Predicting the Most Impactful Strokes Gained Statistics for Sustained Success on the PGA Tour

Devon Anderson, Danny Blumenstein, and Josh Rochlin

March 2, 2024

Contents

1. Abstract

Professional golfers on the PGA Tour are some of the most highly compensated athletes in the world. Because PGA Tour players earn money depending on where they finish in a tournament, consistent performance is key for success. Perhaps no other statistical category has dominated the golf data vernacular over the past decade more than "Strokes Gained." Created by Columbia University professor Mark Broadie, Strokes Gained analyzes each player's performance in comparison to their fellow competitors in a tournament field (Sens, 2018). The injection of Strokes Gained into professional golf has fundamentally changed the way players practice, diagnose their strengths and weaknesses, and select shots during competition. In this paper, we explore which Strokes Gained statistics have the greatest correlation to money earned on the PGA Tour.

2. Introduction

Golf is an ancient game. From the first-time 15th century Scots knocked pebbles around the sandy dunes of what would become The Old Course at St Andrews, its core traditions have remained. Golf's immunity to the grips of time is strengthened by its ability to transform. Analytics was first introduced to the sport in 2001, when the Professional Golf Association (PGA) created ShotLink, which tracks every shot, by every player, during the PGA Tour season. Strokes Gained statistics are the primary metrics recorded by ShotLink (Cunneff, 2019). Current PGA Tour player may have access to ultramodern equipment, technology-infused golf balls, and robust swing monitoring systems, but true success in professional golf stems from the ability to make strategic decisions under pressure. There is no tried-and-true way to win a golf tournament. A frightening combination of physical, mental, earthly, and heavenly factors can affect what a pro will do when he steps over his ball during competition. However, a pro that has Strokes Gained data in their back pocket can more easily navigate the complex decision-making processes the sport demands.

Strokes Gained is the gold standard for data analysis in golf. Because Strokes Gained accounts for hole distance, shot length, and lie, it provides an all-inclusive perspective of how well a player is performing. These metrics are highly sought-after for discovering what the best aspects of a pro's game are. We are interested in uncovering why pro golfers are embracing Strokes Gained stats to improve performance. For our first bullet, we will use a statistical test to decide if Strokes Gained:Approach is one of the most important metrics in earning money on tour. Our second bullet will figure out if a player should increase their average driving distance from 275 yards to 285 to make more money.

2.1 Motivation

We are awe-inspired by what pro golfers can do. Golf is an agonizing game, mental warfare. The painstaking hours of practice these guys have put in throughout their lives to be the absolute best is admirable. This investigation will help us learn what really makes a pro golfer's game whole. Although pro golfers play the sport an entirely different way from average players like us, evaluating these stats could give us more clarity on the weak parts of our game that have been ignored for far too long.

3. Data Description

The data set ("pgaTourDataNew.csv") provides statistical performance data for professional golfers from 2010-2018. Each row contains stats collected during a particular season for a player. There are 18 variables in the data set:

1. Player.Name (Categorical): Name of the golfer.
2. Rounds (Numerical): The number of rounds played in that year by that player.
3. Fairway.Percentage (Numerical): The percentage of time a player hits a tee shot into the fairway.
4. Year (Numerical): The season the statistics were collected.
5. Avg.Distance (Numerical): The average distance of a tee-shot.
6. gir (Numerical): Green in Regulation - A Green in Regulation is achieved when a player reaches the green in 'par minus 2 strokes.' i.e.: On a Par 3, a player hits the ball on the green during their first stroke.
7. Average.Putts (Numerical): The average number of putts.
8. Average.Scrambling (Numerical): Scrambling occurs when a player misses a green in regulation, but still makes par or better on the hole.
9. Average.Score (Numerical): The average of all scores recorded by a player in that season.
10. Points (Numerical): The number of FedEx Cup points scored in a season. The FedEx Cup is a season long competition on the PGA Tour. Players are awarded points toward the FedEx Cup depending on tournament finish.
11. Wins (Numerical): The number of tournaments a player has won during that season.
12. Top.10 (Numerical): The number of Top 10 finishes in tournaments during that season.
13. Average.SG.Putts (Numerical): How many strokes a player gains or loses on the green.
14. Average.SG.Total (Numerical): All strokes gained stats combined.
15. SG.OTT (Numerical): How many strokes a player gains or loses off the tee on Par 4s and Par 5s.
16. SG.APR (Numerical): How many strokes a player gains or loses approaching the green. An approach shot on a Par 4 and Par 5 are shots that are not off of the tee. On a Par 3, a tee shot is considered an approach shot.
17. SG.ARG (Numerical): How many strokes a player gains or loses around the green. A shot within 30 yards of the edge of the green is considered around the green. Shots taken on the putting green are not included.
18. Money (Numerical): The amount of prize money a player earned in that season.

3.1 Preliminary Analysis

Below is a preliminary analysis of the data set ‘Golf.’ We loaded in the data and took a summary to find any irregularities. There were several missing values in the “Wins” and “Top.10” columns, so they were replaced with zeros. The rest of the NA’s in the statistical categories were omitted. “Wins,” “Top.10,” and “Points” were excluded because winning tournaments and finishing in the top 10, as well as gaining FedEx Cup points, are synonymous with earning money.

3.2 Data Set Summary

```
##  Player.Name            Rounds       Fairway.Percentage      Year
##  Length:2312        Min.   : 45.00   Min.   :43.02      Min.   :2010
##  Class :character   1st Qu.: 69.00   1st Qu.:57.94      1st Qu.:2012
##  Mode  :character   Median : 79.50   Median :61.43      Median :2014
##                     Mean   : 78.71   Mean   :61.44      Mean   :2014
##                     3rd Qu.: 89.00   3rd Qu.:64.91      3rd Qu.:2016
##                     Max.   :120.00   Max.   :76.88      Max.   :2018
##                     NA's   :634      NA's   :634        NA's   :634
##   Avg.Distance        gir        Average.Putts   Average.Scrambling
##  Min.   :266.4   Min.   :53.54   Min.   :27.51   Min.   :44.01
##  1st Qu.:284.9   1st Qu.:63.83   1st Qu.:28.81   1st Qu.:55.90
##  Median :290.6   Median :65.79   Median :29.14   Median :58.27
##  Mean   :290.8   Mean   :65.66   Mean   :29.16   Mean   :58.12
##  3rd Qu.:296.4   3rd Qu.:67.58   3rd Qu.:29.52   3rd Qu.:60.42
##  Max.   :319.7   Max.   :73.52   Max.   :31.00   Max.   :69.33
##  NA's   :634     NA's   :634     NA's   :634     NA's   :634
##  Average.Score       Points            Wins            Top.10
##  Min.   :68.70   Min.   :   3.0   Min.   :0.0000   Min.   : 0.000
##  1st Qu.:70.49   1st Qu.: 322.0   1st Qu.:0.0000   1st Qu.: 1.000
##  Median :70.90   Median : 530.0   Median :0.0000   Median : 2.000
##  Mean   :70.92   Mean   : 631.1   Mean   :0.2062   Mean   : 2.332
##  3rd Qu.:71.34   3rd Qu.: 813.8   3rd Qu.:0.0000   3rd Qu.: 3.000
##  Max.   :74.40   Max.   :4169.0   Max.   :5.0000   Max.   :14.000
##  NA's   :634     NA's   :638      NA's   :634      NA's   :634
##  Average.SG.Putts  Average.SG.Total      SG.OTT            SG.APR
##  Min.   :-1.4750   Min.   :-3.2090   Min.   :-1.7170   Min.   :-1.6800
##  1st Qu.:-0.1870   1st Qu.:-0.2547   1st Qu.:-0.1902   1st Qu.:-0.1808
##  Median : 0.0400   Median : 0.1470   Median : 0.0560   Median : 0.0810
##  Mean   : 0.0256   Mean   : 0.1481   Mean   : 0.0378   Mean   : 0.0650
##  3rd Qu.: 0.2570   3rd Qu.: 0.5685   3rd Qu.: 0.2915   3rd Qu.: 0.3145
##  Max.   : 1.1300   Max.   : 2.4060   Max.   : 1.4850   Max.   : 1.5330
##  NA's   :634       NA's   :634       NA's   :634       NA's   :634
##      SG.ARG           Money
##  Min.   :-0.930   Min.   :   24650
##  1st Qu.:-0.123   1st Qu.:  565641
##  Median : 0.022   Median : 1046144
##  Mean   : 0.020   Mean   : 1488682
##  3rd Qu.: 0.175   3rd Qu.: 1892478
##  Max.   : 0.660   Max.   :12030465
##  NA's   :634      NA's   :638
```

3.3 Cleaned Data Set Summary

```
##  Player.Name           Rounds       Fairway.Percentage      Year
##  Length:1674        Min.   : 45.00   Min.   :43.02      Min.   :2010
##  Class :character   1st Qu.: 69.00   1st Qu.:57.95      1st Qu.:2012
##  Mode  :character   Median : 80.00   Median :61.44      Median :2014
##                     Mean   : 78.77   Mean   :61.45      Mean   :2014
##                     3rd Qu.: 89.00   3rd Qu.:64.91      3rd Qu.:2016
##                     Max.   :120.00   Max.   :76.88      Max.   :2018
##   Avg.Distance        gir        Average.Putts   Average.Scrambling
##  Min.   :266.4   Min.   :53.54   Min.   :27.51   Min.   :44.01
##  1st Qu.:284.9   1st Qu.:63.83   1st Qu.:28.80   1st Qu.:55.90
##  Median :290.5   Median :65.79   Median :29.14   Median :58.29
##  Mean   :290.8   Mean   :65.67   Mean   :29.16   Mean   :58.12
##  3rd Qu.:296.4   3rd Qu.:67.59   3rd Qu.:29.52   3rd Qu.:60.42
##  Max.   :319.7   Max.   :73.52   Max.   :31.00   Max.   :69.33
##  Average.Score   Average.SG.Putts   Average.SG.Total      SG.OTT
##  Min.   :68.70   Min.   :-1.47500   Min.   :-3.2090   Min.   :-1.71700
##  1st Qu.:70.49   1st Qu.:-0.18775   1st Qu.:-0.2602   1st Qu.:-0.19025
##  Median :70.90   Median : 0.04000   Median : 0.1470   Median : 0.05500
##  Mean   :70.92   Mean   : 0.02541   Mean   : 0.1475   Mean   : 0.03702
##  3rd Qu.:71.34   3rd Qu.: 0.25850   3rd Qu.: 0.5685   3rd Qu.: 0.28775
##  Max.   :74.40   Max.   : 1.13000   Max.   : 2.4060   Max.   : 1.48500
##      SG.APR             SG.ARG             Money
##  Min.   :-1.68000   Min.   :-0.93000   Min.   :   24650
##  1st Qu.:-0.18000   1st Qu.:-0.12300   1st Qu.:  565641
##  Median : 0.08100   Median : 0.02250   Median : 1046144
##  Mean   : 0.06519   Mean   : 0.02019   Mean   : 1488682
##  3rd Qu.: 0.31450   3rd Qu.: 0.17575   3rd Qu.: 1892478
##  Max.   : 1.53300   Max.   : 0.66000   Max.   :12030465
```

3.4 Data Set Structure

```
## ’data.frame':    1674 obs. of  15 variables:
##  $ Player.Name       : chr  "Danny Lee" "Blake Adams" "Kyle Reifers" "Brendon de Jonge" ...
##  $ Rounds            : int  120 116 115 115 114 113 113 112 111 111 ...
##  $ Fairway.Percentage: num  63.9 65.8 65.3 63.4 65.5 ...
##  $ Year              : int  2015 2011 2016 2012 2010 2016 2011 2017 2014 2014 ...
##  $ Avg.Distance      : num  283 292 287 289 287 ...
##  $ gir               : num  65.7 65.5 67.6 69.2 70.1 ...
##  $ Average.Putts     : num  28.3 28.7 29.3 29.4 29.5 ...
##  $ Average.Scrambling: num  60.9 60.6 56 58.6 55.5 ...
##  $ Average.Score     : num  70.4 70.7 71 70.1 70.3 ...
##  $ Average.SG.Putts  : num  0.374 0.475 0.102 0.144 0 -0.109 0.04 -0.012 0.107 0.113 ...
##  $ Average.SG.Total  : num  0.729 0.411 0.148 1.083 0.762 ...
##  $ SG.OTT            : num  -0.301 0.241 0.112 0.185 0.16 0.336 0.097 0.586 0.008 -0.014 ...
##  $ SG.APR            : num  0.718 -0.453 0.061 0.754 0.721 -0.131 0.508 0.349 0.322 0.296 ...
##  $ SG.ARG            : num  -0.062 0.148 -0.127 0 -0.125 0.403 0.202 -0.031 0.23 0.121 ...
##  $ Money             : int  3965933 1100558 1539578 2015252 2167978 3086369 2320038 4161008 2169723 2
##  - attr(*, "na.action")= 'omit' Named int [1:638] 1532 1564 1604 1632 1679 1680 1681 1682 1683 1684
##   ..- attr(*, "names")= chr [1:638] "1532" "1564" "1604" "1632" ...
```

From the `str` function, we see that the variables are correctly categorized.

3.5 Raw Model

We provide a summary of the raw model. “Player.Name” and “Year” are not used in the raw model because they are not statistical measurements.

```
##
## Call:
## lm(formula = Money ~ . - Player.Name - Year, data = Golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1962732  -570215  -162150   369454  7258232
##
## Coefficients:
##                     Estimate Std. Error t value          Pr(>|t|)
## (Intercept)         64405650    8827794   7.296 0.000000000000458 ***
## Rounds                  4922       1664   2.958          0.003145 **
## Fairway.Percentage      8218       8381   0.981          0.326934
## Avg.Distance           21524       5649   3.810          0.000144 ***
## gir                   -15778      17737  -0.890          0.373807
## Average.Putts        -525830     100738  -5.220 0.000000201690471 ***
## Average.Scrambling    -57207      10562  -5.416 0.000000069763819 ***
## Average.Score        -711906     117100  -6.079 0.000000001492868 ***
## Average.SG.Putts     2775066    1028284   2.699          0.007031 **
## Average.SG.Total    -2289496    1023213  -2.238          0.025382 *
## SG.OTT               3064425    1031608   2.971          0.003016 **
## SG.APR               3199953    1031913   3.101          0.001961 **
## SG.ARG               3128661    1027777   3.044          0.002370 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 885300 on 1661 degrees of freedom
## Multiple R-squared:  0.6088, Adjusted R-squared:  0.6059
## F-statistic: 215.4 on 12 and 1661 DF,  p-value: < 0.00000000000000022
```

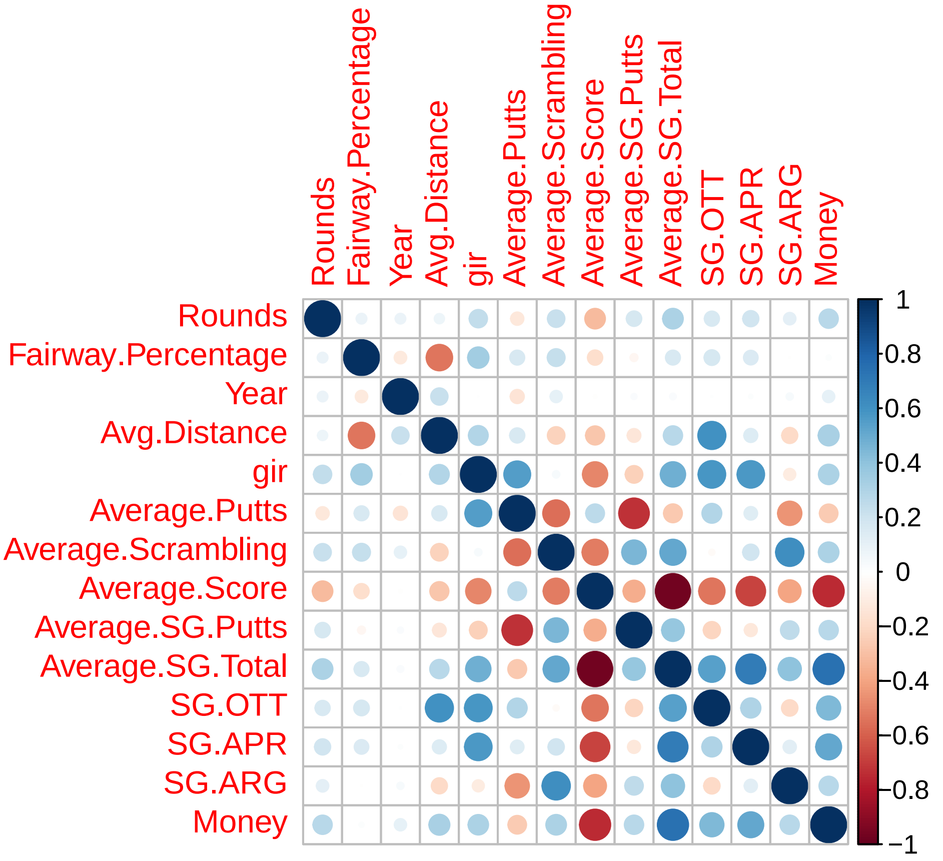
We are satisfied with the performance of our raw model, but we can do much better. The adjusted R2 is 60.59%. From the model, we see that all of the Strokes Gained metrics are statistically significant in earning money. “gir” and “Fairway.Percentage” are the only variables that are not statistically significant.

4. Descriptive Analysis

4.1 Corrpolot

We visualize the relationships between the numerical variables using the `corrplot` function.

```
## corrplot 0.92 loaded
```



From the corrplot, we can make several interpretations about the relationships between the numerical variables. Let's dive into some noticeable findings:

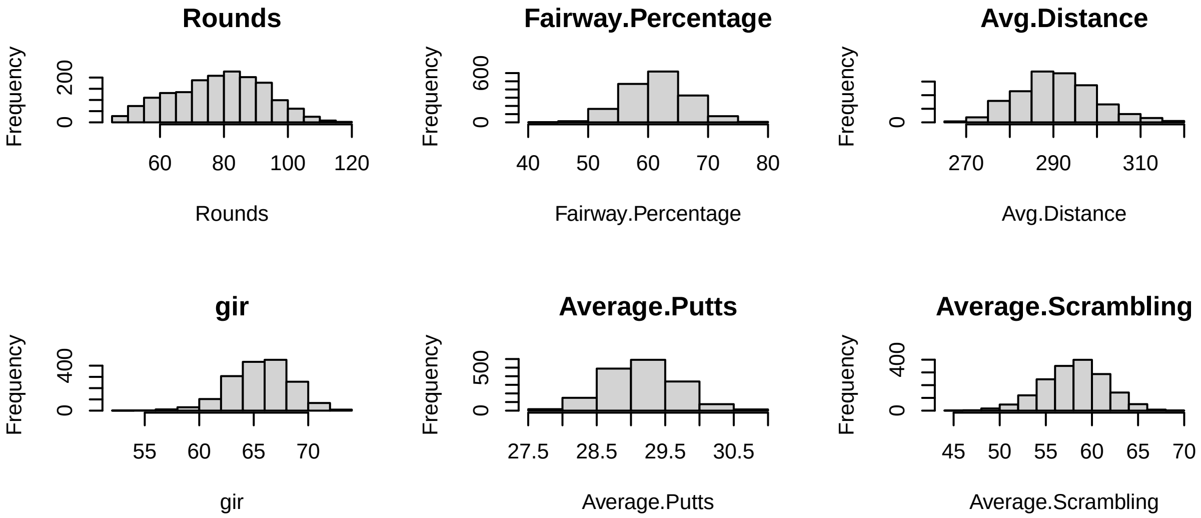
Average.Score and Money: There seems to be a strong, negative correlation between "Average.Score" and "Money." As scoring average increases, money earned tends to decrease. This finding makes sense because lowering score is the ultimate goal in golf. But, as we all know, that does not come easy.

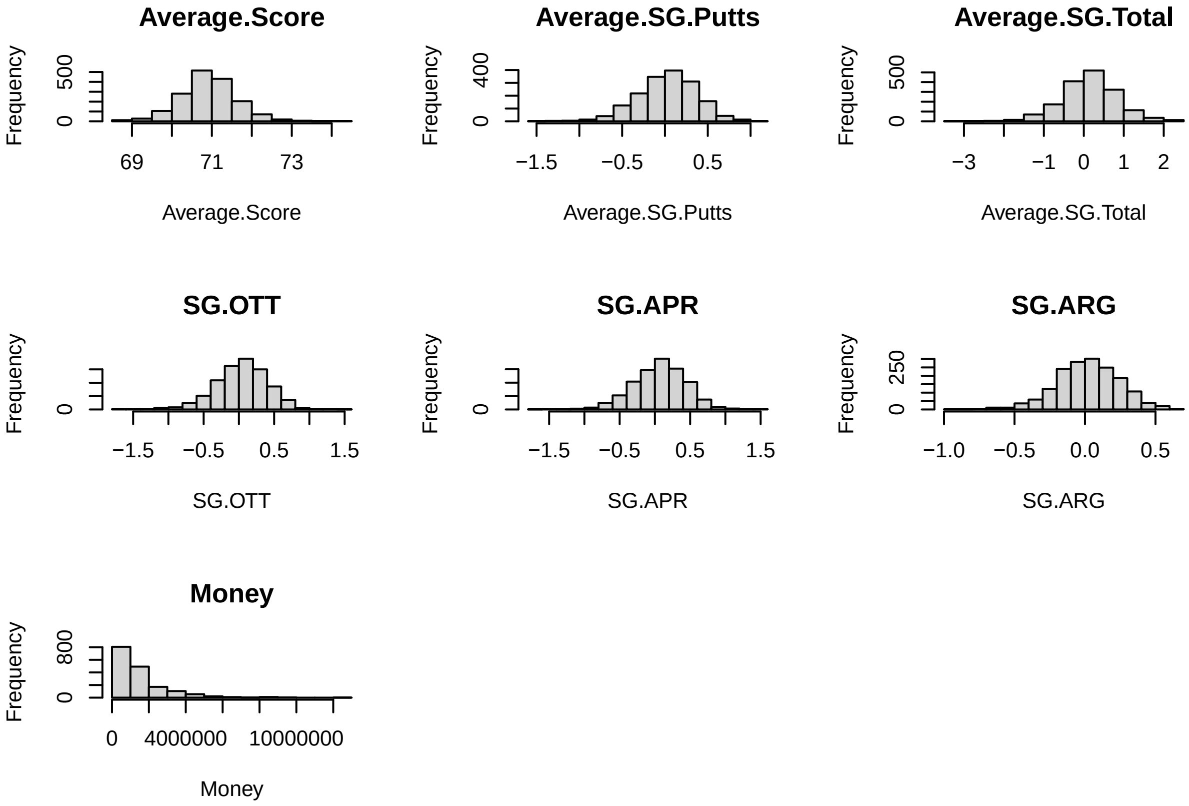
Avg.Distance and gir: There is a strong, positive correlation between average driving distance and greens-in-regulation. This finding indicates that longer drives increase the chance of hitting the green in par minus two strokes. Being farther away from the green diminishes accuracy. Golf courses are adjusting to rapidly increasing distance across the best players in the world. According to a study by the United States Golf Association (USGA), the median course length in the United States increased almost 20 percent from the 1910s through 2010 (Klein, 2021). But interestingly, driving distance on the PGA Tour did not increase dramatically throughout the early 21st century. A study by The R&A found that driving distance on the PGA Tour and the European Tour increased by about .7 percent between 2003 and 2015 (Taylor, 2016).

Average.Putts and Average.Score: There is a weak, positive correlation between "Average.Putts" and "Average.Score." Although putting well is not significant in lowering score for the pros, its importance should not be diminished for amateurs. Putting is a delicate part of the game, and if focused on through dedicated practice, it can provide an amateur with a more well-rounded skill set. A pros time for practice is not limited. For the average golfer, the only thing limited is time.

Average.SG.Total and Average.Score: There is an extremely strong, negative correlation between Average.SG.Total and Average.Score. This relationship highlights the power of Strokes Gained. Players with lowering scoring averages are likely to have higher Strokes Gained Total numbers.

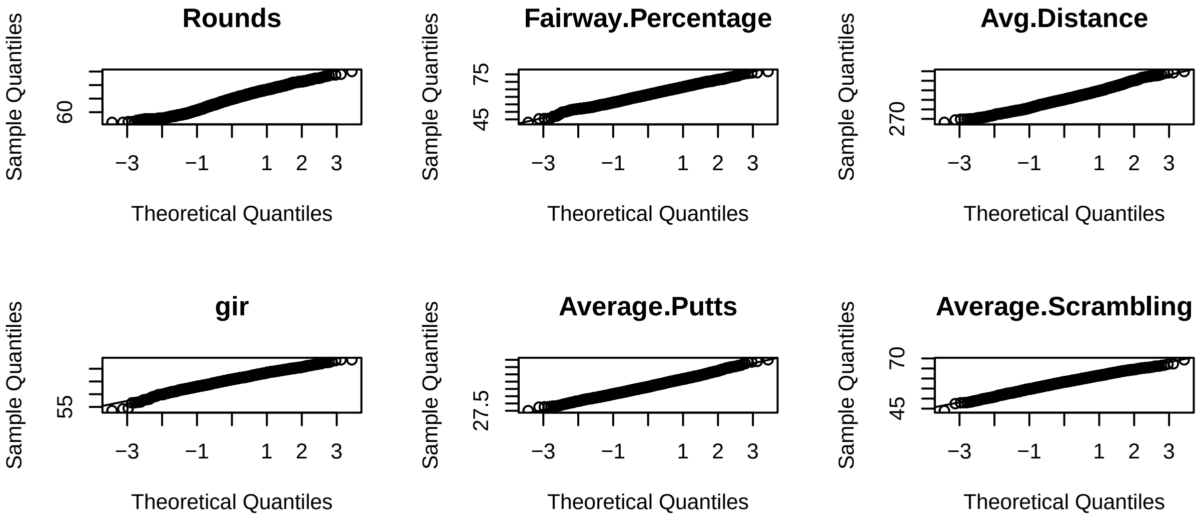
4.2 Histograms

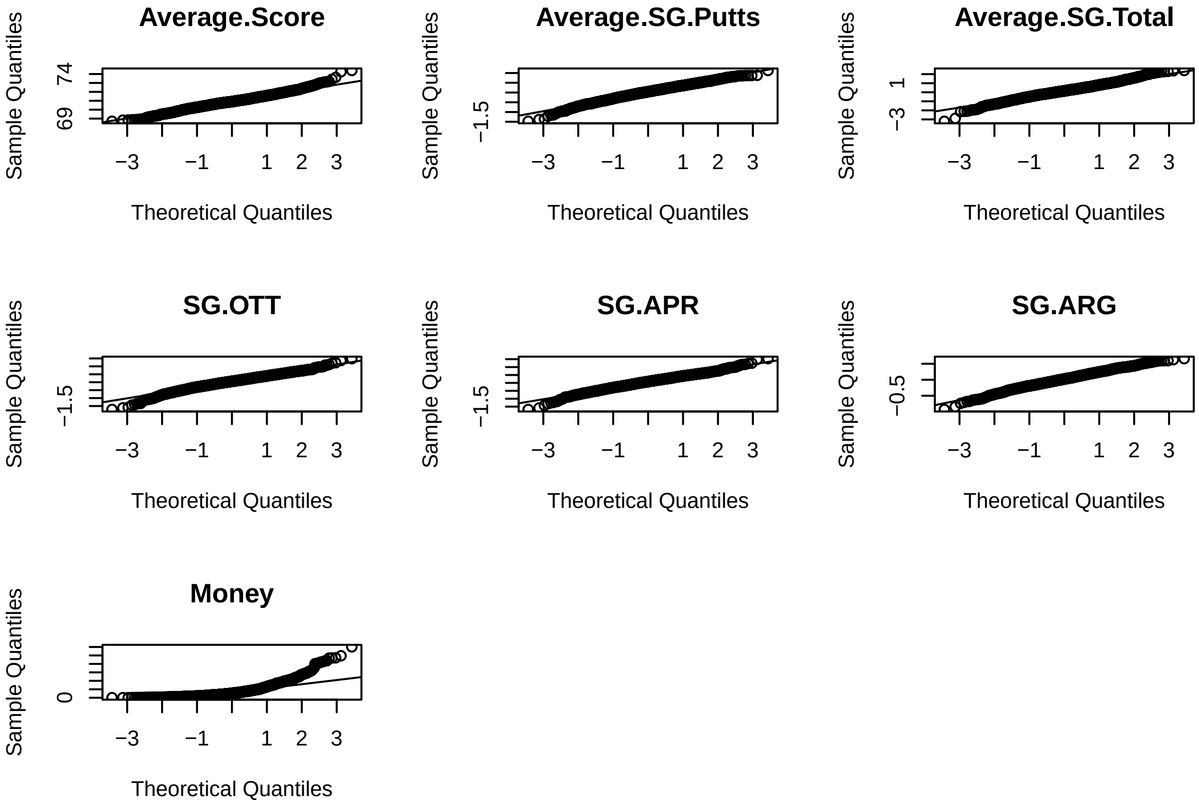




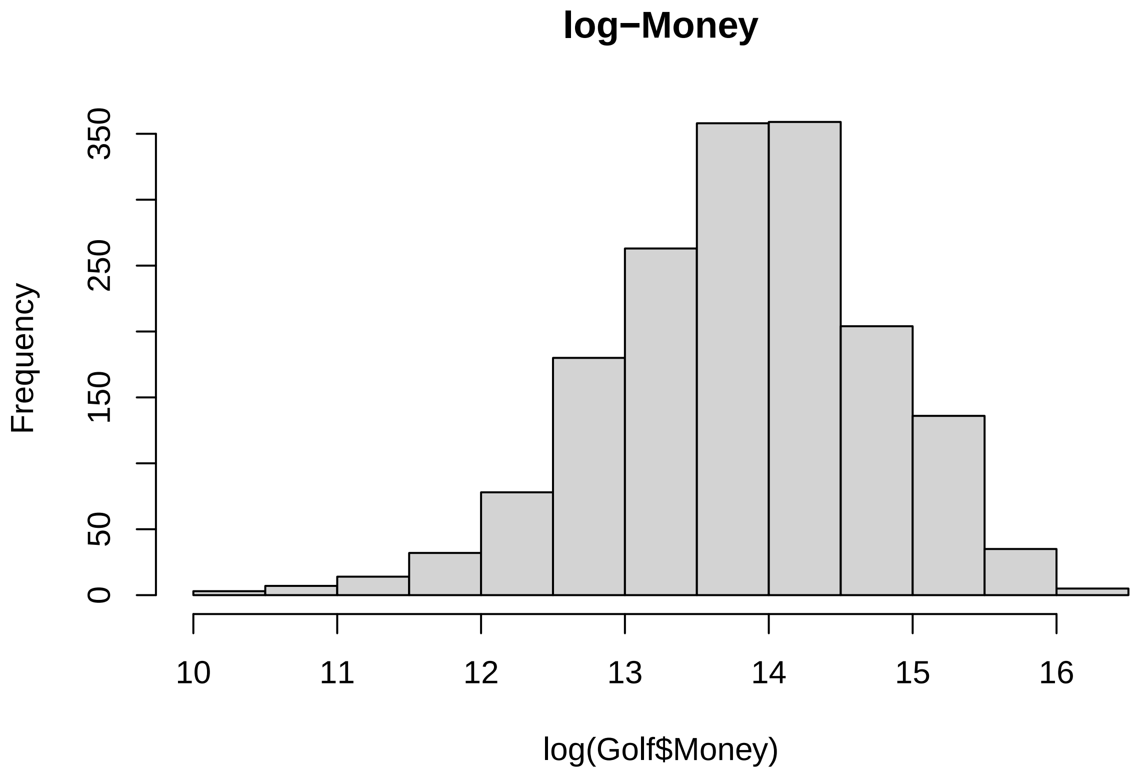
From the histograms, we see that all of the Strokes Gained stats are normally distributed. “gir” is right-skewed. “Money” is heavily right-skewed.

4.3 QQ-Plots


Rounds
Sample Quantiles
60
−3
−1
1
2
3
Theoretical Quantiles
Fairway.Percentage
Sample Quantiles
45
75
−3
−1
1
2
3
Theoretical Quantiles
Avg.Distance
Sample Quantiles
270
−3
−1
1
2
3
Theoretical Quantiles
gir
Sample Quantiles
55
−3
−1
1
2
3
Theoretical Quantiles
Average.Putts
Sample Quantiles
27.5
−3
−1
1
2
3
Theoretical Quantiles
Average.Scrambling
Sample Quantiles
45
70
−3
−1
1
2
3
Theoretical Quantiles

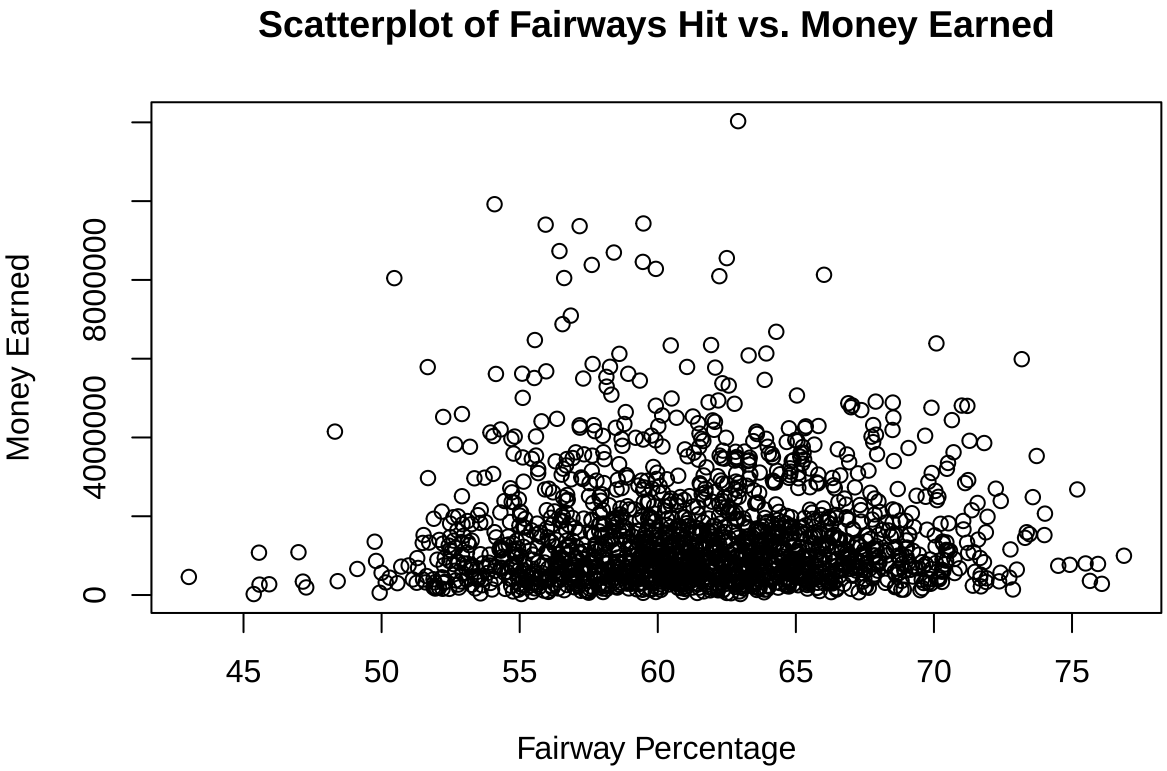


According to the QQ-Plots, all of the variables except “Money” are normally distributed. We can take a log transformation of “Money to make it more normal.

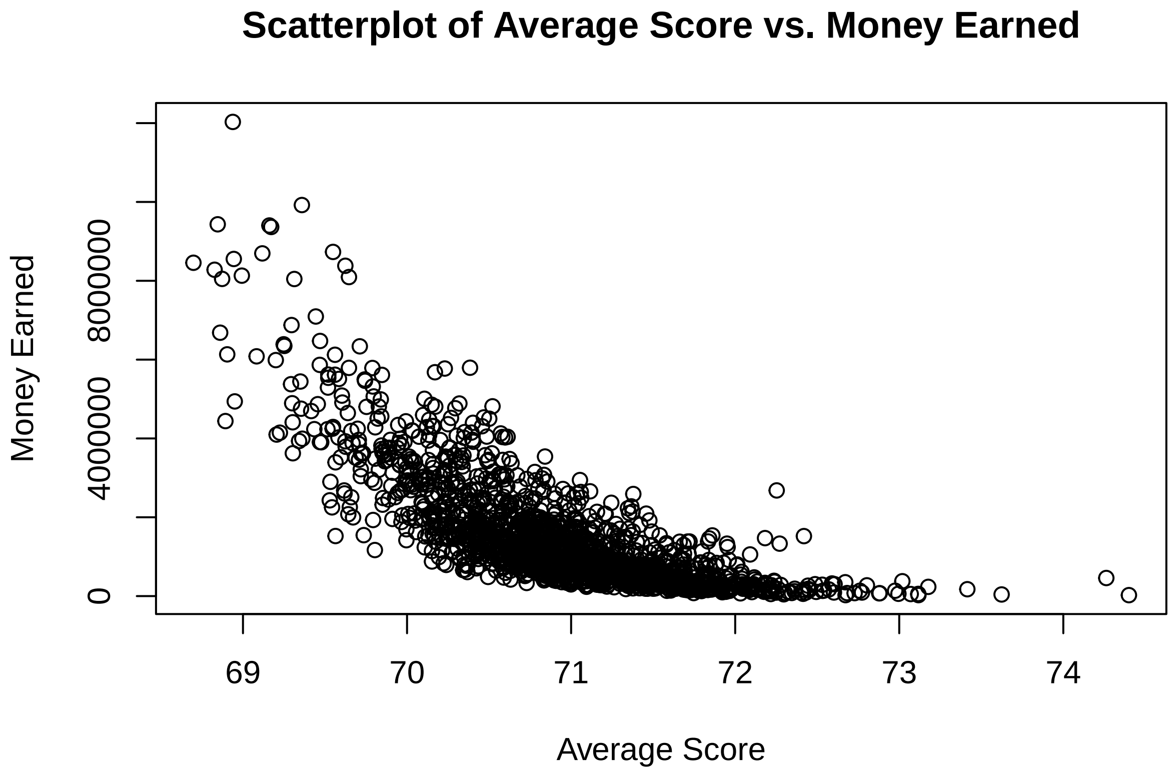

log−Money
Frequency
0
50
150
250
350
10
11
12
13
14
15
16
log(Golf$Money)

4.4 Scatterplots

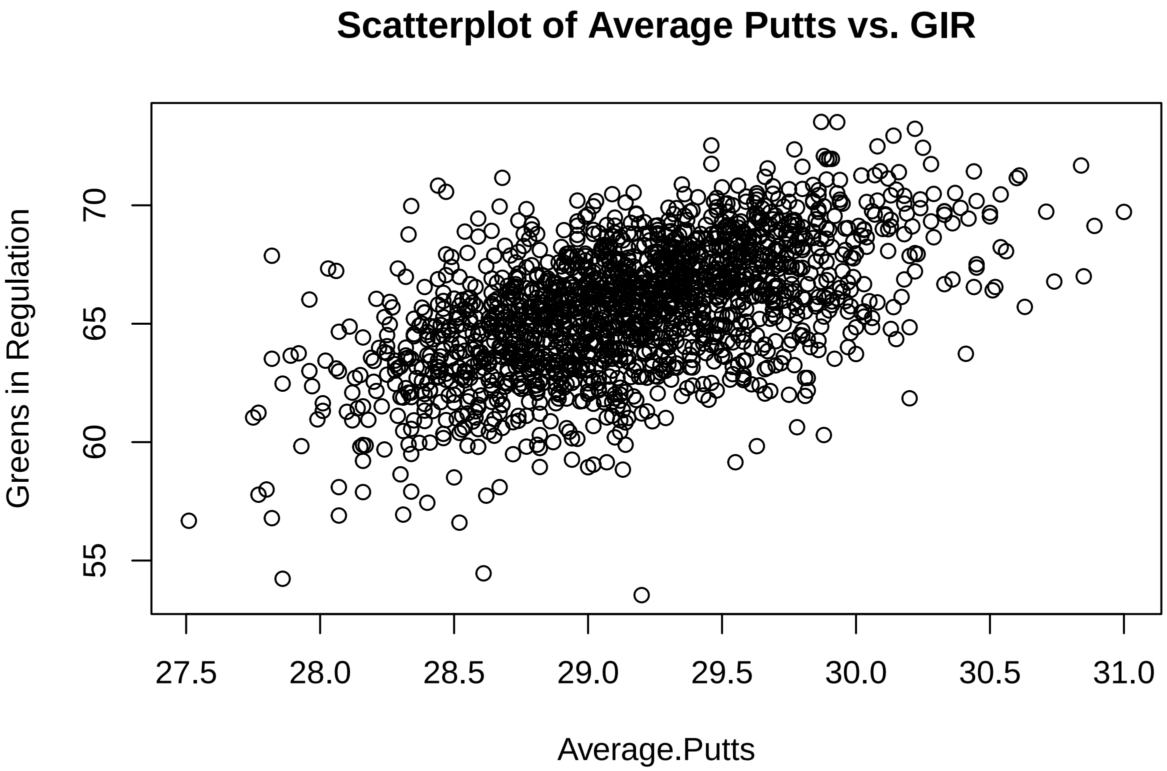
We also investigate the relationships between pertinent metrics through scatterplots.



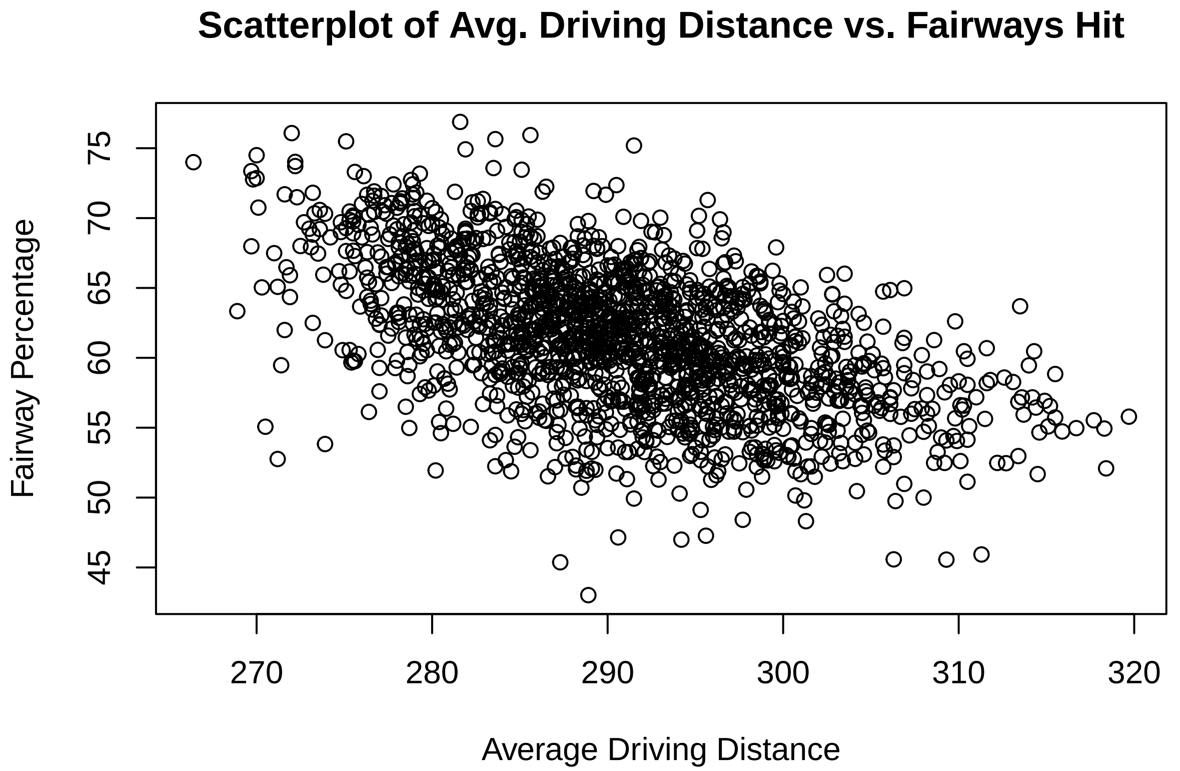
From the scatterplot, it is interesting to see how hitting a high amount of fairways off the tee does not necessarily equate to success. The big earners on tour are bombers of the golf ball and have an average to below-average fairway hit rate. To pros, maximizing distance and having a close proximity to the green are key, no matter the lie.



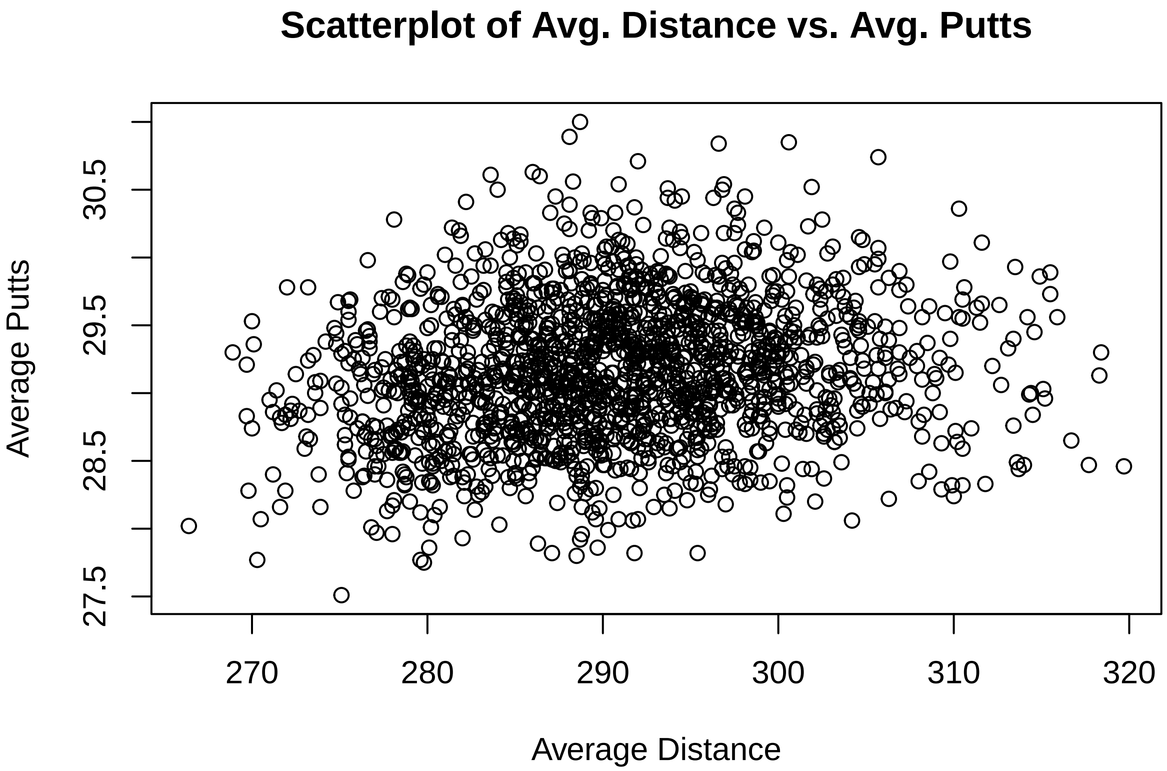
This scatterplot makes a lot of sense. Lower scores contribute to more success. But with all the technological improvements in equipment, has scoring on the PGA Tour changed all that much? Surprisingly, average golfers are becoming better players faster than the pros. According to the USGA, between 1991 and 2016, the average USGA handicap for an average male player dropped from 16.3 to 14.4, almost two full strokes (Stachura, 2017). Two strokes is a critical difference. For tour players, the average score in 1991 was 71.50. In 2016, it only increased .5% to 71.12 (Stachura, 2017). With all of the obstacles average golfers have in the way of actually practicing the sport, we are doing pretty well.



There is a strong, positive correlation between the average number of putts per round and greens in regulation. This relationship suggests that pros need to be accurate iron players and good putters in order to sustain success.



Hitting the ball 310+ yards off the tee may provide a pro with a more forgiving club to hit into the green, but that comes at the expense of accuracy. Thick rough will make the average golfer shiver, but the pros will welcome any lie, as long as the next shot to the green is unobstructed. The pros know the right places to miss a tee shot and minimize danger. The average player's hole is often ruined if they hit their drive into risky areas.



There is a moderately-positive correlation between average driving distance and the average number of putts per round. Although driving the ball far provides pros with closer proximity to the green, the hole is not finished until the ball is in the bottom of the cup. In fact, the average tour player lands the ball 20 feet from the pin from 100-125 yards in the fairway (Sherman, 2023). That is not very close. And to give the average golfer a sigh of relief, pros only make 50% of their putts from inside of eight feet (Sherman, 2023). We average golfers should be easier on ourselves, but we think we can be perfect.

5. Bullet 1

From 2004 to 2021, nearly 40% of winners on tour were ranked in the top 5 in SG:APR (Lack, 2021). According to datagolf.com, about 35% of the scoring dispersion on the average PGA Tour course can be explained by SG:APR (Powers, 2020). Can we refute the claim that SG:APR is one of the most important metrics in earning money on tour?

```
##
## Call:
## lm(formula = Money ~ SG.APR, data = Golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2329347  -740144  -213299   435198  9496366
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  1365397      30075   45.40 <0.0000000000000002 ***
## SG.APR       1891104      77851   24.29 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1213000 on 1672 degrees of freedom
## Multiple R-squared:  0.2609, Adjusted R-squared:  0.2604
## F-statistic: 590.1 on 1 and 1672 DF,  p-value: < 0.00000000000000022
```

From the linear model predicting money won by strokes gained through approach shots, we can see that gaining strokes is a significant predictor.

We can establish a t-test confidence interval of 99% using an alpha of 0.01, which results in t-value bounds of ±3.169. We use the following hypotheses:

Ho: Beta1 equals 0 Ha: Beta1 does not equal 0

The regression model summary shows that the t-value returned for SG.APR is far more extreme than ±3.169. Therefore, we reject the null hypothesis that there is no statistical relationship between Money and SG.APR. As there is a correlation between wins and money earned we can see that strokes gained from approach shots has a significant impact on winning tournaments on the PGA Tour.

Using the regression model above, we get the following equation that we can use to predict winnings from Strokes Gained:APR:

$$\hat{y} = 1365397 + 1891104x$$

For every one unit increase in Strokes Gained:APR, earnings are expected to increase by $1,891,104.

Because Tour pros can incur costs for travel, lodging, food, caddying, and coaching of $150,000 per year, we can derive our break-even point in terms of shots gained from approach shots:

Breakeven =

$$(-1365397 + 150000)/1891104 = -0.6427$$

From this equation, we observe that tour pros who gain fewer than -0.6427 strokes with their approach shots are likely making no money or even losing money after considering the costs of being a pro golfer.

6. Bullet 2

A good drive in golf is far and accurate. In recent years, players have been developing more athletic swings to create more force on the golf ball. The most reliable swings in golf are rooted deep in the core. Utilizing core strength to make a full, unrestricted hip turn is a formula for long and powerful drives. Every pro wants to maximize distance, but does hitting the ball just a little farther off the tee make a substantial difference in money earned?

```
##
## Call:
## lm(formula = Money ~ Avg.Distance, data = Golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2285051  -858450  -339558   495985 10488925
##
## Coefficients:
##                Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  -13670696    1063551  -12.85 <0.0000000000000002 ***
## Avg.Distance     52132       3656   14.26 <0.0000000000000002 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1332000 on 1672 degrees of freedom
## Multiple R-squared:  0.1084, Adjusted R-squared:  0.1079
## F-statistic: 203.4 on 1 and 1672 DF,  p-value: < 0.00000000000000022
```

From the linear model predicting money won by average drive distance, we can quickly see that there is a premium on extra driver yardage off the tee.

We can establish a t-test confidence interval of 99% using an alpha of 0.01, which results in t-value bounds of ±3.169.

We use the following hypotheses:

Ho: Beta1 equals 0

Ha: Beta1 does not equal 0

As shown in the summary of the regression model created above, we observe that t-value returned for Avg.Distance is far more extreme than ±3.169. Therefore, we reject the null hypothesis that there is no statistical relationship between Money and Avg.Distance.

Below is the equation we can use to predict winnings from average drive distance:

$$\hat{y} = -13670696 + 52132x$$

We can interpret the coefficient very simply - each additional yard off the tee when hitting driver is expected to increase earnings that year by $52,132.

Considering tour pros can incur travel, lodging, food, caddying, and coaching costs of $150,000 per year, we can derive our break-even point in terms of yards:

Breakeven =

$$(13679696 + 150000)/52132 = 265.11$$

Therefore, we observe that tour pros hitting less than 266 yards with their driver are likely making no money after considering expenses, and in many cases, losing money.

Next, we examine the expected earnings of a player hitting their driver 275 yards on average:

Earnings=

$$-13670696 + (275 * 52132) = 665604$$

Let's compare this finding to the player that hits their driver 285 yards on average:

Earnings=

$$-13670696 + (285 * 52132) = 1186924$$

As we can see, a player increasing their average driving distance from 275 to 285 can expect to nearly double their earnings from 665,604 to 1,186,924. This jump in earnings is quite remarkable, and demonstrates that a player would be wise to invest in additional coaching and physical training to achieve further driving distance off the tee.

7. Parsimonious Model Building

7.1 Backward Elimination Model

We build a model using backward elimination. We do not use “Player.Name” in our model because it is a categorical variable. “Year” is not used because it is not a statistical metric.

```
## 
## Call:
## lm(formula = Money ~ Rounds + Avg.Distance + Average.Putts + 
##     Average.Scrambling + Average.Score + Average.SG.Putts + Average.SG.Total + 
##     SG.OTT + SG.APR + SG.ARG, data = Golf)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1946247  -571711  -163411   374522  7212113 
## 
## Coefficients:
##                    Estimate Std. Error t value            Pr(>|t|)    
## (Intercept)        66354324    8621503   7.696 0.000000000000023844 ***
## Rounds                 4597       1614   2.848             0.00445 ** 
## Avg.Distance          16901       3251   5.199 0.000000225365237931 ***
## Average.Putts       -573293      69403  -8.260 0.000000000000000293 ***
## Average.Scrambling   -58204      10113  -5.755 0.000000010277830484 ***
## Average.Score       -707203     115928  -6.100 0.000000001313482431 ***
## Average.SG.Putts    2758967    1027868   2.684             0.00734 ** 
## Average.SG.Total   -2297452    1022242  -2.247             0.02474 *  
## SG.OTT              3119295    1030286   3.028             0.00250 ** 
## SG.APR              3177318    1031272   3.081             0.00210 ** 
## SG.ARG              3121295    1026194   3.042             0.00239 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 885200 on 1663 degrees of freedom
## Multiple R-squared:  0.6084, Adjusted R-squared:  0.6061 
## F-statistic: 258.4 on 10 and 1663 DF,  p-value: < 0.00000000000000022
```

The backward elimination model has an adjusted R2 of 60.61% and an AIC of 45856.9. Each statistical variable in the model is significant.

7.2 Foward Selection Model

We build a model using forward selection and produce a summary.

```
##
## Call:
## lm(formula = Money ~ Average.SG.Total + Avg.Distance + Average.Score +
##     Average.Putts + Average.Scrambling + Average.SG.Putts + Rounds,
##     data = Golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1925867  -574375  -164415   368607  7201800
##
## Coefficients:
##                     Estimate Std. Error t value            Pr(>|t|)
## (Intercept)         69052884    8586856   8.042  0.00000000000000166 ***
## Average.SG.Total      817078     117950   6.927  0.00000000000610379 ***
## Avg.Distance           16455       2890   5.693  0.00000001475458604 ***
## Average.Score        -742713     115238  -6.445  0.00000000015093374 ***
## Average.Putts        -572697      66067  -8.668 < 0.0000000000000002 ***
## Average.Scrambling    -59300       9482  -6.254  0.00000000050699860 ***
## Average.SG.Putts     -378566      96808  -3.910  0.00009580352998756 ***
## Rounds                  4618       1614   2.861             0.00427 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 886900 on 1666 degrees of freedom
## Multiple R-squared:  0.6062, Adjusted R-squared:  0.6045
## F-statistic: 366.3 on 7 and 1666 DF,  p-value: < 0.00000000000000022
```

The forward elimination model has an adjusted R2 of 60.45% and an AIC of 45860.55. The backward elimination model has a lower AIC, so it is the leader in the clubhouse.

7.3 Stepwise Selection Model

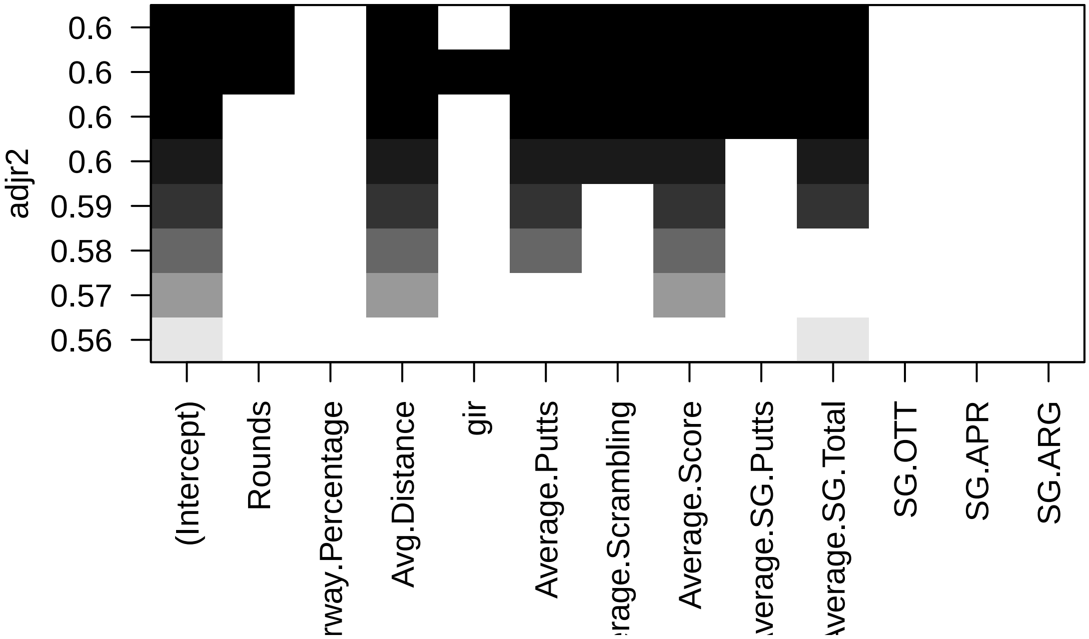
We build a model using stepwise selection and produce a summary.

```
##
## Call:
## lm(formula = Money ~ Average.SG.Total + Avg.Distance + Average.Score +
##     Average.Putts + Average.Scrambling + Average.SG.Putts + Rounds,
##     data = Golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1925867  -574375  -164415   368607  7201800
##
## Coefficients:
##                    Estimate Std. Error t value            Pr(>|t|)
## (Intercept)        69052884    8586856   8.042  0.00000000000000166 ***
## Average.SG.Total     817078     117950   6.927  0.00000000000610379 ***
## Avg.Distance          16455       2890   5.693  0.00000001475458604 ***
## Average.Score       -742713     115238  -6.445  0.00000000015093374 ***
## Average.Putts       -572697      66067  -8.668 < 0.0000000000000002 ***
## Average.Scrambling   -59300       9482  -6.254  0.00000000050699860 ***
## Average.SG.Putts    -378566      96808  -3.910  0.00009580352998756 ***
## Rounds                 4618       1614   2.861              0.00427 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 886900 on 1666 degrees of freedom
## Multiple R-squared:  0.6062, Adjusted R-squared:  0.6045
## F-statistic: 366.3 on 7 and 1666 DF,  p-value: < 0.00000000000000022
```

The stepwise model has the same adjusted R2 as the backward elimination and forward selection models. It has the same AIC as the forward selection model.

7.4 Best Subsets Model

We build a model using best subsets to predict money earned from all other predictor variables.



From the best subsets model, we see that "Avg.Distance," "Average.Putts," "Average.Scrambling," "Average.Score," "Average.SG.Putts," and "Average.SG.Total" should be included in the parsimonious model.

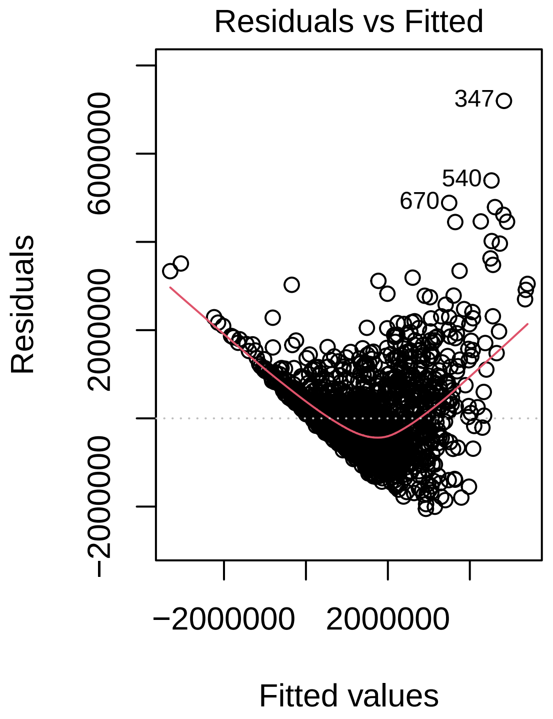
8. Original Parsimonious Model

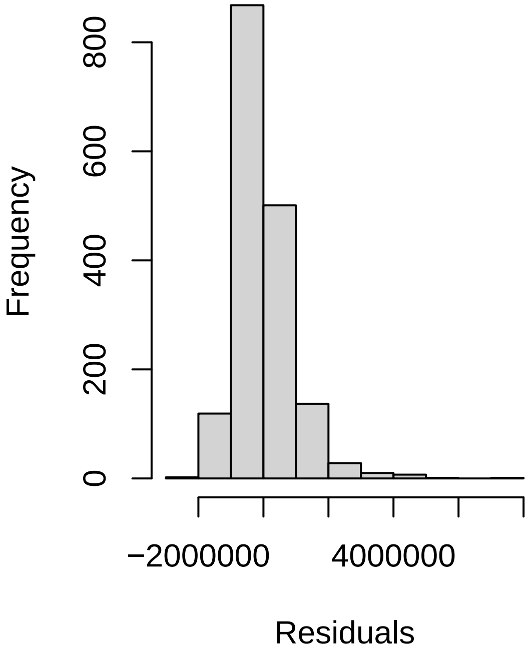
```
##
## Call:
## lm(formula = Money ~ Avg.Distance + Average.Putts + Average.Scrambling +
##     Average.Score + Average.SG.Putts + Average.SG.Total, data = Golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2050409  -572394  -150462   379804  7196889
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        69874246    8600535   8.124 0.000000000000000866 ***
## Avg.Distance          16662       2896   5.754 0.000000010367564645 ***
## Average.Putts       -568917      66196  -8.594 < 0.0000000000000002 ***
## Average.Scrambling   -57584       9483  -6.072 0.000000001558788218 ***
## Average.Score       -753007     115430  -6.523 0.000000000090838240 ***
## Average.SG.Putts    -366337      96921  -3.780             0.000163 ***
## Average.SG.Total     830685     118107   7.033 0.000000000002932891 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 888800 on 1667 degrees of freedom
## Multiple R-squared:  0.6042, Adjusted R-squared:  0.6028
## F-statistic: 424.2 on 6 and 1667 DF,  p-value: < 0.00000000000000022
```

We are satisfied with the performance of our parsimonious model, but we can do much better. We have six predictors in our model, but our adjusted R2 is 60.26, indicating needed improvement. Each variable in the model is statistically significant.

9. Residual Diagnostics for Original Parsimonious Model

We construct a residual plot and a histogram of the residuals of the parsimonious model and discuss if the linear regression model assumptions are met.

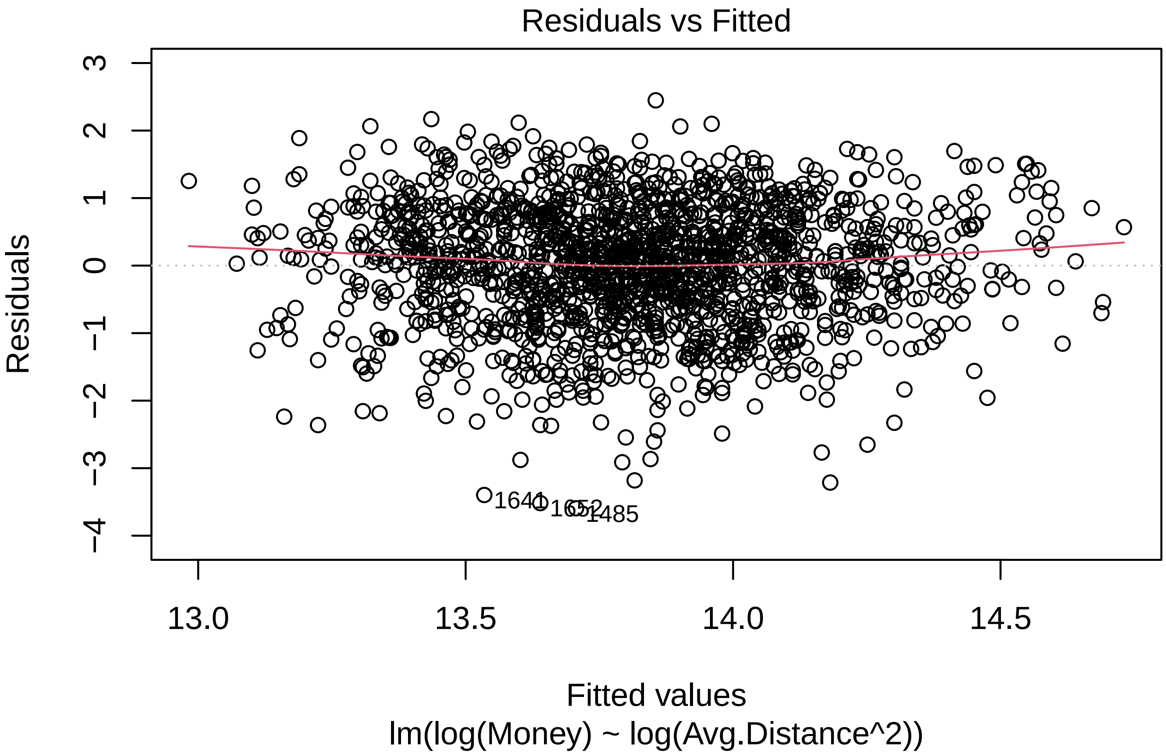




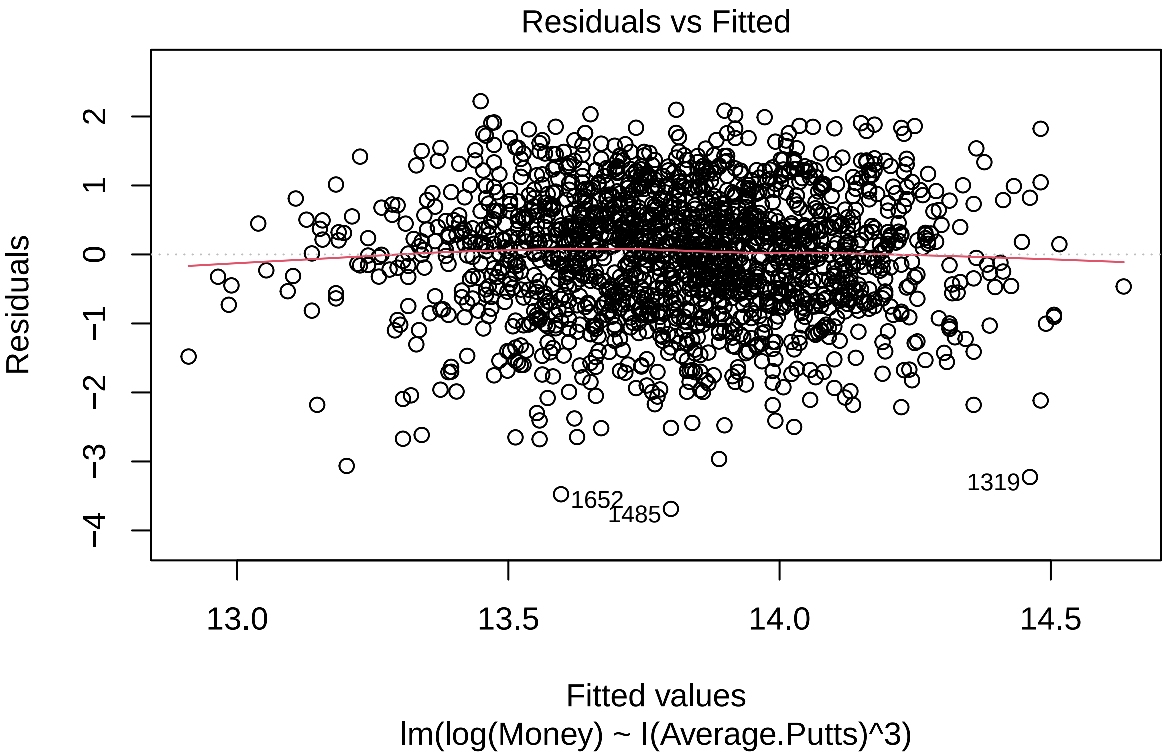
As our parsimonious model stands, we are not meeting the LINE assumptions for a linear regression model. The line is curved, the residuals show a distribution that is right-skewed, and there does not appear to be constant variance. We will adjust the variables in the parsimonious model in order to achieve the assumptions.

9.1 Adjusting Parsimonious Variables

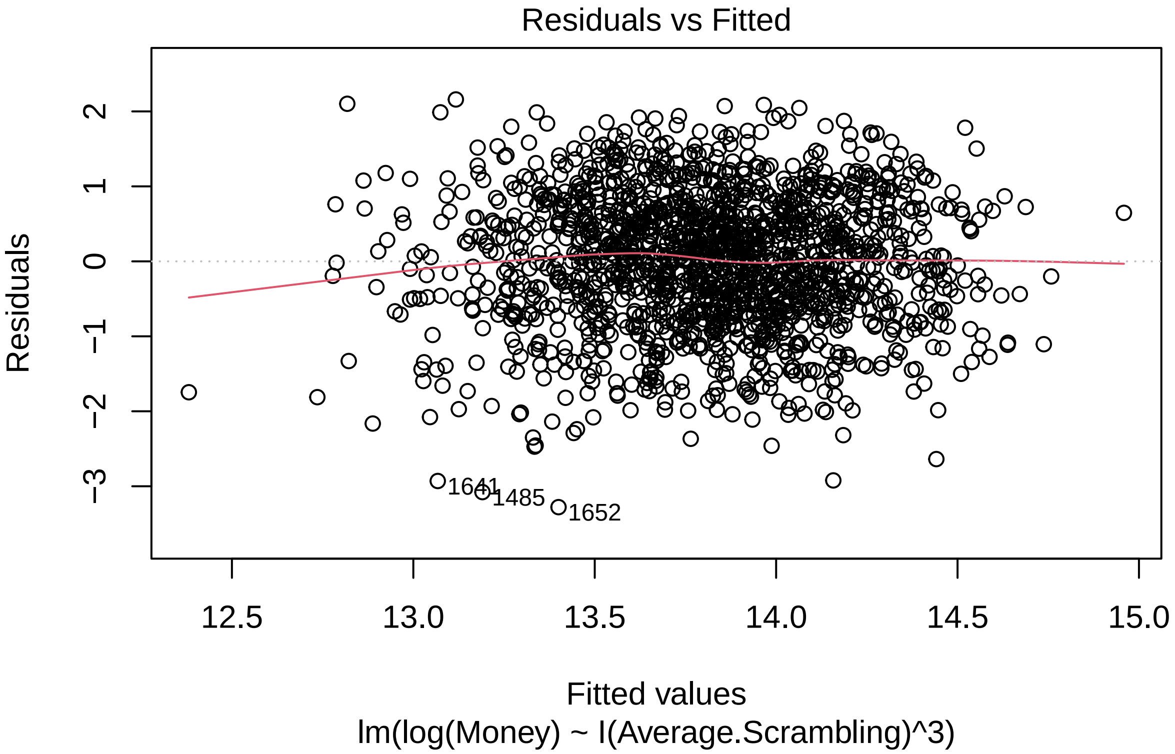
We adjust the parsimonious variables in order to increase their linearity with the response variable.



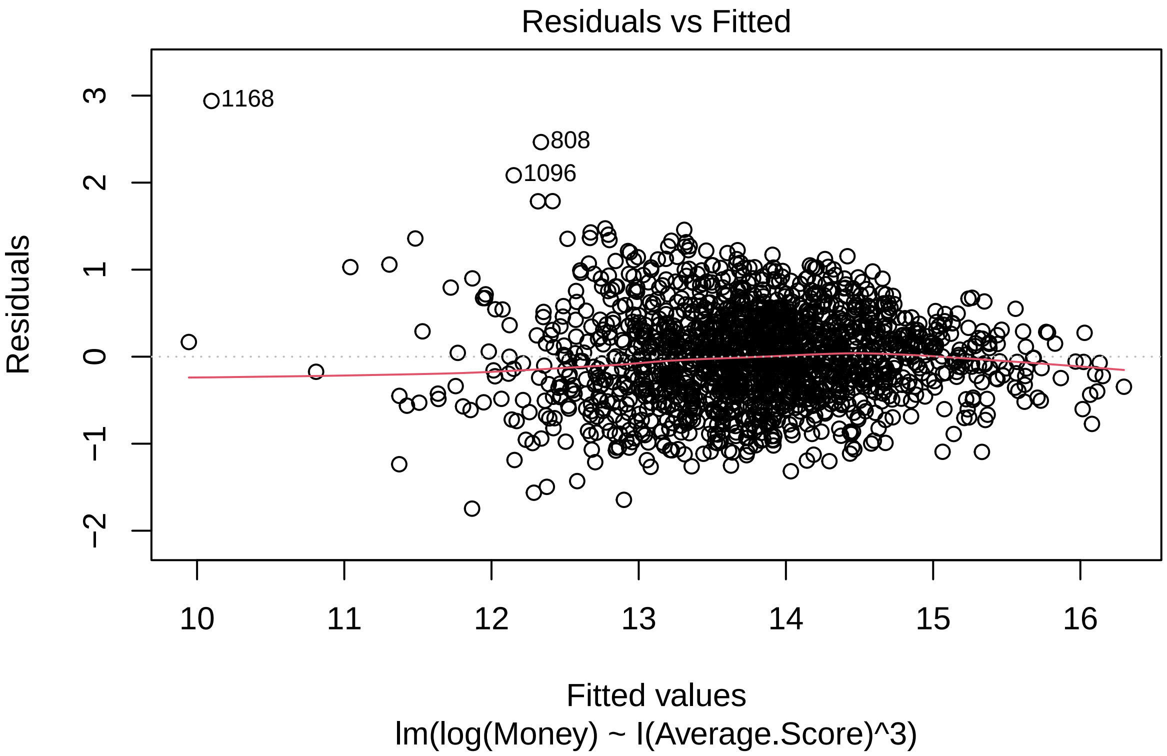
To make the relationship between "Avg.Distance" and "Money" more linear, we take a log transformation of "Avg.Distance" and apply a square to it. Although not perfect, the line is almost centered around zero.



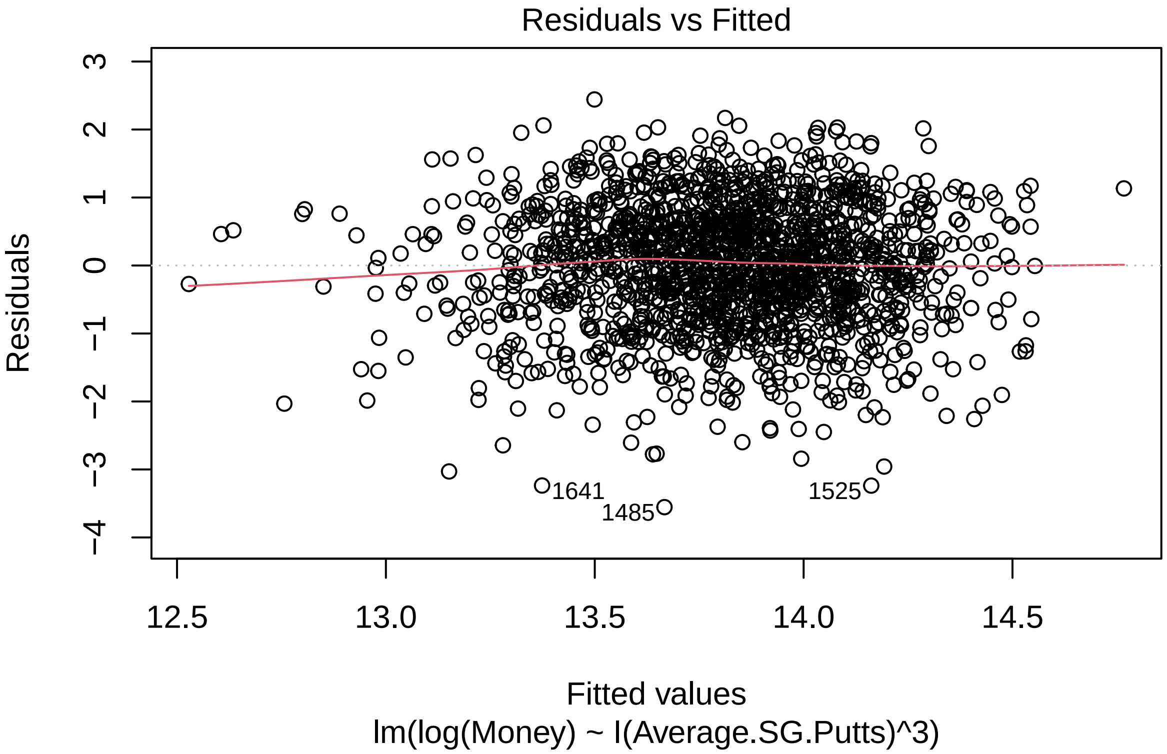
To make the relationship between "Average.Putts" and "Money" more linear, we do a cubic transformation for "Average.Putts." The line is closely centered around zero.



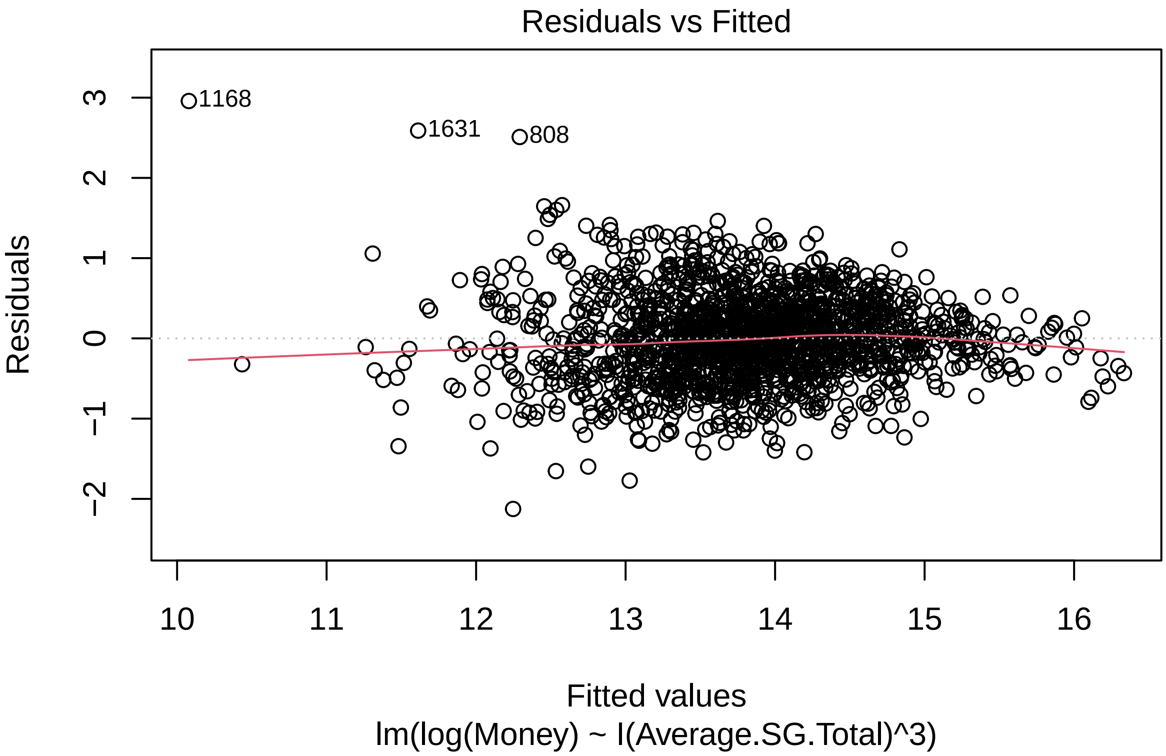
To make the relationship between “Average.Scrambling” and “Money” more linear, we applied a cubic transformation to “Average.Scrambling.” Although not perfect, the line is almost centered around zero.



To make the relationship between "Average.Score" and "Money" more linear, we apply a cubic transformation to "Average.Score." The line is closely centered around zero.



To make the relationship between "Average.SG.Putts" and "Money" more linear, we apply a cubic transformation to "Average.SG.Putts." The line is closely centered around zero.



To make the relationship between “Average.SG.Total” and “Money” more linear, we apply a cubic transformation to “Average.SG.Total.” The line is closely centered around zero.

10. Final Parsimonious Model

```
##
## Call:
## lm(formula = log(Money) ~ log(Avg.Distance^2) + I(Average.Putts)^3 +
##     I(Average.Score)^3 + I(Average.Scrambling)^3 + I(Average.SG.Putts)^3 +
##     I(Average.SG.Total)^3, data = Golf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9394 -0.3079 -0.0182  0.2992  2.6842
##
## Coefficients:
##                         Estimate Std. Error t value             Pr(>|t|)
## (Intercept)            61.324543   5.666929  10.821 < 0.0000000000000002 ***
## log(Avg.Distance^2)     0.912585   0.236694   3.856             0.000120 ***
## I(Average.Putts)       -0.329582   0.037107  -8.882 < 0.0000000000000002 ***
## I(Average.Score)       -0.648936   0.064716 -10.027 < 0.0000000000000002 ***
## I(Average.Scrambling)  -0.039530   0.005318  -7.434 0.000000000000168144 ***
## I(Average.SG.Putts)    -0.200650   0.054324  -3.694             0.000228 ***
## I(Average.SG.Total)     0.537353   0.066218   8.115 0.000000000000000934 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4983 on 1667 degrees of freedom
## Multiple R-squared:  0.7199, Adjusted R-squared:  0.7189
## F-statistic: 714.1 on 6 and 1667 DF,  p-value: < 0.00000000000000022
```

Our final parsimonious model performs much better than our original parsimonious model. 71.99% of the variation in money earned can be explained by the variation in “Avg.Distance,” “Average.Putts,” “Average.Score,” “Average.Scrambling,” “Average.SG.Putts,” and “Average.SG.Total.” Six metrics, including two Strokes Gained metrics, are significant. The adjusted R2 improved to 71.89%. Overall, we are satisfied with this model.

10.1 Interpretation of Final Parsimonious Model Coefficients

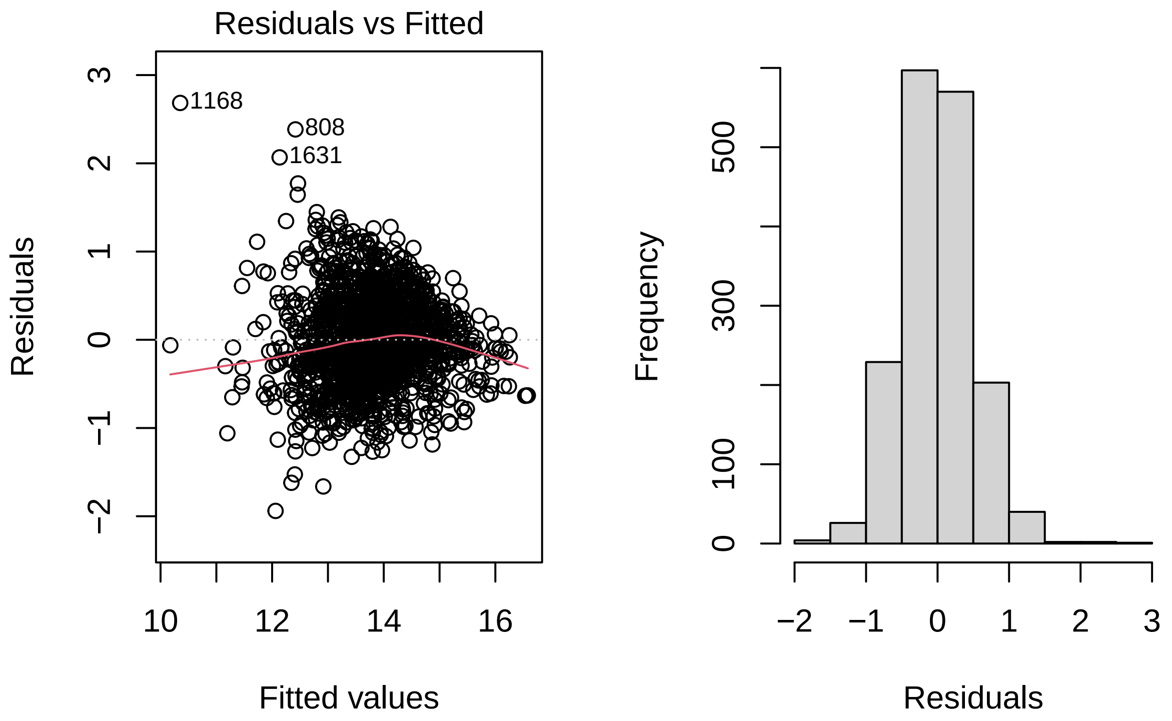
1. Average.Score is the most significant predictor. No matter how well a pro may drive, approach the green, chip, or putt, the whole in golf is greater than the sum of the parts.

2. Average.Distance has a positive relationship with money earned, indicating that as players increase their driving distance, they perform better.

3. Average.SG.Putts and Average.SG.Total are the only two significant Strokes Gained metrics. From 2010-2018, putting well contributed to sustained success on tour. The inclusion of Average.SG.Total in the model suggests that all Strokes Gained stats combined are significant.

4. Average.Putts, Average.Score, Average.Scrambling, and Average.SG.Putts have a negative relationship with money earned. As these metrics decrease, it is expected that money earned will increase.

11. Final Parsimonious Model Residual Diagnotics



From the residual plot for the final parsimonious model, we can see that there is much improvement in the trend. Although it is far from perfect, a significant curve does not exist anymore. However, we would like for the line to be more straight. The linearity assumption is close to being satisfied as the trend in the scatterplot is a line almost centered around 0. There appears to be no clear violation of independence. The residuals show a distribution that is approximately bell-shaped in the histogram. There appears to be constant variance centered around 0 in the residual plot, so the constant variance assumption is met.

12. Conclusion

Playing on the PGA Tour is nearly as difficult financially as it is in terms of skill and ability. Through our analysis we have created a model that PGA professionals can follow to maximize their winnings on tour. Our model says that while driving distance is a critical factor to playing profitably, short game performance is even more so. In our parsimonious model, we see that three out of six of our variables are related to short-game performance (Average.Putts, Average.Scrambling, Average.SG.Putts). This shows that a player who focuses their efforts primarily on improving their short game will see the largest "return on investment" so to speak. Spending money to work with trainers and coaches to improve in this area of the game is money well spent, according to our model.

Our model determined that the most important Strokes Gained metrics in earning money are Average.SG.Putts and Average.SG.Total. It stumps us that Strokes Gained:APR is not featured in our final model, even though it is historically a metric that holds immense value.

Beyond offering training plan insights to tour professionals, our model may also help decide star performers during the PGA season. Given that high yearly earnings inherently require multiple top-10 finishes or better at tournaments, it is reasonable to assume the players showing proficiency in our model's metrics will likely achieve high finishes in competition. In other words, our model may be useful to the sport-betting industry in figuring out odds or giving individuals an educated edge when placing their bets.

Our research team does not promote gambling of any form, nor do we guarantee or ensure successful results by using the model developed in this paper.

13. References

Cunneff, Tom. “The Tour’s evolving ShotLink data is a game changer, and not just for Strokes Gained geeks (you can bet on that).” Golf. September 5, 2019. https://golf.com/gear/pga-tour-shotlink-data-game-changer-you-can-bet-on-it/

Klein, S. Bradley. “Going the Distance.” USGA. April 23, 2021. https://www.usga.org/content/usga/home-page/articles/2021/04/going-the-distance.html

Powers, Christopher. “Want to become a better golf bettor? Here are the stats you should be paying attention to.” GolfDigest. October 29, 2020.
https://www.golfdigest.com/story/stats-you-should-be-paying-attention-to-pga-tour

Sens, Josh. “The Man With Two Brains: Strokes gained guru Mark Broadie’s pioneering analytics have radically altered the game.” Golf. September 11, 2018. https://golf.com/travel/the-man-with-two-brains-stokes-gained-guru-mark-broadies-pioneering-analytics-have-radically-altered-the-game/

Sherman, Jon. Four Foundations of Golf. LinkedIn. June 22, 2023. https://www.linkedin.com/posts/jon-sherman-945497ba_4-pga-tour-stats-that-will-change-your-perspective-activity-7087050047854370817-a6an/

Stachura, Mike. A closer look at handicap data shows just how much golfers have improved in recent years. GolfDigest. February 11, 2017. https://www.golfdigest.com/story/a-closer-look-at-handicap-data-shows-just-how-much-golfers-have-improved-in-recent-years

Taylor, David. “Do We Really Need to Lengthen Golf Courses?” Golf Monthly. August 22, 2016. https://www.golfmonthly.com/features/the-game/really-need-lengthen-golf-courses-113139