Investigating the relationship between scoring average and putts per round average on the PGA Tour

Internal Assessment in Mathematics Applications and Interpretations

Number of pages: 20

Table of contents

Introduction

A golf player myself, I want to investigate importance of putting for golf scores; an exploration into this topic would help me determine what to put emphasis on during my practise sessions. Thus, this exploration will investigate the relationship between scoring average (average total number of strokes per round) and putts per round average (average number of short shots made on the green around the hole) among players of the PGA Tour. The goal of this exploration is to determine the strength and nature of the relationship between scoring average and putts per round average and establish whether a player's putting abilities are independent of his golf scores.

It is expected there will be a weak positive relationship between the variables; the reason for that is that better players, so players with better scoring averages, perform better in a variety of statistics, and a putts per round average is only one of them.

In order to collect data, the list of players who were classified by the PGA Tour was gathered and organised according to their putts per round averages in the season 2019. Their scoring averages in the same season were also collected. A website with the official statistics of the PGA Tour was used as a source of the data.

Procedure of investigation

1. Collecting data: sampling and organising data into a table.
2. Performing the chi-squared test for independence.
3. Plotting a scatter plot, finding the line of the best fit and analysing the Pearson's correlation coefficient of the linear model.
4. Finding the line of quadratic regression and analysing the Pearson's correlation coefficient of the quadratic model.
5. Drawing conclusions.

6. Discussing areas for possible improvement and further research.

Data collection

To collect the data, ranks of players of the PGA Tour by putts per round average and by scoring average in the season 2019 were used. For sampling, the list of players by putts per round average was chosen instead of scoring average because there were 37 players more on the scoring average rank. Avoiding a situation in which a player would have to be replaced because his putts per round average would not be available was desirable, as that would have disrupted the sampling system and would have led to less reliable results. The sample for the investigation was chosen using systematic sampling (meaning that players were selected starting at a random position with a fixed periodic interval). Putts per round averages were available for 190 players and it was decided that a sample of 95 would be chosen, which meant that the interval was two. In numerous instances, there were a few players with the same putts per round average; when that was the case, a player from those with the same putts per round averages was chosen at random. Scoring averages were not taken into consideration when making this decision, because that could have made the sample biased. If, for example, the player with the highest scoring average was always chosen, the sample would underestimate the strength of the relationship if it were positive, and overestimate its strength if the relationship were negative.

When filling in data about scoring averages, two instances in which scoring average was unavailable for a player were encountered. In these cases, deterministic hot deck imputation was applied. Deterministic hot deck imputation is a method used for handling missing data in which a missing value is replaced with an observed value from the most similar unit (Andridge and Little 40). The unit for which data is inserted is called a recipient while that from which the data is being taken is called a donor. To perform deterministic hot deck

imputation, the player whose putts per round average was closest to that of the recipient was taken as donor (in both instances in this investigation that meant choosing another player with the same putts per round average).

Sample entries (top two and bottom two players, and one player for which deterministic hot imputation was performed) are available in table 1 for illustrative purposes. Full dataset used is available in Appendix A.

Table 1

Selected players of the PGA Tour, their scoring averages and putts per round averages in the 2019 season

Note: Blue colour indicates players for which deterministic hot deck imputation was performed. Names of donors are in parenthesis.

Name	Scoring average	Putts per round average
Jordan Spieth	71.46	27.71
Justin Rose	71.82	27.94
Trey Mullinax (J. J. Spaun)	72.24	29.29
John Chin	72.49	30.13
Corey Conners	70.78	30.17

Sources of data: “Putts per round.” *PGA Tour Statistics.* PGA Tour, www.pgatour.com/content/pgatour/stats/stat.119.y2019.eoff.t060.html. Accessed 15 Apr. 2020 and

“Scoring average.” *PGA Tour Statistics.* PGA Tour, www.pgatour.com/stats/stat.120.html. Accessed 15 Apr. 2020

Chi-squared test

Firstly, the chi-squared test was performed to establish whether the correlation between the variables is statistically significant, i.e. unlikely to be a result of chance. For the purposes of this test:

H_0: scoring average is independent of putts per round average.

H_1: scoring average is dependent on putts per round average.

Players were grouped based on means of the two data sets - one of scoring averages and one of putts per round averages (Table 2). This constitutes the observed frequencies – actual number of times each event occurred.

Table 2

Observed frequencies

	Below mean putts per round	Above mean putts per round	Total
Below mean scoring	29	22	51
Above mean scoring	15	29	44
Total	44	51	95

Expected frequencies (how many times each event is expected to occur if there were no relationship between the variables) were calculated by multiplying the observed value in the given scoring category and the given putting category, and then dividing the product by the total number of players. For example, the expected value (f_e) of the number of players with scoring average below mean and putts per round average below mean:

$$f_e = \frac{44 \times 51}{95} = 23.6210 \ldots \approx 23.62 \text{ (2 DP)}$$

Table 3

Expected frequencies

	Below mean putts per round	Above mean putts per round	Total
Below mean scoring	23.62	27.38	51
Above mean scoring	20.38	23.62	44
Total	44	51	95

To interpret the results of a chi-squared test, one needs to know the number of degrees of freedom and decide what significance level they will use. The significance level represents the probability of rejecting a null hypothesis that is true (for example because the result of the chi-squared test obtained is a result of chance alone). A significance level of 0.05 was assumed. This is a fairly low probability of rejecting a true null hypothesis and therefore a significance level of 0.05 was deemed appropriate.

To interpret the results of the chi-squared test, the number of degrees of freedom was calculated using the following formula:

Degrees of freedom = (number of columns-1) × (number of rows-1)

In the present case:

Number of columns - 2

Number of rows - 2

Therefore, the number of degrees of freedom equals 1. Thus, Yates's correction for continuity will be applied in the chi-squared test, as it allows for a more precise interpretation of the results of the test in small datasets, such as if the number of degrees of freedom is 1 (Yates 217). The formula for the chi-squared with Yates's correction for continuity test is as follows:

$$\chi^2_{Yates} = \sum_{i=1}^{n} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Where:

O_i - observed frequency

E_i - expected frequency

n - number of distinct events

After inserting the data into the above formula and doing the initial calculations, the following equation was received:

$$\chi^2 = 1.01+0.87+1.17+1.01 = 4.06$$

Comparing the results to data from a table of critical values (Appendix B), it can be seen that χ^2_{calc} is greater than χ^2_{crit}. This means that the null hypothesis should be rejected, which leads to the acceptance of the alternative hypothesis - that scoring average is dependent on putts per round average. However, if the chi-squared test were to be conducted at a 0.025 significance level, χ^2 would be smaller than χ^2_{crit}, which means the null hypothesis that scoring

average is independent from putts per round average would have to be accepted. Since the alternative hypothesis can only be accepted at a 0.05 significance level, but not at a 0.025 one, it can be concluded that there is a correlation between scoring average and putts per round average, but it is not very strong.

After the relationship had been confirmed to be statistically significant, its nature was investigated.

Analysing the relationship

Line of best fit

A scatter plot with a line of best fit was created using Google Sheets to see whether there was a visible relationship between the two variables (figure 1).

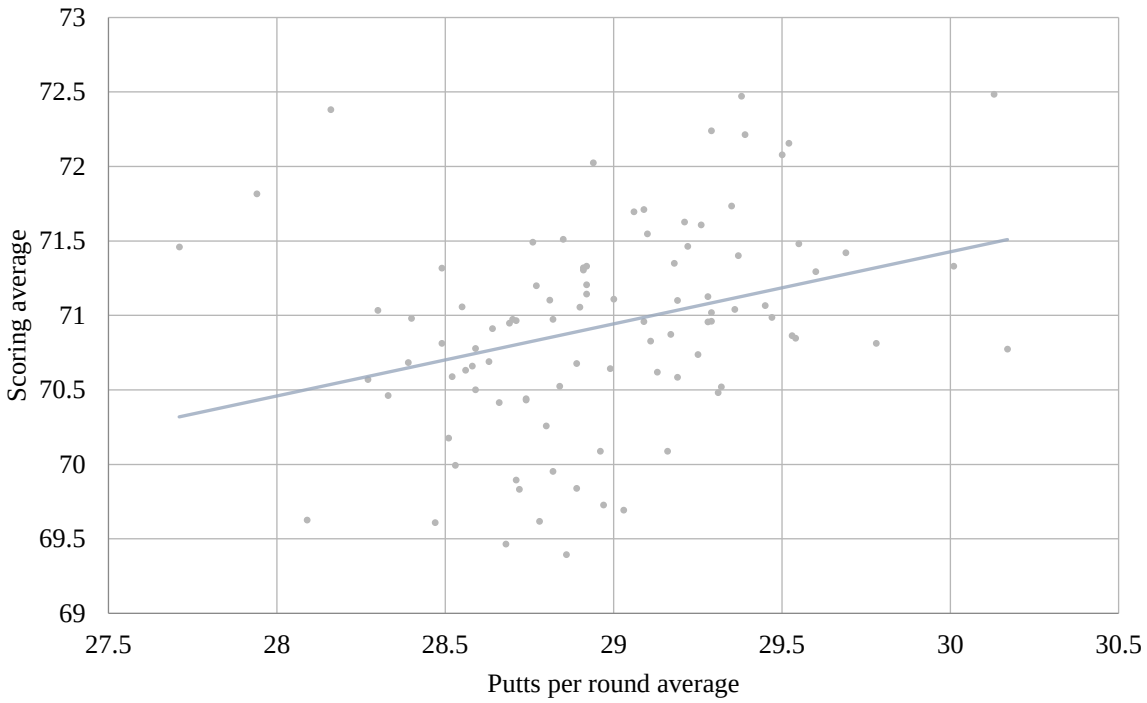

Figure 1. Scatter plot with linear regression

The equation for the line of best fit is y = 0.48x + 56.91, which confirms the hypothesis that the relationship is positive. The Pearson's correlation coefficient was calculated to establish the strength of the relationship. The formula is as follows:

$$r = \frac{n \times (\sum xy) - (\sum x) \times (\sum y)}{\sqrt{[n \times (\sum x^2) - (\sum x)^2] \times [n \times (\sum y^2) - (\sum y)^2]}}$$

Where:

r - Pearson's correlation coefficient

n - sample size, in the present case 95

$\sum xy$ - sum of products of the product of pairs of x and y values, in the present case 195169.52 (2 DP[1])

$\sum x$ - sum of x values, in the present case 2751.61 (2 DP)

$\sum y$ - sum of y values, in the present case 6737.96 (2 DP)

$\sum x^2$ - sum of squares of x values, in the present case 79717.47 (2 DP)

$\sum y^2$ - sum of squares of y values, in the present case 477940.39 (2 DP)

After inserting the data into the formula, the following equation was received:

$$r = \frac{95 \times 195169.52 - 2751.61 \times 6737.96}{\sqrt{(95 \times 79717.47 - 2751.61^2) \times (95 \times 477940.39 - 6737.96^2)}}$$

$$= 0.3148564452 \ldots \approx 0.315$$

[1] Decimal places

Since the absolute value of r is greater than 0.25 but smaller than or equal to 0.5, and r is positive, there is a weak positive correlation between the two variables (Chang Wathall et al. 272).

The value of r^2, which shows how close data points are to the line of best fit, is, to three significant figures, 0.0992. This is a very low value, as it indicates that only 9.92% of the variance of the y-variable can be explained with the x-variable. Because of how low value of r^2 is, quadratic regression was calculated. Although was is difficult to say whether quadratic regression would be more accurate in representing the dataset when looking at the graph with a bare eye, there was a possibility of it being a more appropriate fit, and thus it was calculated.

Quadratic regression

The scatter plot with quadratic regression is presented in figure 2.

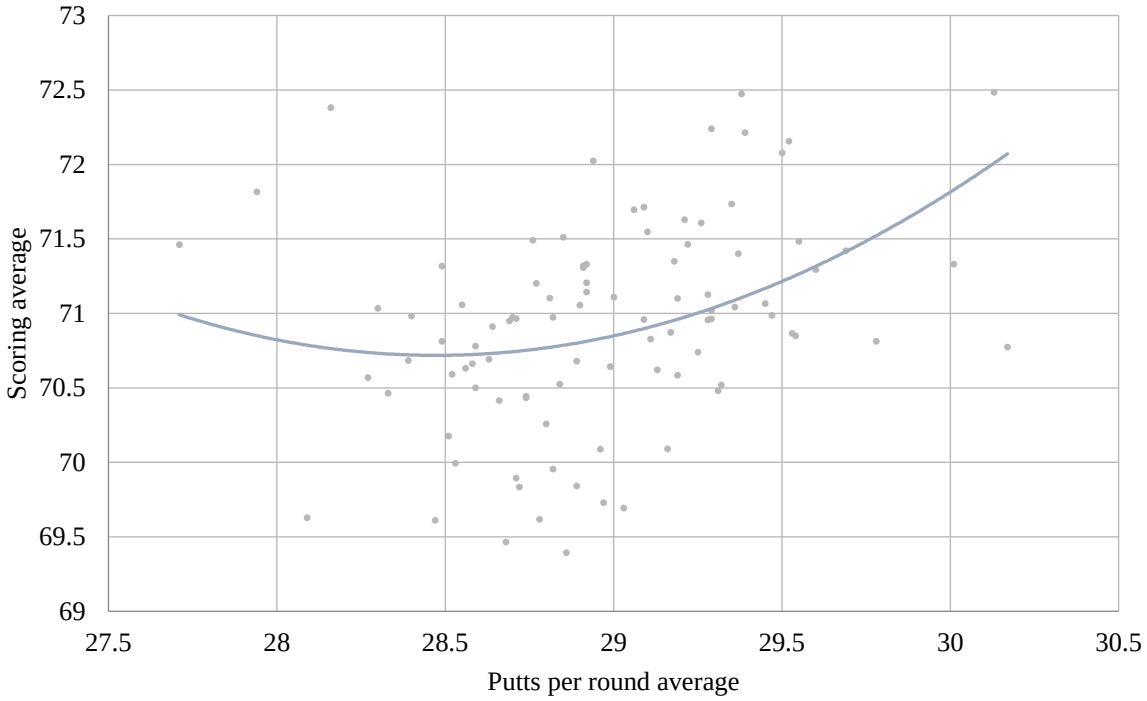

Figure 2. Scatter plot with quadratic regression

The equation of the line of quadratic regression is $y = 0.47x^2 - 26.72x + 451.15$, which also indicates a positive relationship. The r^2 was calculated using a TI-84 Plus graphing calculator to establish whether the quadratic model represents the dataset more accurately than the linear model. The value of r^2 is, to three significant figures, 0.143, which means the quadratic regression represents the trend in the data set better than the linear regression.

The fact that the vertex of the curve located at (28.47, 70.71) is interesting, as that suggests that players with putting averages lower than that have increasing scoring averages. Although this is true for the data set, it seems counterintuitive empirically. This effect can be explained by the low availability of players with putting averages lower than the x-value of the vertex. Only 9 players in the sample have putting averages lower than 28.47 (the x-coordinate of the vertex), which is a small number, and that could have led to the distortion of the quadratic regression model. Thus, the linear model can be deemed more realistic.

Conclusion

As expected, the relationship between putts per round averages and scoring averages turned out to be weak and positive; this was already visible on the scatter plot and calculations (the chi-squared test and both regressions) proved it. Three statistical operations performed (the chi-squared test and calculating the r^2 value for a linear and quadratic model) all showed that there is a weak correlation between the two variables. The Pearson's Correlation Coefficient was equal to 0.315, which means there is a weak positive correlation. The chi-squared test conducted at a 0.05 significance level resulted in accepting the alternative hypothesis that the variables are dependent on each other. The fact that if the chi-squared test were conducted at a 0.025 significance level, the result would be the opposite strengthens the conclusion that the relationship between the two variables is not strong.

As stated in the introduction, there is a wide range of aspects that determine a player's result in the form of his scoring average. Putting, measured by putts per round average, is only one of them. This can be a cause of why the relationship is only weak. It follows from the results of this exploration that practising putting, albeit important, is alone not enough to make one an outstanding player.

Mathematical processes were conducted several times, using Google Sheets, a graphing calculator (TI-84 Plus) and by hand, which served to ensure that the results are correct. Furthermore, the two methods of analysis (the chi-squared test and regression analyses) that were conducted both showed the same results, confirming that the variables under investigation are related, albeit weakly. All this means that the results of the investigation can be accepted with a high degree of confidence.

Evaluation

In order to improve the design of the study, the players for whom scoring averages were unavailable could have been crossed out before sampling was carried out from the ranking of putts per round averages (the list which was used for sampling). While the putts per rounds averages of the donor players were the same as those of the recipients, crossing out players for whom data were unavailable would have made the investigation easier to carry out, as there would have been no need to perform deterministic hot deck imputation.

To further improve the design of the investigation, putts per green in regulation instead of putts per round could have been used. On greens on regulation, which are approached from a larger distance, putts are, usually, longer than when greens are entered with a chip – from a short distance. As putts per round average is also influenced by the length of putts, using putts per green in regulation have better reflected to what extent a player's putting abilities influences his score. A player that misses many greens might have a low putts per round

average not because his putting is exceptional, but because he plays many putts after chips, which tend to be shorter, and thus easier to make, than an average putt. Such a player would have a low putts per round average, even if his putting were not that good compared to others'. Putts per round average, therefore, did not isolate putting as a variable as well as putts per green in regulation would have had.

A further limitation of the study is that it only investigated male players. There are significant differences in the style of play of men and women, including that women average more putts (Rudy 2010). Further developments of this investigation might look at the trends investigated in this exploration among female players.

Appendix

Appendix A - Selected players of the PGA Tour, their scoring averages and putts per round averages in the 2019 season.

Note: Blue colour indicates players for which deterministic hot deck imputation was performed. Names of donors are in parenthesis.

Name	Scoring average	Putts per round average
Jordan Spieth	71.46	27.71
Justin Rose	71.82	27.94
Webb Simpson	69.63	28.09
Si Woo Kim	72.38	28.16
Vaughn Taylor	70.57	28.27
Andrew Putnam	71.04	28.30
Brandt Snedeker	70.46	28.33
Zach Johnson	70.69	28.39
Brian Stuard	70.98	28.40
Patrick Reed	69.61	28.47
Sam Burns	71.32	28.49
Cameron Smith	70.81	28.49
Bryson DeChambeau	70.18	28.51
Graeme McDowell	70.59	28.52
Rickie Fowler	70,00	28.53
Pat Perez	71.06	28.55
J. T. Poston	70.63	28.56
Michael Thompson	70.66	28.58
Ian Poulter	70.50	28.59
Rafa Cabrera Bello	70.78	28.59
Nate Lashley	70.69	28.63
Sam Ryder	70.91	28.64
Kevin Kisner	70.42	28.66

Justin Thomas	69.47	28.68
Troy Merritt	70.95	28.69
Francesco Molinari	70.97	28.70
Dustin Johnson	69.90	28.71
C. T. Pan	71.00	28.71
Xander Schauffele	69.83	28.72
Mark Leishman	70.43	28.74
Daniel Berger	70.44	28.74
Sam Saunders	71.49	28.76
Austin Cook	71.20	28.77
Jon Rahm	69.62	28.78
Louis Oosthuizen	70.26	28.80
Richy Werensky	71.10	28.81
Ryan Moore	70.97	28.82
Tony Finau	69.96	28.82
Matthew Fitzpatrick	70.53	28.84
Peter Uiheiln	71.51	28.85
Brooks Koepka	69.40	28.86
Hideki Matsuyama	69.841	28.89
Ryan Palmer	70.68	28.89
Bill Haas	71.06	28.90
Danny Lee	71.31	28.91
Ryan Armour	71.32	28.91
Phil Mickelson	71.33	28.92
Kevin Na	71.14	28.92
Roberto Díaz	71.21	28.92
Seamus Power	72.03	28.94
Henrik Stenson	70.09	28.96
Tommy Fleetwood	69.73	28.97
Billy Horschel	70.64	28.99

Max Homa	71.11	29.00
Adam Scott	69.69	29.03
David Hearn	71.68	29.06
Scott Brown	70.96	29.09
Joey Garber	71.71	29.09
Ernie Els	71.55	29.10
Danny Willett	70.83	29.11
Joaquin Niemann	70.62	29.13
Jim Furyk	70.09	29.16
Russel Henley	70.87	29.17
Hudson Swafford	71.35	29.18
Sung Kang	71.10	29.19
Tyrrell Hatton	70.59	29.19
Robert Streb	71.63	29.21
Adam Long	71.46	29.22
Aaron Wise	70.74	29.25
Tom Hoge	71.61	29.26
Roberto Castro	71.13	29.28
Shawn Stefani	70.96	29.28
Trey Mullinax (J. J. Spaun)	72.24	29.29
Keith Mitchell	71.02	29.29
Michael Kim	70.96	29.29
Charles Howell III	70.48	29.31
Russell Knox	70.52	29.32
J. B. Holmes	71.74	29.35
Adam Svensson	71.04	29.36
Charley Hoffman	71.40	29.37
Whee Kim	72.47	29.38
Ted Potter Jr.	72.21	29.39
Brice Garnett	71.07	29.45

Sepp Straka	70.99	29.47
Kyle Jones	72.08	29.50
Seth Reeves	72.16	29.52
Martin Laird	70.87	29.53
Hunter Mahan (Keegan Bradley)	70.85	29.54
Jim Knous	71.48	29.55
Bronson Burgoon	71.30	29.60
Branden Grace	71.42	29.69
Emiliano Grillo	70.81	29.78
Alex Prugh	71.33	30.01
John Chin	72.49	30.13
Corey Conners	70.78	30.17

Sources of data: “Putts per round.” *PGA Tour Statistics.* PGA Tour, www.pgatour.com/content/pgatour/stats/stat.119.y2019.eoff.t060.html. Accessed 15 Apr. 2020 and “Scoring average.” *PGA Tour Statistics.* PGA Tour, www.pgatour.com/stats/stat.120.html. Accessed 15 Apr. 2020

Appendix B - Table of critical values for the chi-squared test with one degree of freedom

Degrees of freedom	P = 0.05	P = 0.025
1	3.84	5.02

Source: “Critical Values of the Chi-Square Distribution.” *Engineering Statistics Handbook*, Information Technology Laboratory. www3.med.unipmn.it/~magnani/pdf/Tavole_chi-quadrato.pdf. Accessed 24 Apr. 2020

Works cited

Andridge, Rebecca R., Little, Roderick J.A. "A Review of Hot Deck Imputation for Survey Non-response." *International Statistical Review*, vol. 78, no. 1, 2010, pp. 40–64. www.ncbi.nlm.nih.gov/pmc/articles/PMC3130338/#. Accessed 23 Apr. 2020

"Critical Values of the Chi-Square Distribution." *Engineering Statistics Handbook*, Information Technology Laboratory. www3.med.unipmn.it/~magnani/pdf/Tavole_chi-quadrato.pdf. Accessed 24 Apr. 2020

Chang Wathall, Jennifer, et al. *Mathematics: Applications and Interpretations Standard Level Course Companion*. Oxford, 2019

"Putts per round." *PGA Tour Statistics.* PGA Tour, www.pgatour.com/content/pgatour/stats/stat.119.y2019.eoff.t060.html. Accessed 15 Apr. 2020

Rudy, Matthew. "Why Women Putt Worse Than Men." *Golf* Digest, 11 Aug. 2010. https://www.golfdigest.com/story/putting-matthew-rudy. Accessed 28 Jan. 2021

"Scoring average." *PGA Tour Statistics.* PGA Tour, www.pgatour.com/stats/stat.120.html. Accessed 15 Apr. 2020

Yates, Frank. "Contingency Tables Involving Small Numbers and the χ^2 Test." *Supplement to the Journal of the Royal Statistical Society*, vol. 1, no. 2, 1934, pp. 217–235. *JSTOR*, www.jstor.org/stable/2983604. Accessed 2 Feb. 2021.