The Effects of Climate Memory on Expert Golfers in PGA Tour Events

A Regression Analysis by Jacob W. Hunter

# 

# Introduction

### Effects of Climate on Golf Play

A seasoned golfer carries many tools. They have the obvious set of 14 clubs, but their arsenal may include more unusual tools as well. They might carry notes they’ve taken about the course, with special care to remind themselves of past traps they’ve fallen into. They could carry advanced measuring devices to get precise distances to the flag. Perhaps the golfer has a knowledgeable caddie, one that knows exactly the right words to say in order to speak calm sense to them in the midst of a risky decision. Clearly a golfer carries many tools- but what else do they bring with them to the course? What else can affect their performance? What of their history, their tendencies, and their biases?

There is a near-ancient adage about the game: “Golf is ninety percent mental.” This saying is applied across many aspects of play. It is often used as a proof that an especially skilled golfer is mentally “tougher” than their competitors. Other times it is used to encourage belief in one’s self, as if to say that a majority of achievement relies just in the inward conviction that something can, and will, be done.

In a post-Sabermetrics world it is easy to scoff at these kinds of philosophies. It is nearly impossible to quantify something like “mental toughness,” and even if you could, what about all other elements under the “mental” umbrella? Nobody has ever performed rigorous studies on these kinds of things, so there is little reason to believe the mind has such a profound effect on the body and athletic performance. “Ninety percent!” Absurd!

And yet, any amateur golfer knows the panic that sets in when faced with a long tee shot over water. Experience tells us that comfort, to a tangible degree, affects the game. This comfort zone is perhaps the most active, and yet most nebulous, effect that a golfer brings with them onto the course. These are the kinds of things that merit an analytical study.

This paper seeks to see through some small part of the fog that is golf’s mental game. It does so by analyzing whether or not a sensible and comprehensive measure of “climate memory” increases the explanatory power of scoring models for professional golfers. That is to ask, “do pro golfers play differently in more familiar climates?” If this is true, then there may be more reason to believe golf is mostly “mental” in nature, and that golfers able to ignore the discomfort of an unfamiliar climate may have an advantage.

The study will be conducted as follows:

1. Find a suitable model for predicting scores based on performance statistics
2. Develop a metric which captures a golfer’s “climate memory”
3. Add that metric to the model from step 1
4. Observe any changes in explanatory power

# Studies

### Finley and Halsey

Authors Peter S. Finley and J. Jason Halsey of University of Northern Colorado have performed a comprehensive and intuitive study on predicting pro golf scores. Their model is a multiple regression that incorporates performance statistics:

* Driving Distance
* Driving Accuracy
* Greens in Regulation
* Sand Saves
* Putts Per Round

They published more updated models that included statistics like Scrambling and Bounce Back, but they ultimately found that Greens in Regulation and Putts Per Round were the greatest in determining scoring for a pro golfer. The model uses the inputs to predict a simple scoring average for an entire season.

### Other Models

Other studies have been performed in attempts to explain scoring variation. Davidson & Templin of Purdue performed a similar study to Finley and Halsey in 1986 with emphasis on the distinction between the “short game” and the “long game.” Scott Callan, Ph.D. and Janet Thomas, Ph.D. performed a regression study that analyzed NCAA golfers using academic statistics in addition to athletic ones. Christian Drappi and Lance Co Ting Keh of Duke University developed a predicting model on the shot-by-shot level. There are several more studies that need not be mentioned at this time, but the point stands that current research has had a bed of support for years.

### Why choose Finley and Halsey?

The Finley and Halsey model was chosen as the base for this study because of its intuitiveness, simplicity, and ease of recreation given the available datasets. It is recent enough to have captured the developments in golf technology, and it is within the computing means available.

# Methodology

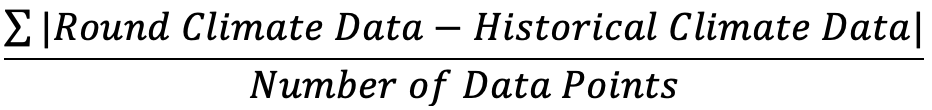
### Climate Memory Metric

Climate has pressing effects on a golf shot. Before the club is even swung, any moisture in the ground will reduce the ball’s spin after impact, reducing distance and increasing variance caused by wind. Air density can cause impressive differences in driving distance. Heat allows a golf ball to compress more at impact, increasing its elastic action off of a club face and allowing for more distance and spin. Needless to say, some discretion had to be used in the design of this climate memory variable. The initial design was to include the following points of data:

* Temperature
* Elevation (a proxy for air density)
* Humidity
* Wind speed
* Precipitation

The data collection process proved to cause an unexpected limitation, and will be addressed later.

The metric is designed to be a normalized difference between a golfer’s collective climate memory and the weather experienced during any given round they play. The metric is expressed by the following function:



where data points are normalized within their own dimensions to the scale of [0,1]. The Round Climate Data is specific to the day of the round in question. The historical climate data is an average specific to the hometown of the golfer in question. For additional specificity, the only historical climate data compared in the numerator is specific to the month in which the round in question was played. For example, for Jordan Spieth’s 2015 Sunday round at the Masters Tournament, the metric is comparing the weather in Augusta, Georgia on April 12 to the average April weather of Dallas, Texas (Spieth’s hometown). This produces the climate memory metric for that one round.

Recall that the Finley and Halsey model predicts a simple scoring average. Because of this, the climate memory metric must eventually be averaged over each golfer’s season to be factored into the regression.

# Data

There are two branches of data for this project: golf related and climate related.

### Sources

The golf data (statistics and past results) were gathered from PGATour.com. This website maintains a modest collection of statistics for pro golfers based on their performance solely on the PGA Tour.

The climate data had several sources. Elevation data is widely available, but data like wind speed and humidity are not. Historical data was collected from weatherbase.com, and daily data was collected from timeanddate.com. No formal assessment of these data sources was performed, as there was no alternative source of data at the most granular levels.

### Nature of the Data

All of the data is considerably structured.

The performance statistics on PGATour.com were aggregated on the seasonal level, which was disappointing as it did not allow for any analysis of specific metrics as climate changed.

The weather data was structured, but highly unorganized. Most of this effort was designated to scraping the web for this data. In scraping for the data it became apparent that there was no consistent and trustworthy way of collecting precipitation data, so that factor was no longer considered in the model.

### Collection Methods

After attempting to use two powerful web scraping tools (Excel and R) to no avail, it was decided that much of the data needed to be entered by hand. The data was compiled into 3 tables:

* A “Hometowns” table with one row for each golfer in the study
  + This table held the historical climate data for each of those golfers’ hometowns
  + This results in over 50 columns, as each historical climate data point had to be broken into monthly levels
* A “Finley and Halsey” table with one row for each season each golfer in the study participated in between the years 2015 and 2018, inclusive
  + This table held the performance statistics mentioned above
  + For example, there were 4 records for player Justin Rose, as his performance statistics were different between his 2015 and 2016 seasons, so on to 2018
* A “Rounds” table with one row for each round of golf examined in the study
  + This table held the climate data for the days those rounds were held

The players selected for the study were among the top tier of world rankings, chosen for a wide range of hometown climates. The rounds selected for the study were high caliber tournaments, as they were more likely to have data on the entire set of golfers.

### Cleaning Methods

The data went through a modest cleaning process before processing. Each table was importing into R with null rows, which were easily removed with one line of code.

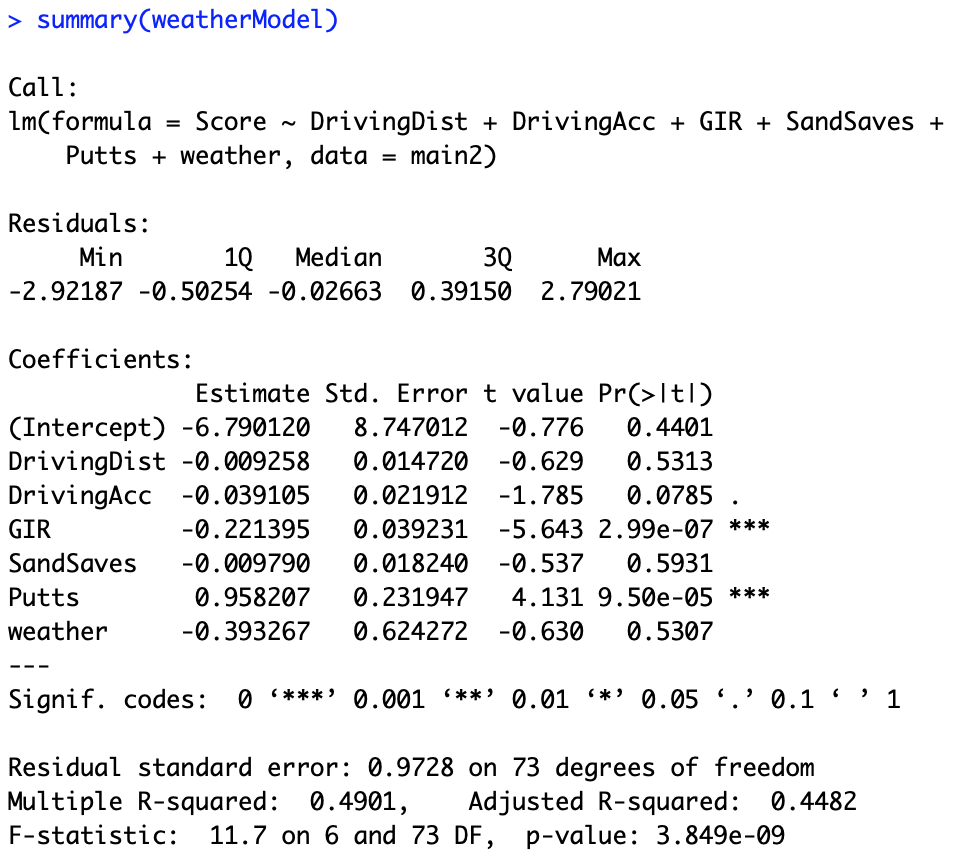
One singular dataframe was constructed by left joining climate data and removing unnecessary data over and over again. This was good practice for mastering the left\_join() function in R, among other cleaning and aggregation functions.

# R and Regression Analysis

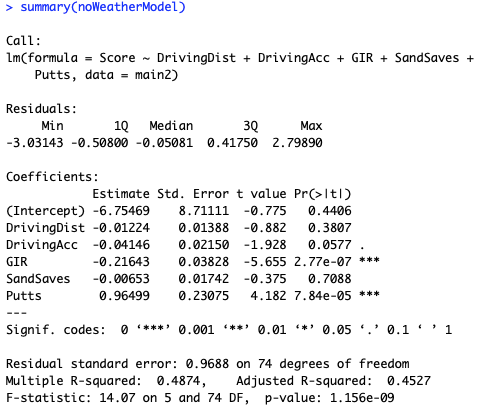
After many hours of wrangling, the data was finally ready for analysis. Several regressions were conducted in this process.

### Initial Regression

This was the first idea for the model. It included all variables from the Finley and Halsey model as well as the aggregated climate memory metric. The results are as follows:



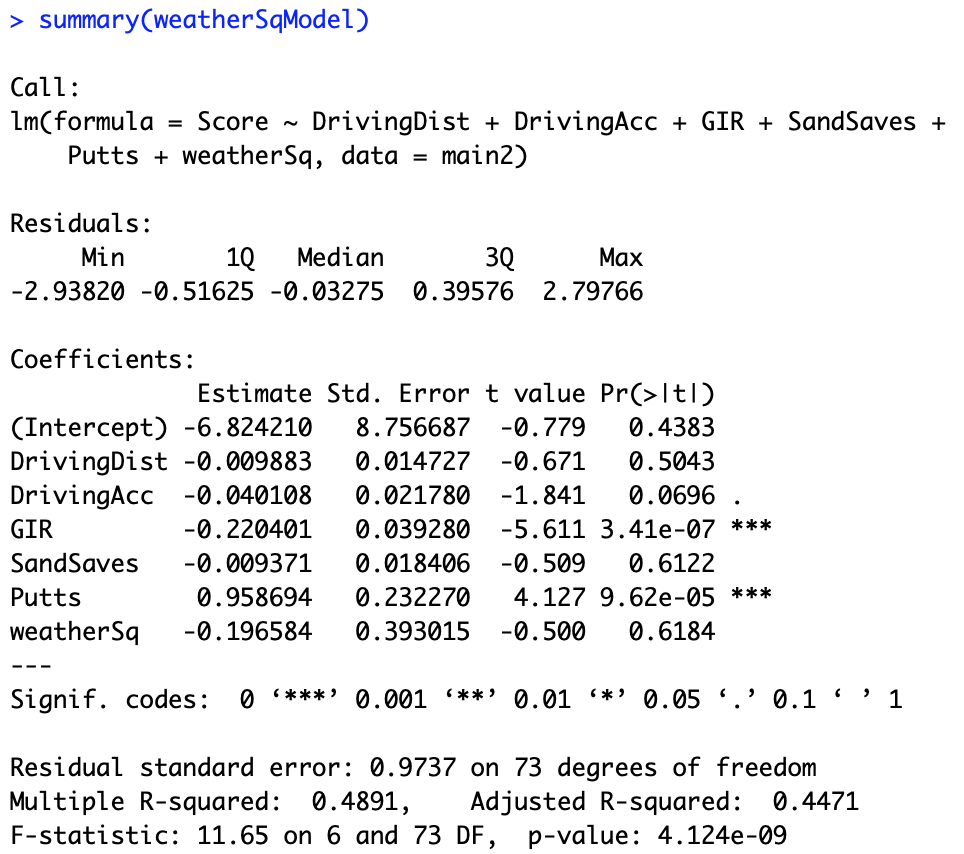
“weather” is the name representing the climate memory metric. Here we can notice a few things. First is that the Finley and Halsey model is not as explanatory as we would have expected. Perhaps there were differences in the data collection not being captured in this practice, but they do not have anywhere close to the R2 that the research would suggest we expect. We also see that the climate memory metric has a p-value of .5307, showing that there is no reason at all to reject the null hypothesis which claims weather has no explanatory power. For good measure, we compare to the Finley and Halsey model with no climate memory metric.



The Adjusted R2 has, in fact, increased. The climate memory metric is adversely affecting explanatory power.

### Climate Memory Enhanced

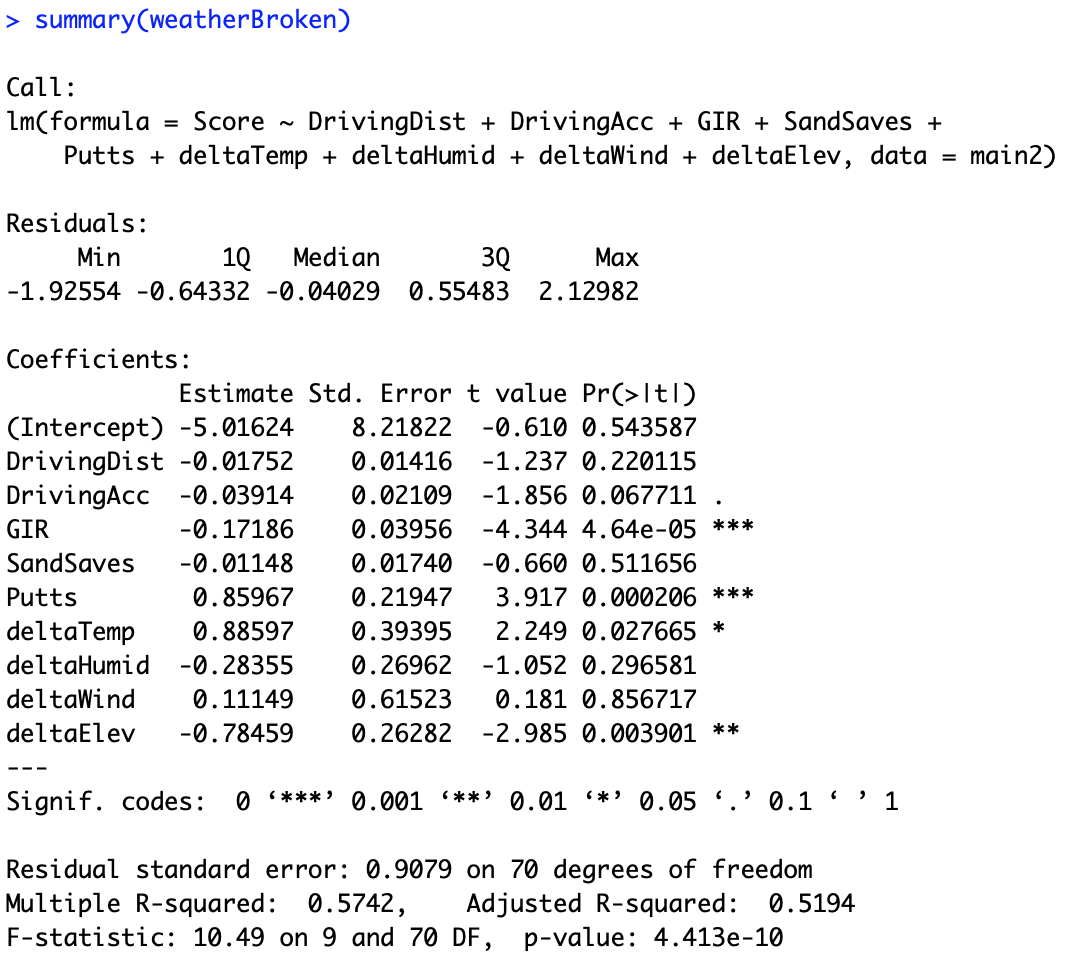
With the hunch that the climate memory could be augmented in some meaningful way, a regression was run that squared the climate memory metric. This would emphasize the effects of rounds that were in a climate very different from a golfer’s hometown. This is the following output:



The Adjusted R2 has, in fact, decreased once again. The squared climate memory metric is adversely affecting explanatory power.

### Broken Weather Climate Memory Metric

For the sake of deeper understanding, this model breaks apart the climate memory metric, including each climate factor separately. The output was more encouraging:



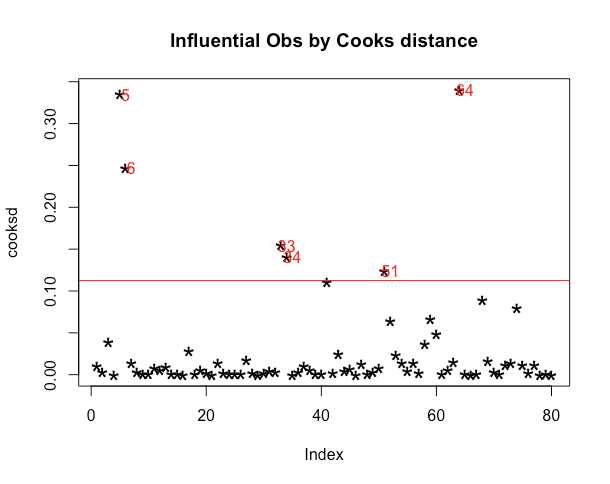
There is finally an increase in explained variation. We can see that there is actually some part of the climate memory that can be a contributing factor to explaining golf scoring. It is seen in the low p-values of deltaTemp (memory of temperature) and deltaElev (memory of air density, effectively). It is surprising that these variables have more significance than any performance statistics, but it is interesting to see that they rose above the other climate memory factors.

Temperature Memory has a positive estimate, so it asserts that a golfer’s scores will increase with any large difference in comfortable temperature. This is intuitive, as temperature can be a major source of discomfort if someone has not been acclimated. However, the Air Density Memory has a negative estimate. This would imply that as changes in air density increase, a golfer’s score gets worse. This is counterintuitive, as one would expect to perform best in a setting they are most comfortable in. The model cannot show us this, but perhaps most changes in air density are caused by *decreases*. A golf ball flies much farther with low air density, so that increased performance may be perceived and capitalized better with a golfer unfamiliar to it.

I will perform a more thorough analysis of this regression, since it seems to bear the most promise.

#### Outlier Treatment

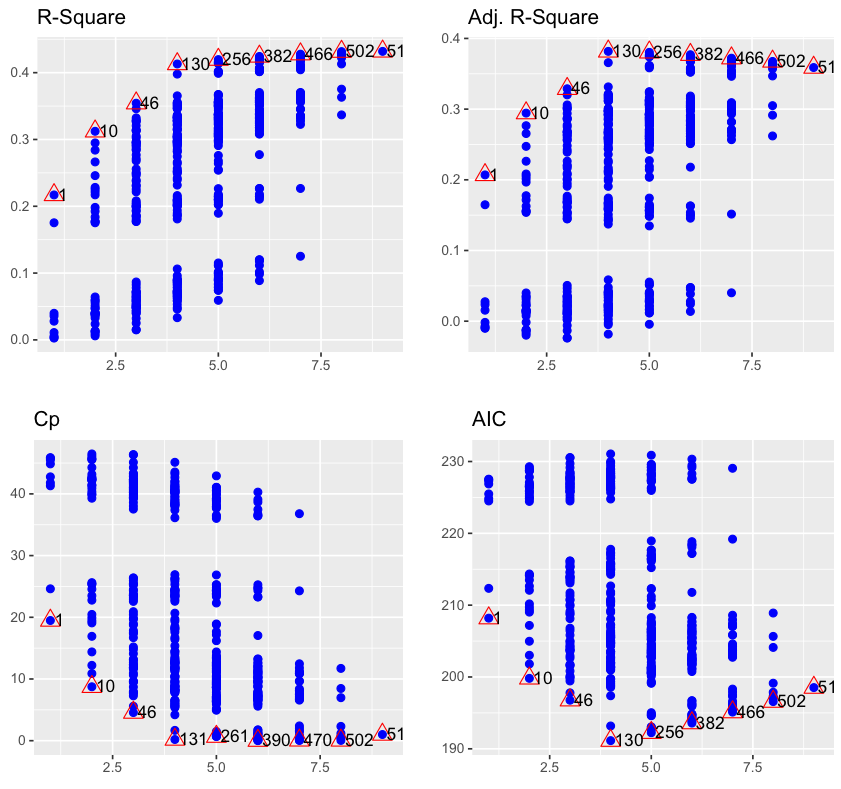
The plot below is a test of Cook’s distance for the observations. This is a test for observations that can be reasonably considered outliers with too much influence on the model.



Since the sample size is relatively small for this study, I exchanged all outliers with the mean output.

#### Stepwise Variable Selection

This process is made possible with Aravind Hebbali’s olsrr package. This function performs regression on all possible combinations of independent variables and generates visuals, noting well performing combinations.



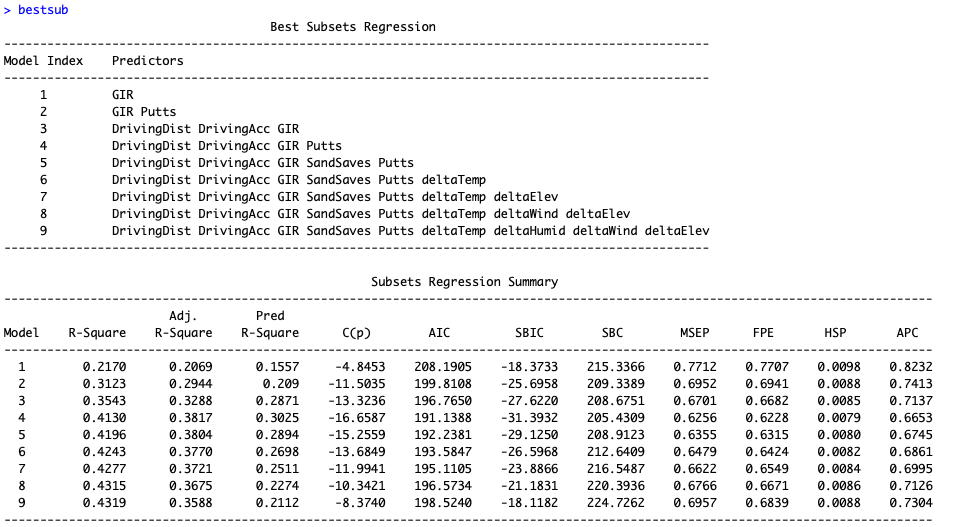
In these plots, each point represents a linear regression with a unique combination of variables from the ones provided in the Broken Weather Climate Memory Metric model. The x axis represents the number of variables in the model.

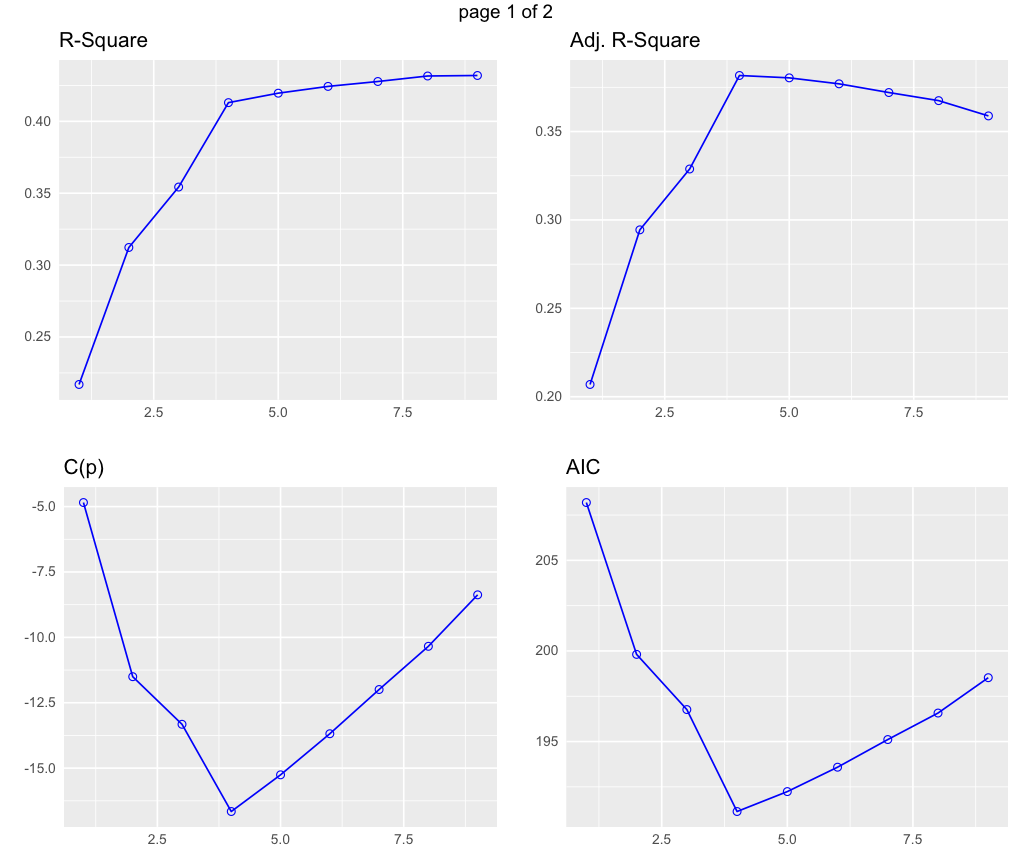
This shows that, as we increase the number of variables past 4 or 5 we gain no perceptible performance in the AIC or Cp statistics in addition to Adjusted R2. The 46th regression model included the following variables:

* Driving Distance
* Driving Accuracy
* Greens In Regulation

The 130th combination included Driving Accuracy.

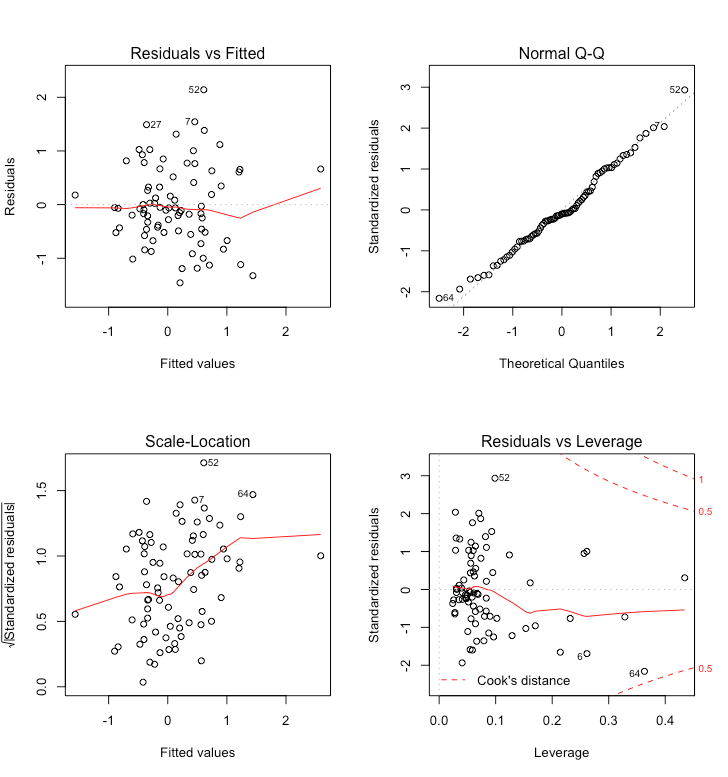
As another method of variable selection, I used the ols\_step\_best\_subset function from the olsrr package. This generates the best subsets of variables based on a number of goodness of fit metrics and plots visuals to show good comparisons between the models.





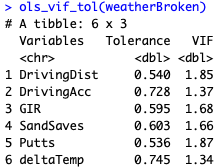
Here we can see each meaningful addition of variables. Note that after 4 variables, the AIC and Cp variables begin to increase, showing that it is the point that complexity is negatively affecting the model. These results are encouraging because, at least, we see that the Finley and Halsey model’s explanatory power is rising to the top. This implies that the model may be some reasonable representation of reality. For the next portion of the problem, the Broken Weather Climate Memory Metric model is consolidated to the best six factor subset (Driving Distance, Driving Accuracy, Greens In Regulation, Sand Saves, Putts, and Temperature Memory). This denies the model with the best explanatory power and simplicity, but some inclusion of climate memory should be tested for regression assumption diagnostics.

#### Assumptions analysis



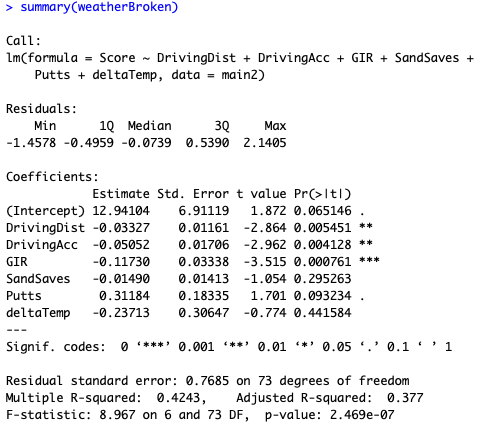
Here we see a few of the standard plots for analyzing the model as it pertains to the assumptions necessary for multiple linear regression. In the top left we see a residuals plot that assesses the assumption of a linear relationship between the independent variables and the output. This output shows that the model meets the linear relationship requirement. The Normal Q-Q plot tests the assumption that all variables are multivariate normal (i.e., the residuals are normally distributed. We should see the points centered on the y=x diagonal line, which is true here. The Scale-Location plot is a test for heteroscedasticity. We should see a reasonably straight line, noting that the variance does not change throughout the studied range of outcomes. This is not strictly true, according to the graphed output, but this is not a terribly bad result. The Residuals vs Leverage plot visualizes the influence of each data point. The red line should be relatively horizontal, showing that influence is equally spread. The dashed line represents Cook’s Distance. We can see that record 64 is upon that threshold, but not passing. The model passes this test with note of some reasonably influential points.

The last diagnostic is a test for multicollinearity. I used the ols\_vif\_tol function to generate the following variance inflation factors and tolerance statistics:



Here the tolerance statistics show that each variable is explaining at least half of its variance uniquely. Additionally, the variance inflation factors are reasonably close to 1, showing that there is little collinearity between the variables and their counterparts.

These diagnostics show that the model reasonably meets the demands for a multiple linear regression. For finality, the output for the Finley and Halsey model, temperature memory included, is presented below:



One item of note is the negative coefficient estimate for temperature memory- this is counterintuitive, as increased temperature difference would seem to be detrimental to scores. This model implies that an unfamiliarly cold or hot climate would drive scores down. Everything else seems to make sense- as Driving Distance, Driving Accuracy, Greens In Regulation, and Sand Saves increase, the score will *decrease* (hence the negative coefficient estimate). Putts is a measure of total putting strokes in a round, so an increase in that variable will lead to an increase in scoring.

# Limitations, Conclusions, and Further Research

Candidly, the author admits that this study is rife with limitations. Data was extremely difficult to collect, therefore the sample size was lacking. Additionally, the selection was only from the most elite golfers. Perhaps a less skilled sample of golfers to analyze would introduce more meaningful variation in scores.

Mentioned previously, this model was forced to not include precipitation data. That topic might be worth a study of its own. Additionally, there is little reason to believe that the historical weather data is viable.

Finally, there were clearly some issues with the reconstruction of the Finley and Halsey method. The factors included did not explain nearly enough of the variation in scoring. The most likely cause of this was the season-wide aggregate of the PGA Tour statistics. Were the scores analyzed on a round-by-round basis it is nearly assured that there would be more explanatory power.

In spite of all of this, there is some reason to believe that temperature memory could have some explanation for scoring. This could give some quantifiable weight to the heralded “mental” aspect of golf, but it depends on the psychological explanation for those climate memories.

With adequate resources, more research could easily be done on these effects. A set of performance statistics on a round-by-round basis would be invaluable.