A Statistical Examination of Golf Strokes to Predict Scoring Average*

Driving or Putting: which has a larger impact on overall game?

Arye Santosh

October 29, 2025

This paper uses simple linear regression to test whether driving or putting is a stronger linear predictor of scoring average on the PGA Tour. Using aggregated 2022 player data, scoring average is modeled as a function of average strokes gained off-the-tee and average strokes gained putting. The analysis reveals that while both predictors have a statistically significant negative relationship with scoring average, the model with putting performance has a higher R-squared than driving performance and explains more of the variability. From the data and analysis, it can be concluded that putting performance is a stronger linear predictor of average score for professional golfers.

Introduction

In the world of golf, the age old adage goes "Drive for show, putt for dough." It suggests that while long tee shots (drives) are impressive, it's a skilled short game (putting) that secures more match victories and consequently, more financial success. This is a long held belief in golf, understandably influencing which aspect people think of as more difficult or more profitable in some cases. Professional Golfers' Association (PGA) tour performance data can assist in substantiating such a claim (Tour 2025).

While the relationship between certain golf skills is often discussed as tee time banter, this paper aims to address this area which is seldom backed by quantitative analysis. Using actual PGA metrics such as strokes gained, it can be determined whether power off the tee or expertise on the green is more strongly (linearly) associated with a professional golfers performance.

As aforementioned, actual PGA tour data will be utilized in this analysis. A set of simple linear regression models will be used to model a golfer's scoring average (`avgScore`) based

*Project repository available at: https://github.com/arye69/MATH261A-p1.

on the two key performance metrics: strokes gained off the tee (`driveSG`) and strokes gained putting (`puttsSG`). The heart of this analysis lies in fitting and comparing these two models to see which is a more effective linear predictor of scoring. The paper will also examine the validity of a linear model in this context.

The remainder of the paper is structured as follows. The **Data** section goes over the PGA dataset in detail including its source, relevant variables, and an explanation of pertinent metrics. The **Methods** section defines the simple linear regression technique used and describes the process of evaluating model fit and assumption checks. Finally, the **Results** section presents and assesses the findings of the fitted models, including summaries of the data and diagnostic plots.

Data

The data for this analysis was obtained from the SCORE Sports Data Repository which sourced its data from the official PGA Tour stats website after the 2022 tournament season (Alyssa Bigness and PGA Tour 2023).

The observational unit in the raw dataset (`PGA2022`) is a player-tournament combination. Each of the 1387 observations represents a single players' stats from a single tournament they played in. Each tournament consists of four rounds and it's possible for golfers to be eliminated after as few as two rounds. A caveat of the data collection process is that only players that made it *all four rounds* were taken as observations. This means that only the most skilled of all competing professional golfers were included in this data. The data covers 19 tournaments and 280 unique golfers in total.

Relevant variables

The following variables are relevant:

- `avgScore` (Scoring average): The response variable denoting the average number of strokes a golfer used for a standard 18-hole. It's calculated as the total number of strokes for all holes divided by the total tournament rounds (usually four). Lower is better.
- `puttsSG` (Strokes gained putting): A predictor that measures how many strokes a golfer gains or loses compared to the PGA average on the green. For example, if the average strokes required to make a putt on the green is 1.5 and a golfer makes it in one stroke, they've gained 0.5 strokes. A positive value indicates better than average putting ability.

- `driveSG` (Strokes gained driving): A predictor that measures how many strokes a golfer gains or loses compared to the PGA average off-the-tee on all par 4s and 5s. The metric is calculated exactly how `puttsSG` is and accounts for both distance and accuracy. Again, a positive value indicates better than average driving ability.

A key limitation to note is that the average score is prone to artificial inflation and deflation. Certain tournaments and courses are simply more difficult than others, and factors like changes in weather and conditions play a role as well. A player who plays in friendlier tournaments like The Sentry or The American Express, where scores are consistently under par, is likely to have a lower average score than a player who competes in difficult tournaments like the Honda Classic or Arnold Palmer Invitational, which are intensely technical.

While the `avgScore` variable is affected by this course difference, both predictors (of the Strokes Gained metric) are not. This is due to the nature of the Strokes Gained metric, which is both a course and situation specific measurement. This is possible due to the massive amount of data collected by the PGA. For any given shot on any course within a PGA Tour sponsored tournament, they know the average number of shots it takes a pro golfer to get the ball in the hole from the *exact* location and situation. This makes both the `puttsSg` and `driveSG` variables reliable predictors due to their inherent capability to control for course difficulty (Shotscope 2025).

Data cleaning

With over 1300 observations and only 280 unique golfers, it's clear that to measure any kind of linear relationship without introducing a false correlation effect, the data must be reduced down to a maximum of 280 observations, with only a single golfer per observation. To do this, the original dataframe was aggregated into only the averages of all pertinent metrics using R (R Core Team 2024). The result is a dataframe with the new variables `avg_Score` (average score, response), `avg_puttsSG` (average strokes gained putting, predictor), and `avg_driveSG` (average strokes gained off-the-tee, predictor). All of these are simply averages of the original stats by player, which reduces the bias of treating each tournament as an independent event.

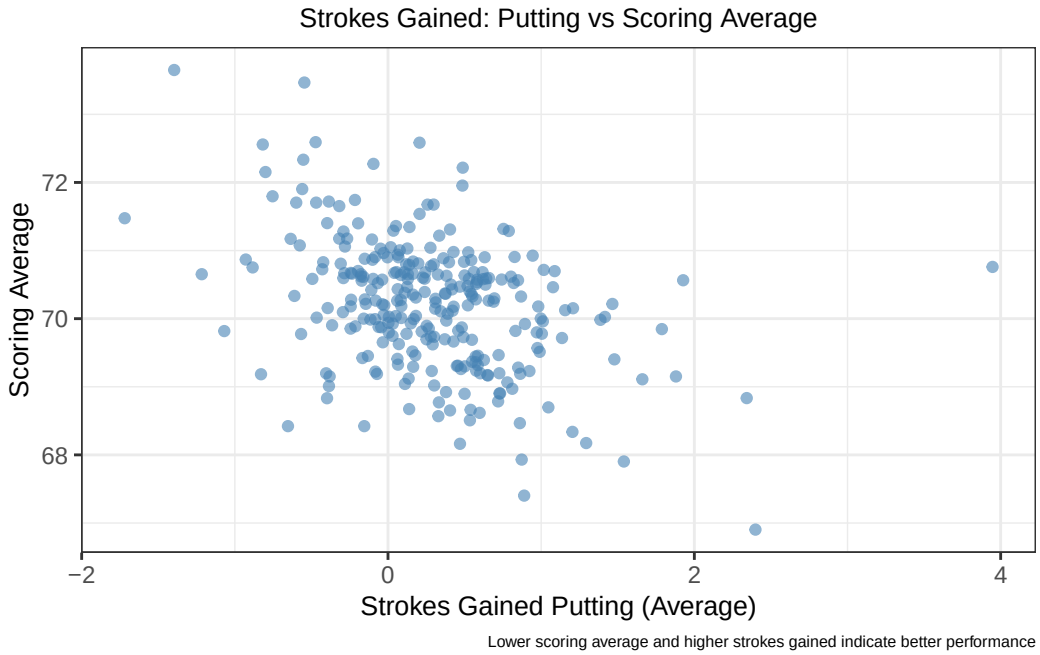
Two observations for players D.J. Trahan and David Hearn had empty stats in the original dataframe and were omitted from the cleaned data, resulting in a total 278 unique golfers with complete stats.

Some useful summary statistics about the pertinent variables can be found below, compiled in Table 1 using Kable (Zhu 2024).

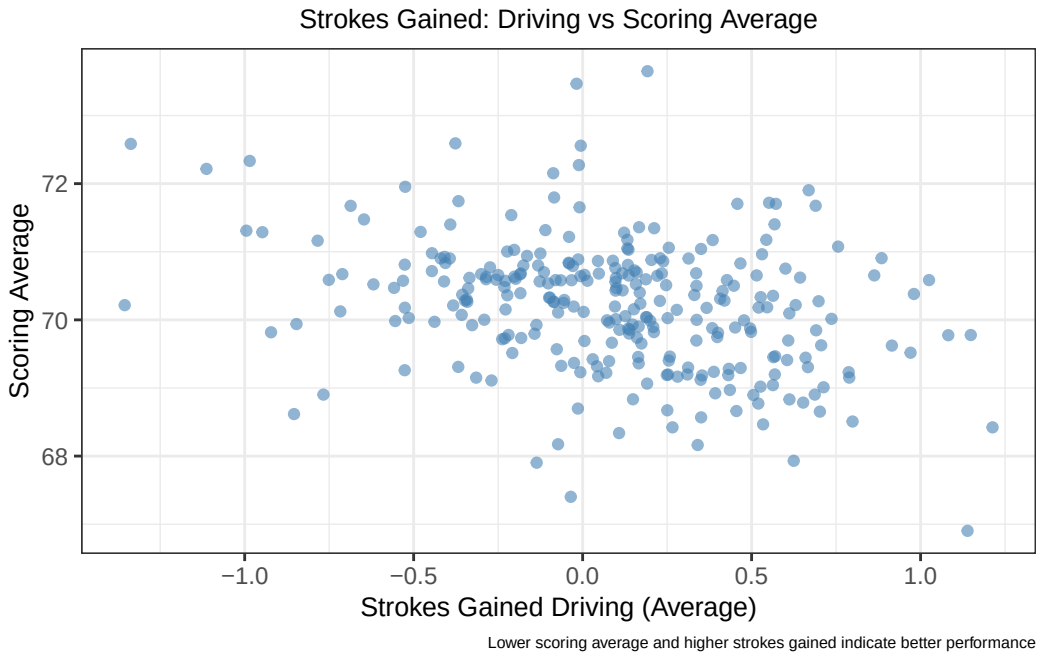
Find a scatter plot of `avg_puttsSg` vs. `avg_Score` below. All plots were created using ggplot2 (Wickham 2016).

Table 1: Summary Statistics of Key Variables (n = 278 golfers)

Variable	Mean	SD	Min	Max
avg_Score	70.21	0.95	66.90	73.65
avg_puttsSG	0.28	0.61	-1.72	3.95
avg_driveSG	0.09	0.44	-1.35	1.21



Find a scatter plot of `avg_driveSG` vs. `avg_Score` below.



The plots suggest that negative linear relationships might exist between both strokes gained metrics and scoring average, which aligns with the nature of the game: gaining strokes will likely improve scoring average.

Limitations

A potential limitation of the data is measurement error. While the PGA Tour's ShotLink system is leagues ahead of what shot tracking used to be, it's still prone to factors like weather, shot variability, and equipment limitation. Additionally, the "Strokes Gained" metric is useful for measuring relative performance, but doesn't necessarily capture all aspects of how a golfer plays and relies heavily on the accuracy of the shot tracking equipment.

As stated earlier, the tournament independence biased was mitigated as best as possible by using an aggregation of player stats. However, it should be noted that aggregation removes the ability to model the effects of specific course and tournament conditions. Hence, conclusions about such effects won't be discussed in this analysis.

Finally, the limitation of varying course difficulty with regards to `avgScore` is one that cannot be resolved without further data collection. The idea of gauging course difficulty is one that isn't feasible since while there are certainly objectively difficult courses, there's a subjective aspect involved as well. Since the response variable is biased in this sense, it introduces some additional noise, but it's reasonable to assume that the trends will be preserved. See the **Relevant Variables** section for more on course difficulty and its effects.

Methods

To model a linear relationship between `avg_Score` and each of the predictors (`avg_puttsSg` and `avg_driveSG`), a simple linear regression (SLR) model will be used, with the goal of determining which metric is a stronger linear predictor of average score. The two models used are as follows.

Model 1: Putting performance

$$\text{avg\_Score} = \beta_0 + \beta_1(\text{avg\_puttsSG}) + \varepsilon$$

Model 2: Driving performance

$$\text{avg\_Score} = \beta_0 + \beta_1(\text{avg\_driveSG}) + \varepsilon$$

Where

- `avg_Score` is the response variable, scoring average
- `avg_puttsSG` and `avg_driveSG` are the predictors in the two separate models. These are the average strokes gained putting and average strokes gained off-the-tee, discussed in greater detail in the **Data** section.
- β_0 is the intercept term for each model, representing the average score when each predictor is equal to 0. Note that although the term β_0 is used for both models, the numeric value for this parameter will likely be different across models.
- β_1 is the slope term for each model, representing the expected change in average score given a unit increase in each predictor. Again, note that while the term is used for both models, it's likely the value will not be the same across models.
- ε is the error term for each model, denoting the variation not captured by the model.

Variable selection and justification

The response variable `avg_Score` was selected due to its straightforward measure of tournament success. There were other success metrics available including money and points, but certain tournaments offer different cash prizes and point levels, resulting in the need to standardize across tournaments in order to get meaningful results. For this reason, the average player score worked the best and yields meaningful results without many complications. Meanwhile, the predictor variables were the best choice since they're both Strokes Gained metrics. Primarily, both predictors being in terms of the same metric removed an element of bias across putting and driving analysis. Additionally, Strokes Gained is a tried and true metric within golf, further strengthening the use of the chosen predictors.

With regards to the chosen method, SLR seemed like a viable choice after visually inspecting the plots (in the **Data** section). From a visual standpoint, there's enough evidence to say that the true relationship could possibly be linear.

Model Validation

A SLR analysis assumes linearity, independence of observations, normality of residuals, and equal variances (homoscedasticity). These assumptions were verified using the following.

- Examination of plots of both predictors during exploratory analysis (see **Data**)
- Inspection of relevant diagnostic plots (residuals plots and QQ plots in **Results**)
- Aggregation of player stats to ensure all provided data was encompassed while keeping players independent (see **Data**)

The fit of both models was evaluated using the coefficient of determination R^2 and the predictor significance was evaluated using the inbuilt slope parameter t-tests in R with a significance level of $\alpha = 0.05$.

Software

All statistical analysis was done in R (R Core Team 2024). The here package was used to import and load files (Müller 2025). The lm() function in base R was utilized to fit the linear regression model and further diagnostic checks were generated from the same function output. Data manipulation and cleaning was performed using the Tidyverse package (Wickham et al. 2019). All plots and visualizations were created using ggplot2 (Wickham 2016) unless otherwise stated.

Limitations and extensions

A critical limitation of the structure of the data in this analysis is that course/tournament-specific effects will not be accounted for, as well as the variability within players' performances. This is due to the aggregation and is discussed in greater detail in the **Data** section.

A logical extension to this work could be a multiple linear regression model using both `avg_puttsSG` and `avg_driveSG` as predictors. It's possible that using both predictors could improve average score prediction. Another extension could be seeing whether the same predictors could be manipulated to work at predicting either financial success or overall tournament points. Finally, if it seems that the models in fact are not linear after SLR models are fitted, nonlinear transformations of the response would also be a possible extension.

Results

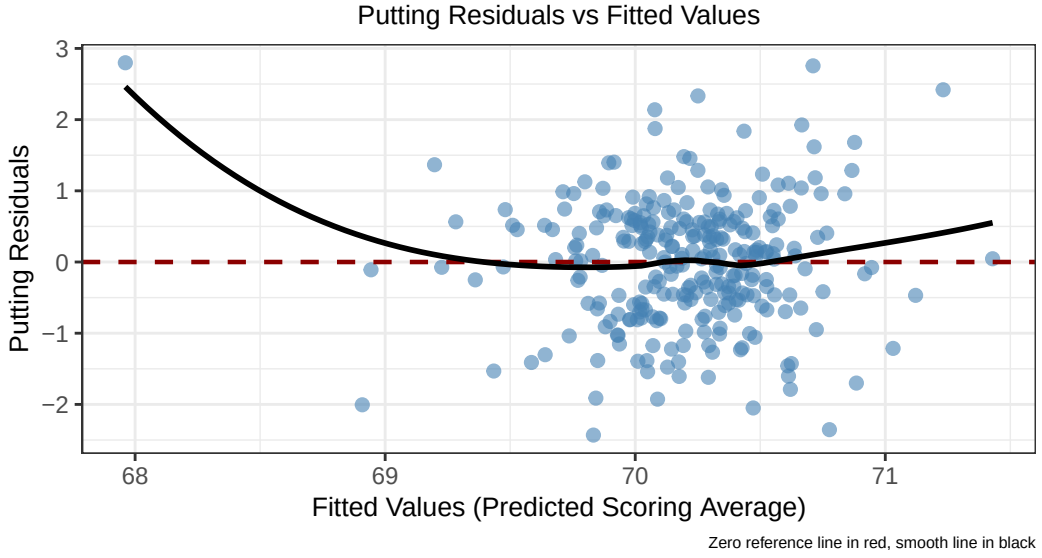
This section contains the results of the previously described analysis performed on each of the two proposed models. It starts with a check of all necessary assumptions for inference upon a linear regression model. After the assumptions check, both models are analyzed and a comparison is performed to address the central research question.

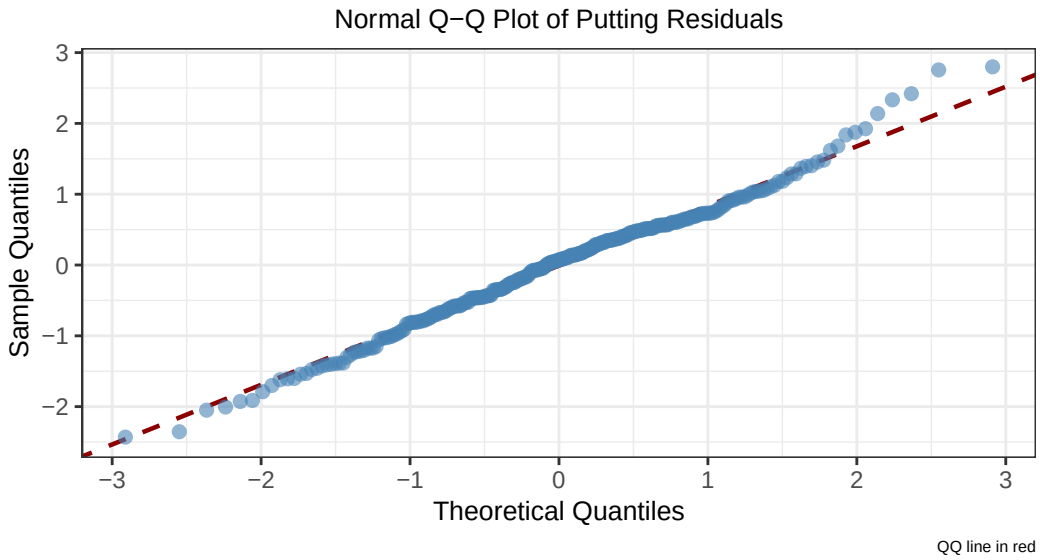
Assumptions

The assumptions for performing valid inference on a Normal error regression model are as follows:

1. Linearity: The relationship between variables must reasonably linear. The exploratory scatterplots show this is the case (see the **Data Cleaning** section).
2. Independence: Since the data consists of one observation per professional golfer (see **Data Cleaning**), it is assumed that golfers' performances are independent of each other.
3. Normality of errors: The residuals for both models should be approximately normally distributed. A QQ plot of the residuals will be examined further along in this section.
4. Equal variances: The variance of residuals for both models should be relatively constant. Residual plots are examined further in this section.

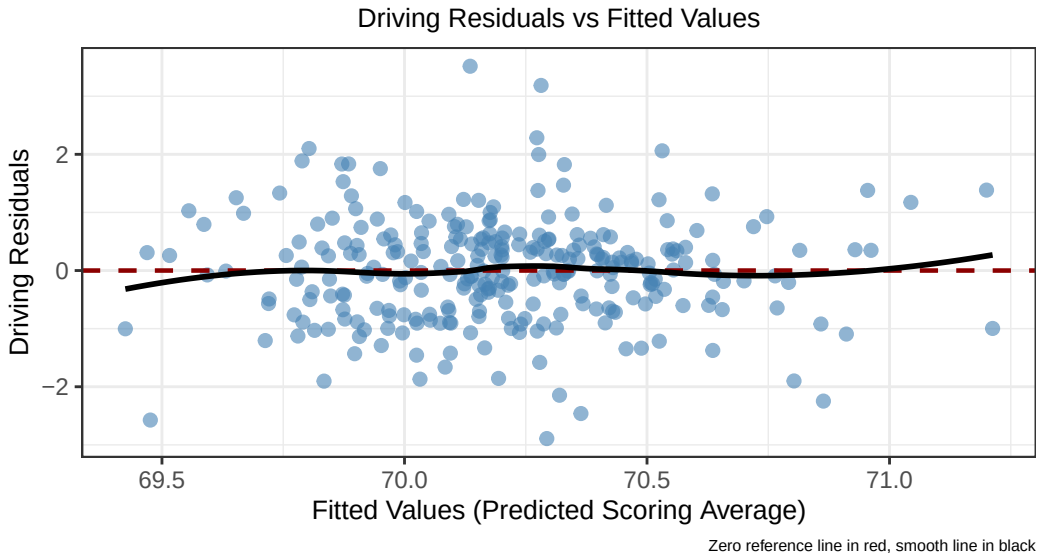
The residuals and QQ plots for both the putting model can be found below:

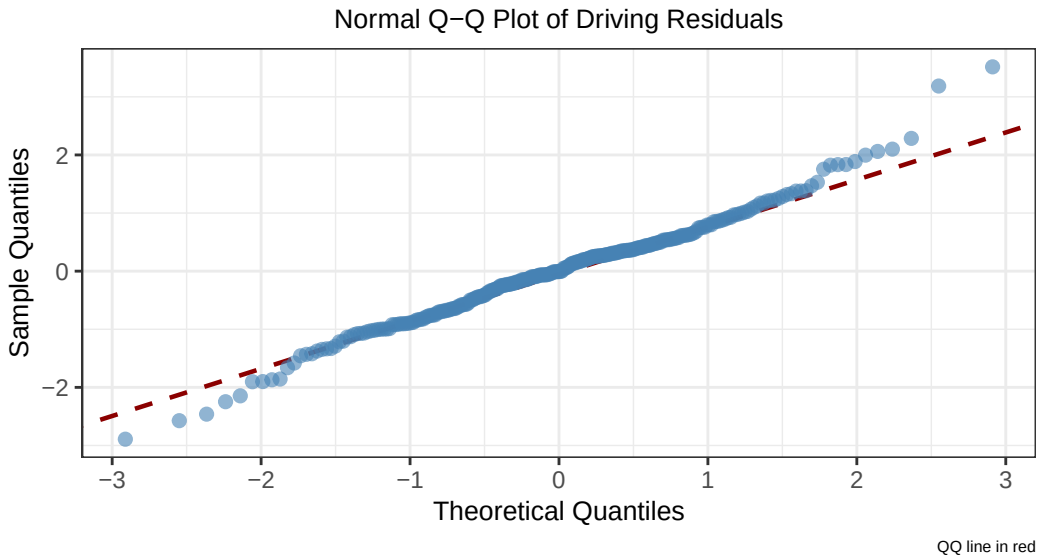




From the plots, it can be seen that the residuals, while heavily clustered, have similar variance. There's no 'cone shape', which is encouraging, although observation 246 appears to be a heavy outlier. The QQ plot is also relatively good, showing that the residuals are likely normal.

Below are the residuals and QQ plots for the driving model:





The driving residuals plot looks even better than the one for putting, with only a few outliers, but mostly constant variance. The QQ plot is again very good with almost all observations falling on the line, indicating passable error normality.

From the diagnostic plots, it can be concluded that the data for both models reasonably passes the final two assumptions (equal variances and error normality), validating the inferences in the following section.

Model 1: Putting performance and scoring average

Fitting an SLR model between strokes gained putting and scoring average yields the following table created using Stargazer (Hlavac 2022):

The hypotheses for the test on the slope coefficient β_1 were:

- $H_0 : \beta_1 = 0$ (There is no linear relationship between average strokes gained putting and average score)
- $H_a : \beta_1 \neq 0$ (There is a linear relationship between average strokes gained putting and average score)

As can be seen from the table, the regression revealed a statistically significant negative relationship between putting performance and average score with a b_1 p-value of $1.1146095 \times 10^{-11}$. This p-value is for the t-test of the slope coefficient and is less than the common alpha level of 0.05, leading us to reject H_0. Find a discussion of the pertinent assumptions for such a test in the **Assumptions** section. The fitted regression equation is

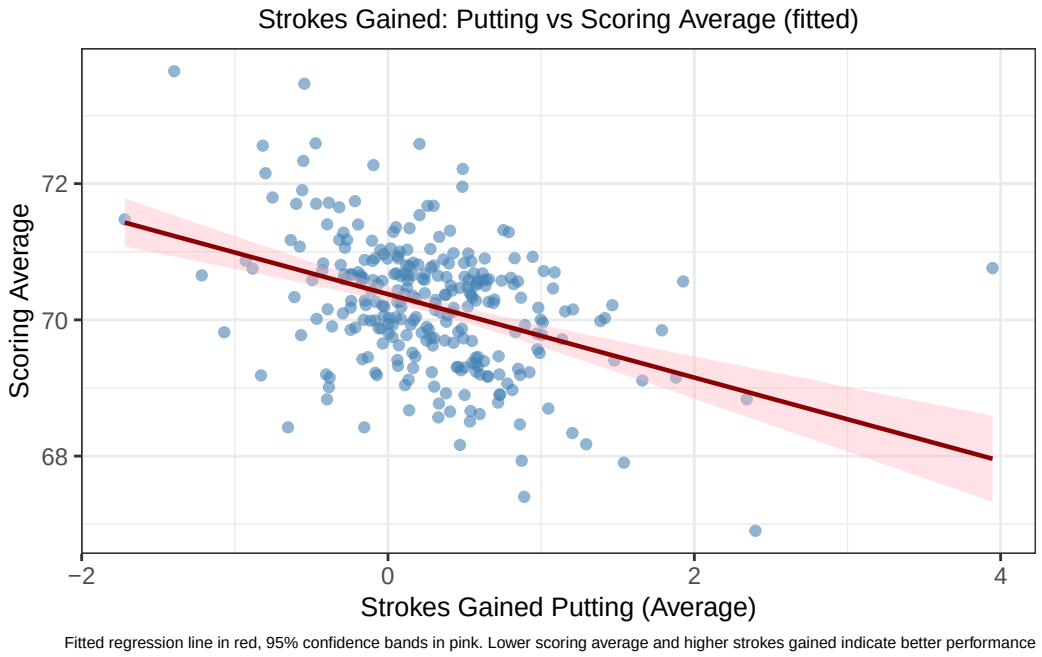
Table 2: Regression of scoring average on putting performance

	Dependent variable:
	avg_Score
avg_puttsSG	-0.612^{***}
	(0.086)
Constant	70.377^{***}
	(0.058)
Observations	278
R^2	0.154
Adjusted R^2	0.151
Residual Std. Error	0.880 (df = 276)
F Statistic	50.290^{***} (df = 1; 276)
Note:	*p<0.1; **p<0.05; ***p<0.01

$$\text{avg\_Score} = 70.377 - 0.612(\text{avg\_puttsSG})$$

The slope coefficient of -0.612 means that for every stroke gained against the PGA average on the green, a professional golfers scoring average is expected to decrease by 0.612 strokes. The R^2 value of the model is 0.154, which means that putting performance only accounted for about 15.4% of the variation in average score. This translates to average strokes gained putting being a weak predictor of average score.

Below is a plot of `avg_Score` vs. `avg_puttsSG` with a fitted regression line and 95% confidence interval shading around it.



Model 2: Driving performance and scoring average

The SLR model of scoring average on strokes gained off-the-tee produced the following output table:

Table 3: Regression of scoring average on driving performance

	Dependent variable:
	avg_Score
avg_driveSG	−0.696***
	(0.123)
Constant	70.269***
	(0.055)
Observations	278
R^2	0.104
Adjusted R^2	0.101
Residual Std. Error	0.905 (df = 276)
F Statistic	31.994*** (df = 1; 276)
Note:	*p<0.1; **p<0.05; ***p<0.01

The hypotheses for the test on the slope coefficient β_1 were:

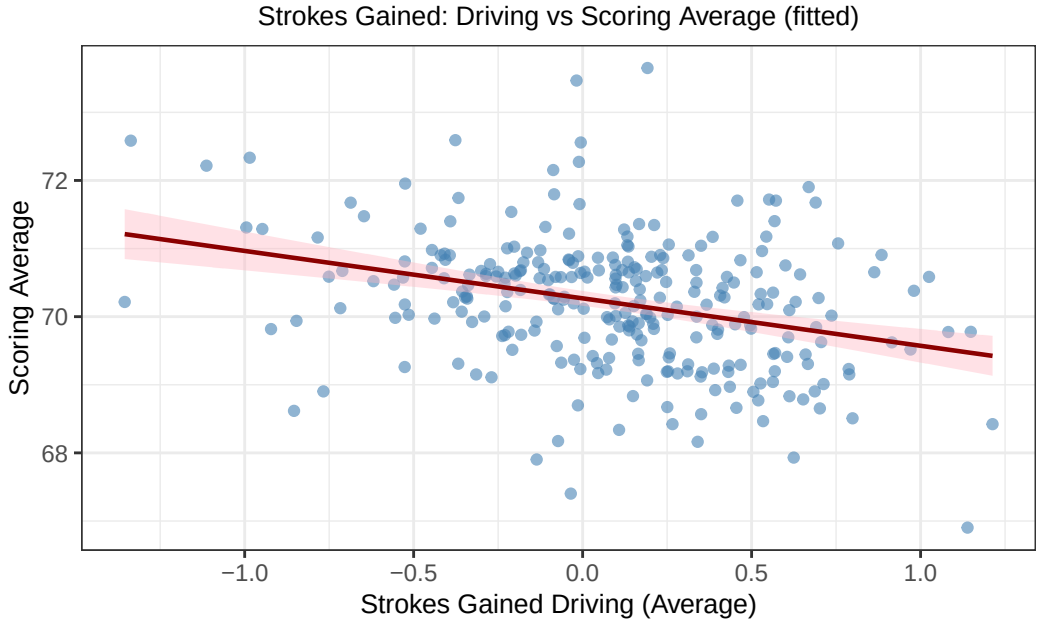
- $H_0 : \beta_1 = 0$ (There is no linear relationship between average strokes gained driving and average score)
- $H_a : \beta_1 \neq 0$ (There is a linear relationship between average strokes gained driving and average score)

As expected, this regression also yielded a statistically signficant negative relationship between driving performance and scoring average. The p-value for the slope of this model is 3.8631421×10^{-8}, leading us to reject the null hypothesis again at significance level 0.05. Find a discussion of the pertinent assumptions for the β_1 t-test in the **Assumptions** section. The fitted regression equation is

$$\text{avg\_Score} = 70.269 - 0.696(\text{avg\_driveSG})$$

Similar to the putting performance, the -0.696 slope coefficient denotes that for every one stroke increase in strokes gained over the PGA average, a professional golfer's average score is expected to decrease by 0.696 points. Despite the more extreme slope coefficient between average strokes gained off-the-tee and average score, the R^2 is even lower than that of the previous model, at 0.104. This means only 10.4% of the variation in average score is explained by `avg_driveSG`, making it a weak predictor.

Below is a plot of `avg_Score` vs. `avg_puttsSG` with a fitted regression line and 95% confidence interval shading.



Conclusions

From the SLR analysis of both models, it was found that both strokes gained metrics had a statistically significant negative relationship with average score. This was predicted, since doing better than the average amount of strokes on any golf course will translate to a lower (better) score.

The analysis also found that the model with average strokes gained putting as a predictor had an R^2 of 0.154 while the model with average strokes gained driving had an R^2 of 0.104. From the limited scope of this analysis, the data suggests that to answer the initial question, putting is a stronger linear predictor of PGA tournament success.

Although the suggestion that putting is a stronger predictor of score than driving can be derived from this analysis, it's important to note that both models represented less than a quarter of the overall variation in average score. This speaks to the complicated nature of scores even in something as seemingly straightforward as golf. There are likely better predictors for score, although they may or may not be linear. They not even be measurable aspects, such as the mental state and focus of a golfer within a high pressure tournament.

While this study led to the interesting result that putting performance might be the more reliable path to success, driving and other aspects are fundamental in constituting a good score. The most successful golfers are those who strive to optimize all parts of their game.

References

Alyssa Bigness, and PGA Tour. 2023. "Statistics of PGA Tournaments During the 2022 Season." https://www.score.com/data/pga2022; SCORE Sports Data Repository.

Hlavac, Marek. 2022. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Social Policy Institute. https://CRAN.R-project.org/package=stargazer.

Müller, Kirill. 2025. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Shotscope. 2025. "Understanding Strokes Gained: How It Works." https://shotscope.com/blog/practice-green/stats-and-data/understanding-strokes-gained/; Shotscope.

Tour, PGA. 2025. "PGA Tour Stats." https://www.pgatour.com/stats; PGA Tour.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.