THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS

WHICH AREA OF THE GOLF GAME IS MOST IMPORTANT FOR SUCCESS? A STATISTICAL ANALYSIS

THOMAS JOSEPH CLARKE
FALL 2013

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree in Science
with honors in Statistics

Reviewed and approved* by the following:

David R. Hunter
Department Head of Statistics
Thesis Supervisor & Honors Adviser

Linda Clark
Assistant Professor, Research Associate, Department of Statistics
Faculty Reader

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Within the golfing community, the debate over which aspect of the game is most important has raged since the sport's creation. Numerous studies within the past couple decades have used the large number of statistics available from the PGA Tour to determine the areas of the game which are most important to achieving success. Success in golf can be measured in numerous ways; some studies assert *Money Earned* or *Top 10 Finishes* are the best indicators, others believe shooting low scores is best and use *Scoring Average* as the indicator. Similarly, a vast array of statistics have been used to measure the various areas of the golf game and their effect on success. The current study differs from previous studies in two major ways. First, it benefits from the advances in technology made over the past decade which have allowed the PGA's data warehouse to grow to over 450 statistics, with virtually every aspect of the game being measured. Secondly, this increase in specificity results in statistics that can isolate and measure only the areas of the game which they are intended to measure. With the advent of new statistics like *Fairway Proximity* and *Strokes Gained-Putting*, these areas of the game can be isolated and measured and thus give a better diagnosis of the true importance of these areas on success. Using a standardized multiple regression approach, the current study found that driving distance had the largest impact on scoring average, followed by driving accuracy and putting, while iron play was found to be least important.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank the following people for making this thesis possible.

Dr. Linda Clark, who served as my faculty reader and selflessly dedicated countless hours of her time guiding this thesis through all its various stages. Thank you for taking the time to educate me along the way. Without you, this thesis would not have been possible.

Dr. David Hunter, who served as my thesis supervisor and honors adviser. Thank you for pushing the content of this thesis beyond its initial scope and for ensuring its validity.

My father, who introduced me to the game of golf at a young age. Thank you for teaching me everything I know and for all the great times we've had on the course together. It is my hope you'll be able to use the information contained in this thesis to make our matches a little bit more competitive.

My mother, who read through this thesis numerous times and contributed significantly to its development. Thank you for always supporting me.

My sister, Michelle, who has been a constant source of support and a tremendous friend throughout my college career.

Chapter 1

Introduction

The PGA Tour

The PGA, or Professional Golf Association, Tour is the organizer of the main men's professional golf tours in the United States and North America and is one of the most popular and competitive tours in golf. The 2012 schedule consisted of 45 official money events running from early January to November[1]. The PGA Tour website houses a database of over 450 statistics on 191 players, ranging from something as specific as *Approach Shot Proximity to the Pin from 250-275 Yards from the Rough* which is 38' 6" for Frederik Jacobson for example, to something as broad as the percentage of fairways hit.

There has always been considerable debate over which aspect of the golf game is most important and subsequently demands the most amount of practice time. Some constituencies argue that the short game, chipping and putting, is the most important part of golf while others advocate that being able to hit the ball long and straight off the tee presents a bigger advantage [2] [3]. Many studies in the past have used this warehouse of statistics to try to determine the relative importance of these areas.

Course/Terminology Overview

Courses consist of three types of holes, Par 3s, Par 4s and Par 5s. A traditional course has a total par of 72, consisting of four par 3 holes, ten par 4s, and four par 5s.

"Par" refers to the expected number of strokes it takes to finish a hole. Par 3s allow for one shot onto the green and two putts. Par 4s allow for an initial drive, a shot onto the green (known as the approach shot), and two putts. Par 5s allow for an initial drive, a layup shot which positions the player for the approach shot, an approach shot, and two putts. In other words, if a hole has a par of *n*, the player has $n - 2$ strokes to hit the ball onto the green with the $n - 2$ shot being the approach shot.

A player's initial shot on par 4s and 5s is called a "drive" and this shot is usually hit with a player's longest club, known as the driver, in an effort to get the ball as close to the hole as possible to set up for the shortest, and therefore most accurate, approach shot. How far the player hits the ball is known as the driving distance, and how often the player hits the fairway with his drive is known as his driving accuracy. When a player approaches the green, he typically uses an iron or a fairway wood. "Long" irons and fairway woods are used when the player is a long distance from the green, usually over 200 yards away for a professional golfer. "Short" irons are used inside 200 yards. Typically, the farther a player is from the green, the longer the iron he uses, and less accurate the shot usually is. If a player misses the green with the approach shot, or $n - 2$ shot, he is said to be "scrambling." A player then chips the ball onto the green and tries to make the subsequent putt in order to make par. The best players in the world consistently shoot under par on some of the hardest courses in the world.

Chapter 2

Literature Review

Much of the research done on the topic of analyzing golf game proficiency in an area of the game and its impact on performance was done many years ago. In 2004, during one of the greatest eras in golf history, Robert Quinn of the University of Nevada examined the correlations between *Driving Distance* vs. *Driving Accuracy*, *Putts per Round* vs. *Greens in Regulation*, *Putts per Round* vs. *Putting Average* and all of these variables impact on scoring. Quinn found a strong negative correlation between the two driving variables (-.61), meaning the longer one hits the ball, the less likely he is to hit the fairway. He also found a strong correlation between *Putts per Round* vs. *Greens in Regulation* (.41), meaning a player will hit more putts if he hits the green more. A player that misses a green in regulation takes one shot to chip the ball close to the hole and then putts the ball. The subsequent putt is usually of much shorter distance than the putts that are taken when a green is hit in regulation and is why someone who misses so many greens may end up taking many fewer putts than someone who does hit the green in regulation. This *Putts per Round* may actually be more indicative of someone's inability to hit a green than of someone's putting ability. This analysis reveals one of the major flaws in putting statistics that were used in past studies. All too often, statistics used to analyze putting ability are not realistic indicators of this area of the game. Quinn found that *Driving Distance* was the least correlated with *Weighted Scoring Average* (-.05) while the most highly correlated were *Putting Average* (.63) and *Greens in Regulation* (-

.62). *Putting Average* attempts to remove the effects of chipping the ball close after a green is missed in regulation. It is the number of putts that are taken on greens reached in regulation. While Quinn points out that this is a purer measure of putting ability, it still does not take into account that certain players can hit their approach shots much closer to the hole than others, and thus putt from much shorter distances on average. It should also be noted that the PGA Tour statistic *Driving Distance*, by definition, is only measured on two holes per round, and is not very representative of a player's average driving distance off the tee. He also notes how *Greens in Regulation* has nearly the largest impact on scoring average. This may be the case, but this is a result of using a variable that encompasses multiple aspects of the game. The *Greens in Regulation* variable is impacted by a player's driving ability and iron ability. In other words, this statistic measures a player's prowess in all areas of the game except putting and scrambling. If another statistic was created that somehow measured a player's driving, iron, and putting ability, it would likely lead to an even higher correlation with *Scoring Average*. While *Greens in Regulation* is highly correlated with *Scoring Average*, no meaningful conclusions can be drawn from this other than that driving ability and iron ability have a significant impact on *Scoring Average*[4].

In 2001, Robert Ketscher and Trevor Ringrose did a similar analysis using 1998-2001 summary statistics of the PGA European Tour. The pair eloquently describes why *Scoring Average* is the best indicator of a golfer's success over the course of a season. Prize money won is staggered depending on a player's finish within the tournament, but actual amounts differ from tournament to tournament. Comparing average finishing positions, they explain, is a much better unweighted indicator, but a player who is runner-

up shooting six shots adrift is treated the same as a player losing in a play-off. The study uses an unadjusted scoring average statistic, and notes how this is not perfect because certain courses have a higher par, or a higher difficulty with the same par, than others and players enter different combinations of tournaments. This brings to light the benefits of using more current and advanced statistics. A new statistic, *Weighted Scoring Average*, was developed after this study was done and weights a player's score for the round against the field average. In this way, the statistic controls for the difficulty and conditions of the course for that day. The study attempted to measure components of the golf game that impact scoring average but concludes that only *Driving Distance* and *Driving Accuracy* are true measures of one aspect of the game while other figures are representatives of mixtures of skills. Ketscher and Ringrose say that driving data is relatively easy to collect since, at a given hole, every player starts in virtually the same place. The data for iron shots, they continue, would need to record start as well as finish locations, not to mention positions relative to the green, since all analyses would need to be conditional on where the drive ended to distinguish between driving and iron skills. In their proposals for future work in regards to putting, they mention an interesting figure to collect would be to record the distance from the hole both before and after every putt a player makes, giving a more direct measure of a player's putting ability. They conclude that a greater variety of data, at a greater level of detail, needs to be collected in order to achieve practically useful results. This current thesis will benefit from advances in data variety and detail, and will even expound upon Ketscher and Ringroses's final thoughts on ideal iron and putting figures[5].

Edgar Shields and Nathan Tomasini, of the University of North Carolina, did a study that sought to develop a statistical profile of the PGA golfer for the 2003 season and to determine the relative importance of subparts of the game to the average 18-hole score. They collected 19 statistics on 190 players, and stratified the players into a Top Group, composed of the 50 players who had the lowest mean score per round, and a Bottom Group, composed of the 50 players who had the highest mean score per round. They then used t-tests to evaluate the differences between the two groups and relationships between average scores per round and 18 other selected statistics. The statistics selected encompassed many areas of the game, but also included statistics such as *Birdie Average,* which is inherently correlated with a low score. The only statistic where there was not a significant difference between the two groups was *Driving Distance* while *Driving Accuracy*, *Greens in Regulation*, *Scrambling* and *Putting* all emerged as significant. While this study claims to determine the relative importance of each area, it uses statistics that encompass more than one area of the game. As was mentioned before, *Greens in Regulation* not only measures a player's approach shot ability but also his driving ability. *Scrambling Percentage* (the percentage of time a player makes par or better after his approach shot misses the green) not only measures a player's chipping ability around the green but also his putting ability. In this way, this study fails to reach any meaningful conclusions other than that there is no significant difference between the top players and bottom players in how far they drive the ball[6].

Several studies have sought to evaluate the validity of the most popular adage in golf: "Drive for show, putt for dough." Donald Alexander and William Kern, of Western Michigan University, analyzed whether or not this adage held true during their lengthy

analysis over the 1992-2001 tour seasons. In their model, they used inflation adjusted earnings from winnings in official U.S. PGA Tour events (not the European PGA tour, for example). The study attempts to account for the previously discussed way in which certain measures are contaminated by others by regressing *Greens in Regulation* by *Driving Distance* and *Driving Accuracy* to develop the *IRON* statistic, and also by regressing *Putts per Green in Regulation* by this *IRON* statistic to account for the distance from the hole that these putts are made. The article ultimately found some limited support that the popular adage "Drive for show, putt for dough" may no longer be relevant with the advancement in technology in player equipment and subsequent lengthening of courses. There was a small increase in the marginal value of *Driving Distance* over the years observed while that of putting declined. However, while all statistics were deemed significant determinants of the earnings of PGA Tour golfers, putting ability was still the single most important determinant of earnings. The article concludes that professional golfers would do well to heed the message to keep practicing their putting, because improvement in that skill remains the fastest way to increase their earnings[2].

D.S. Belkin, Et Al found similar results in statistics two decades prior while analyzing statistics from 1986 to 1988. When analyzing the effects of *Driving Distance, Driving Accuracy, Greens in Regulation, Sand Saves %,* and *Putts per Round,* he found that *Greens in Regulation* had the largest impact on *Scoring Average,* while *Putts per Round* came in a close second. Interestingly, a noticeable increase in the importance of *Driving Distance* and *Driving Accuracy* occurred over the three year span, which corroborates the findings of Alexander and Kern above. The study also noted that the

correlation between *Scoring Average* and *Money Earned* during these three years was very significant, .77, and expands on the downside of using *Money Earned* as a measure of success. While the correlation is high, the lack of complete correspondence between *Scoring Average* and *Money Earned* suggests that *Scoring Average* does not necessarily reflect how well a golfer plays in one particular tournament or the money he earns over the course of a year. A professional with a low overall scoring average may not earn a great deal of money unless he places very high in several tournaments while a poor overall performer may be masked by a couple high finishes given that the distribution of money is heavily weighted toward the top finishers[7]. This is the same logic which will support the use of *Scoring Average* in this thesis.

George Engelhardt, of the Sports Performance Institute, tested a similar, lesser known adage "It's not how you drive, it's how you arrive". He found a significant positive correlation between the top ten money winners and their total driving skills. The correlation between the rankings of money leaders and *Greens in Regulation* was insignificant. Each year, those players ranked as the top ten in *Total Driving* won more money per tournament than the players ranked in the top ten in *Greens in Regulation*, casting doubt on the adage and suggesting that driving skills contribute most to winning money. The study examined performance statistics from the 1993 and 1994 PGA Tour, and only observed *Total Driving*, which is a player's combined ranking of *Driving Distance* and *Driving Accuracy*, and *Greens in Regulation*. Despite its limited scope, it does seem to corroborate the trend towards an increase in the importance of driving ability identified in Belkin et al's analysis of 1986-1988 statistics[3]. *Total Driving* is a PGA Tour measure which is meant to incorporate the effects of both *Driving Distance*

and *Driving Accuracy*. This statistic is recorded by combining a player's rank in these two areas, with the player having the lowest combined ranking being the top ranked in *Total Driving*.

The later study by Wiseman and Chatterjee points out that this measure is obtained by adding together the ranks of the two individual driving measures; ranked data are nonnumerical, ordinal-scaled data which cannot be meaningfully summed. Frederick Wiseman and Sangit Chatterjee differed from previous studies by creating their own total driving statistic and by evaluating all players over a 15 year period from 1990-2004. They calculated their *Total Driving* statistic by multiplying the *Driving Distance* by the *Driving Accuracy* percentage. Throughout most of the fifteen year period, the *Total Driving* measure increased, but, in the final four years of the study, the measure declined as the increase in *Driving Distance* led to a decrease in accuracy. Their analysis indicated relatively stable results over time with *Greens in Regulation* and *Putting Average* as the two measures explaining most of the variance in *Scoring Average*. The *Total Driving* measure indicated that the joint effects of driving have apparently diminished in recent years, especially in relation to *Greens in Regulation* and *Putting Average*. One reason for this is the increased negative correlation between *Driving Distance* and *Driving Accuracy*[8]. These conclusions actually run contrary to several of the studies mentioned above. However, previous studies have not used the same calculated *Total Driving* statistic as Wiseman and Chatterjee to measure driving ability's effect on scoring average.

Thomas Dorsel and Rob Rotunda analyzed 1990 PGA Tour statistics *Driving Distance, Driving Accuracy, Greens in Regulation, and Putting Average* in relation to

their impact on the performance outcomes *Scoring Average, Top 10 Finishes*, and *Money Won*. They only analyzed the 42 golfers who ranked among the top 130 statistical leaders in each of the four primary skill categories and three performance outcome variables, denoting these players as the "Elite 42". The study varied in the results of what statistic was most important depending on whether they used a multiple regression or simple correlation analysis. However, a key takeaway and trend seen in this study is that *Driving Accuracy* was one of the top determinants in all of the performance outcomes while *Driving Distance* hardly correlated with any of the performance variables examined. The study qualifies this finding by noting that this does not mean that driving distance is not important. All professional golfers hit the ball a reasonable distance. Perhaps once that reasonable length has been achieved, they surmise, one may gain more in overall performance by focusing on accuracy rather than trying to add more distance. All in all, this study took an interesting approach by only analyzing the top 42 golfers and the effects of the four broad variables. This study also makes the mistake of correlating *Greens in Regulation* to iron play, and, again, neglects the fact that this variable is also a reflection of a golfer's driving ability[9].

It appears many studies that use earnings as a measure of success tend to conclude that putting ability is the most important area of the game. In John Watkins extensive statistical analysis of 2006 statistics on earnings, he found that the driving statistics had a negligible effect on earnings while *Greens in Regulation* and putting ability, as measured by *Putting Average*, are the two most important determinants of player earnings[10]. Leo Kahane, of the International Journal of Sport Finance, sought to better handle the skewness and outlier values found in PGA earnings data by using quantile regression,

using data from the PGA for the years 2004 to 2007. This quantile regression approach found that *Greens in Regulation, Putting Average,* and *Save Percentage* (the percentage of time a player makes par or better when missing the green in regulation) had an extremely significant impact on earnings, in that order, while *Driving Accuracy* and *Driving Distance* were barely significant.[11].

If there is one common theme throughout all these studies, it is that there is no consensus on which area of the game is most important. Methodologies among these papers differ, from including all golfers on the PGA tour, to stratifying a select number into top and bottom groups, to examining the only 42 golfers who rank in the top 130 of several skill and performance categories. Studies also differ in their measure of success. Some studies use *Scoring Average* because a player only controls his score and always wants to lower it, while some use *Money Won* because, at the end of the day, the players want to win more money. For the most part, however, the past studies suggest that *Greens in Regulation* and *Putting Average* have the largest impact on success, while driving ability might be becoming increasingly important. One possible area for concern is that many of these studies relate statistics to areas of the game which they do not necessarily represent. Several studies have found *Greens in Regulation* to be significant, and subsequently concluded that iron play was the most important area of the game, without acknowledging that one's ability to hit a green with an approach shot is heavily influenced by that shot's originating position. It is very possible for a player to hit many greens in regulation because his driving ability positions him so well for the approach shot, and not necessarily because he is just a skilled iron player. Similar alternative explanations are possible for the *Scrambling Percentage* statistic.

This study will differ from previous studies in several ways. The most recent of all the aforementioned studies used 2007 data, while this study will use data from the 2012 PGA Tour season. Over those five years, the PGA Tour has vastly expanded the number, specificity, and sophistication of statistics measured. Perhaps the greatest advancement has been in putting statistics. Previous studies have used *Putting Average* or *Putts per Round*, which fail to take into account from how far away each putt is taken. This study, however, will use the recently developed *Strokes Gained* statistic which takes into account the length from which each putt is taken. While this study has similar goals to those mentioned above, it will accomplish those goals by choosing statistics that truly isolate each area of the game, and can do so because of the advancement in the specificity of statistics recorded by the PGA. Instead of using *Greens in Regulation* to measure players' supposed iron efficiency (even though it is impacted by driving skill), this study will use *Fairway Proximity* to measure just how close players hit the ball to the pin when hitting from similar environments. Instead of using *Scrambling Percentage* to measure a player's supposed short game efficiency (even though it is clearly impacted by putting skill), this study will use *Scrambling Average Distance to the Pin* to remove the putting element from this measure. Similar driving statistics will be used, but again, this study will benefit from advancements in the statistics tracked by the PGA and use a statistic that measures driving distances on all holes as opposed to just two like in the studies observed. While the goal is the same, the methodology and variables used will be different than past studies.

Chapter 3

Selection of Variables

Use of a Regression Model

The goal of this thesis is to use specific variables to predict a player's weighted scoring average, and then analyze which variables have the largest impact on the scoring average. Because this is the goal, a linear regression model will be used. Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data[12]. The dependent, or response variable, is the variable which is attempting to be predicted from the other variables. The explanatory, or independent, variables are the multiple variables which will be used to explain the variability in the response variable.

Variable Selection Rationale

By investigating which areas of the golf game have the largest impact on scoring average, players will be able to identify the areas in which the most successful players typically excel. While this observational study cannot establish a causal relationship between success and these areas of the game, it is worth noting as some element of causation may exist. This section will outline the rationale behind the selection of the variables included in the theory based regression model. The emphasis will be on choosing and manipulating variables that are the best indicators of each aspect of the game. The aspects of the game that will be observed are driving distance, driving

accuracy, short iron play, long iron play, scrambling, and putting. As was mentioned in the literature review, many past studies used indicators that did not adequately isolate the areas of the game. For example, *Greens in Regulation* was often used to measure a player's iron prowess. However, this conclusion does not take into account that a player's ability to hit greens in regulation is not only dependent on his iron play ability, but also on his driving ability, as driving ability is what positions the player for the approach shot. *Greens in Regulation*, therefore, is not only an indicator of iron shot skill but also of a player's driving ability. A player cannot focus on improving the *Greens in Regulation* aspect of his game, rather the player must focus on driving the ball straighter and longer off the tee, and then on hitting his irons more accurately into the pin. By isolating these different shots, more useful conclusions can be made about the areas of the golf game that must be emphasized during practice.

As was outlined in the Literature Review, much of the rationale for the selection of variables used in past studies to predict player performance contains several assumptions that may be faulty. This study will use different variables that are more representative indicators of a player's proficiency in certain areas of the game. Because the purpose of this study is to determine which types of shots have the biggest impact on scoring average, the sole focus will be on six different areas that encompass each aspect of the game.

Driving Distance-All Drives

The first of these statistics is driving distance. Simply put, this is the average distance a player hits the ball on his first shot on Par 4s and Par 5s. Whether or not the

ball lands in the fairway is irrelevant. There are two variables that measure this driving distance. The first is named *Driving Distance* on the PGA Tour website. The website states,

> The average number of yards per measured drive. These drives are measured on two holes per round. Care is taken to select two holes which face in opposite directions to counteract the effect of wind. Drives are measured to the point at which they come to rest regardless of whether they are in the fairway or not[13].

A big problem with this variable is that the "measured drive" is only for two holes per round. The two holes are selected by the hosting tournament committee. The chosen holes are holes in which the player will most likely use their driver so as to indicate how far the player hits the ball when using the club he can hit the farthest. In this way, this variable measures the average distance a player can hit the ball with his longest club. This variable fails to take into account how far the player hits the ball on average. For example, a player may be able to hit the ball the longest on tour with his driver, but be so erratic that most of the time he prefers to tee off with his more accurate 3-wood. This player may rank at the top of the *Driving Distance* statistic, but may actually be hitting the ball much shorter than his competition on all of the other holes in which the drives are not measured. Because of these reasons, the second driving distance statistic, *Driving Distance-All Drives*, may be a better statistic, and the one which will be used in the model.

Driving Distance-All Drives is defined as, "The average number of yards per drive for all drives where the distance was measured by a laser[14]". This variable encompasses all drives on Par 4s and Par 5s, and is not subject to the selection of the tournament committee. The result is a statistic which is much more representative of how

far a player actually hits the ball on average. The players who hit the ball very far with their drivers, but use them infrequently, will be taken into account with the "all drives" statistic. These players who use 3-woods, or low irons off the tee will not be hitting the ball as far as their driver-using counterparts, and their driving distance with this statistic will suffer as a result. The difference between these two variables is evident when observing their means. The mean distance for the *Driving Distance* variable is 290.5 yards, more than seven yards longer than the 282.76 yards mean of *Driving Distance-All Drives*. This is because the *Driving Distance* variable includes drives made mostly using the driver, while the *Driving Distance-All Drives* accounts for every drive on Par 4s and 5s. *Driving Distance-All Drives* will be used in the model because it gives a more representative measure of how far the player usually hits the ball off the tee.

Driving Accuracy

The use of this statistic is rather straightforward. It is defined as "The percentage of time a tee shot comes to rest in the fairway (regardless of club)." This variable will indicate how often a player hits the ball into the fairway off the tee on Par 4s and Par 5s. Players who are less accurate off the tee are usually punished because they must hit their next shot from the "rough", or grass that is longer, thicker and less groomed than the fairway. Also, missing the fairway brings into account the possibility of incurring penalty strokes for hitting the ball out of bounds (a two shot penalty) or into water (a one shot penalty). The average *Driving Accuracy* percentage was 61.029% with the most accurate driver, Jerry Kelly, hitting 73% of his fairways and the least accurate driver, Daniel Chopra, hitting only 47% of his fairways[15].

Another possible option for determining this driving accuracy was the *Distance from the Edge of Fairway* statistic. This statistic measured how far a player was from the edge of the fairway when the player missed the fairway. The rationale behind this statistic's early inclusion in the model was that it would take into account *how severely* a player missed the fairway on his drives. It is possible a player could miss the fairway a large percentage of the time, but only miss the fairway by a couple feet when he does miss. Another player may hit a larger percentage of the fairways, but miss badly when he does miss. This statistic would penalize the player for bad misses. In the end, however, these statistics were very highly correlated so *Driving Accuracy* was selected as the only one to be used in the model.

Fairway Proximity (Inches)

After a player's initial drive, the next shot is often toward the green. This approach to the green is an important shot in golf, and the *Fairway Proximity* variable captures this proficiency well. The statistic is defined as:

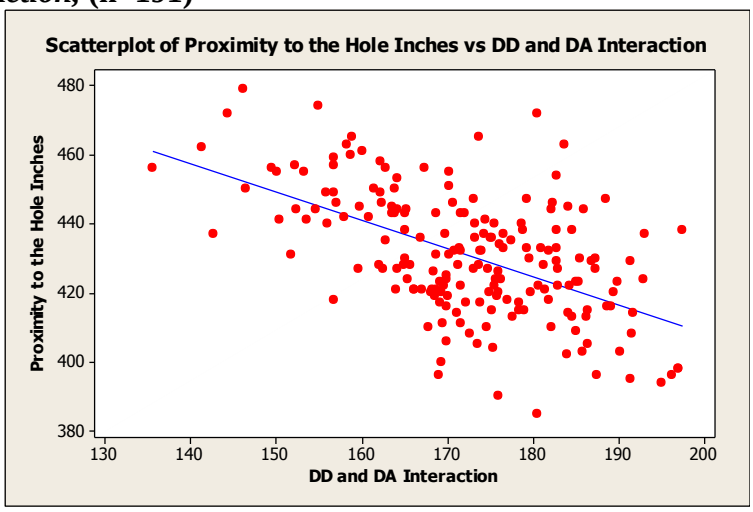
> The average distance remaining to the hole for all approach shots hit from the fairway (or the tee box on a par 3). The approach shot distance must be determined by a laser, and the shot must not originate from on or around the green. The shot also must end on or around the green or in the hole. Note: 'Around the green' indicates the ball is within 30 yards of the edge of the green[16].

The definition states that the ball's proximity to the hole is measured when the ball originates from the fairway from greater than 30 yards from the edge of the green. The goal of this statistic is to measure the player's approach shot ability. Some may argue that this statistic fails to incorporate how well a player hits the ball out of the rough, which is true. However, by using *Fairway Proximity*, all players are originating their shots from a

near identical setting. In other words, when the players' balls are resting in the fairway, this statistic shows how close they hit the ball to the pin. *Rough Proximity* was also an option to assess this approach shot proficiency, but there would be a lack of consistency in how the players' balls laid at rest in the rough. In golf, the setting of the ball in relation to the ground is referred to as the ball's "lie". A good "lie" would consist of any ball at rest in the fairway, or even perhaps a ball that was sitting up in the rough where the iron could easily make solid contact with it. Conversely, a bad "lie" refers to any ball where solid contact may be impeded by the ball's surroundings, such as when the ball is sitting down in thick rough, or behind a tree. By using this *Fairway Proximity* variable, these "lies" are being controlled for because it is assumed that the vast majority of balls that come to rest in the fairway have good "lies".

Some may believe that *Proximity to the Hole* would be a better statistic to use, in that it would encompass a player's ability to hit the ball from both the rough and fairway into the green. *Proximity to the Hole* has the same definition as *Fairway Proximity*, except the ball can originate from anywhere, including the rough. This statistic may seem more representative at first because it includes a player's ability to hit from the rough, but it is excluded from the model because it is too heavily influenced by a player's driving ability. A player who misses the fairway a lot will be playing a larger majority of his shots into the green from the rough, and thus, from worse "lies". As a result, this player is likely to have a larger (worse) *Proximity to the Hole* score because he hits more of his shots from the rough. This can be demonstrated with a simple correlation plot of *Proximity to the Hole* vs. *Driving Accuracy and Driving Distance Interaction* shown in Figure 1-1.

Figure 1-1: Scatterplot of *Proximity to the Hole* vs. *Driving Accuracy and Driving Distance Interaction*, (n=191)



Pearson correlation of DD and DA Interaction and Proximity to the Hole Inches = -0.548
P-Value = 0.000*

**Note: A p-value of less than .05 means that the two variables are statistically significantly correlated. A p-value above .05 means the variables are not significantly correlated.*

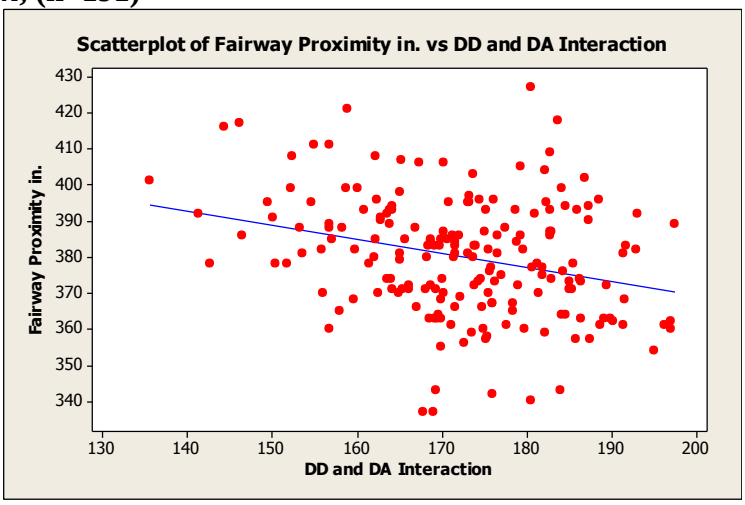
Assumptions: There are four assumptions that are made in the calculation of a p-value. (1) The two variables are measured on an interval or ratio scale (the variables are continuous). (2) There is a linear relationship between the two variables. (3) There are no significant outliers. (4) The variables are approximately normally distributed.

(1) Proximity to the Hole and Driving Accuracy and Driving Distance Interaction are continuous variables.
(2) The scatterplot demonstrates a linear relationship.
(3) While there are some outliers, none appear to be extreme.
(4) Both variables are approximately normally distributed.
The 4 p-value calculation assumptions are met.

The *Driving Accuracy and Driving Distance Interaction* is calculated by multiplying the *Driving Distance* statistic by the *Driving Accuracy* statistic. The interaction then gives a variable that includes all aspects of driving skill. A player who drives the ball far and accurately will have a large interaction value, while a player who drives the bar equally as far but less accurately will be penalized with a smaller interaction value. As is demonstrated in Figure 1, the correlation coefficient of -.548, and the highly significant p-value of 0.000, *Proximity to the Hole* is significantly negatively

impacted by these driving statistics. In this manner, *Proximity to the Hole* is not only dependent on a player's approach shot capability, but also on his driving ability. Including this variable in the model would defeat the purpose, because the variable is not measuring solely iron play ability, but also driving ability. The scatterplot of *Fairway Proximity* vs. *Driving Accuracy and Driving Distance Interaction* and corresponding correlation coefficient demonstrate how using *Fairway Proximity* instead controls for the approach shot's originating position.

Figure 1-2: Scatterplot of *Fairway Proximity* vs. *Driving Accuracy and Driving Distance Interaction*, (n=191)



Pearson correlation of DD and DA Interaction and Fairway Proximity in. = -0.288
P-Value = 0.000
(1) Fairway Proximity and Driving Accuracy and Driving Distance Interaction are continuous variables.
(2) The scatterplot demonstrates a linear relationship.
(3) While there are some outliers, none appear to be extreme.
(4) Both variables are approximately normally distributed.
The 4 p-value calculation assumptions are met.

From examining Figure 1 and Figure 2, it appears that there is a weaker relationship between *Fairway Proximity* and driving ability than between *Proximity to the Hole* and driving ability. Because the goal is to isolate each area of the game and use variables that

are not impacted by other variables already accounted for in the model, *Fairway Proximity* will be used.

Long Iron Calc

The *Long Iron Calc* variable was calculated using the *Average Distance after Going for It Shot* and dividing it by the *Average Going for It Shot Distance (yds)* variable. The *Average Distance after Going for It Shot* variable is defined as follows: "The average distance (in yards) remaining to the hole after the going for it shot. Going for it shots are the first shots on par 4's that land on or around the green and second shots on par 5's that land on or around the green or in the water[17]." This variable effectively tests how well players hit their longer irons and fairway woods into a green. When "Going for It" a player is usually much farther away from the green than when approaching the green with his regulation shot (2^{nd} shot on Par 4, 3^{rd} shot on Par 5.) This can be demonstrated by observing the top ranked players' *Average Going for it Shot Distance* and *Average Approach Shot Distance*. In 2012, Adam Scott has the shortest *Average Going for it Shot Distance* with an average of 236 yards while the player with the shortest *Average Approach Shot Distance*, David Toms, had an average of 157 yards. Therefore, it is fair to say that the *Average Distance after Going for It Shot* is an indicator of a player's long iron/fairway wood efficiency because the shot originates from a much larger distance where long irons and fairway woods must be used.

This variable alone, however, is not enough. In the same way that *Proximity to Hole* could not be used to assess short iron proficiency because it was heavily dependent on driving proficiency, *Average Distance after Going for It Shot* is dependent on a

player's position off the tee. A player's position off the tee is determined by the driving distance, and whether the ball lands in the fairway. The correlation between *Average Distance after Going for It Shot* and *Driving Distance-All Drives* and *Driving Accuracy* are below.

> Pearson correlation of Driving Distance-All Drives Avg and Avg Distance after Going for it Shot = -0.490
> P-Value = 0.000
>
> Pearson correlation of Driving Accuracy Percentage and Average Distance after Going for it Shot = 0.033
> P-Value = 0.650

Because there is only a significant correlation between *Driving Distance-All Drives* and *Average Distance after Going for It Shot*, the *Average Distance after Going for It Shot* variable will be adjusted to account for those players who hit their "going for it" shots from a shorter distance. The *Average Distance after Going for It Shot* variable will be divided by the *Average Going for It Shot Distance (yds)*. This makes sense. Take a player, referred to as Player 1, whose *Average Distance after Going for It Shot* is 20 feet and whose *Average Going for It Shot Distance* is 250yds. His *Long Iron Calc* will equal .08. Now take a player, Player 2, who has the same *Average Distance after Going for It Shot* of 20 feet, but whose *Average Going for It Shot Distance* is 300yds. Player 2 can be said to have a better long iron/fairway wood game than Player 1 because even though he hits his going for it shots from much farther away, he hits the ball just as close to the pin. A smaller *Long Iron Calc* score is better, because it means the player has a smaller distance to the pin after his "going for it" shot. In this case, Player 2 is rewarded with a *Long Iron Calc* score of .0667. Similarly, if Player 1 and Player 2 were to have the same *Average Going for It Shot Distance*, but Player 2 hit the ball closer to the pin on average

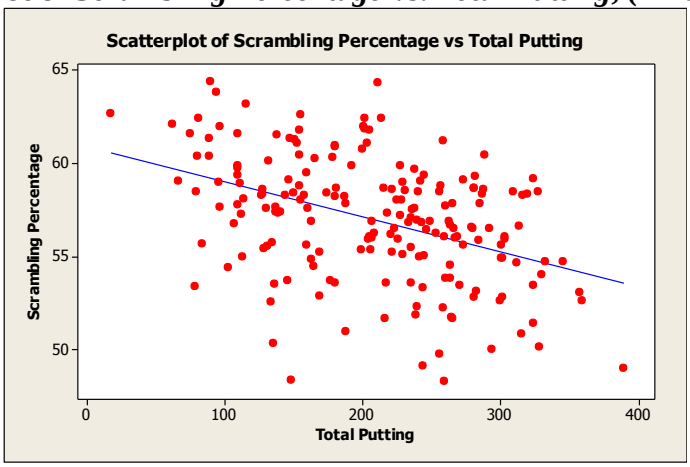
than Player 1, Player 2 would be rewarded with a lower *Long Iron Calc* score. Examining the correlation coefficient of *Long Iron Calc* vs. *Driving Distance-All Drives* also shows how these variables are now less correlated than when *Average Distance after Going for It Shot* was not adjusted.

> Pearson correlation of Driving Distance-All Drives Avg and Long Iron Calc = -0.463
> For these reasons, *Long Iron Calc* will be included in the model.

Scrambling Average Distance to the Hole (Inches)

The term "scrambling" refers to anytime that a player misses the surface of the green with his regulation approach shot. This regulation approach shot is the first shot on a par 3, the second shot on a par 4, and the third shot on a par 5. After this player misses the green with this regulation shot, he is said to be "scrambling." A very popular statistic referenced on the PGA Tour is *Scrambling Percentage*, which is the percentage of the time that the player makes par or better after missing the regulation approach shot. Effectively, this is the percentage of time that the player can chip the ball onto the putting surface and make the putt, or more rarely, chip the ball into the hole for birdie. This statistic is often referenced to denote the player's short game and chipping proficiency, but in reality is heavily dependent on the player's putting ability, which Figure 1-3 scatterplot and correlation demonstrates.

Figure 1-3: Scatterplot of *Scrambling Percentage* vs. *Total Putting*, (n=191)



Pearson correlation of Total Putting and Scrambling Percentage = -0.425
P-Value = 0.000
(1) Scrambling Percentage and Total Putting are continuous variables.
(2) The scatterplot demonstrates a linear relationship.
(3) While there are some outliers, none appear to be extreme.
(4) Both variables are approximately normally distributed.
The 4 p-value calculation assumptions are met.

The scatterplot and correlation coefficient of -.425 shows how strongly correlated the *Scrambling Percentage* statistic is to putting. As was mentioned before, the main goal of this study is to isolate the different areas of the game to determine the relative importance of each area. This *Scrambling Percentage* statistic encompasses two aspects of the game, how close the player hits the ball to the pin when scrambling, and putting ability.

To avoid this combination of skills, the *Scrambling Average Distance to the Hole* variable will be used. This variable is defined as "The average distance the ball comes to rest from the hole (in feet) after the birdie stroke when the player misses the green in regulation[18]." This statistic effectively removes the putting aspect of the popular *Scrambling Percentage* statistic so that it only measures the player's ability to hit the ball close with his scrambling shot. The resulting p-value and correlation coefficient shows how *Scrambling Average Distance to the Hole* is not significantly correlated with putting.

Pearson correlation of Total Putting and Scrambling Average Distance to = 0.135
P-Value = 0.062

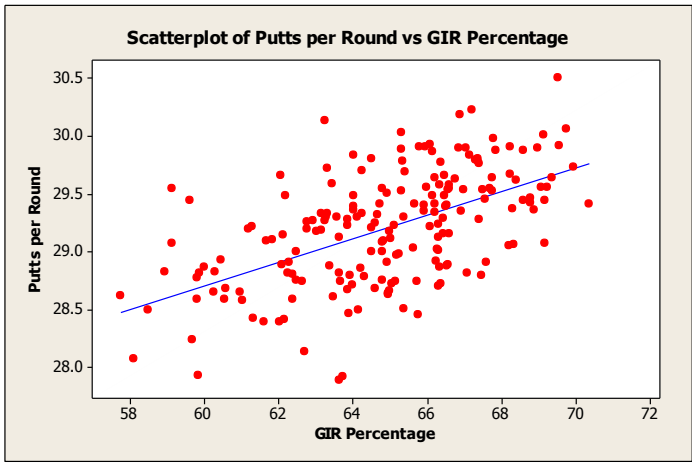
One could argue, however, that the *Scrambling Average Distance to the Hole* statistic is impacted by how close to the pin the player's scrambling shot originates. This is a valid point, but the correlation coefficient of .200 between *Scrambling Average Distance to the Hole* and *Proximity to the Hole* is not severe enough to warrant the transformation of this variable. The difference between the smallest and largest average *Proximity to Hole* is a mere 7' 10", or slightly less than 3 yards. The average *Proximity to the Hole* is just under 35'. A difference of less than 3 yards between the best and worst player in the *Proximity to the Hole* is small and likely has no measurable effect on how close to the pin a player hit his scrambling shot[19]. For this reason, *Scrambling Average Distance to the Hole* is determined to be the best statistic for indicating scrambling ability and will be used in the model.

Strokes Gained-Putting

Putting statistics have developed significantly over time. Advances in math and methodology have led to new statistics, particularly in the putting area, that are much more indicative of a player's putting ability. Popular putting statistics used throughout the Literature Review in studies done multiple years ago were *Putts per Round* and *Putting Average*. The problem with these statistics is that they fail to take into account from how far away the putt originates. Players who miss the green frequently and chip close to the pin are more likely to one putt than a player who hits the green in regulation but is 25 feet

away. This relationship can be demonstrated with a simple correlation between *Putts per Round* and *Green in Regulation Percentage*.

Figure 1-4: Scatterplot of *Putts per Round* vs. *Green in Regulation Percentage*, (n=191)



Pearson correlation of GIR Percentage and Putts per Round = 0.563
P-Value = 0.000
(1) Putts per Round and Green in Regulation Percentage are continuous variables.
(2) The scatterplot demonstrates a linear relationship.
(3) While there are some outliers, none appear to be extreme.
(4) Both variables are approximately normally distributed.
The 4 p-value calculation assumptions are met.

Clearly, there is a high, positive correlation between these two variables. This means that the more frequently the player hits the green in regulation over the course of a round, the more likely he is to have a greater number of putts throughout that round. A statistic is needed that measures a player's ability to make a putt from a given distance, and the *Strokes Gained-Putting* statistic does just that. The *Strokes Gained-Putting* statistic is a mathematically complex statistic defined as:

> The number of putts a player takes from a specific distance is measured against a statistical baseline to determine the player's strokes gained or lost on a hole. The sum of the values for all holes played in a round minus the field average strokes gained/lost for the round is the player's Strokes gained/lost for that round. The sum of strokes gained for each round are divided by total rounds played[20].

The formula below captures how this statistic is calculated:

$$\textit{Strokes Gained/Round} = \frac{\Sigma\,(\text{Expected Number of Putts} - \text{Putts Actually Taken on Hole})}{18\ \text{Holes per Round}}$$

Take one hole, for example. A player's ball is at rest 6 feet from the hole. The tour average percentage of putts made from 6 feet is 65%, with the expected number of putts from that distance to be 1.357 putts[21]. If a player makes the putt, a 1 is subtracted from the expected number of putts of 1.357. The result is +0.357 strokes gained for this putt. If he were to miss the putt and make the second, the result would be 1.357 – 2 = -0.643 strokes gained. These strokes gained are totaled over the course of the round and then averaged over the course of the year. This statistic perfectly accounts for putting distance when evaluating a player's putting ability. A player who misses a very easy putt from two feet is heavily penalized by losing 0.991 strokes (because the expected number of putts from 2 feet is 1.009)[21]. Conversely, a player who makes an outstanding putt from 46 feet gains 1.102 strokes. Because of this statistic's ability to measure a player's putting ability from any distance, it is the best indicator of total putting ability and will be used in the model.

Summary:

Table 1-1: Summary of Variables in Theory Based Model

Area of the Game	Variable Used	Variable Type, Units	Range
Driving Distance	*Driving Distance-All Drives*	Continuous, Yards	(265.8, 309.2) Better = Larger
Driving Accuracy	*Driving Accuracy Percentage*	Continuous, Percentage	(47.27, 73) Better = Larger
Short Iron Play	*Fairway Proximity*	Continuous, Inches	(337, 427) Better = Smaller
Long Iron Play	*Long Iron Calc*	Continuous, None	(.07198, .13247) Better = Smaller
Scrambling Ability	*Scrambling Average Distance to the Pin*	Continuous, Inches	(80, 149) Better = Smaller
Putting	*Strokes Gained-Putting*	Continuous, Strokes	(-1.177, .86) Better = Larger
Performance Indicator*	*Weighted Scoring Average*	Continuous, Strokes	(68.873, 72.995) Better = Smaller

**To be discussed in Section IV: Method*

Automatic Search Procedures for Model Selection

While the above selection of variables have been selected using logic and reasoning, it is also important to consider statistical techniques that can be used to determine, from a given set of predictors, the combination of variables that yields the best model. The best model, in this case, would mean the regression model whose independent variables account for the most variation in the dependent variable, *Weighted Scoring Average*. Based off of the prevalence of certain variables throughout the literature review, the following twelve variables will be considered the set of predictors from which the statistical model selection techniques will draw on to determine the best model:

Variables Considered
Driving Distance
Driving Distance-All Drives
Driving Accuracy
Distance from Edge of Fairway
Proximity to the Hole
Fairway Proximity
Average Distance After Going for It Shot
Long Iron Calc
Scrambling Average Distance to the Hole
Strokes Gained
Total Putting
Overall Putting Average

The number of possible models which could be generated from the above twelve variables is 2^{p-1} or 2^{11} = 2,048 models[21]. Thankfully, there are a variety of computer search procedures which simplify the task of generating and assessing the fit of these models. The two most common approaches are the *stepwise* and *best subsets* regression procedures. In order to assess which regression technique is best, we will use the model with the greatest adjusted R^2 value. R^2 is a measure of how well the model fits the data, expressed as a percentage. The adjusted R^2 takes the R^2 metric and adjusts it depending on how many variables the model contains. It is calculated as:

$$R^2_{adj} = R^2 - (1 - R^2)p/(n-p-1)$$

Where:

p is the total number of regressors
n is the sample size

As can be seen from the formula, adjusted R^2 will always be less than R^2. While R^2 will not decrease as more explanatory variables are added to the model, adjusted R^2 will decrease if the added regressors do not significantly add to the variability explained by the model. A model which included hundreds of variables may fit the data near perfectly

and thus have a large R^2, but the adjusted R^2 would be much lower because of the number of variables included in the model[24].

Data Splitting

Before running the various statistical model selection procedures, it is important to note the use of a very valuable method for validating regression models, called data splitting. The preferred method to validate a regression model is through the collection of new data, but in this study, this is not feasible due to the constraints of time. However, because the number of observations in this dataset is large (n=191), the data set can be split into two separate data sets. The first set is called the *model-building* or *training sample*, and is used to develop the model. The second data set is called the *validation* or *prediction set*, and is used to evaluate the predictive ability of the model generated by the *training sample*. In this way, the validation data set is used in the same way as when new data are collected. The regression coefficients can be reestimated for the selected model and then compared for consistency with the coefficients obtained from the training data set[22]. As is customary, data splitting will only be used to validate the variables which are to be included in the model, while the final regression model will be generated from the entire data set. The observations in this data set will be randomly split into a training data set of 99 observations, and a validation data set of 92 observations. The final model will include all 191 observations.

Indicators of Model Fit: PRESS and MSE

Before examining the output from the stepwise and best subsets, it is important to note two key indicators of model fit, the *PRESS* criterion and mean squared error, or MSE. These are measures of how well the use of fitted values for a subset model can predict the observed responses in the dependent variable. The *PRESS* (prediction sum of squares) criterion is obtained by deleting the *i*th case from the data set, estimating the regression function for the subset model from the remaining $n - 1$ cases, and then using the fitted regression function to obtain the predicted value $\hat{Y}_{i(i)}$ for the *i*th case. $\hat{Y}_{i(i)}$ is the fitted value, by indication of the first subscript *i*, that it was a predicted value for the *i*th case and, by the second subscript *(i)*, that the *i*th case was omitted when the regression function was fitted.

The *PRESS* prediction error for the *i*th case is:

$$Y_i - \hat{Y}_{i(i)}$$

and the *PRESS* criterion is the sum of the squared prediction errors over all *n* cases:

$$PRESS = \sum_{i=0}^{n} (Y_i - \hat{Y}_{i(i)})^2$$

Models with small *PRESS* values are considered good candidate models because when the prediction errors $Y_i - \hat{Y}_{i(i)}$ are small, so are the squared prediction errors and the sum of the squared prediction errors[22].

Similarly, models with a smaller MSE fit the data better than those with larger MSE values. The MSE, or Mean Squared Error, is calculated by summing all of the squared error terms and dividing by the degrees of freedom:

$$MSE = \frac{\sum_{i=0}^{n} (Yi - \hat{Y}i)2}{n-1}$$

where:

Y_i is the value of the response variable in the *i*th case
$\hat{Y}_i$ is the predicted value of the response variable
$n - 1$ is the degrees of freedom, where n is the number of observations in the data set

Therefore, the closer the predicted values of the model are to the observed values in the data set, the smaller the sum of squared errors will be. This means that a model which fits the data better than another model will have smaller prediction errors and, thus, a smaller MSE[22].

Stepwise

Figure 1-5 shows the SPSS output for the stepwise regression technique, using the training data set.

Figure 1-5: SPSS Output, Stepwise Technique, Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.497[a]	.247	.239	.64146
2	.674[b]	.455	.443	.54871
3	.740[c]	.548	.534	.50226
4	.790[d]	.624	.608	.46070
5	.847[e]	.717	.701	.40194
6	.861[f]	.742	.725	.38594
7	.868[g]	.754	.735	.37894

As can be seen in the output, the 7th model has the highest adjusted R^2 value, .735. The variables selected in this 7th model are *Driving Distance-All Drives*, *Driving Accuracy %*, *Fairway Proximity*, *Average Distance After Going for it Shot*, *Scrambling Average Distance to the Hole*, *Strokes Gained-Putting,* and *Overall Putting Average*. Table 1-2 contains the statistics corresponding to each variable, as well as important measures of model fit, for both the training and validation data sets.

Table 1-2: Consolidated Stepwise Model Statistics Output

	Stepwise Training	P-Value	Stepwise Validation	P-Value
B_0	83.672	.000	81.952	.000
B_1: *Driving Distance-All Drives*	-.045571	.000	-.044527	.000
B_2: *Driving Accuracy %*	-.06227	.000	-0.05820	.000
B_3: *Fairway Proximity*	.00804	.000	0.010239	.001
B_4: *Average Distance After Going for It Shot*	.06550	.003	.06116	.002
B_5: *Scrambling Average Distance to Hole*	.02240	.007	.019294	.000
B_6: *Strokes Gained-Putting*	-.9321	.000	-0.8476	.000
B_7: *Overall Putting Average*	-1.9138	.038	-1.460	.224*
R^2 (adj)	73.5%		65.7%	
PRESS	16.1233		15.6162	
MSE	.1436		.1597	

**Note: Overall Putting Average is insignificant in the validation model.*

Best Subsets

Figure 1-6 shows the Minitab output for the best subsets regression technique, using the training data set.

Figure 1-6: Minitab Output, Best Subsets Technique, Model Summary

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	1	2	3	4	5	6	7	8	9	10	11	12
1	24.7	23.9	185.2	0.64146							X					
1	22.2	21.4	194.6	0.65218									X			
2	45.5	44.3	109.9	0.54871							X		X			
2	42.1	40.9	122.6	0.56563								X	X			
3	57.8	56.5	65.9	0.48510	X		X									X
3	57.1	55.8	68.5	0.48906		X	X									X
4	68.0	66.7	29.9	0.42457	X		X						X			X
4	66.4	65.0	36.0	0.43524		X	X						X			X
5	72.7	71.2	14.5	0.39438	X		X			X			X			X
5	72.1	70.6	16.9	0.39893		X	X			X			X			X
6	74.2	72.5	11.0	0.38557	X		X			X			X	X		X
6	74.2	72.5	11.2	0.38594		X	X			X	X		X			X
7	75.6	73.7	7.9	0.37714	X		X			X	X		X	X		X
7	75.4	73.5	8.7	0.37894		X	X			X	X		X	X		X
8	76.1	74.0	8.0	0.37532	X		X			X	X		X	X	X	X
8	75.9	73.8	8.6	0.37656	X	X	X			X	X		X	X		X
9	76.4	74.0	8.8	0.37484	X	X	X			X	X		X	X	X	X
9	76.3	73.9	9.1	0.37543		X	X			X	X	X	X	X	X	X
10	76.7	74.1	9.7	0.37460	X	X	X			X	X	X	X	X	X	X
10	76.6	73.9	10.2	0.37566	X	X	X	X		X	X		X	X	X	X
11	76.8	73.9	11.2	0.37575	X	X	X	X		X	X	X	X	X	X	X
11	76.8	73.8	11.5	0.37628	X	X	X		X	X	X	X	X	X	X	X
12	76.9	73.7	13.0	0.37743	X	X	X	X	X	X	X	X	X	X	X	X

Note: Because Adj R^2 are so close in the models, the model with the lowest Mallows C_p statistic will be used.

As can be seen in the output, the models containing more than seven variables have adjusted R^2 values that are very similar. Because of this, the model with the lowest Mallows C_p statistic will be used. This criterion is concerned with the *total mean squared error* of the *n* fitted values for each subset regression model. Models with a smaller C_p value are said to fit the data better. In this case, the model with the lowest Mallows C_p value contains the variables *Driving Distance, Driving Accuracy %, Fairway Proximity, Average Distance After Going for it Shot, Scrambling Average Distance to the Hole, Strokes Gained-Putting,* and *Overall Putting Average.* Table 1-3 shows the statistics corresponding to each variable, as well as important measures of model fit, for both the training and validation data sets[22].

Table 1-3: Consolidated Best Subsets Statistics Output

	Best Subsets Training	**P-Value**	**Best Subsets Validation**	**P-Value**
B_0	83.489	.000	81.404	.000
B_1: *Driving Distance*	-0.040819	.000	-0.041031	.000
B_2: *Driving Accuracy %*	-0.06709	.000	-0.06866	.000
B_3: *Fairway Proximity*	0.007846	.004	0.009778	.002
B_4: *Average Distance After Going for It Shot*	0.05524	.025	0.06034	.002
B_5: *Scrambling Average Distance to Hole*	0.023180	.000	0.020332	.000
B_6: *Strokes Gained-Putting*	-0.9545	.000	-0.8833	.000
B_7: *Overall Putting Average*	-2.1126	.020	-1.104	.347*
R^2 (adj)	73.7%		67.7%	
PRESS	15.9078		14.8026	
MSE	.1422		.1508	

**Note: Overall Putting Average is insignificant in the validation model.*

Theory Based Model

Table 1-4 contains the statistics corresponding to each variable chosen for the theory based model, as well as important measures of model fit, for both the training and validation data sets.

Table 1-4: Consolidated Theory Based Model Statistics Output

	Theory Based Training	**P-Value**	**Theory Based Validation**	**P-Value**
B_0	82.974	.000	80.535	.000
B_1: *Driving Distance-All Drives*	-0.053025	.000	-0.048619	.000
B_2: *Driving Accuracy %*	-0.06979	.000	-0.06151	.000
B_3: *Fairway Proximity*	0.008965	.001	0.011063	.001
B_4: *Long Iron Calc*	14.441	.029	15.411	.004
B_5: *Scrambling Average Distance to Hole*	0.023305	.000	0.020715	.000
B_6: *Strokes Gained-Putting*	-0.8716	.000	-0.7817	.000
R^2 (adj)	71.8%		64.8%	
PRESS	16.8082		15.8785	
MSE	.1527		.1639	

Final Model Selection

The stepwise, best subsets, and theory based models' variables are in Table 1-5:

Table 1-5: List of Variables Chosen by Model

Stepwise	**Best Subsets**	**Theory Based**
Driving Distance-All Drives	Driving Distance	Driving Distance-All Drives
Driving Accuracy %	Driving Accuracy %	Driving Accuracy %
Fairway Proximity	Fairway Proximity	Fairway Proximity
Average Distance After Going for it Shot	Average Distance After Going for it Shot	Long Iron Calc
Scrambling Average Distance to Hole	Scrambling Average Distance to Hole	Scrambling Average Distance to Hole
Strokes Gained-Putting	Strokes Gained-Putting	Strokes Gained-Putting
Overall Putting Average	Overall Putting Average	

In all of the models, the training model has a higher adjusted R^2 than the validation model, with the best subsets having the largest adjusted R^2 value of 73.7%. While all models are very close in these values of fit, the model has an adjusted R^2 which is smaller by less than 2%. The PRESS and MSE values are all also very close, although the theory based model's PRESS value is slightly higher. In the training models, all variables are also significant, although to varying degrees. All three models are very similar in terms of the variables selected, with all having at least five variables in common.

Ultimately, the theory based model will be chosen for several reasons. The primary reason is that in both the training and the validation data sets, all variables in the theory based model are significant. While all variables in the training models are significant for the stepwise and best subsets models, *Overall Putting Average* is insignificant ($p > .05$) in the validation models of each procedure. The p-value of *Overall Putting Average* in the stepwise validation model is .224 and is .347 in the best subsets validation model. This inconsistency in variable significance is not found in the theory based model, meaning the theory based model has greater validity. Another advantage of the theory based model is that it contains one fewer variable, six, than the stepwise and best subset models, which contain seven. This means that the theory based model is more parsimonious than the other two models. In regression, whenever an option is presented between two models which have comparable measures of fit, the simpler model is often considered better. Because of the consistency in variable significance and its parsimonious nature, the theory based model will be used in this regression analysis.

Chapter 4

Method

Response Variable: Weighted Scoring Average

The response variable for this model will be *Weighted Scoring Average*. This variable is defined as:

> The weighted scoring average takes the stroke average of the field into account. It is computed by adding a player's total strokes to an adjustment and dividing by the total rounds played. The adjustment is computed by determining the stroke average of the field for each round played. This average is subtracted from par to create an adjustment for each round. A player accumulates these adjustments for each round played[23].

Because players play in different tournaments and at different courses throughout the golf season, their scores are subject to the conditions of that day and the difficulty of the courses. Player 1 may play in a tournament at an extremely difficult course on a windy and rainy day leading him to shoot a higher score. Player 2, on the other hand, may play an easy course with wide open fairways on a warm and sunny day. Player 2 will probably have a lower scoring average over the course of the tournament, which is why using an unadjusted *Scoring Average* statistic would be inappropriate to use to compare the golfers. *Weighted Scoring Average,* on the other hand, takes the stroke average of the field into account for each round, and adjusts the player's *Scoring Average* accordingly. Referring to the prior example: Player 1 may shoot 75 in the brutal conditions, but the rest of the field may shoot over 80, in which case his score would be very good. Player 2 may shoot a 69, but if the rest of the field takes advantage of the conditions and shoots a 65, his score really is not all that impressive. *Weighted Scoring Average* accounts for this

inconsistency in course difficulty and scoring conditions and is the reason why it is the response variable in the model.

One could argue that earnings would be a more important indicator of a player's success than scoring average; being able to shoot low scores does not matter if a player cannot win the tournament and succeed under pressure. This is a valid argument; however, the methodology a player would use to go about increasing his earnings would be to decrease his scores. At the end of the day, even if the player's objective is to earn more money and win more events, he goes about doing so by working on his game in order to lower his scores which are reflected in his subsequent scoring average.

Initial Regression Model

Using the Minitab Software, the selected variables *Driving Distance-All Drives, Driving Accuracy, Fairway Proximity, Long Iron Calc, Scrambling Average Distance to the Hole,* and *Strokes Gained-Putting* will be entered into the regression model with *Weighted Scoring Average* as the response variable. The model will look as follows:

$$\textit{Weighted Scoring Average} = \beta_0 + \beta_1 \textit{Driving Distance-All Drives} + \beta_2 \textit{Driving Accuracy} + \beta_3 \textit{Fairway Proximity} + \beta_4 \textit{Long Iron Calc} + \beta_5 \textit{Scrambling Average Distance to the Hole} + \beta_6 \textit{Strokes Gained-Putting}$$

Chapter 5

Results

The Minitab output:

```
The regression equation is
Weighted Scoring Average = 81.9 - 0.0509 Driving Distance-All Drives Avg
                           - 0.0654 Driving Accuracy %
                           + 0.00986 Fairway Proximity (in.)
                           + 14.6 Long Iron Calc
                           + 0.0220 Scrambling Avg Distance to Hole
                           - 0.853 Strokes Gained - Putting


Predictor                            Coef   SE Coef      T      P    VIF
Constant                           81.866     2.259  36.24  0.000
Driving Distance-All Drives Avg  -0.050929  0.005527  -9.21  0.000  1.761
Driving Accuracy %               -0.065434  0.007357  -8.89  0.000  1.498
Fairway Proximity (in.)           0.009858  0.001950   5.05  0.000  1.254
Long Iron Calc                      14.551     3.906   3.73  0.000  1.411
Scrambling Avg Distance to Hole   0.022035  0.002961   7.44  0.000  1.097
Strokes Gained - Putting          -0.85339   0.08962  -9.52  0.000  1.132


S = 0.392634   R-Sq = 70.3%   R-Sq(adj) = 69.3%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        6  67.028  11.171  72.46  0.000
Residual Error  184  28.366   0.154
Total           190  95.394


Source                           DF  Seq SS
Driving Distance-All Drives Avg   1   9.825
Driving Accuracy %                1  24.005
Fairway Proximity (in.)           1   2.616
Long Iron Calc                    1   2.707
Scrambling Avg Distance to Hole   1  13.895
Strokes Gained - Putting          1  13.979
```

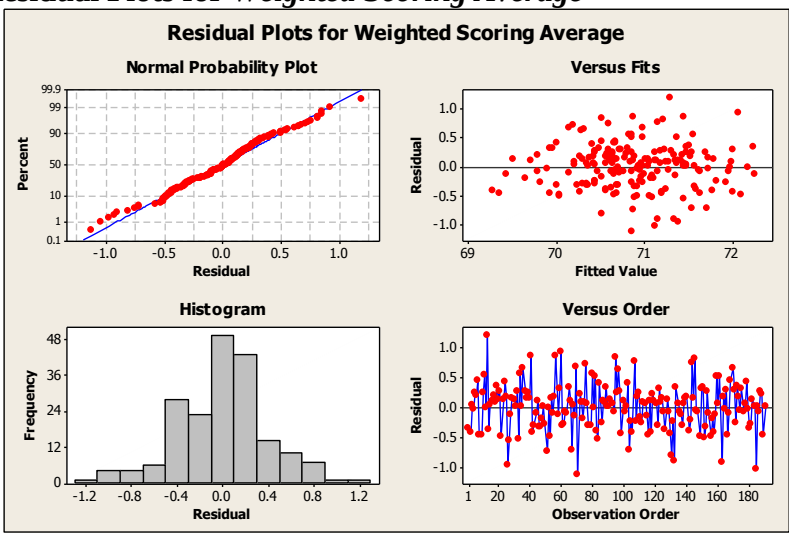
By observing the p-values for all of the explanatory variables, it is evident that all of these aspects of the game have a significant impact on *Weighted Scoring Average*. The R^2, or coefficient of determination, is a determinant of how well the regression model fits the data. This value of 70.3% means that 70.3% of the variation observed in the response

variable, *Weighted Scoring Average*, is explained by the regressors, or explanatory variables, included in the model[24].

Multicollinearity is an undesirable situation where the correlations among the independent variables in the model are strong. If a multiple regression model exhibits multicollinearity, it is difficult to determine the estimate of one variable's impact on the dependent variable while controlling for the others. The Variance Inflation Factors are also low, which indicates there is little multicollinearity among the variables included in the model[25]. Many regard VIFs greater than 10 and VIFs greater than 2.5 in weaker models as indicating multicollinearity. But, because the VIFs for this regression model are all close to 1, which is the VIFs lower bound, it can be determined there is little multicollinearity in this model[26].

Residual Analysis

Figure 2-1: Residual Plots for *Weighted Scoring Average*



The four assumptions of residuals that need to be satisfied for linear regression are normality, linearity, independence, and equal variance. Because time is not a factor in this study, independence of residuals will be assumed. The normal probability plot in the

top left corner above indicates that the residuals exhibit normality. The versus fits graph in the top right indicates that the residuals exhibit equal variance and linearity. All residual assumptions are met.

Problems with Interpretation

While this model accounts for much of the variation in *Weighted Scoring Average*, its interpretations are limited. Due to the varying nature of the units in which each of these explanatory variables are measured (e.g. *Driving Accuracy*=%, *Driving Distance-All Drives* = yards, *Fairway Proximity* = inches), one cannot easily determine or say that the variable with the largest corresponding β has the largest impact on *Weighted Scoring Average*. For example, had *Driving Distance-All Drives* been measured in feet instead of yards, its β would have been three times as large, but surely the change in units does not change the importance of the variable. Even if the p-values were not all 0.000 and measured on the same scale, a small coefficient that can be estimated precisely will have a small p-value, while a large coefficient that cannot be estimated precisely will have a large p-value[27]. In an attempt to solve these issues, a standardized regression model will be used.

The Standardized Regression Model

Standardizing a regression model includes centering and scaling all variables in the model. *Centering* involves taking the difference between each observation and the mean of all the observations for the variable. *Scaling* involves expressing the centered observations in units of the standard deviation of the observations for the variable. Thus,

the standardizations for response variable Y (*Weighted Scoring Average*) and the predictor variables X_1, X_2, … X_6 are as follows[27]:

$$\frac{Y_i - \bar{Y}}{S_y}$$

$$\frac{X_i - \bar{X}}{S_X}$$

After standardizing these variables, the variables will all have a mean of 0 and standard deviation of 1. The Standardized Regression Model will now include all of the variables. Because these variables all have the same mean and standard deviation, observing their coefficients will show which variables have the largest impact on scoring average[28].

The Minitab output of the Standardized Regression Model is below:

```
      The regression equation is
      Weight Scor Avg Std. = 0.0007 - 0.491 DD All Drives Std. - 0.437 Drive
Acc Std.
                           + 0.227 Fairway Prox. Std. + 0.178 Long Iron Calc
Std.
                           + 0.313 Scrambling Avg Dist Std.
                           - 0.407 Strokes Gained Std.

      Predictor                    Coef  SE Coef      T      P    VIF
      Constant                  0.00070  0.04007   0.02  0.986
      DD All Drives Std.       -0.49133  0.05332  -9.21  0.000  1.761
      Drive Acc Std.           -0.43737  0.04918  -8.89  0.000  1.498
      Fairway Prox. Std.        0.22748  0.04500   5.05  0.000  1.254
      Long Iron Calc Std.       0.17778  0.04771   3.73  0.000  1.411
      Scrambling Avg Dist Std.  0.31296  0.04206   7.44  0.000  1.097
      Strokes Gained Std.      -0.40707  0.04275  -9.52  0.000  1.132

      S = 0.553786   R-Sq = 70.3%   R-Sq(adj) = 69.3%

      Analysis of Variance

      Source           DF       SS      MS      F      P
      Regression        6  133.341  22.223  72.46  0.000
      Residual Error  184   56.429   0.307
      Total           190  189.770

      Source                   DF  Seq SS
      DD All Drives Std.        1  19.546
      Drive Acc Std.            1  47.754
      Fairway Prox. Std.        1   5.204
      Long Iron Calc Std.       1   5.385
```

```
Scrambling Avg Dist Std.    1  27.643
Strokes Gained Std.         1  27.809
```

Weighted Scoring Average Standardized =.0007 -.491*Driving Distance Standardized* - .437*Driving Accuracy Standardized* + .227 *Fairway Proximity Standardized* + .178*Long Iron Calc Standardized* + .313*Scrambling Avg Distance to Hole Standardized* - .407*Strokes Gained Standardized*

The variables with the largest coefficients can be said to explain the largest amount of variability in *Weighted Scoring Average* and thus, can be considered the most important[28]. It is important to note that the magnitude, not direction, of the coefficient is all that matters. For example, *Driving Distance-All Drives* has a negative coefficient because driving the ball farther usually results in a lower scoring average. Similarly, *Fairway Proximity* has a positive coefficient because hitting the ball greater distances from the pin usually results in higher scores. The standard error of the coefficient should also be considered. These values reflect the variability of the corresponding regression coefficients and give insight into how likely it is that the observed ranking is simply due to randomness in the sample.

$$Standard\ Error = \frac{Standard\ Deviation}{\sqrt{n}}$$

A summarized ranking of the coefficients is given below:

Ranking	Predictor	Coefficient	SE of Coef.
1	*Driving Distance-All Drives*	-.49133	.05332
2	*Driving Accuracy %*	-.43737	.04918
3	*Strokes Gained-Putting*	-.40707	.04275
4	*Scrambling Avg Dist to Hole*	.31296	.04206
5	*Fairway Proximity*	.22748	.04500
6	*Long Iron Calc.*	.17778	.04771

Interpretations

The interpretations that the above coefficient rankings lead to are below:

Ranking	Area
1	*Driving Distance*
2	*Driving Accuracy*
3	*Putting*
4	*Scrambling*
5	*Short Iron Play*
6	*Long Iron/Fairway Wood Play*

These rankings are based on the magnitude of the betas. While this is useful information, it is also of interest to determine if any of these variables are more or less influential than others. To accomplish this, a significance test of the betas, using a *t-test*, can be conducted. This test has a null hypothesis that the coefficients are equal in magnitude. The *t-test* is calculated by taking the difference of the betas being compared and dividing by the standard error of the difference. The *t-statistics* for several pairwise comparisons are below.

Variables	t-statistic
Driving Distance-All Drives and *Driving Accuracy %*	-1.0646
Driving Accuracy % and *Strokes Gained-Putting*	-0.4873
Strokes Gained-Putting and *Scrambling Avg. Dist. To Hole*	-13.3918
Scrambling Avg. Dist. To Hole and *Fairway Proximity*	1.3234
Fairway Proximity and *Long Iron Calc.*	0.8919

From observing the *t-statistic*, it is evident that only one pair of betas, *Strokes Gained-Putting* and *Scrambling Avg. Dist. To Hole*, is significantly different. It is important to note that, while not all the conducted pairwise comparisons yielded significant differences, all of these coefficients are still significant predictors.

As with any multiple *t-test*, it is important to remember to take the Type 1 error inflation into account. This error occurs because, when multiple tests are being run, there is actually a larger than .05 probability of rejecting the null in at least one case, even if the null hypothesis is true in all cases. There are multiple ways to handle this multiple testing problem; one of the most prominent is the Bonferroni Correction. With a *t-statistic* as large in magnitude as 13.4, this multiple comparison inflation error will not impact the conclusion that *Strokes Gained-Putting* is a significantly larger beta, and therefore is significantly more influential on performance, than is *Scrambling Average Distance to the Hole*. Although only five *t-tests* were run, it is possible to conduct *t-tests* for every pairwise combination of variables. For example, because the coefficient for *Driving Distance-All Drives* is larger than the coefficient for *Strokes Gained-Putting*, it is very likely that a *t-test* would yield a significant difference between the coefficients of *Driving Distance-All Drives* and *Scrambling Avg. Dist. to Hole*. Conducting the rest of these various *t-tests* would add to the validity of the aforementioned rankings.

Chapter 6

Conclusions

The output of the model shows that driving distance has the greatest impact on low scoring average, followed by driving accuracy, putting ability, scrambling ability, short iron play, and long iron or fairway wood play.

Discussion

There are several possible explanations for the rankings found above. The two driving statistics appear to be the most important, and this is probably because these two areas of the game are what position the player for the rest of his shots on a hole. If a player hits the ball very far, he has a shorter shot into the green and is able to use a shorter iron to hit a shorter shot with which he is probably much more accurate. Similarly, a player who hits the ball into the rough frequently will be hitting his approach shots from thick rough which will negatively impact his ability to hit his approach shot close to the pin, resulting in longer putts or more scrambling shots and higher scores. Furthermore, if this player hits the ball so far from the fairway that the ball ends up behind a tree or out of bounds, this will clearly have a profound, negative impact on the player's ability to hit a good second shot and on the player's score. These two variables are what position the player for his shot into the green and is probably why these two driving variables are the most important.

Another explanation that could explain the trends observed above is the frequency with which each area of the game is utilized. There is a relatively large gap in the coefficients of the first three areas (*Driving Distance, Driving Accuracy, Putting*) and (*Scrambling, Short Iron Play, Long Iron/Fairway Wood Play*). On a regulation, 18-hole course, there are usually four par 3 holes, ten par 4 holes, and four par 5 holes. Because driving distance and accuracy are measured on all par 4 and par 5 holes, these two areas of the game are utilized on 14 of the 18 holes. Scrambling ability, on the other hand, is utilized less frequently. As was mentioned before, *Greens in Regulation*, is the percentage of greens that are hit in regulation (1 shot on par 3, 2 shots on par 4, 3 shots on par 5). The mean *Greens in Regulation Percentage* for the 2012 PGA Tour is 64.88%. This means that tour players hit about 65% of the greens and miss about 35% of the greens. The players must use their scrambling ability on these 35% of greens missed, which comes out to 6.3 holes per round. If a player is only using this scrambling ability on a little over 6.3 holes per round, it would make sense that driving ability would be more important because it is used on 14 of the 18 holes. The last of the top 3 areas is putting. The average number of *Putts per Round* is 29.2 putts. The mean *Scoring Average (Actual)*, which is unadjusted and is the actual number of strokes per round, is 71.07 strokes, meaning putting strokes account for a whopping 41.8% of all strokes taken during a round. Proficiency with a certain stroke is undoubtedly important if it is used on such a large percentage of shots taken throughout a round. Long irons and fairway woods, in contrast, may only be used on the four Par 5s when the player attempts to go for the green in two shots. Driving and putting ability appear to have a larger impact on

scoring average than do scrambling and iron play, and the frequency with which these areas of the game are utilized may be the reason why this is the case.

Limitations

One important limitation of the conclusions that can be drawn from linear regression is that correlation does not necessarily imply causation. This means that simply because someone can drive the ball extremely far, does not mean that him driving the ball far leads to his low scoring average. This player may be so inaccurate with his drives that this power actually hurts his score, and that rather it is his incredible putting ability that leads to his low scores. This means that some caution should be used before concluding that a player's proficiency in a certain area is the reason for a low scoring average.

While these coefficients can be ranked as so above, it should be noted that no test has been done to determine whether these coefficients are significantly different than each other. It is possible that the rankings found may be due to random variation, and some caution should be used before claiming that driving distance is, without a doubt, the most important factor in determining one's score.

Furthermore, it should be reinforced that the data set used to reach these conclusions was from the 2012 PGA Tour Golf season. The 191 golfers in the dataset are among the best in the world, and one cannot simply extrapolate these findings to the average golfer. The courses professionals play are hundreds of yards longer than those amateurs play. The greens are faster, the breaks more severe, the hazards more aplenty and the fairways more narrow. While the above findings indicate that driving distance has

the largest impact on scoring average for professional golfers, it is quite possible that another variable, such as driving accuracy or putting, may be more important for an amateur player to focus on in order to lower their scores. Players on tour hit the ball phenomenally straight as it is, and the difference between a 69 and a 70 may be approaching the green with a shorter club because of a longer drive. For amateur golfers, a long driving distance may do more harm than good, as amateurs miss more fairways and miss them by larger margins, bringing in to play out of bounds areas, woods, and other obstacles that will increase their score. Similarly, the average golfer misses more greens in regulation than professional golfers, meaning they scramble more. For these reasons, in regards to the amateur contingency, it is quite possible for driving accuracy or scrambling ability to have a more significant impact on their scoring average than driving distance.

Appendix A

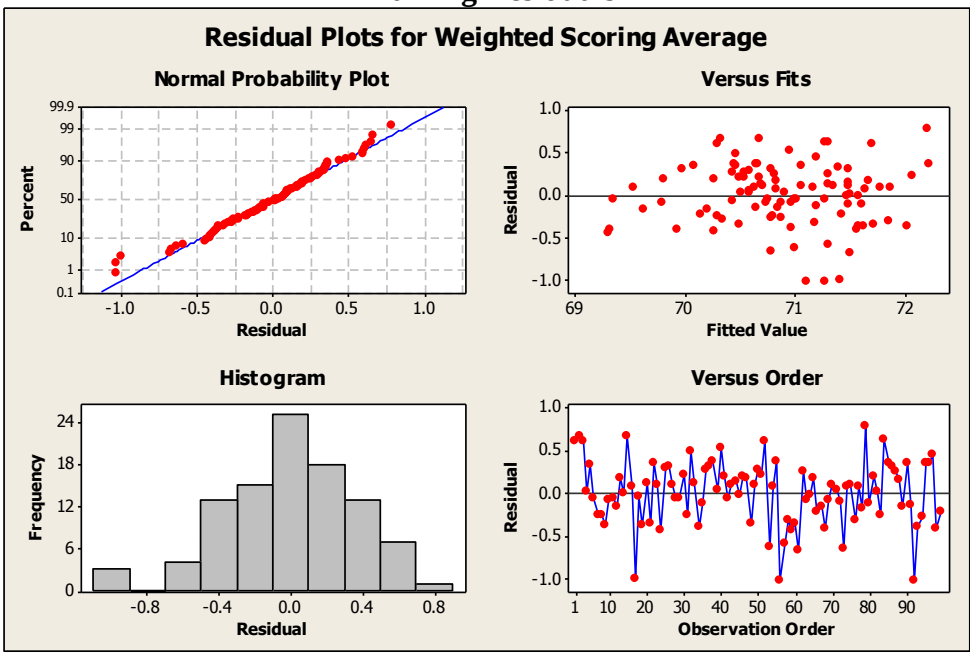
Regression Output

Stepwise Model Minitab Output

Training (n=99)

```
The regression equation is
Weighted Scoring Average = 83.7 - 0.0456 Driving Distance-All Drives Avg
                           - 0.0623 Driving Accuracy %
                           + 0.00804 Fairway Proximity (in.)
                           + 0.0655 Average Distance after Going fo
                           + 0.0224 Scrambling Avg Distance to Hole
                           - 0.932 Strokes Gained - Putting
                           - 1.91 Overall Putting Average (Putts)


Predictor                            Coef   SE Coef      T      P    VIF
Constant                           83.672     3.162  26.46  0.000
Driving Distance-All Drives Avg  -0.045571  0.008325  -5.47  0.000  2.069
Driving Accuracy %                -0.06227   0.01063  -5.86  0.000  1.573
Fairway Proximity (in.)           0.008042  0.002645   3.04  0.003  1.488
Average Distance after Going fo    0.06550   0.02355   2.78  0.007  1.792
Scrambling Avg Distance to Hole   0.022395  0.004195   5.34  0.000  1.141
Strokes Gained - Putting           -0.9321    0.1355  -6.88  0.000  1.434
Overall Putting Average (Putts)    -1.9138    0.9090  -2.11  0.038  1.271


S = 0.378938   R-Sq = 75.4%   R-Sq(adj) = 73.5%

PRESS = 16.1233   R-Sq(pred) = 69.59%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       7  39.9468  5.7067  39.74  0.000
Residual Error  91  13.0670  0.1436
Total           98  53.0138


Source                           DF   Seq SS
Driving Distance-All Drives Avg   1   4.9063
Driving Accuracy %                1  14.7758
Fairway Proximity (in.)           1   0.8488
Average Distance after Going fo   1   3.5945
Scrambling Avg Distance to Hole   1   9.0159
Strokes Gained - Putting          1   6.1689
Overall Putting Average (Putts)   1   0.6366
```

Training Residuals



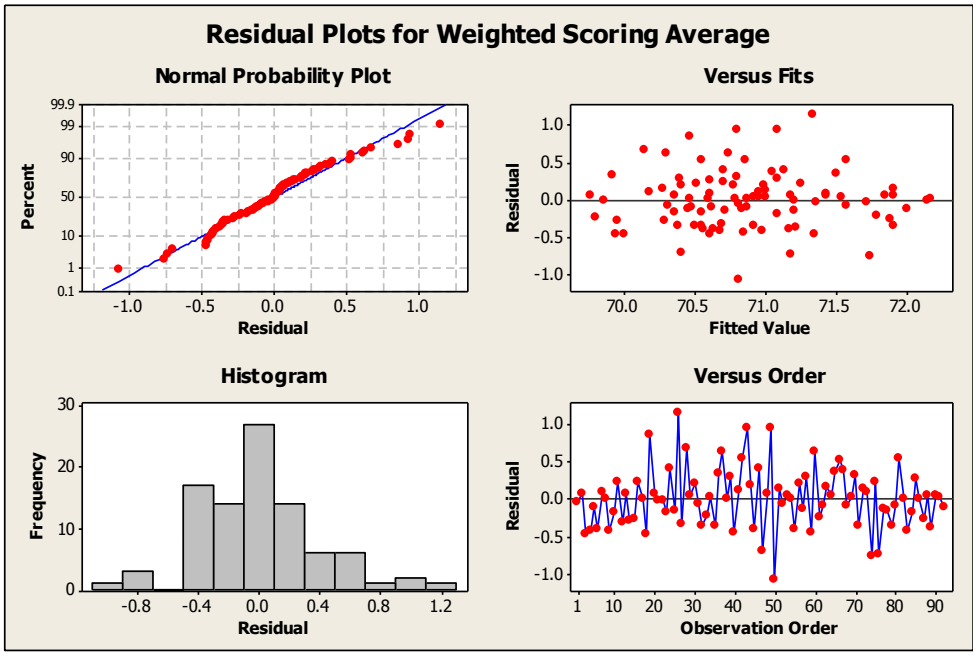
The four assumptions of residuals that need to be satisfied for linear regression are normality, linearity, independence, and equal variance. Because time is not a factor in this study, independence of residuals will be assumed. The normal probability plot in the top left corner above indicates that the residuals exhibit normality. The versus fits graph in the top right indicates that the residuals exhibit equal variance and linearity. All residual assumptions are met.

Validation (n=92)

```
The regression equation is
Weighted Scoring Average = 82.0 - 0.0445 Driving Distance-All Drives Avg
                           - 0.0582 Driving Accuracy %
                           + 0.0102 Fairway Proximity (in.)
                           + 0.0612 Average Distance after Going fo
                           + 0.0193 Scrambling Avg Distance to Hole
                           - 0.848 Strokes Gained - Putting
                           - 1.46 Overall Putting Average (Putts)


Predictor                          Coef    SE Coef      T      P    VIF
Constant                         81.952      3.591  22.82  0.000
Driving Distance-All Drives Avg -0.044527  0.008058  -5.53  0.000  1.870
Driving Accuracy %               -0.05820   0.01116  -5.22  0.000  1.717
Fairway Proximity (in.)          0.010239  0.003110   3.29  0.001  1.229
```

```
Average Distance after Going fo    0.06116   0.01950   3.14  0.002  1.419
Scrambling Avg Distance to Hole   0.019294  0.004253   4.54  0.000  1.129
Strokes Gained - Putting           -0.8476    0.1428  -5.94  0.000  1.300
Overall Putting Average (Putts)     -1.460     1.191  -1.23  0.224  1.451


S = 0.399565   R-Sq = 68.3%   R-Sq(adj) = 65.7%

PRESS = 15.6162   R-Sq(pred) = 63.14%

Analysis of Variance

Source           DF       SS      MS      F      P
Regression        7  28.9530  4.1361  25.91  0.000
Residual Error   84  13.4108  0.1597
Total            91  42.3638


Source                              DF   Seq SS
Driving Distance-All Drives Avg      1   4.9208
Driving Accuracy %                   1  10.1901
Fairway Proximity (in.)              1   1.9667
Average Distance after Going fo      1   1.5797
Scrambling Avg Distance to Hole      1   4.4519
Strokes Gained - Putting             1   5.6042
Overall Putting Average (Putts)      1   0.2396
```

Validation Residuals



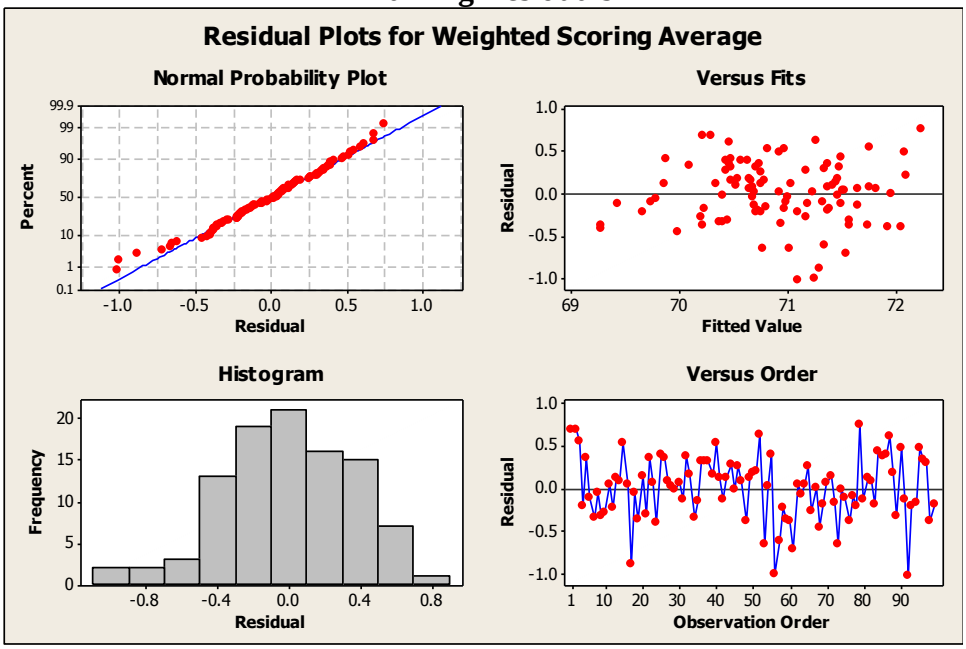
The residuals assumptions of normality, linearity, independence, and equal variance are met.

Best Subsets Minitab Output

Training (n=99)

```
The regression equation is
Weighted Scoring Average = 83.5 - 0.0408 Driving Distance Avg.
                           - 0.0671 Driving Accuracy %
                           + 0.00785 Fairway Proximity (in.)
                           + 0.0552 Average Distance after Going fo
                           + 0.0232 Scrambling Avg Distance to Hole
                           - 2.11 Overall Putting Average (Putts)
                           - 0.954 Strokes Gained - Putting


Predictor                             Coef   SE Coef      T      P    VIF
Constant                            83.489     3.091  27.01  0.000
Driving Distance Avg.            -0.040819  0.007317  -5.58  0.000  2.366
Driving Accuracy %                -0.06709   0.01100  -6.10  0.000  1.701
Fairway Proximity (in.)           0.007846  0.002633   2.98  0.004  1.488
Average Distance after Going fo    0.05524   0.02431   2.27  0.025  1.929
Scrambling Avg Distance to Hole   0.023180  0.004176   5.55  0.000  1.142
Strokes Gained - Putting           -0.9545    0.1352  -7.06  0.000  1.442
Overall Putting Average (Putts)    -2.1126    0.8946  -2.36  0.020  1.243

S = 0.377136   R-Sq = 75.6%   R-Sq(adj) = 73.7%

PRESS = 15.9078   R-Sq(pred) = 69.99%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       7  40.0708  5.7244  40.25  0.000
Residual Error  91  12.9431  0.1422
Total           98  53.0138

Source                            DF   Seq SS
Driving Distance Avg.              1   3.9841
Driving Accuracy %                 1  15.5525
Fairway Proximity (in.)            1   0.6436
Average Distance after Going fo    1   3.2991
Scrambling Avg Distance to Hole    1   9.4740
Overall Putting Average (Putts)    1   0.0322
Strokes Gained - Putting           1   7.0852
```

Training Residuals



The residuals assumptions of normality, linearity, independence, and equal variance are met.

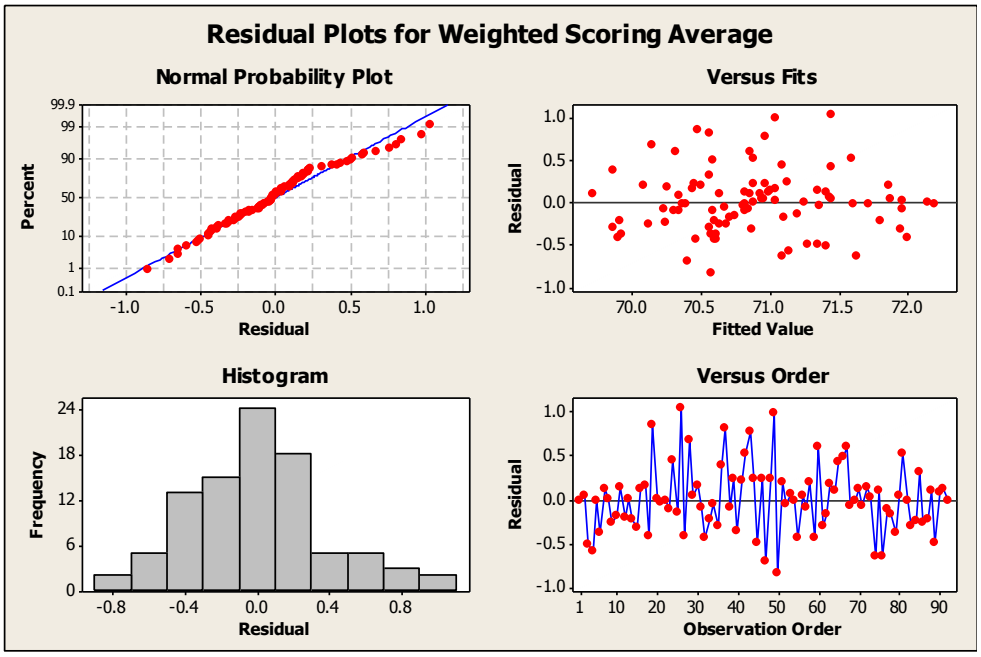
Validation (n=92)

```
The regression equation is
Weighted Scoring Average = 81.4 - 0.0410 Driving Distance Avg.
                           - 0.0687 Driving Accuracy %
                           + 0.00978 Fairway Proximity (in.)
                           + 0.0603 Average Distance after Going fo
                           + 0.0203 Scrambling Avg Distance to Hole
                           - 0.883 Strokes Gained - Putting
                           - 1.10 Overall Putting Average (Putts)

Predictor                            Coef   SE Coef      T      P    VIF
Constant                           81.404     3.308  24.61  0.000
Driving Distance Avg.           -0.041031  0.006721  -6.11  0.000  2.112
Driving Accuracy %               -0.06866   0.01161  -5.91  0.000  1.967
Fairway Proximity (in.)          0.009778  0.003029   3.23  0.002  1.234
Average Distance after Going fo   0.06034   0.01876   3.22  0.002  1.391
Scrambling Avg Distance to Hole  0.020332  0.004145   4.91  0.000  1.136
Strokes Gained - Putting          -0.8833    0.1390  -6.36  0.000  1.304
Overall Putting Average (Putts)    -1.104     1.167  -0.95  0.347  1.474

S = 0.388308   R-Sq = 70.1%   R-Sq(adj) = 67.6%


PRESS = 14.8026   R-Sq(pred) = 65.06%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       7  29.6980  4.2426  28.14  0.000
```

```
Residual Error   84   12.6658   0.1508
Total            91   42.3638
```

Source	DF	Seq SS
Driving Distance Avg.	1	3.6993
Driving Accuracy %	1	11.1796
Fairway Proximity (in.)	1	1.8330
Average Distance after Going fo	1	1.7659
Scrambling Avg Distance to Hole	1	4.7730
Strokes Gained - Putting	1	6.3121
Overall Putting Average (Putts)	1	0.1350

Validation Residuals



The residuals assumptions of normality, linearity, independence, and equal variance are met.

Theory Based Model Minitab Output

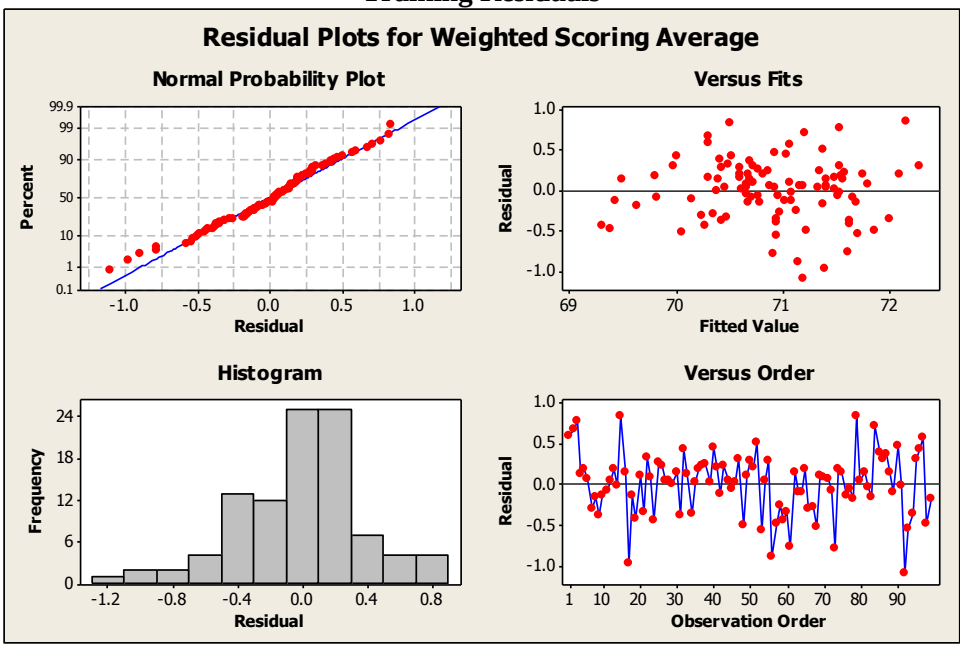
Training: (n=99)

```
The regression equation is
Weighted Scoring Average = 83.0 - 0.0530 Driving Distance-All Drives Avg
                           - 0.0698 Driving Accuracy %
                           + 0.00897 Fairway Proximity (in.)
                           + 14.4 Long Iron Calc
                           + 0.0233 Scrambling Avg Distance to Hole
                           - 0.872 Strokes Gained - Putting
```

Predictor	Coef	SE Coef	T	P	VIF
Constant	82.974	3.208	25.87	0.000	
Driving Distance-All Drives Avg	-0.053025	0.008108	-6.54	0.000	1.846

```
Driving Accuracy %                -0.06979   0.01044  -6.68  0.000  1.428
Fairway Proximity (in.)           0.008965  0.002655   3.38  0.001  1.410
Long Iron Calc                      14.441     6.520   2.21  0.029  1.640
Scrambling Avg Distance to Hole   0.023305  0.004317   5.40  0.000  1.136
Strokes Gained - Putting           -0.8716    0.1344  -6.48  0.000  1.327


S = 0.390804   R-Sq = 73.5%   R-Sq(adj) = 71.8%

PRESS = 16.8082   R-Sq(pred) = 68.29%

Analysis of Variance

Source           DF       SS       MS      F      P
Regression        6  38.9629   6.4938  42.52  0.000
Residual Error   92  14.0510   0.1527
Total            98  53.0138


Source                            DF   Seq SS
Driving Distance-All Drives Avg    1   4.9063
Driving Accuracy %                 1  14.7758
Fairway Proximity (in.)            1   0.8488
Long Iron Calc                     1   2.6355
Scrambling Avg Distance to Hole    1   9.3737
Strokes Gained - Putting           1   6.4227
```

Training Residuals



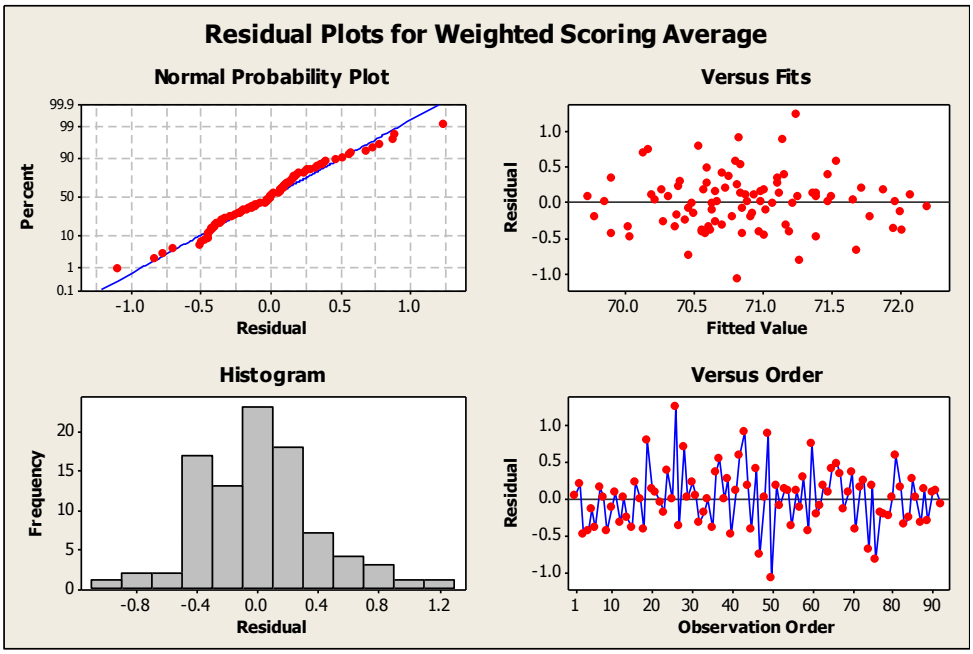
The residuals assumptions of normality, linearity, independence, and equal variance are met.

Validation: (n=92)

```
The regression equation is
Weighted Scoring Average = 80.5 - 0.0486 Driving Distance-All Drives Avg
                           - 0.0615 Driving Accuracy %
```

```
                          + 0.0111 Fairway Proximity (in.)
                          + 15.4 Long Iron Calc
                          + 0.0207 Scrambling Avg Distance to Hole
                          - 0.782 Strokes Gained - Putting


Predictor                             Coef   SE Coef      T      P    VIF
Constant                            80.535     3.324  24.23  0.000
Driving Distance-All Drives Avg  -0.048619  0.007834  -6.21  0.000  1.722
Driving Accuracy %               -0.06151   0.01107  -5.56  0.000  1.646
Fairway Proximity (in.)          0.011063  0.003125   3.54  0.001  1.209
Long Iron Calc                     15.411     5.163   2.98  0.004  1.333
Scrambling Avg Distance to Hole  0.020715  0.004300   4.82  0.000  1.125
Strokes Gained - Putting          -0.7817    0.1334  -5.86  0.000  1.105


S = 0.404857   R-Sq = 67.1%   R-Sq(adj) = 64.8%

PRESS = 15.8785   R-Sq(pred) = 62.52%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       6  28.4315  4.7386  28.91  0.000
Residual Error  85  13.9323  0.1639
Total           91  42.3638


Source                            DF   Seq SS
Driving Distance-All Drives Avg    1   4.9208
Driving Accuracy %                 1  10.1901
Fairway Proximity (in.)            1   1.9667
Long Iron Calc                     1   1.0255
Scrambling Avg Distance to Hole    1   4.6979
Strokes Gained - Putting           1   5.6306
```

Validation Residuals



The residuals assumptions of normality, linearity, independence, and equal variance are met.

BIBLIOGRAPHY

[1] "Official Schedule of the PGA Tour." *PGATour.com*. PGA Tour. Web. 20 May 2013. <http://www.pgatour.com/tournaments/schedule.html>.

[2]Alexander, Donald, and William Kern. "Drive for Show and Putt for Dough?: An Analysis of the Earnings of PGA Tour Golfers." *Journal of Sports Economics*. 6.1 (2005): 46-60. Web. 20 May. 2013. <http://jse.sagepub.com/content/6/1/46.full.pdf html>.

[3] Engelhardt, George M. "It's Not How You Arrive, It's How Your Arrive: The Myth." *Perceptual and Motor Skills*. 80. (1995): 1135-1138. Print.

[4] Quinn, Robert J. "Exploring Correlation Coefficients with Golf Statistics." *Teaching Statistics*. 28.1 (2006): 10-13. Web. 20 May. 2013. <http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9639.2006.00229.x/abstract>.

[5] Ketzscher, Robert, and Trevor J. Ringrose. "Exploratory Analysis of European Professional Golf Association Statistics." *Statistician*. 51.2 (2002): 215-228. Web. 20 May. 2013. <http://www.jstor.org/stable/3650321?seq=13>.

[6] Shields, Edgar, and Nathan Tomasini. "Golf...What is Important?." *Research Quarterly for Exercise and Sport*. 76.1 (2005): 129-130. Web. 20 May. 2013. <http://search.proquest.com.ezaccess.libraries.psu.edu/docview/218508645/fulltextPDF?accountid=13158>.

[7]Belkin, David S. "Predictability and Stability of Professional Golf Association Tour Statistics." *Perceptual and Motor Skills*. 78. (1994): 1275-1280. Print.

[8] Wiseman, Frederick, and Sangit Chatterjee. "Comprehensive Analysis of Golf Performance on the PGA Tour: 1990-2004." *Perceptual and Motor Skills*. 102. (2006): 109-117. Print.

[9] Dorsel, Thomas N., and Rob J. Rotunda. "Low Scores, Top 10 Finishes, and Big Money: An Analysis of Professional Golf Association Tour Statistics and How These Relate to Overall Performance." *Perceptual and Motor Skills*. 92. (2001): 575-585. Print.

[10] Watkins, John R. "Drive for Show, Putt for Dough: Rates of Return to Golf Skills, Events Played, and Age on the PGA Tour." *Michigan Journal of Business*. 1.1 (2008): 35-59. Print.

[11] Kahane, Leo. "Returns to Skill in Professional Golf: A Quantile Regression Approach." *Internation Journal of Sport Finance*. 5.3 (2010): 167-180. Print.

[12] "Multiple Linear Regression." Yale Statistics. Yale University. Web. 1 Sep 2013. <http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm>.

[13] "Statistics: Driving Distance." *PGA Tour*. PGA Tour Inc., 30 Dec 2012. Web. 20 May 2013. <http://www.pgatour.com/stats/stat.101.html

[14] "Statistics: Driving Distance-All Drives." *PGA Tour*. PGA Tour Inc., 30 Dec 2012. Web. 20 May 2013. <http://www.pgatour.com/content/pgatour/stats/stat.317.html>.

[15] "Statistics: Driving Accuracy Percentage." *PGA Tour*. PGA Tour Inc., 30 Dec 2012. Web. 20 May 2013. <http://www.pgatour.com/stats/stat.102.html

[16] "Statistics: Fairway Proximity." *PGA Tour*. PGA Tour Inc., 30 Dec 2012. Web. 20 May 2013. <http://www.pgatour.com/content/pgatour/stats/stat.431.html>.

[17] "Statistics: Average Distance After Going for It Shot." *PGA Tour*. PGA Tour Inc., 30 Dec 2012. Web. 20 May 2013. <http://www.pgatour.com/content/pgatour/stats/stat.02431.html

[18] "Statistics: Scrambling Average Distance to Hole." *PGA Tour*. PGA Tour Inc., 30 Dec 2012. Web. 20 May 2013. <http://www.pgatour.com/content/pgatour/stats/stat.481.html

[19] "Statistics: Proximity to Hole." *PGA Tour*. PGA Tour Inc., 30 Dec 2012. Web. 20 May 2013. <http://www.pgatour.com/content/pgatour/stats/stat.331.html

[20] "Statistics: Strokes Gained-Putting." *PGA Tour*. PGA Tour Inc., 30 Dec 2012. Web. 20 May 2013. <http://www.pgatour.com/content/pgatour/stats/stat.02564.html

[21] "Strokes Gained-Putting: 2010 Baseline Probabilities." *PGA Tour*. PGA Tour Inc., 30 Dec 2010. Web. 20 May 2013. <http://www.pgatour.com/r/strokes-gained-putting-baseline/index.html>.

[22]Kutner, Michael. "Standardized Multiple Regression Model." *Applied Linear Statistical Models*. 5th Edition, Boston: 2005: p271-278. Print.

[23] "Statistics: Scoring Average." *PGA Tour*. PGA Tour Inc., 30 Dec 2010. Web. 20 May 2013. <http://www.pgatour.com/content/pgatour/stats/stat.108.html

[24]Henry, Neil. "Adjusted R Square." *R-square and Standardization in Regression*. Virginia Commonwealth University. Web. 18 Nov 2013. < http://www.people.vcu.edu/~nhenry/Rsq.htm >.

[25]Fattah, Moataz. "Multicollinearity." Introduction to Empirical Methods. Central Michigan University. Web. 1 Sep 2013. <http://www.chsbs.cmich.edu/fattah/courses/empirical/multicollinearity.html>.

[26] "The Variance Inflation Factor (VIF)." *Identifying Multicollinearity in Multiple Regression Statistics Help for Dissertation Students and Researchers*. ResearchConsultation.com, n.d. Web. 20 May 2013. <http://www.researchconsultation.com/multicollinearity-regression-spss-collinearity-diagnostics-vif.asp>.

[27] Chatterjee, S., Hadi, A.S. and Price, B. (2000). Regression Analysis by Example, 3rd Edition, A Wiley-Interscience Publication,John Wiley and Sons.

[28]Dallal, Gerard E.. "Which Predictors Are More Important?." . JerryDallal.com, 22 May 2012. Web. 20 May 2013. <http://www.jerrydallal.com/LHSP/importnt.htm>.

ACADEMIC VITA

Thomas Joseph Clarke
tjclarke16@gmail.com

Education

Bachelor of Science Degree in Science, Penn State University, Fall 2013
Master of Business Administration, Penn State University, Spring 2015
Honors in Statistics
Thesis Title: Which Areas of the Golf Game are Most Important for Success? A Statistical Analysis
Thesis Supervisor: David Hunter
Faculty Reader: Linda Clark

Honors and Awards

Schreyer Honors College Academic Excellence Scholarship
Phi Beta Kappa Honor Society
Thomas J. Watson IBM National Memorial Scholarship
Science BS/MBA James Balog Scholarship
Alpha Phi Delta Foundation National Scholarship
Dean's List: 6/6 semesters

Professional Experience

IBM Global Business Services, Commercial Consulting Project Management Intern, Summer 2013
Johnson & Johnson Business Process Revenue Analyst Co-op, Summer/Fall 2012
IBM Systems and Technology Group IT Financial Analyst Intern, Summer 2011

Activities

Vice President, Phi Gamma Nu Professional Business Fraternity
Standards Board, Phi Gamma Nu Professional Business Fraternity
THON Financial Chair, Phi Gamma Nu Professional Business Fraternity
Hospitality Committee Member, Penn State Pan-Hellenic Dance Marathon
Finance Committee Member, Penn State Pan-Hellenic Dance Marathon