# The Relationship Between Score and Driving Distance on the Professional Golfer's Association (PGA) Tour

**Abstract**

This paper investigates the relationship between average driving distance and performance for PGA Tour golfers. Using a dataset containing information on more than 400 different golfers over the course of nine years, we performed a best subsets analysis on the 7 most relevant explanatory variables (fairway percentage, number of rounds played, greens in regulation percentage, average putts per round, average scrambling percentage, number of wins, and number of top 10 finishes) to create a reduced multiple regression model. We then added average drive distance as an explanatory variable to the model and performed an Extra Sum of Squares test to compare the full and reduced models. Based on this ESS test and the fact that the full model (the model including average driving distance) had a higher adjusted R-squared value than the reduced model, we concluded that higher average driving distances are significantly correlated with increased average scores. Implications of this finding are discussed.

## Introduction

Golf is a challenging sport, rewarding precision and accuracy in players. Players must continually adapt, as the game and its strategies are constantly evolving. Over the past 100 years, average driving distance has increased along with the length of golf courses. Often, coaches often emphasize that players should increase their driving distance to improve, though there are concerns that this focus draws away from other crucial parts of the game (USGA, 2020). Moreover, from a purely mathematical perspective, there is a tradeoff between driving distance and accuracy. If a player tends to miss their target on either side by about 10% of their driving distance, for example, that 10% becomes larger as the driving distance increases. As such, it is unclear as to whether or not increasing driving distance genuinely improves golf performance.

At the professional level, the discourse concerning the relationship between driving distance and score has proven influential. One example of a PGA Tour golfer following the trend to increase driving distance is Bryson DeChambeau, who spent years adding yards to his drives (Golf, 2019). In fact, his current average driving distance of 322 yards leads the 2020-2021 PGA Tour season, with the average Tour player only driving the ball 295.3 yards (PGA Tour, 2021). This topic goes beyond DeChambeau, however: it has caught the attention of the United States Golf Association (USGA), which produces and interprets the rules of golf that PGA Tour players must follow. The USGA released a report stating that "increased driving distance can begin to undermine the core principle that the challenge of golf is about needing to demonstrate a broad range of skills to be successful" (USGA, 2020).

Given this background, we wondered if higher average driving distance could be related to lower average scores (which are desirable in golf) on the PGA Tour. Further, we wanted to examine the nature and extent of the relationship. We predicted that increased driving distance would be correlated with lower (better) average scores.

## Methods

### *Data Preparation*

The dataset used in our analysis was obtained from Kaggle and includes information on key golf metrics for 439 unique PGA Tour players between the years of 2010 and 2018 (Jong, 2019). The data was scraped from the PGA Tour website (Jong, 2019). Out of the 2,312 original combinations of years and players, 634 rows only contained data on four of the sixteen variables. Those observations were dropped. The final dataset contained 1678 observations and 19 variables. All analysis was completed in R Studio.

### *Variables and Exploratory Analysis*

Before we began our analysis, we decided to not consider each of the variables relating to shots gained because they directly explain average score. Furthermore, we did not consider variables relating to earnings or points, as these are, in part, reliant on average score. Thus, the variables remaining included fairway percentage, number of rounds played, average driving distance, greens in regulation percentage, average putts per round, average scrambling percentage, number of wins, and number of top 10 finishes. Each row represented a year for each player from 2010 to 2018. As such, all averages indicated the average for a given player during a given year of competition, and all percentages indicated a given percentage of the player's professional play for a given year. Additionally, each row for average score indicated the average score for a given player, for a given year across PGA Tour tournaments that the player competed in.

An exploratory data analysis revealed that each variable was roughly normally distributed, excluding year (Appendix A). This pattern of normal distribution was expected given that our dataset is representative of players on the PGA Tour.

*Creating the reduced model*

We first created a matrix correlation plot using our 8 possible explanatory variables to examine any potential multicollinearity in our model (Figure 1). Although there is some correlation between explanatory variables, our focus on average distance means we don't need to be too concerned with any multicollinearity not involving it. We then performed a best subsets analysis using the remaining explanatory variables, excluding average distance. We examined residual plots, normal probability plots, and residual vs. fit plots for all of our models to determine whether data transformations would be beneficial. These plots suggested that no transformations would improve the model, which we confirmed by testing various transformations with no positive result. Additionally, we examined whether interactions would improve the model and found that two interactions did (the interaction between rounds, average putts, gir, and top 10, and the interaction between rounds, average putts, gir, top 10, and average scrambling). Our final reduced model has an adjusted R^2 value of 0.779 and a residual plot analysis did not raise any flags for concern. The reduced model summary and residual plots are located in Appendix B and Appendix C, respectively.

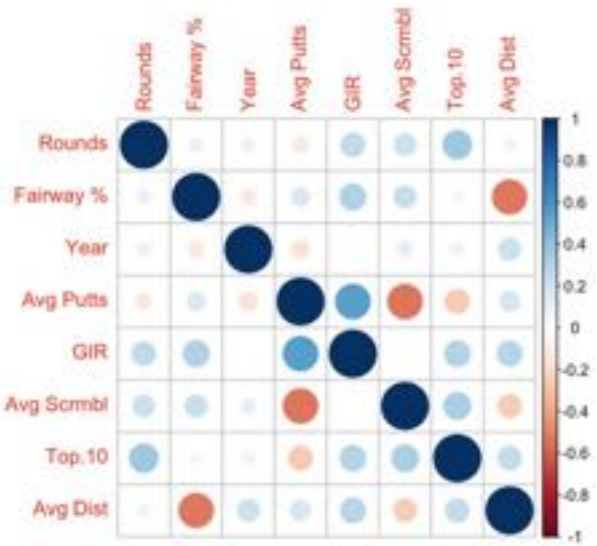


Figure 1. Matrix correlation plot for the possible explanatory variables in our model.

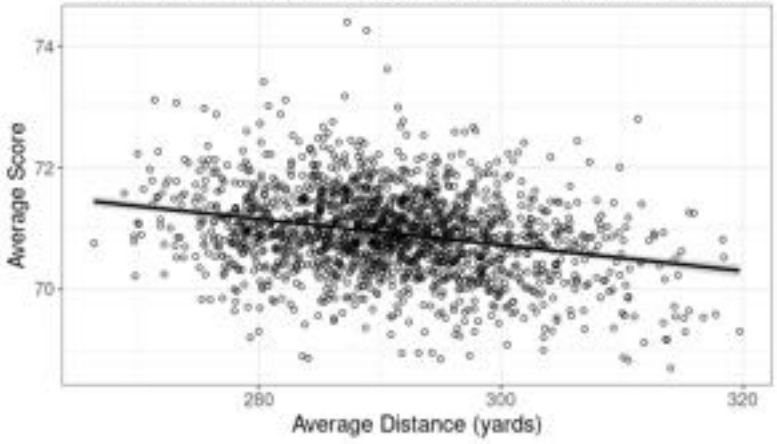
*Testing for the significance of average distance*

We created a full model that was identical to the reduced model, except that the average distance variable was included. No interaction terms or transformations involving average distance were included as a part of the full model. We then conducted an Extra Sum of Squares test between the full model and the reduced model to see if adding average distance significantly improves the overall fit of the model.

**Results**

The final reduced model contains the two interaction terms outlined above and no transformations to either explanatory or response variables. It has an adjusted $R^2$ value of 0.772, suggesting that it explains most of the variation in the data. Moreover, the reduced model summary indicated that each variable improved the model (Appendix B). The ESS test of the full model versus the reduced model resulted in a p-value of less than 0.00001, indicating a very significant difference in goodness-of-fit between the models (Appendix D). Thus, we found evidence that average distance should be included in the full model. Furthermore, the full model has an adjusted $R^2$ value of 0.794, indicating that the model improved due to the addition of the average distance variable and that each variable improves the model (Appendix E). Analysis of the residual plots indicates that no further transformations are necessary and that the model assumptions about the residuals are well met (Appendix F).

Figure 2. The Effect of Average Distance on Average Score



The coefficient of the average distance variable in the final model is -0.019 (p-value = 0.015). Interpretation of this coefficient suggests that a one yard increase in average driving distance, while holding all other factors constant, corresponds to an average score 0.019 strokes lower. This analysis suggests that there is a significant relationship between average distance and average score of the golfers in this dataset.

**Discussion**

The findings of this analysis indicate, as stated above, a significant correlation between a player's average driving distance and their average score. The findings do not indicate causation between the variables. The correlative relationship we found is nevertheless interesting for potential use in analysis of golf players' performance in the PGA.

Since golf is scored in a way that is inversely proportional to the number of drives, it makes sense that golfers who make longer distances for each drive would need fewer drives overall to reach the hole and therefore earn a better score. However, since longer shots may also be less accurate, there is cause to believe that golfers who take longer shots may also miss their desired target more often, resulting in a higher number of total shots and therefore a worse score. As a result, it is interesting that the model suggests that average distance correlates with score. That result implies that the drives of better golfers are, on average, both longer and more accurate than those of lesser golfers, as logic dictates that simply having either a long driving distance or good accuracy may not be enough to make a top-tier golfer.

Our findings are limited by a few sources of error. One is the size of the dataset. While the dataset contains useful information about many players in the PGA, its 1,678 usable observations are not near the scale of some 50,000+-case datasets used to generalize predictions to large populations. Further, the ability to analyze professional golf players is limited by the total number of people who have ever played professional golf, and limited further if we wish only to generalize results to players in the PGA as opposed to all professional golfers. Furthermore, our dataset contains data from only 2010 to 2018, so the results we found may only be applicable to current pro golfers.

For future research on this topic, it would be interesting to perform the same analysis on different golf datasets, such as those containing the statistics of amateur or college golf players, to see if the findings are also present at other skill levels. It would also be worthwhile to conducted the same test on data striated by year or decade to see if the impact of average distance on score has changed over time, as historical context suggests (USGA, 2020).

Potential error sources notwithstanding, using data to analyze and predict golf scores is a practice performed by many sports journalists (Porter 2010, for example), and resultingly even smaller-scale analyses that can be generalized only to a given league are still useful for predictive and explanatory purposes. As a result, the information gleaned from our analysis about the correlation between average driving distance and average score of the golfers in the dataset could very possibly be useful in the ongoing attempt to determine which skills and characteristics make or break players in professional golf.

**References**

*Golf Stat and Records: PGA TOUR*. PGATour. (n.d.). https://www.pgatour.com/stats.html.

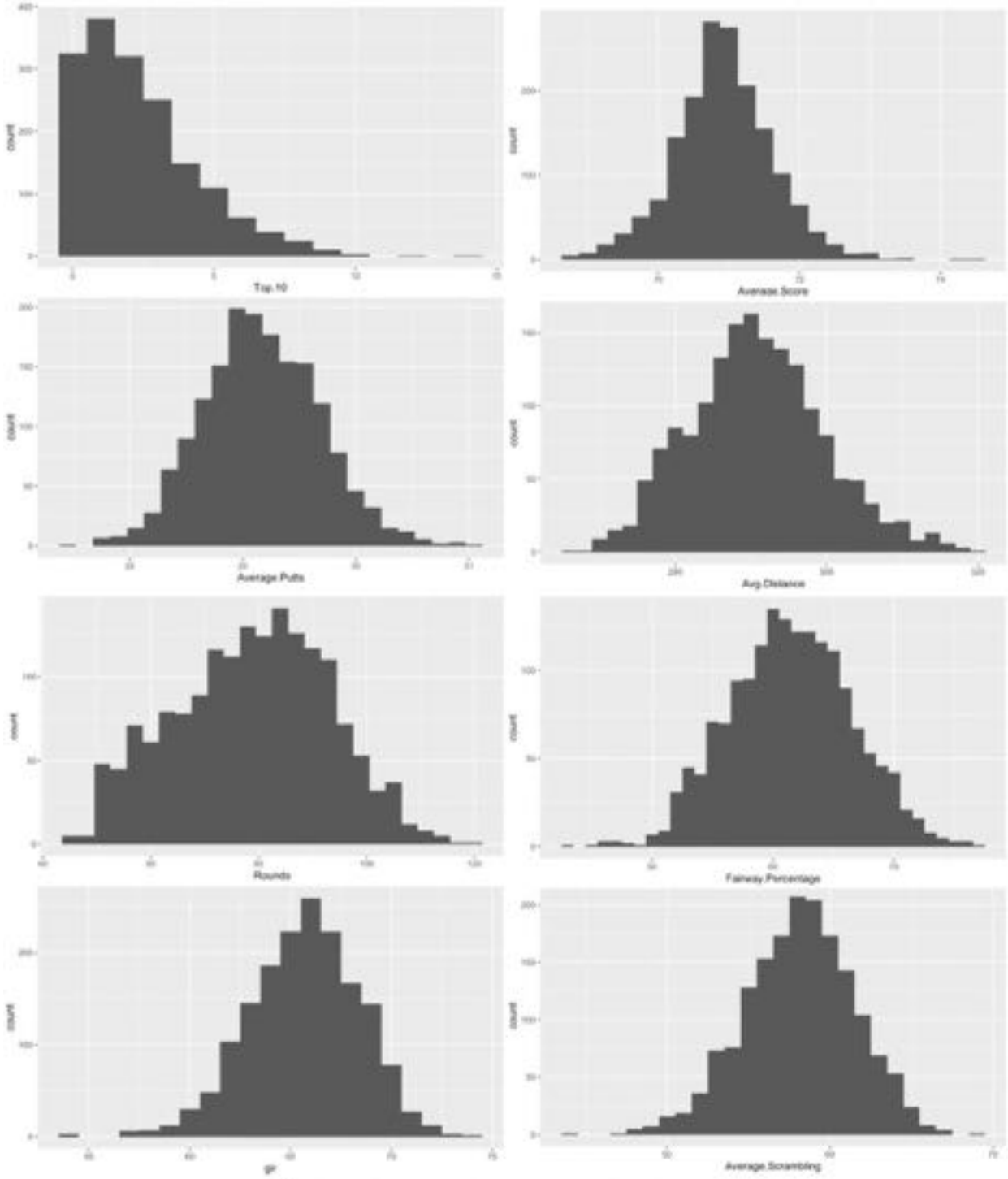
*'I'm going to become massive': Why Bryson's bulk-up is just getting started*. Golf. (n.d.). https://golf.com/news/features/bryson-dechambeau-massive-workout-routine/.

Jong. (2019, April 30). *PGA Tour Data*. Kaggle. https://www.kaggle.com/jmpark746/pga-tour-data-2010-2018.

Porter, Kyle. *2021 Zurich Classic picks, format, field grade, odds, best bets, predictions at TPC Louisiana*. CBS Sports (2021, April 22). https://www.cbssports.com/golf/news/2021-zurich-classic-picks-format-field-grade-odds-best-bets-predictions-at-tpc-louisiana/

*Key Findings of Distance Insights Project Released*. USGA. (2020, February 19). https://www.usga.org/content/usga/home-page/articles/2020/02/key-findings-distance-insights-usga.html.

## Appendix




Appendix A: Histograms of each variable considered in the models, each demonstrating roughly normal distribution aside from 'Top.10' ('Year' not included, but there are roughly the same number of observations across each of the years).

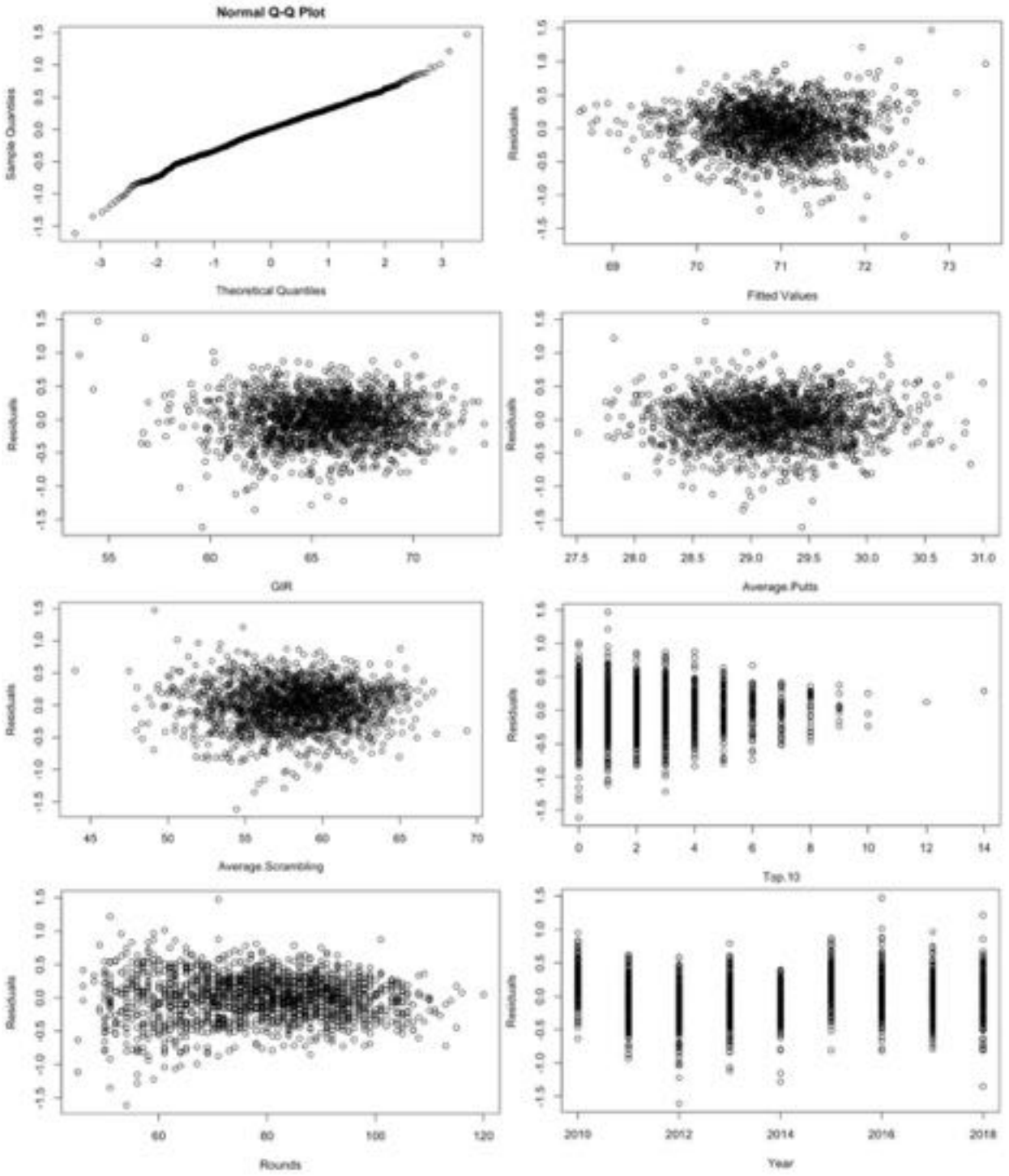
```
Call:
lm(formula = pga$Average.Score ~ pga$Rounds + pga$Year + pga$Average.Putts +
    pga$gir + pga$Average.Scrambling + pga$Top.10 + pga$interaction1 +
    pga$Fairway.Percentage + pga$interaction2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.61095 -0.21188  0.01464  0.22437  1.47317

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.042e+01  6.612e+00   3.088  0.00205 **
pga$Rounds             -4.119e-03  8.984e-04  -4.585 4.89e-06 ***
pga$Year                2.417e-02  3.216e-03   7.517 9.12e-14 ***
pga$Average.Putts       4.988e-01  3.041e-02  16.402  < 2e-16 ***
pga$gir                -1.521e-01  5.068e-03 -30.017  < 2e-16 ***
pga$Average.Scrambling -4.069e-02  4.258e-03  -9.556  < 2e-16 ***
pga$Top.10             -3.869e-01  2.355e-02 -16.428  < 2e-16 ***
pga$interaction1        3.708e-08  4.682e-07   0.079  0.93689
pga$Fairway.Percentage  4.445e-03  1.818e-03   2.446  0.01456 *
pga$interaction2        2.585e-08  7.170e-09   3.606  0.00032 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3333 on 1668 degrees of freedom
Multiple R-squared:  0.7734,    Adjusted R-squared:  0.7722
F-statistic: 632.5 on 9 and 1668 DF,  p-value: < 2.2e-16
```

Appendix B: Final Reduced Model Summary. 'Interaction1' represents the interaction between 'Rounds','Average.Putts','gir', and 'Top.10'. 'Interaction2' represents the interaction between 'Rounds','Average.Putts','gir','Average.Scrambling', and 'Top.10'.




Appendix C: Residual plots for the final reduced model, none of which invoked concern about the model assumptions and composition.

```
Analysis of Variance Table

Model 1: pga$Average.Score ~ pga$Rounds + pga$Year + pga$Average.Putts +
    pga$gir + pga$Average.Scrambling + pga$Top.10 + pga$interaction1 +
    pga$Fairway.Percentage + pga$interaction2
Model 2: pga$Average.Score ~ pga$Rounds + pga$Year + pga$Average.Putts +
    pga$gir + pga$Average.Scrambling + pga$Top.10 + pga$interaction1 +
    pga$interaction2 + pga$Avg.Distance + pga$Fairway.Percentage
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1668 185.31
2   1667 167.32  1    17.991 179.24 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix D: Results from the ESS test of the reduced model versus the full model.

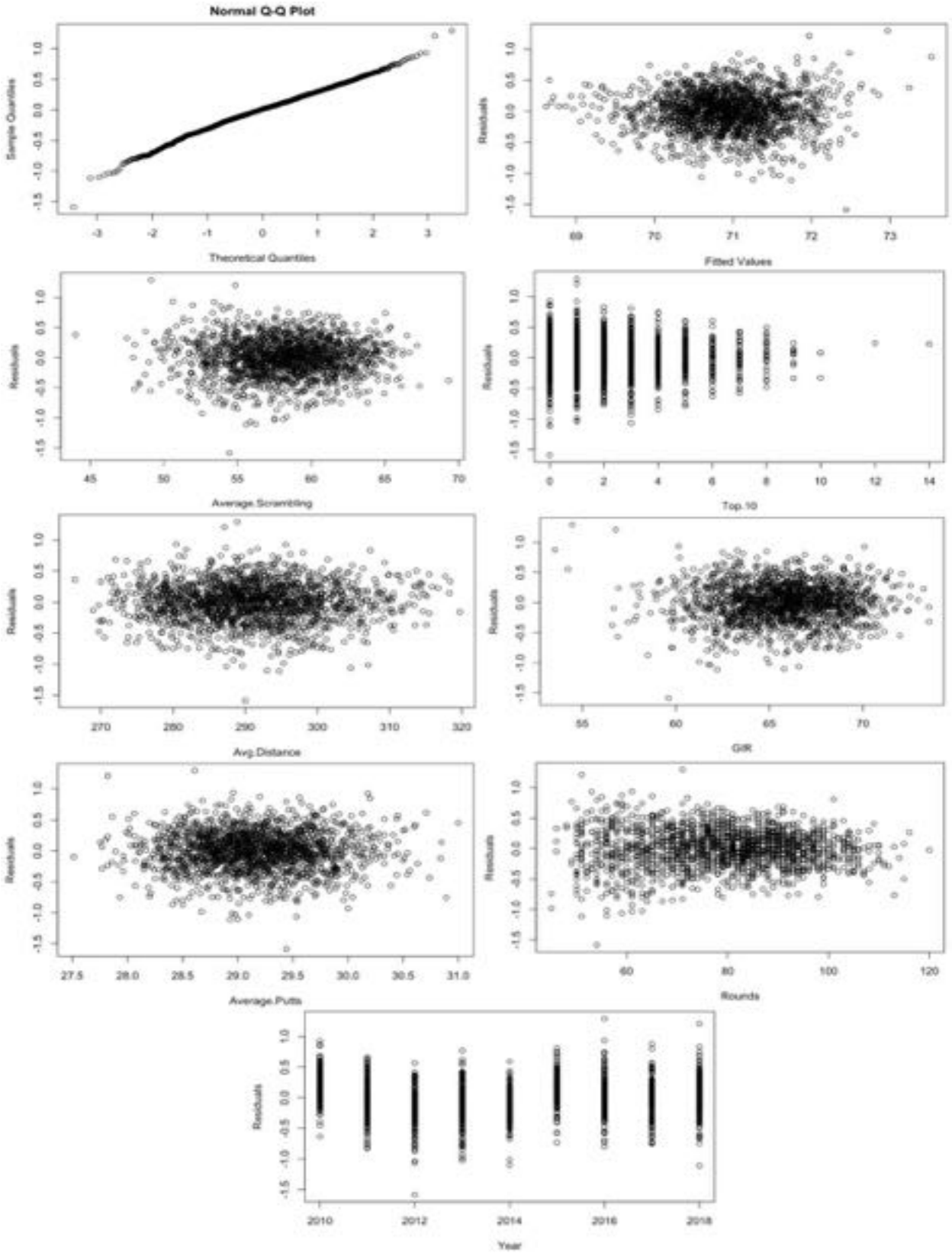
```
Call:
lm(formula = pga$Average.Score ~ pga$Rounds + pga$Year + pga$Average.Putts +
    pga$gir + pga$Average.Scrambling + pga$Top.10 + pga$interaction1 +
    pga$interaction2 + pga$Avg.Distance + pga$Fairway.Percentage)

Residuals:
     Min       1Q   Median       3Q      Max
-1.58680 -0.18330  0.01237  0.20955  1.29237

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               5.015e+00  6.389e+00   0.785  0.43258
pga$Rounds               -3.819e-03  8.542e-04  -4.471 8.33e-06 ***
pga$Year                  3.435e-02  3.150e-03  10.906  < 2e-16 ***
pga$Average.Putts         5.034e-01  2.891e-02  17.415  < 2e-16 ***
pga$gir                  -1.242e-01  5.249e-03 -23.657  < 2e-16 ***
pga$Average.Scrambling   -4.773e-02  4.081e-03 -11.695  < 2e-16 ***
pga$Top.10               -3.400e-01  2.266e-02 -15.008  < 2e-16 ***
pga$interaction1          2.118e-07  4.452e-07   0.476  0.63439
pga$interaction2          1.981e-08  6.830e-09   2.900  0.00378 **
pga$Avg.Distance         -1.867e-02  1.395e-03 -13.388  < 2e-16 ***
pga$Fairway.Percentage   -1.657e-02  2.334e-03  -7.099 1.85e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3168 on 1667 degrees of freedom
Multiple R-squared:  0.7954,    Adjusted R-squared:  0.7942
F-statistic:   648 on 10 and 1667 DF,  p-value: < 2.2e-16
```

Appendix E: Full Model Summary with Avg.Distance included. 'Interaction1' represents the interaction between 'Rounds','Average.Putts','gir', and 'Top.10'. 'Interaction2' represents the interaction between 'Rounds','Average.Putts','gir','Average.Scrambling', and 'Top.10'.




Appendix F: Residual plots for the full model, none of which invoked concern about the model assumptions and composition.