A Statistical Examination of Golf Strokes to Predict Scoring Average*

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

Arye Santosh

September 21, 2025

Abstract here.

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. PGA tour performance data can assist in substantiating such a claim.

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 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

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

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.

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.

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. 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.

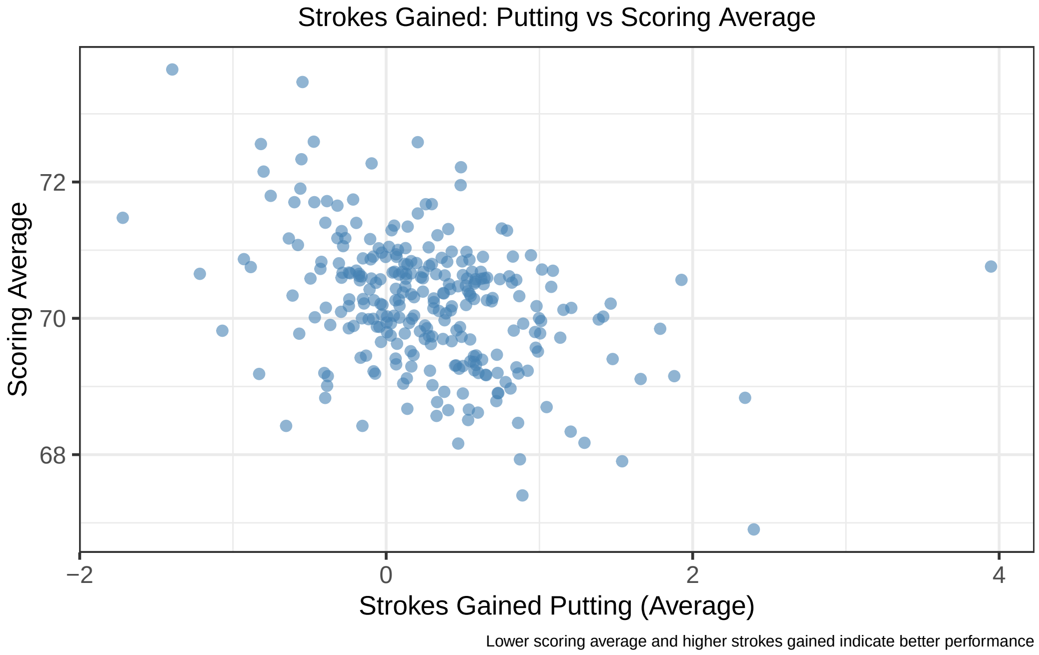
Two observations for players D.J. Trahan and David Hearn had empty stats in the original dataframe and were omitted from the cleaned data, resulting in a total 278 unique golfers with complete stats.

Some useful summary statistics about the pertinent variables can be found below, in Table 1.

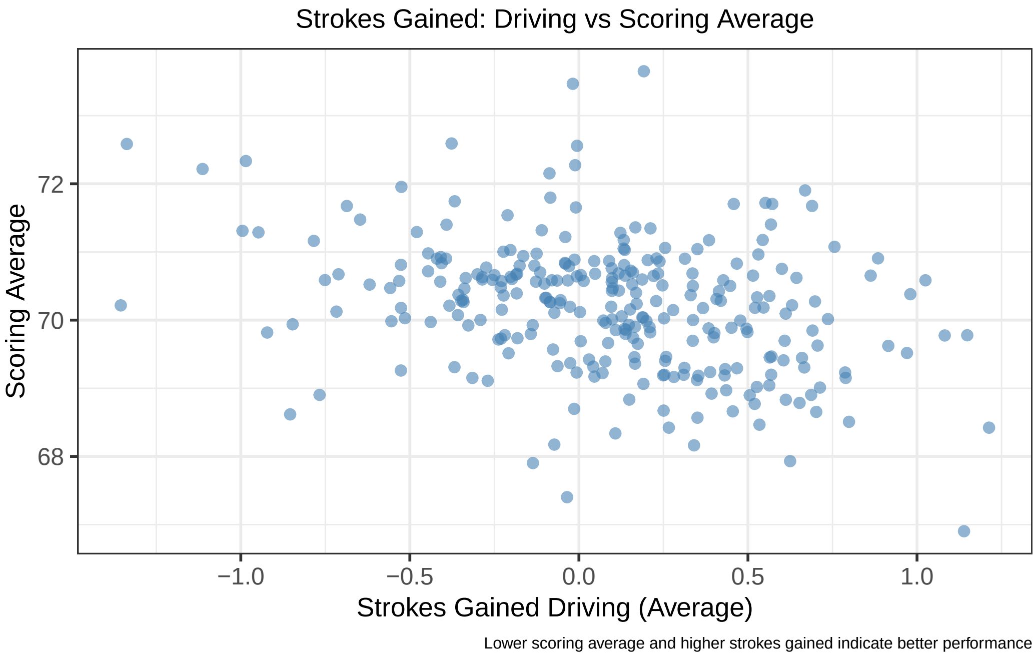
Table 1: Summary Statistics of Key Variables (n = 278 golfers)

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_puttsSg` vs. `avg_Score` below.



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.

Methods

Results

References