the model, the average distance does affect the average score, during the actual competition, increasing the average distance might not contribute greatly to reducing the average score.

Average Scrambling:
The Average Scrambling rate indicates the performance of the golfer after missing the green in regulation (when they failed to hit the ball the the putting area with at least two strokes fewer than par). A high average scrambling rate shows a good performance on chipping and putting. Accordingly, regardless of whether the player did not effectively accomplish par on, they had the option to make a high scrambling rate by adopting a strategy shot.

Average Stroke Gained Putts:
Compared to the driving distance (300 yards), the putting distance is usually within 26 yards. So, putting accuracy plays an important role during the competition. Compare the highest and the lowest Average Stroke Gained Putts, 2.6 points of Average Stroke Gained Putts would lead to 2.4 strokes difference on Average Score. For this reason, putting should be given more importance and attention even at the professional level.

Stroke Gained Off The Tee/Approach The Green/Around the green:
During the PGA Tour, the penalty for driving the ball out of bound could be high. Hitting the golf ball straight and long with the driver is very important. A driving distance between 100-225 yards could lead to minus 1.7 of the stroke gained off the tee. Based on the dataset, the min average distance is 266 yards, which means the most players might not lose the stroke gained off the tee due to the distance. However, since the mean fairway percentage is only 61% (refer to table 1), many players might lose the stroke gained by hitting the ball onto a rough area or out of bound. For the mature golfers, as they are already able to hit long, the accuracy should be given more importance. Excellent second and third shots(usually with iron) would increase the stroke gained approach the green, and would improve the average score greatly. The stroke gained around the green is similar to scrambling rate, it measures the performance of player's chipping techniques. Great chipping skills would improve the average score significantly.

In conclusion, although the average distance and the fairway percentage would affect the average score for the PGA golfers, they might not influence the average score greatly. The stroked gained during each round of the competition widen the difference of the average score of each players.

5.3 Limitations:

1. The wind cannot be controlled in each game. When the ball penetrate through the air, the distance might be influenced, and the score would be affected accordingly.
2. The dataset contains only 1678 full observations, which may reduce the variability and ability of the model.
3. The dataset contains the average data of each player. However, the player played many rounds (45-120 rounds), if the player did not play very nicely on some of the rounds, the Average Score and other factors might be influenced greatly.

5.4 Next Steps:

1. Since there were many missing data, and the dataset is relatively small, completing the dataset would help to build a more precise model.
2. A new dataset could be made with the data records from each round played by each players. Instead of the value of the factors on average, the data from each rounds may shows the weaknesses and strengths of each player. If the model is build on the dataset based on each round, the model might be more precise, and the variability would increase.
3. The generalized linear mixed model could be build if more than one record for each players are used to build the model. In this case, people would be able to observe the difference between each round the player played, and people may learn the excellent technique from different professional players.
4. By comparing the different factors that influenced golf score between amateur players and professional golfers, people could build a more complete training plan and focus on different factors on different level.

Bibliography

1. Daignault, Katherine. “STA302W3B_slide.” STA302/1001: Methods of Data Analysis 1, Department of Statistical Sciences University of Toronto.
   This Class slide from STA302 states the assumptions of the linear regression model, and methods for checking those assumptions.
2. Daignault, Katherine. “STA302W6B_slide.” STA302/1001: Methods of Data Analysis 1, Department of Statistical Sciences University of Toronto.
   This class slides from STA302 clarifies the criteria for selecting models, and the automated stepwise selection methods.

3. Jong. (2019, April 30). PGA Tour Data. Retrieved December 15, 2020, from https://www.kaggle.com/jmpark746/pga-tour-data-2010-2018?select=pgaTourData.csv
This is the dataset used in this report, which contains the data showing the performance of the professional golf players during the PGA Tour 2010-2018.

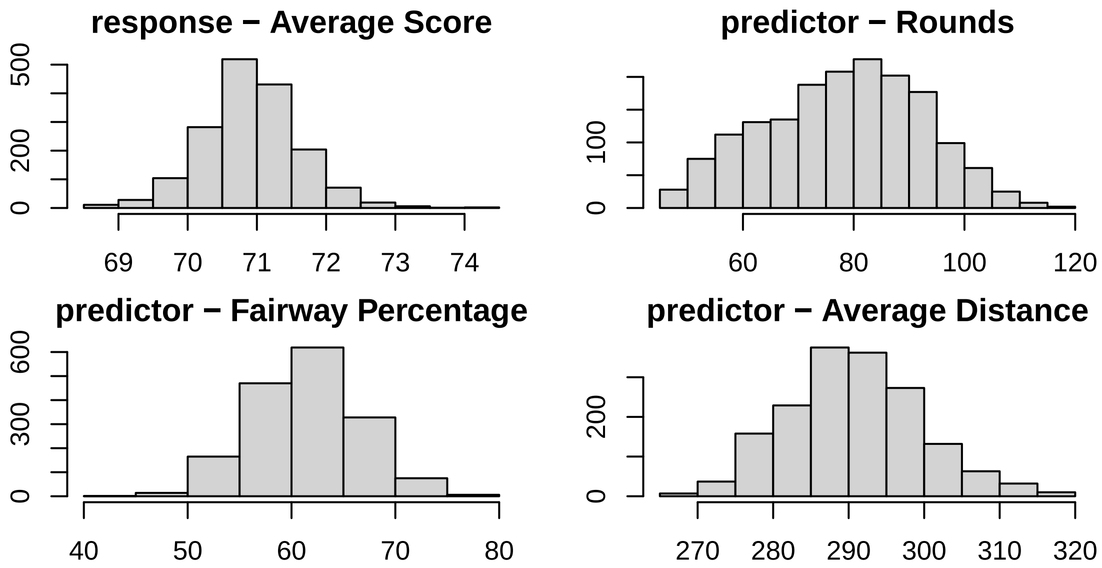
4. Pgatour.com. (2016, June 03). Strokes gained: How it works. Retrieved December 15, 2020, from https://www.pgatour.com/news/2016/05/31/strokes-gained-defined.html
This article states how the strokes gained worked and calculated in golf.

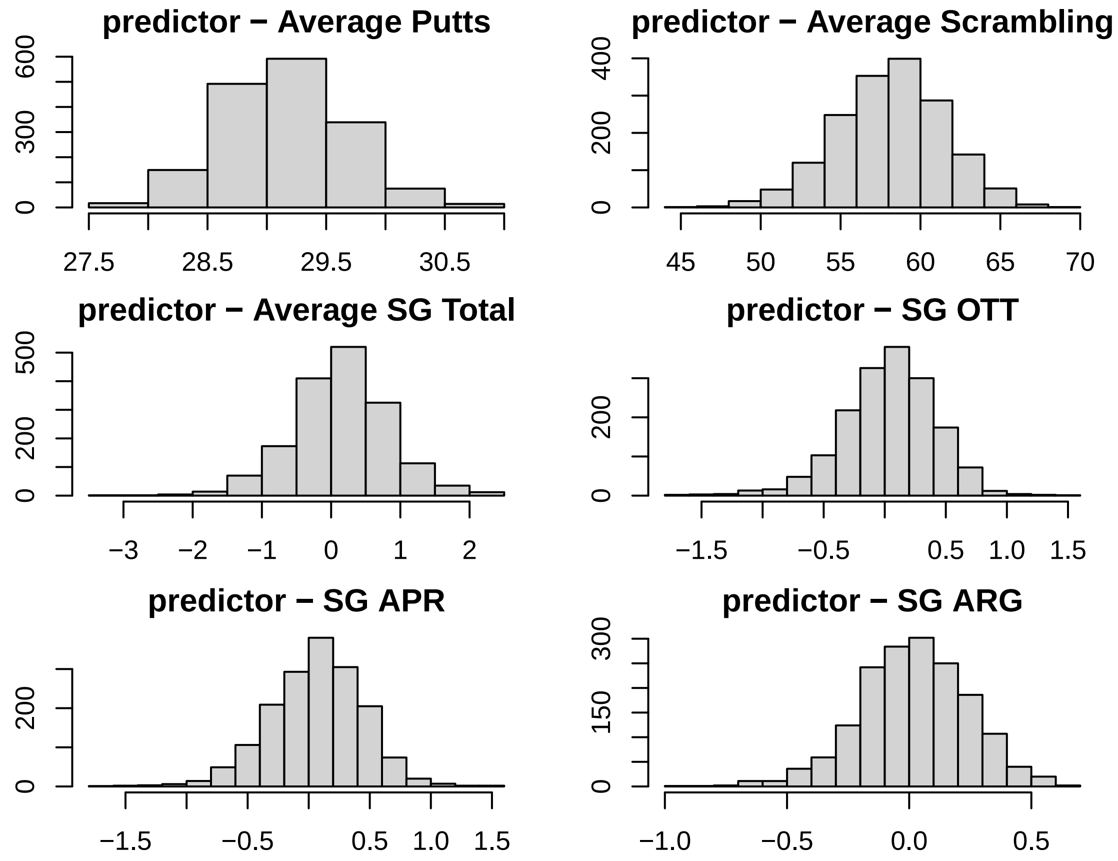
5. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
RStudio is the software used to generate result of this study.

6. Suzuki, T., Sheahan, J.P., Okuda, I., & Ichikawa, D. (2020). Investigating factors that improve golf scores by comparing statistics of amateur golfers in repeat scramble strokes and one-ball conditions. Journal of Human Sport and Exercise, in press. doi: https://doi.org/10.14198/jhse.2021.164.09
This article states important factor that influence the score for amateur golfers.

Appendix

Appendix 1: histograms of each variable




predictor − Average Putts
0
300
600
27.5
28.5
29.5
30.5
predictor − Average Scrambling
0
200
400
45
50
55
60
65
70
predictor − Average SG Total
0
200
500
−3
−2
−1
0
1
2
predictor − SG OTT
0
200
−1.5
−0.5
0.5
1.0
1.5
predictor − SG APR
0
200
−1.5
−0.5
0.5
1.0
1.5
predictor − SG ARG
0
150
300
−1.0
−0.5
0.0
0.5