The Truth behind PGA Tour Player Scores

Sukhyun Sean Park, Dong Kyun Kim, Ilsung Lee

May 7, 2016

Abstract

The main aim of this project is to analyze the variation in a dataset that is obtained from the PGA tour website. The data has been merged according to the player names, and formatted to render it for effective procedures in data analysis. In the recent years where computational power is no longer a hindrance, predicting player performance has been integral to player selection and sports planning. Along with modern equipment supplied for the playground, increased data recording and storage has facilitated better predictive analytics of players. This project aims to demonstrate the efficacy and usefulness of such techniques to some extent.

Introduction: Exploratory Data Analysis

The dataset is pertaining to PGA tour of different players up to the current week. The structure of the raw data did not facilitate analysis right away, however after merging the according to player names over the 9 variables we created an interpretable dataset consisting of 200 players, with monitored data over 9 different indicators,

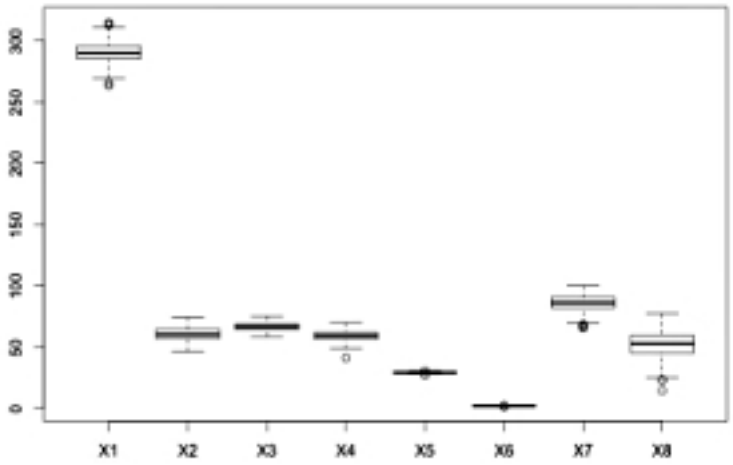
1. Y Scoring Average,
2. X_1 Driving Distance Average,
3. X_2 Driving Accuracy,
4. X_3 Greens in Regulation,
5. X_4 Scrambling,
6. X_5 Total Putting per Round,
7. X_6 Average putting per hole,

8. X_7 Scramble from 10yds,

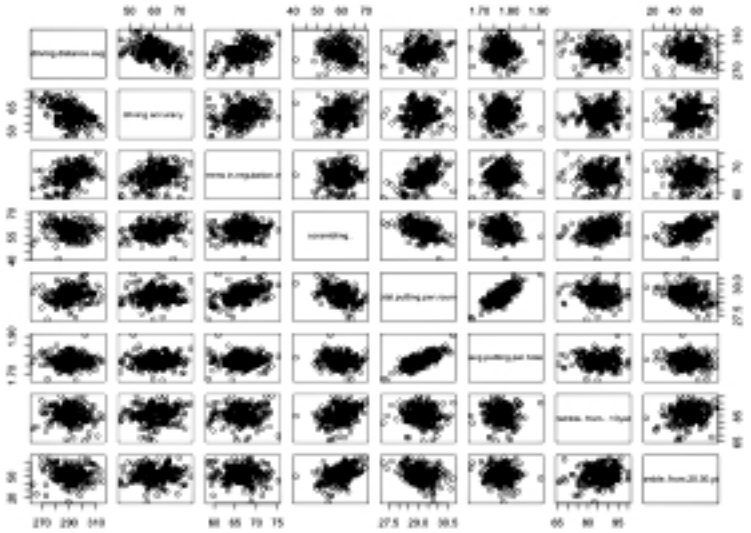
9. X_8 Scramble from 20-30yds.

Naturally, we would like to analyze the variation in Y, that is the scoring average of the players based on the other factors, $\boldsymbol{X} = (X_1, \ldots, X_8)$. Before proceeding directly into modeling the data at hand, we first take a short digression into making a thorough exploratory analysis of the dataset. Referring to [1], we use the common methods, histogram , boxplots and the pairwise correlation plot, to determine the nature of the predicted variables.

Referring to the respective plots shown below, adhered to in the previous paragraph, we see that $\boldsymbol{X}$ are mostly multimodal, although the degree and significance of multimodality is to be tested further. The table below shows the descriptive statistics for $\boldsymbol{X}$. From the histograms we also note that there is a slight skewness in the data. Moreover, referring to the boxplots, we notice the significant shift in scoring average. Referring to the correlation plots we see that there is significant correlation between the variables in $\boldsymbol{X}$. Therefore, we should be expecting multicollinearity in the data while fitting a linear model for prediction.



The rest of the discussion is laid out as follows, firstly due to the multivariate nature of the data, we fit a factor analysis model, referring to [2], [3] for the details of modeling, this allows us to visualize the data on two components and show the impact of clustering thereby. Next, we proceed to carry out cluster analysis, using hierarchical as well as parametric procedures, to determine trends in the data. Again visualization is sought through PCA. Finally, we carry out a regression analysis (multivariate) for building a predictive model for the data.



	X1	X2	X3	X4	X5	X6	X7	X8
Min.	263.4	46.50	58.59	41.12	27.23	1.672	65.52	14.29
Q_1	285.4	56.91	64.39	56.31	28.73	1.753	81.82	45.38
Q_2	290.2	60.26	66.67	59.37	29.14	1.773	86.21	52.88
$\bar{X}$	290.6	60.46	66.53	59.17	29.15	1.772	85.76	52.44
Q_3	296.0	64.50	68.75	61.88	29.53	1.789	90.63	59.44
Max.	314.9	74.18	75.12	69.74	30.77	1.905	100.00	77.14

Factor Analysis: PCA

For this part, we mainly borrow the theoretical background from [2] and [3]. The model fitted to the data, is basically a factor analysis model, which focuses on identifying orthogonal components via linear transformations of the dataset. The modeled variable here is Y, while the factors are constructed using $\boldsymbol{X}$ as follows,

$$\begin{aligned} Y_1 &= \sum_{i=1}^{p} \beta_{1i} X_i, \\ Y_2 &= \sum_{i=1}^{p} \beta_{2i} X_i, \\ &\vdots \\ Y_{p_1} &= \sum_{i=1}^{p} \beta_{p_1 i} X_i. \end{aligned}$$

In matrix notation we transform the data as follows, $\boldsymbol{Y} = \boldsymbol{\beta X}$, where $\boldsymbol{\beta}$ is a rotation matrix that constitutes the loadings. The main goal that is achieved by this model is that

previously correlated X_i s are transformed linearly (orthogonally) to an uncorrelated subset $Y_1, \ldots, Y_{p_1}$ of predictors for the data. Then we have the generalized setup for the model,

$$\boldsymbol{X} - \boldsymbol{\mu} \;=\; \boldsymbol{LF} + \boldsymbol{\epsilon} \tag{1}$$

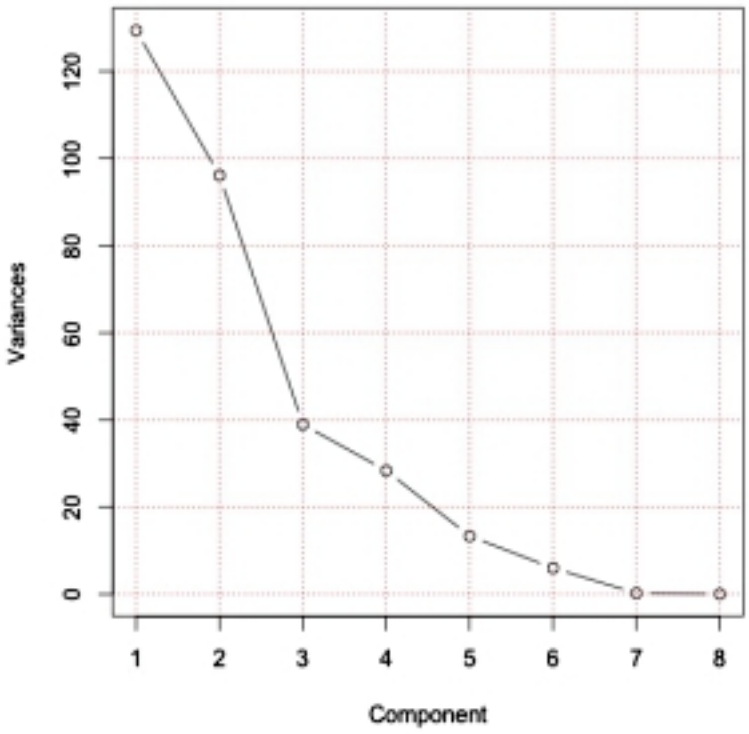
with $\boldsymbol{\epsilon} \sim WN(0, \sigma^2)$. The assumptions for the model are as follows,

1. $\boldsymbol{F}$ and $\boldsymbol{\epsilon}$ are independent,
2. $E(F) = 0$,
3. $Cov(F) = I_{p_1}$.

Under these assumptions we fit model (1), to the data. The summary of the fit is explained in terms of the decreasing order of variance of components, as follows, note that the scree plot shows the diagrammatic representation as well,

Table 1: Variance for the 8 components of PCA.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Std. Dev.	11.37	9.80	6.24	5.32	3.64	2.44	0.42	0.01
Prop.Var.	0.41	0.31	0.12	0.09	0.04	0.02	0.00	0.00
Cum.Prop.	0.41	0.72	0.85	0.94	0.98	1.00	1.00	1.00



We note that selecting the first 3 components account for 85% of the variability in the data. We should be careful at the trade-off we make in between the number of components and the percentage of variance that we are looking for to be explained by the factor model that is fitted to the data. To be on the fair-side we include the 4 components maximum, which accounts for 94% of the variability in the data. It is important to note here that the components fitted, $Y_1, \ldots, Y_4$, cause sufficient dimension reduction for the data, as we have seen that the original data, has 8 predictor variables for the model. This is one of the principal reasons why we consider factor analysis.

The loadings matrix for the 4 principal components selected are given by,

Table 2: Loadings for the 4 components of PCA.

X	Comp.1	Comp.2	Comp.3	Comp.4
driving distance avg	0.33	0.86	0.16	-0.26
driving accuracy	-0.14	-0.34	0.16	-0.79
greens in regulation in	0.04	0.08	0.14	-0.42
scrambling	-0.21	0.02	0.19	-0.23
total putting per round	0.02	0.00	0.00	-0.02
avg.putting per hole	0.00	-0.00	-0.00	0.00
scramble from (10yards)	-0.19	-0.03	0.92	0.29
scramble from (20-30 yards)	-0.89	0.37	-0.20	0.00

The loadings matrix $\boldsymbol{L}$, in equation (1) denotes the degree of rotation that is imposed on $\boldsymbol{Y}$, to introduce orthogonality in the transformed variables, $Y_1, \ldots, Y_4$. In practice we see that the first two principal components of the PCA that is,

$$\begin{aligned} Y_1 &= \sum_{i=1}^{8} \beta_{1i} X_i, \\ Y_2 &= \sum_{i=1}^{8} \beta_{2i} X_i. \end{aligned}$$

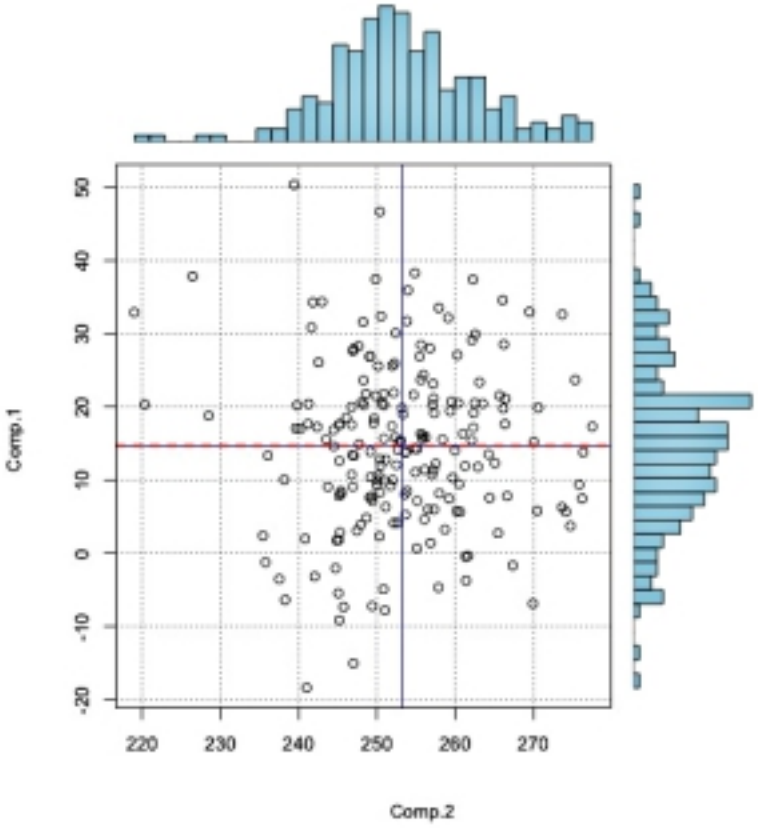
prove useful in projecting the data that is currently in $\mathbb{R}^8$, to $\mathbb{R}^2$.

Analyzing the first 2 principal components

The scatter plot for the projected data on first two components is as shown below, we see that in this case the new variables Y_1, and Y_2 are independent of each other.

Table 3: Loadings for the first 2 PC

X	Comp.1	Comp.2
driving distance avg	0.3269373	0.8588868
driving accuracy	-0.1393701	-0.3383757
greens in regulation in	0.0446734	0.0831210
scrambling	-0.2126148	0.0191585
total. putting per round	0.0198506	0.0003671
avg. putting per hole	0.0002626	-0.0002390
scramble from (10 yards)	-0.1890584	-0.0258099
scramble from (20-30 yards)	-0.8890163	0.3739958



First Component: The first component is characterized by significant scrambling from 20-30 yards, which affects the component negatively, if we note carefully we see that the signs of the co-efficients affect the projections, that is,

- X_2 Driving accuracy,
- X_4 Average Putting per hole,
- X_7 Scramble (10 yards),

- X_8 Scramble (20-30 yards)

affect the first components negatively, while the others affect the first component positively creating the first orthogonal component Y_1, which is responsible for the component explaining the 41% of the variation in the data.

Second Component: The second component is characterized by significant driving distance average, which affects the component positively, if we note carefully we see that the signs of the co-efficients affect the projections, that is,

- X_2 Driving accuracy,
- X_6 Average Putting per hole,
- X_7 Scramble (10 yards)

affect the second components negatively, while the others affect the second component positively creating the second orthogonal component Y_2, which is responsible for the component explaining the 31% of the variation in the data.

We shall use these two components for projecting the data to inspect the data for clustering, and locate X's using which the clustering is implemented on the data.

Cluster Analysis: Hierarchical

The agglomerative hierarchical clustering algorithms build a cluster hierarchy that is commonly displayed as a tree diagram called a dendrogram. They begin with each object in a separate cluster. At each step, the two clusters that are most similar are joined into a single new cluster. Once fused, objects are never separated.

The joining of the clusters depends on the nature of linkage that is defined in between the points in the data. As is rightly inferred we understand that the clustering of the data points are highly dependent on the type of linkage that is used. Clustering is essentially done using a bottom-up approach. Explicitly stating we set a cut point to obtain a crude estimate of the number of clusters in the data. Note that, the latter discussion is based assuming that we have a fixed linkage to evaluate herarchical clustering in the data. More explicit methods of clustering are obtained by considering model based clustering approaches, which are more intricate. For the sake of simplicity we keep this analysis restricted to two most commonly used linkages, the "average linkage", and the complete linkage.

While describing the linkages in between sets of points we use the concept of a distance function. The distance function satisfies the following properties.

A function $d(a, b)$, is said to be a distance function if,

1. For points $a \in A$, and $b \in B$, $d(a, b) \geq 0$,
2. For points $a \in A$, and $b \in B$, $d(a, b) = d(b, a)$,
3. For points $a \in A$, $b \in B$ and $c \in C$, $d(a, b) \geq d(a, c) + d(c, b)$.

Average Linkage:

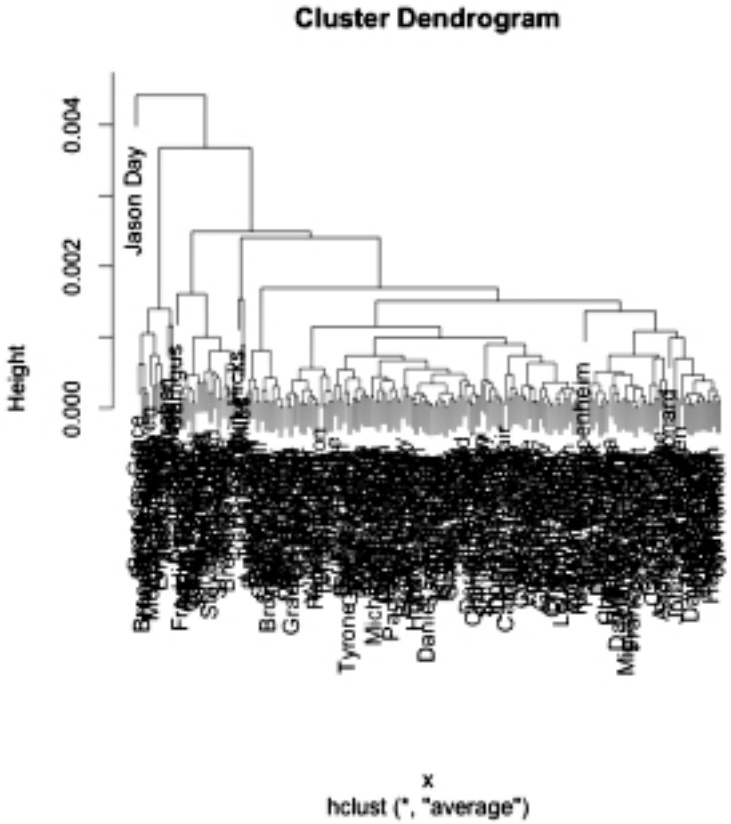
The average linkage uses the following model for linking points in two different clusters A and B through,

$$L_{avg} = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

under a suitable metric usually taken to be the statistical distance between two sets,

$$d(a, b) = \sqrt{(a-b)^T S^{-1} (a-b)}$$

The latter distance function satisfies all of the above properties and also stays consistent with the properties of adhered to above. This is also commonly known as the Mahalanobis distance. Below we have the dendrogram that is created.

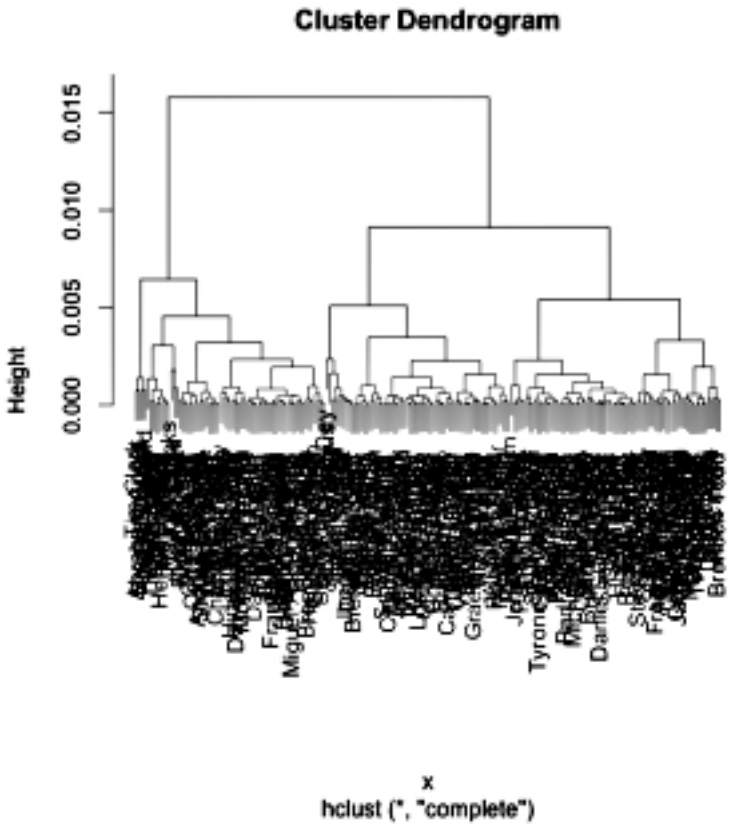


Complete Linkage:

In case of the complete linkage we use the following linkage function,

$$\max\{d(a,b)\Big|a \in A, b \in B\}$$

where d serves as the usual distance function between the sets A and B.



The dendrogram that is created to indicate clustering in the data requires a suitable cutoff point that needs to be selected for obtaining suitable number of clusters in the data. We see that from the figure attached, the tree is plotted bottom up. Therefore, in the PGA data we use the complete linkage that allows us to look for 4 clusters in the data, as a suitable option for plotting the data. The linkage that is used is the complete linkage.

Visualization for the multivariate data becomes increasingly difficult due to the multivariate nature of the data. We use the 2 principal components that have been obtained in the previous section, that account for the largest variability in the dataset to show the properties in clustering through the effect exerted on the two different principal components. Having obtained an intuitive idea regarding the hierarchical clustering in the data, we proceed to go for the more parametric robust methods in clustering, by implementing the K-Means clustering over the dataset.

Cluster Analysis: K-Means

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into $k(\leq n)$ sets $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) (sum of distance functions of each point in the cluster to the K center). In other words, its objective is to find:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

where μ_i is the mean of points in S_i.

The parameter k is selected accordingly as we minimize the WSS, for the data at hand we initialize the algorithm starting from the number of clusters that are located for the hierarchical clustering case. We see that the initial number of 4 clusters, we proceed to 10 clusters,

$$k \;=\; (4, 5, \ldots, 10)$$

The implemented algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the k-means algorithm; it is also referred to as Lloyd's algorithm.

Given an initial set of k means $m_1^{(1)}, \ldots, m_k^{(1)}$ as shown below, the algorithm proceeds by alternating between two steps,

Assignment step: Assign each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). Since the sum of squares is the squared Euclidean distance, this is intuitively the "nearest" mean. Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means.

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \; \forall j, 1 \leq j \leq k \right\},$$

where each x_p is assigned to exactly one $S^{(t)}$, even if it could be assigned to two or more of them.

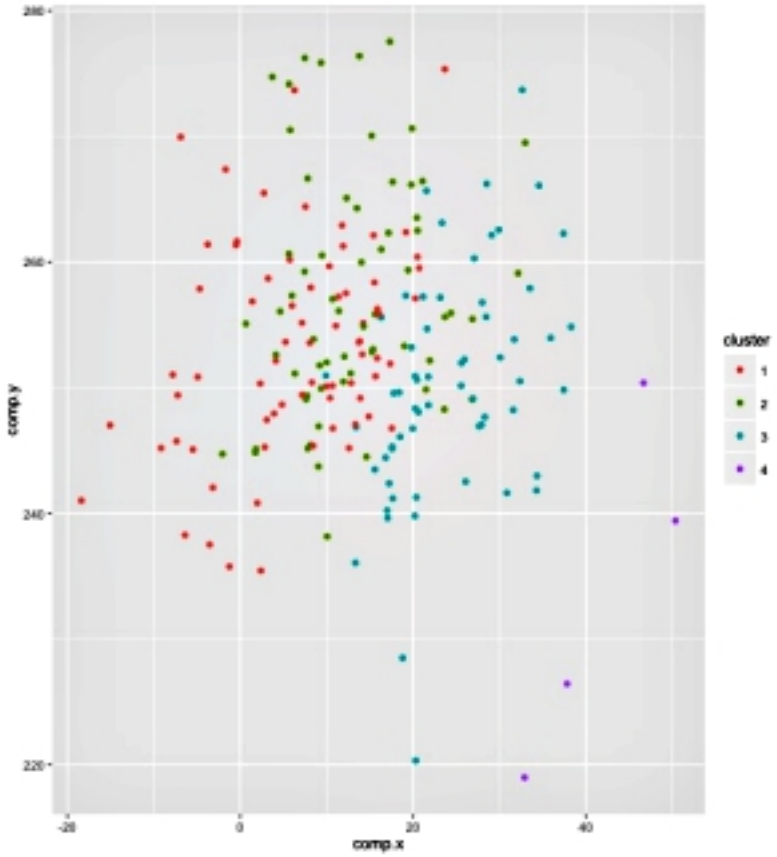
Update step: Calculate the new means to be the centroids of the observations in the new clusters.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Since the arithmetic mean is a least-squares estimator, this also minimizes the within-cluster sum of squares (WCSS) objective. The algorithm has converged when the assignments no longer change. Since both steps optimize the WCSS objective, and there only exists a finite number of such partitioning, the algorithm must converge to a (local) optimum. There is no guarantee that the global optimum is found using this algorithm.

The algorithm is often presented as assigning objects to the nearest cluster by distance. The standard algorithm aims at minimizing the WCSS objective, and thus assigns by "least sum of squares", which is exactly equivalent to assigning by the smallest Euclidean distance. Using a different distance function other than (squared) Euclidean distance may stop the algorithm from converging. Various modifications of k-means such as spherical k-means and k-medoids have been proposed to allow using other distance measures.

The clusters located are projected onto PCA first component and shown below,



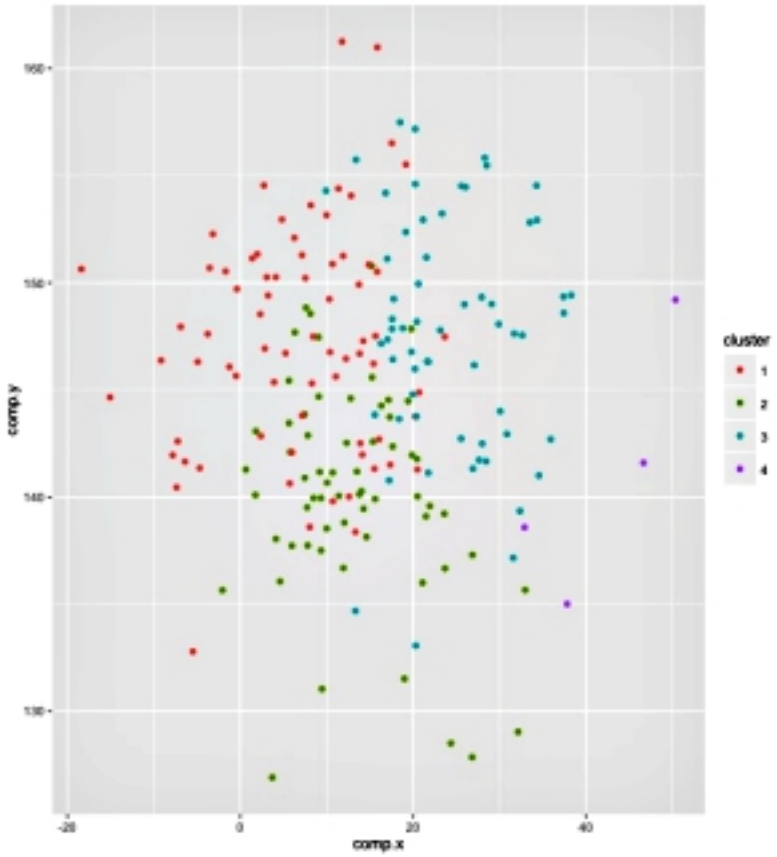
For the problem of visualizing the data, we again use the 2 principal components that have been located in the previous section. We observe that clustering located 4 main clusters in the PGA data. The within and the between sum of squares for the data is shown in the table below,

We conclude that there are 4 clusters in the data by also noting the between sum of squares (BSS) that are located in the data, with $k = 4$ clusters. The plots showing the clustering in the data using the first 2 PCAs has been shown below.

	$k=4$	$k=5$	$k=6$	$k=7$	$k=9$	$k=10$
WSS	33951.35	30461.92	28055.06	25860.70	22564.80	21170.02
BSS	28470.51	31959.95	34366.80	36561.16	39857.06	41251.84

It is imperative to note that the number of clusters located, is sensitive to the norm that is selected for the objective function. We select the L_2-norm, or the Euclidean distance, to evaluate the distance between the two points in the dataset. As it can be seen above that the BSS, is minimum in the case of 4 selected clusters, and therefore we conclude that there are four effective clusters in the data.

The clusters located are projected onto PCA second component and shown below,



Regression Analysis: Linear

In this section we look at linear relationships between Y, and $\boldsymbol{X}$, through the following equation,

$$Y \;=\; \boldsymbol{X}\beta + \epsilon$$

where the underlying assumptions are as follows,

1. $X_i \in \boldsymbol{X}$ are independent of each other.

2. $\epsilon \overset{iid}{\sim} N_n(0, \sigma^2 I_n)$, that shows that the errors are independent of each other.

where $0 < \sigma^2 < \infty$ is the residual error variance and serves as the standard error for the model. It is integral to the reliability of the inference from the model that the above assumptions are verified.

We inspect the linear relationship by inspecting the model coefficients for the linear model specified above. The coefficients are displayed below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	67.2346093	2.3294362	28.8630391	0.0000000
driving.distance.avg	-0.0246449	0.0043382	-5.6809298	0.0000000
driving.accuracy..	-0.0155864	0.0069286	-2.2495908	0.0256161
greens.in.regulation.in..	-0.1697189	0.0191694	-8.8536348	0.0000000
scrambling..	-0.0333482	0.0119415	-2.7926342	0.0057601
total.putting.per.round	0.6991790	0.1885771	3.7076560	0.0002740
avg.putting.per.hole	2.8872139	2.5474925	1.1333552	0.2584857
scramble..from...10yards	0.0011016	0.0048131	0.2288749	0.8192110
scramble..from.20.30.yards	-0.0033570	0.0030965	-1.0841225	0.2796771

We see that the model coefficients above show that for the unrestricted model that is implemented we have the following structure that is assigned for the model,

$$\begin{aligned} Y &= 67.235 - 0.0245X_1 - 0.0156X_2 - 0.170X_3 - 0.034X_4 + 0.7X_5 + 2.887X_6 \\ &\quad +0.001X_7 - 0.004X_8 + \epsilon. \end{aligned}$$

The inferences drawn on the model without looking at any model checking and selection procedures show that the residual standard error for the model is 0.4119 over 191 degrees of freedom, which makes it suitable for prediction. Moreover, we see that the possibility of multicollinearity, therefore, we proceed to more explicit procedures of subset selection on $\boldsymbol{X}$, predictors.

Inference: The multiple R^2, and the adjusted R^2, show that there is significant efficacy in the linear relationship. The inferences drawn are dependent on the p-values of the model coefficients and R^2. The model efficacy that is R^2, is defined as,

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)}$$

that is the amount of variation in Y, that is explained by the regression equation. The latter formula is affected adversely by the presence of outliers. Therefore, we select the adjusted R^2 to draw our inferences. We note that the adjusted R^2 for the model is 0.7719, which shows that the current equation explains 77.2% of the variability that is present in the data.

Subset Selection: Parsimonious models

For implementing subset selection procedures on the given dataset we first proceed to implement selecting independent variables in $\boldsymbol{X}$. The problem of variable selection is one of the most pervasive model selection problems in statistical applications. Often referred to as the problem of subset selection, it arises when one wants to model the relationship between a variable of interest and a subset of potential explanatory variables or predictors, but there is uncertainty about which subset to use.

We use three different kinds of measures to evaluate the criteria for selecting the variables for the purpose of formulating an optimum prediction equation. The direction, starting variable and method of inclusion and deletion affect the model, which is formulated for the PGA data. Keeping all of the latter things in mind we proceed to construct a valid selection procedure. We use the following structure, which is suitable for creating the most parsimonious models.

Method used for the data: PGA parameters

1. Method of Selection: Backward and Forward Selection,
2. Criteria for evaluation: Akaikae Information criteria (AIC), Bayesian Information Criteria (BIC), Residual sum of squares (RSS),
3. Model structure: Linear relationship.

With the above mentioned parameters we proceed to select the variables to formulate the regression equation for predicting Y, that is the scoring average.

Intermediate Steps for model selection

Step 1:

Here we start with the initial model, and show the AIC and RSS, along with the degrees of freedom for the model under construction with the above mentioned parameters. We start with the initial model and evaluate the residual sum of squares (RSS) and AIC, according to the formulae below, for each of the 8 predictors.

$$\begin{aligned} RSS &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i(\boldsymbol{X})), \\ AIC &= 2k - 2\ln(L), \\ AICc &= AIC + \frac{2k(k+1)}{n-k-1}. \end{aligned}$$

where $AICc$, is the corrected formula for AIC for finite, n-sample sizes.

The AIC and the RSS are calculated for the both ways inclusion and exclusions for each $X_i \in \boldsymbol{X}$ and we proceed to begin the extraction of the variables in the predictor set according to the difference caused in AICs, for the model. Note that the corrected AIC (AICc) is used for this purpose.

Table 4: Step 1: Variable Selection

Srl. No	$\boldsymbol{X}$	Df	SumofSq	RSS	AIC
1	scramble..from...10yards	1	0.0089	32.407	-347.99
2	scramble..from.20.30.yards	1	0.1994	32.597	-346.82
3	avg.putting.per.hole	1	0.2179	32.616	-346.70
4	driving.accuracy..	1	0.8584	33.257	-342.81
5	scrambling..	1	1.3229	33.721	-340.04
6	total.putting.per.round	1	2.3318	34.730	-334.14
7	driving.distance.avg	1	5.4743	37.872	-316.82
8	greens.in.regulation.in..	1	13.2963	45.694	-279.27

Step 2:

We see that from the previous table, we see that the `scramble..from...10yards` predictor has the least AICc amongst all predictor variables, and therefore the variable is excluded from the model and the above procedure is repeated for the remaining variables to result in the output shown below.

Table 5: Step 2: Variable Selection

Srl. No	$\boldsymbol{X}$	Df	SumofSq	RSS	AIC
1	scramble..from.20.30.yards	1	0.1936	32.601	-348.80
2	avg.putting.per.hole	1	0.2201	32.627	-348.63
3	driving.accuracy..	1	0.8625	33.270	-344.74
4	scrambling..	1	1.3165	33.723	-342.02
5	total.putting.per.round	1	2.3281	34.735	-336.11
6	driving.distance.avg	1	5.5490	37.956	-318.38
7	greens.in.regulation.in..	1	13.2915	45.698	-281.25

We note that at each step a single variable is deleted from the model, to reach towards an optimum model structure. Again, from the above shown output we observe that, the predictor `scramble..from.20.30.yard` has the least AICc, which implies that the variable, is not as effective a predictor as the rest of the variables that are still present in the model.

Step 3:

After removing, the predictor `scramble..from.20.30.yard`, we are left with 6 variables, for which the RSS and AICc is recalculated. Note the increasing nature of AICc, in absolute value shows gradual progress towards a parsimonious model.

In this output we see that AICc for `avg.putting.per.hole` is considerably low, in comparison to the rest of the variables. Therefore we remove the predictor, to construct the final stage of Forward and Backward Stepwise variable selection process.

Step 4:

We have the following model, at the end, for which we note that the no AICc, is significantly lower in comparison to all of the variables that are considered.

Table 6: Step 3: Variable Selection

Srl. No	$\boldsymbol{X}$	Df	SumofSq	RSS	AIC
1	avg.putting.per.hole	1	0.1551	32.756	-349.85
2	driving.accuracy..	1	0.8459	33.446	-345.67
3	scrambling..	1	1.4460	34.047	-342.12
4	total.putting.per.round	1	2.7421	35.343	-334.65
5	driving.distance.avg	1	5.5124	38.113	-319.55
6	greens.in.regulation.in..	1	13.9688	46.569	-279.47

Table 7: Step 4: Variable Selection

Srl. No	$\boldsymbol{X}$	Df	SumofSq	RSS	AIC
1	driving.accuracy..	1	0.916	33.672	-346.33
2	scrambling..	1	1.738	34.494	-341.51
3	driving.distance.avg	1	5.521	38.277	-320.69
4	total.putting.per.round	1	29.161	61.917	-224.50
5	greens.in.regulation.in..	1	39.392	72.148	-193.92

The variable selection procedure is concluded with the final selected model being,

$$Y = 66.99 - 0.02X_7 - 0.02X_4 - 0.19X_8 - 0.03X_5 + 0.90X_6 + \epsilon \tag{2}$$

with $\epsilon \sim \mathcal{N}$, which marks a conclusive model for prediction of Y, the scoring average for all players in the data.

Hypothesis Testing for Model Coefficients:

For each the β_i, for $i = 1, \ldots, 8$ we test the following hypothesis,

$$\begin{aligned} H_0 : \beta_i &= 0, \\ H_A : \beta_i &\neq 0. \end{aligned}$$

We see that the asymptotic distribution under normality for the variables is,

$$\hat{\beta}_i \overset{a}{\sim} \mathcal{N}(\beta_i, \sigma^2(X'X)^{-1})$$

using the structure we have our desired test statistic, which is used to figure out the significance of the predictors for the model selected for PGA predictors for player scoring averages.

Table 8: Selected Model

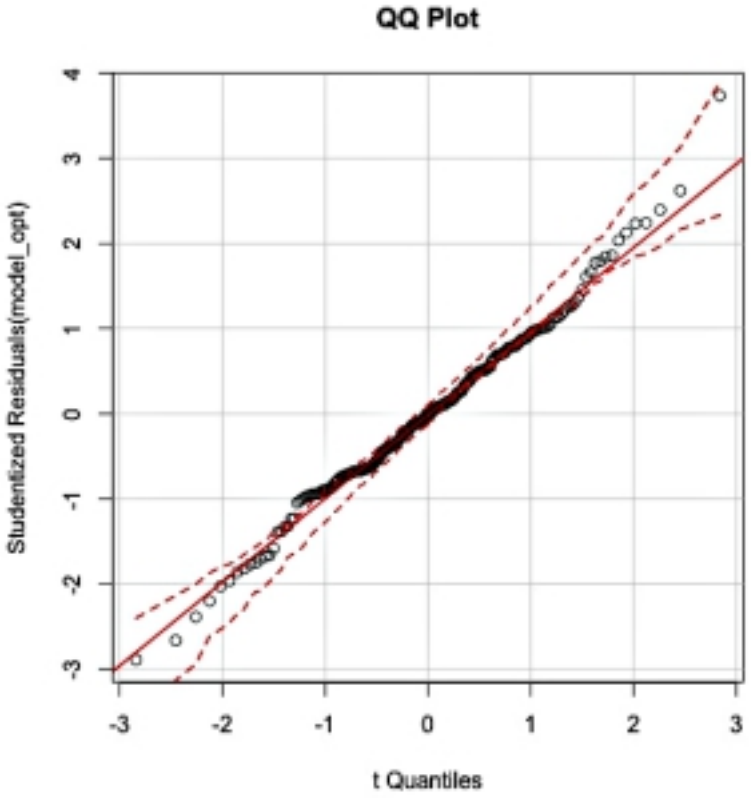
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66.99	2.28	29.33	0.00
driving.distance.avg	-0.02	0.00	-5.72	0.00
driving.accuracy..	-0.02	0.01	-2.33	0.02
greens.in.regulation.in..	-0.19	0.01	-15.27	0.00
scrambling..	-0.03	0.01	-3.21	0.00
total.putting.per.round	0.90	0.07	13.14	0.00

Outlier Testing: Final Model

Now that we have a desirable model for prediction, we proceed to conduct outlier inspection, for influential observations in the data. We first consider the nature of the prediction errors, in terms of tails for prediction errors. The plot below, shows the nature of residuals under the assumption of a t-studentized structure, the dotted lines representing confidence bands. Using naive methods of locating outliers, we see that,

Table 9: Outlier/s

Observation	rstudent	unadjusted p-value	Bonferroni p
Steven Bowditch	3.736688	0.0002455	0.049101



We first establish the definitions for the outliers, leverage and influence with respect to points. Firstly, outliers are observations that are extreme in nature with respect to either location or scale of the entire dataset. Their presence affects inference adversely, in case of parametric estimates, interval or point. Secondly by leverage, we mean the effect or trend, or systematic variation that is provided by each of the predictors. Intuitively, if systematic variation is present, there would be presence of significant trend in the residuals for player scoring averages upon inclusion of that particular predictor variable. Thirdly, influence of an observation refers to the degree of criticalness of position, not only with respect to mean and variance, with respect to the dataset, but as a whole. The difference between the outliers and influence points is that, influence points when removed can cause a significant change in the inference drawn from a dataset.

The commonly used methods that are applied for testing the significance of outliers, influence points and leverage that shall also be used in case of the PGA data are,

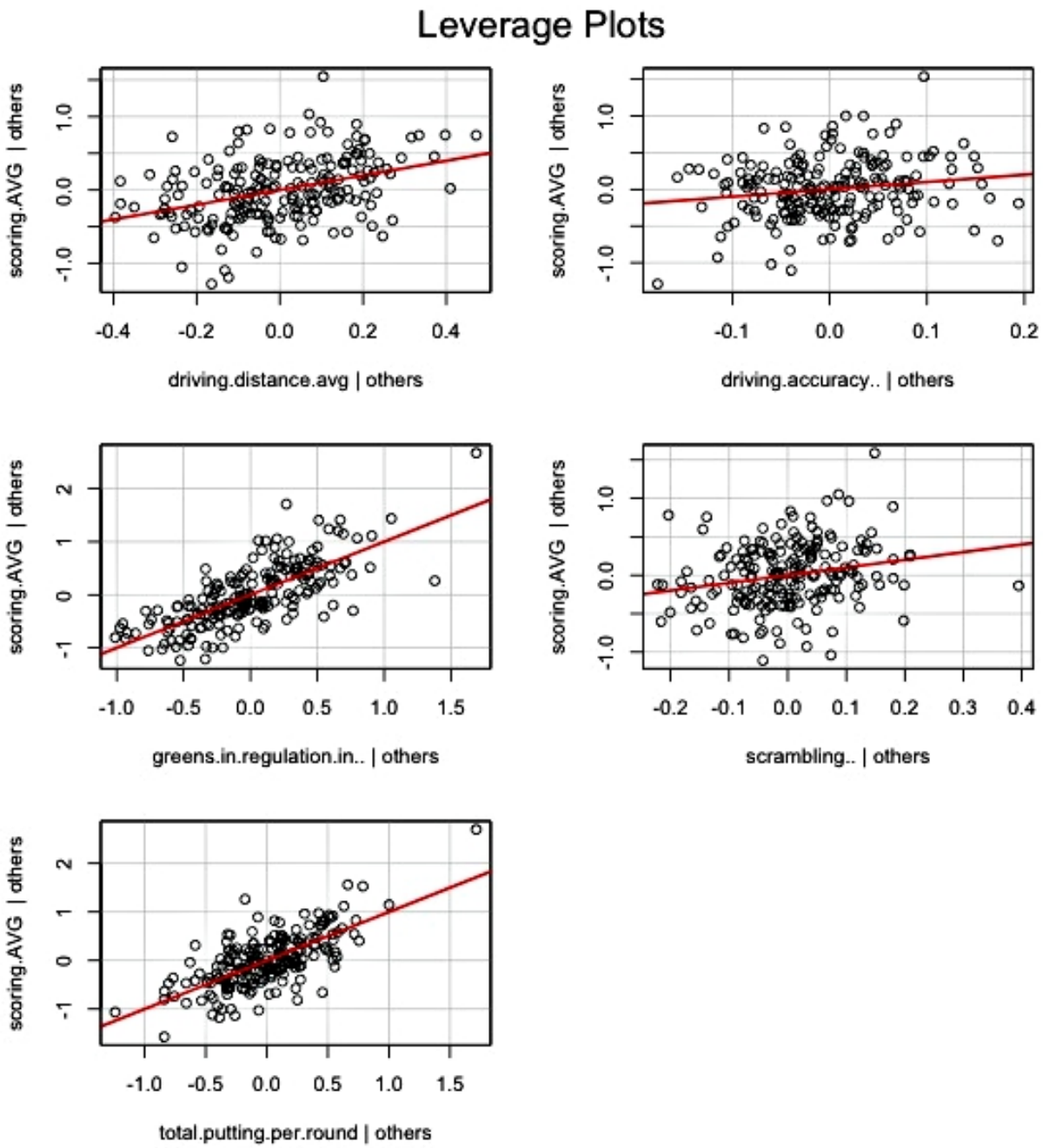
1. QQ-Plots for residuals, checking for tail behavior, and normality of errors,
2. Bonferroni confidence intervals for outlier values,
3. Cook's Distance, Mallows C_p, for testing influence,
4. Hat matrix $H = X(X'X)^{-1}X$, for testing significant outliers and influence points,
5. Model validation and adequacy in terms of standardized and residual plots, to check how correlated the errors are with the fitted values,
6. Autocorrelation in the residuals.

In the ensuing sections we apply each of these techniques for testing for outliers and model adequacy, the model here being the selected parsimonious model given by equation (2),

$$Y = 66.99 - 0.02X_7 - 0.02X_4 - 0.19X_8 - 0.03X_5 + 0.90X_6 + \epsilon$$

Leverage:

While considering the leverage of predictors in the above model, we refer the plots that are generated for the residuals in prediction, on the inclusion and exclusion of the different predictors. We test for significant correlation of a particular selected predictor with other predictors. Referring the plot shown below, we see that the presence of outliers in case of `greens.in.regulation,in..` and `total.putting.per.round` bring about significant change, in the correlation/dependence structure for the predictors. The rest of the plots show fairly independent structures amongst the predictors $\boldsymbol{X}_{(1)} = (X_4, X_5, X_6, X_7, X_8)$ in the equation displayed above.



Hence inspecting the data for outliers and influence points seems integral to the reliability and the inference of the model shown. We aim to locate players that show significantly different behavior in the data, in comparison to the other players in terms of statistics analyzed.

Influence Points

For locating influence points in the dataset, we consider a 2-fold approach towards analyzing the data,

1. Added-Variable Plots,
2. Cook's Distance/ Mallow's C_p plots.

In the PGA data we have the following type of a model,

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_7 + \beta_2 X_4 + \beta_3 X_8 + \beta_4 X_5 + \beta_5 X_6 + \epsilon, \\ &= 66.99 - 0.02 X_7 - 0.02 X_4 - 0.19 X_8 - 0.03 X_5 + 0.90 X_6 + \epsilon. \end{aligned}$$

where the predictor variables X_i and X_j, for $i \neq j$ may be correlated. For instance the slopes indicate by β_1 and β_2 are both negative, we can say that,

1. Y, scoring average for players decreases as with increase in driving distance average X_7, if driving accuracy, X_4 is held constant,
2. Y, scoring average for players decreases as with increase in driving accuracy, X_4 if driving distance average X_7 is held constant.

Since, both β_1, and β_2 are negative.

We have already established in previous sections the importance of interpreting multiple regression coefficients by considering what happens when the other variables are held constant ("ceteris paribus"). For example, if a model with Y against X_4 is regressed with a model $Y = \beta_0' + \beta_1' X_4 + \beta_2' X_7 + \epsilon$. The estimates of the model coefficients may be different in both of the models, consequently there is an adverse effect on the predictive efficacy for the model for scoring average of players. This problem is commonly termed as *omitted-variable* bias.

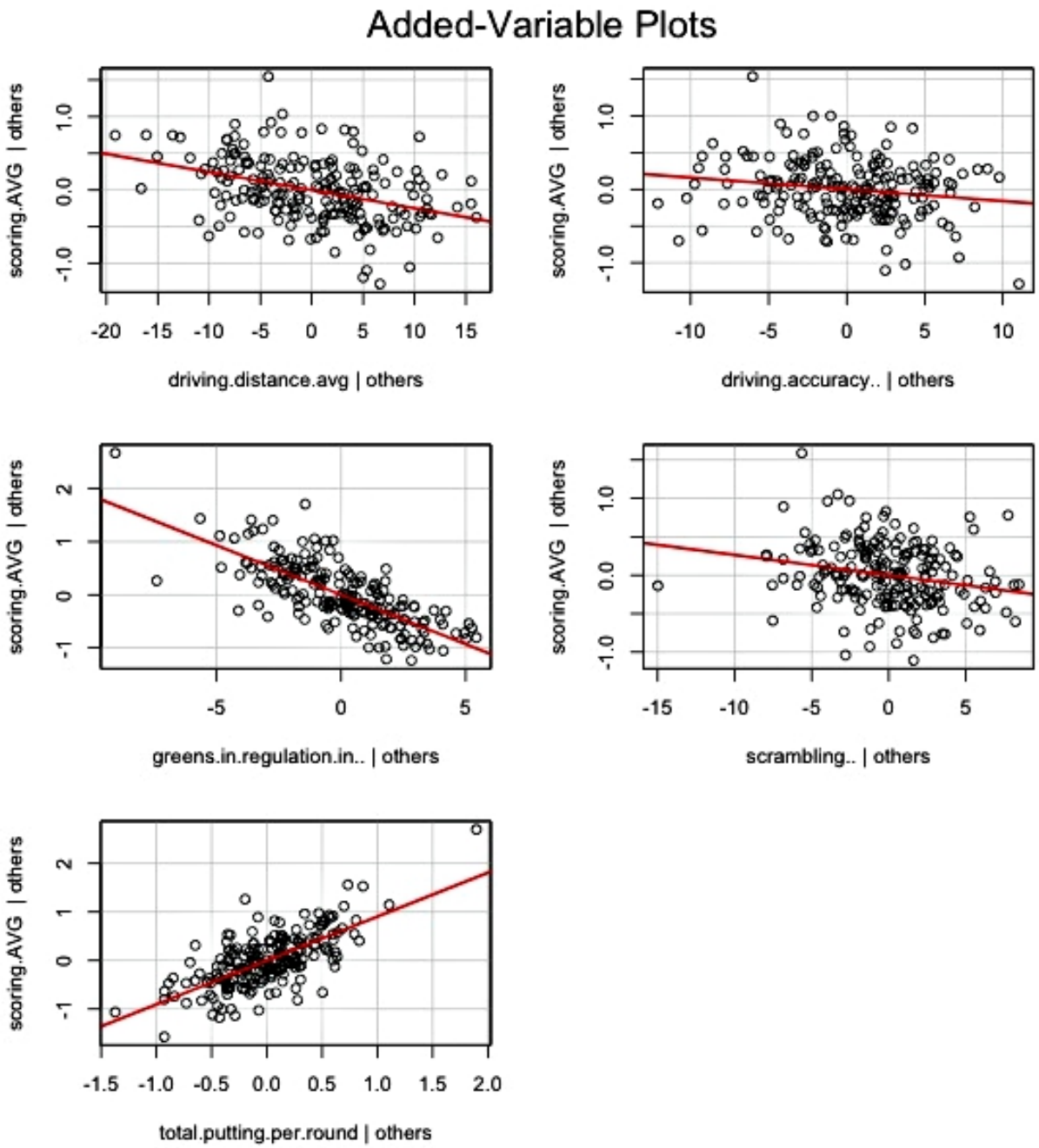
A lot of the value of an added variable plot comes at the regression diagnostic stage, especially since the residuals in the added variable plot are precisely the residuals from the original multiple regression. This means outliers and heteroskedasticity can be identified in a similar way to when looking at the plot of a simple rather than multiple regression model. Influential points can also be seen - this is useful in multiple regression since some influential points are not obvious in the original data before you take the other variables into account.

Hence in short by using added variable plots we delve deeper into inspecting the model selected for further inconsistencies amongst predictors, in terms of the subjective amount of predictive advantage offered by considering one regression over the other. The added variable plots for the predictors $\boldsymbol{X}_{(1)}$ are shown below.

Cook's Distance Plot:

In this section we consider the Cook's distance Plots. Note that we have the following structure,

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{\mathbf{X}} \quad \underset{p\times 1}{\boldsymbol{\beta}} \quad + \quad \underset{n\times 1}{\boldsymbol{\epsilon}}$$



with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and $\boldsymbol{\beta} = [\beta_0\, \beta_1 \ldots \beta_{p-1}]'$. The estimator of $\boldsymbol{\beta}$, under squared-error loss is, $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$. With this the Cook's distance is,

$$D_i = \frac{e_i^2}{s^2 p}\left[\frac{h_i}{(1-h_i)^2}\right],$$

where $s^2 = (n-p)^{-1}\left[\frac{h_i}{(1-h_i)^2}\right]$ is the estimate of the squared error. The acceptable bound that is used for the Cook's distance is

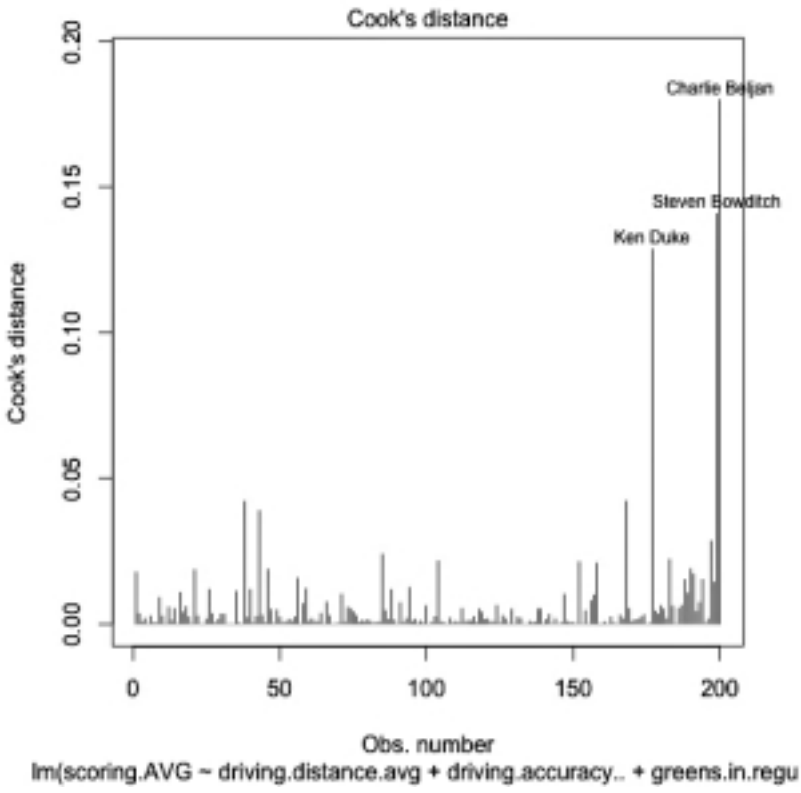
$$\frac{4}{n-p-2}$$

The significance of influence of the points is measured by testing the following hypothesis,

$$\begin{aligned} H_0 : D_i &\leq \frac{4}{n-p-2}, \\ H_A : D_i &> \frac{4}{n-p-2}, \end{aligned}$$

for all points $i = 1, \ldots, n$. From, the plot shown below, we infer the following points that are shown as outliers in the current PGA dataset. We display the total data that is available on the located points. Note that all of the shown players have high scoring averages, low driving distance averages, low driving accuracy, low scrambling, high total putting per round, high average putting per hole, low greens in regulation, and unusually low, scramble from 10/20/30 yards.

The plot showing Cook's Distance for different data points is shown below,



These characteristics are picked up by the Cook's distance and the points shown below are classified are outliers/influence points for the model.

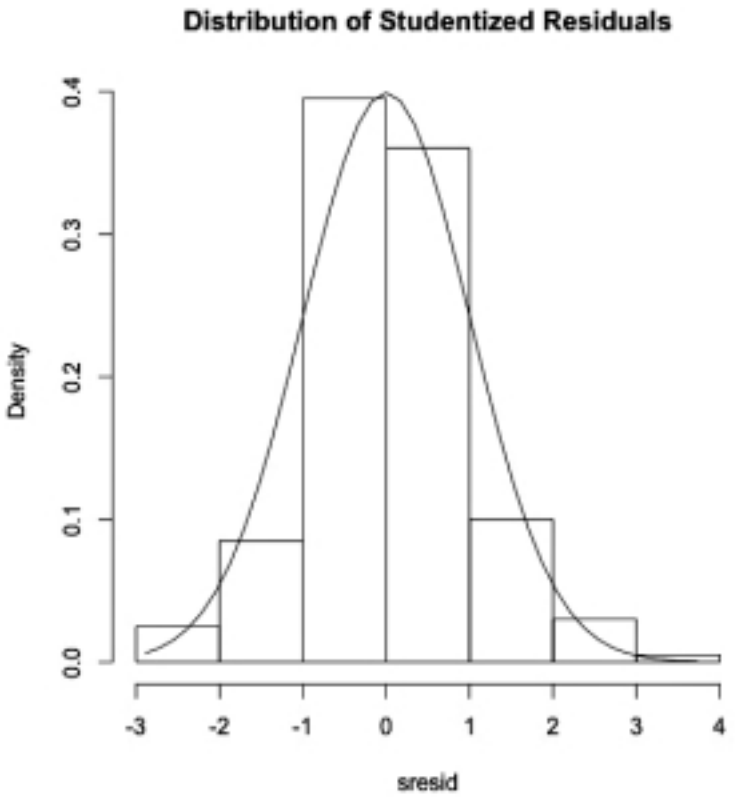
Table 10: Outliers: Cook's Distance

Player Name	Ken Duke	Charlie Beljan	Steven Bowditch
scoring AVG	72.31	74.75	74.34
driving distance avg	276.20	301.80	289.60
driving accuracy	70.24	51.38	46.93
greens in regulation in	58.73	60.90	59.41
scrambling	53.21	55.19	49.43
total putting per round	29.52	30.77	28.92
avg putting per hole	1.83	1.91	1.77
scramble from 10 yards	77.27	94.74	81.82
scramble from 20-30 yards	50.00	47.06	37.14

Normality for residuals:

We look at studentized residuals for the model, through the following diagrammatic representations,

1. QQ-Plot
2. Histogram



The studentized residuals are calculated using,

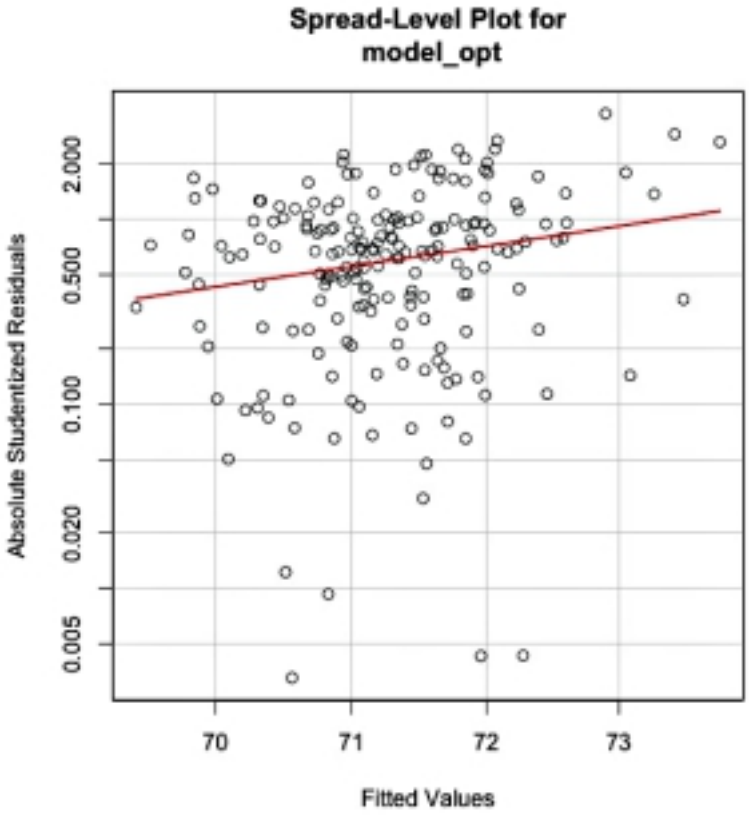
$$\begin{aligned} t &= \frac{e_i - E(e_i)}{S.E.(e_i)} \\ e &= Y - \hat{Y}. \end{aligned}$$

the quantiles are shown for the latter variable.

The histogram and the residual QQ-plot are shown above. The histogram and the residuals show sufficient consistency with normality, which is inferred using the confidence bands and the shape of the histogram, overlaid with the probability curve.

Variance Inflation Factor:

We inspect the correlation structure between the fitted values in the optimum model, (2). Ideally one should expect no significant correlation in the scatter plot. As we can see from the scatter-plot, there is no significant correlation in between the fitted values and residuals. The standardized residuals are used in their absolute value to indicate the relationships in terms of magnitude.



The variance inflation factor, picks up the spike in residual variance due to high correlation being present in between the predictors. The table below shows that there is no significant correlation in between the predictors for the model selected in equation (2).

Table 11: VIF: Significant

$\boldsymbol{X}$	VIF sig.
driving.distance.avg	FALSE
driving.accuracy..	FALSE
greens.in.regulation.in..	FALSE
scrambling..	FALSE
total.putting.per.round	FALSE

CERES Plots , Auto correlated Errors

It is imperative to verify that the errors are not auto-correlated, because if the errors are auto-correlated the assumption of independence for observations, of Y scoring average is violated. We use the Durban-Watson test,

$$\begin{aligned} H_0 &: \text{The errors are independent,} \\ H_A &: \text{The errors show significant dependence.} \end{aligned}$$

Also, a homoscedasticity test is conducted, to check whether there is a presence of significant heteroscedasticity in the model. The CERES plots, and the Durban Watson test show that there is sufficient non-linearity still present in the errors.

Conclusion

We finally conclude the analysis, by suggesting the prediction equation (2), as an optimum model. Also, a cross-validation procedure implemented shows that the out-sample variance for the model, is sufficiently lower than the in-sample variance, indicating reliability of inference drawn. The methods of cluster and regression analysis proved effective in understanding the nature of relationship of the dependence in between the predictors $\boldsymbol{X}_{(1)}$ and Y, scoring average. The outliers for the model are shown in the table (10).

This analysis establishes the following relationship,

$$Y = 66.99 - 0.02X_7 - 0.02X_4 - 0.19X_8 - 0.03X_5 + 0.90X_6 + \epsilon$$

as the governing relationship for scoring averages for players. Using multiple linear regression we located the factors that are significant in terms of their effect on explaining variability in player scoring averages are,

1. Driving Distance Average,
2. Driving Accuracy,

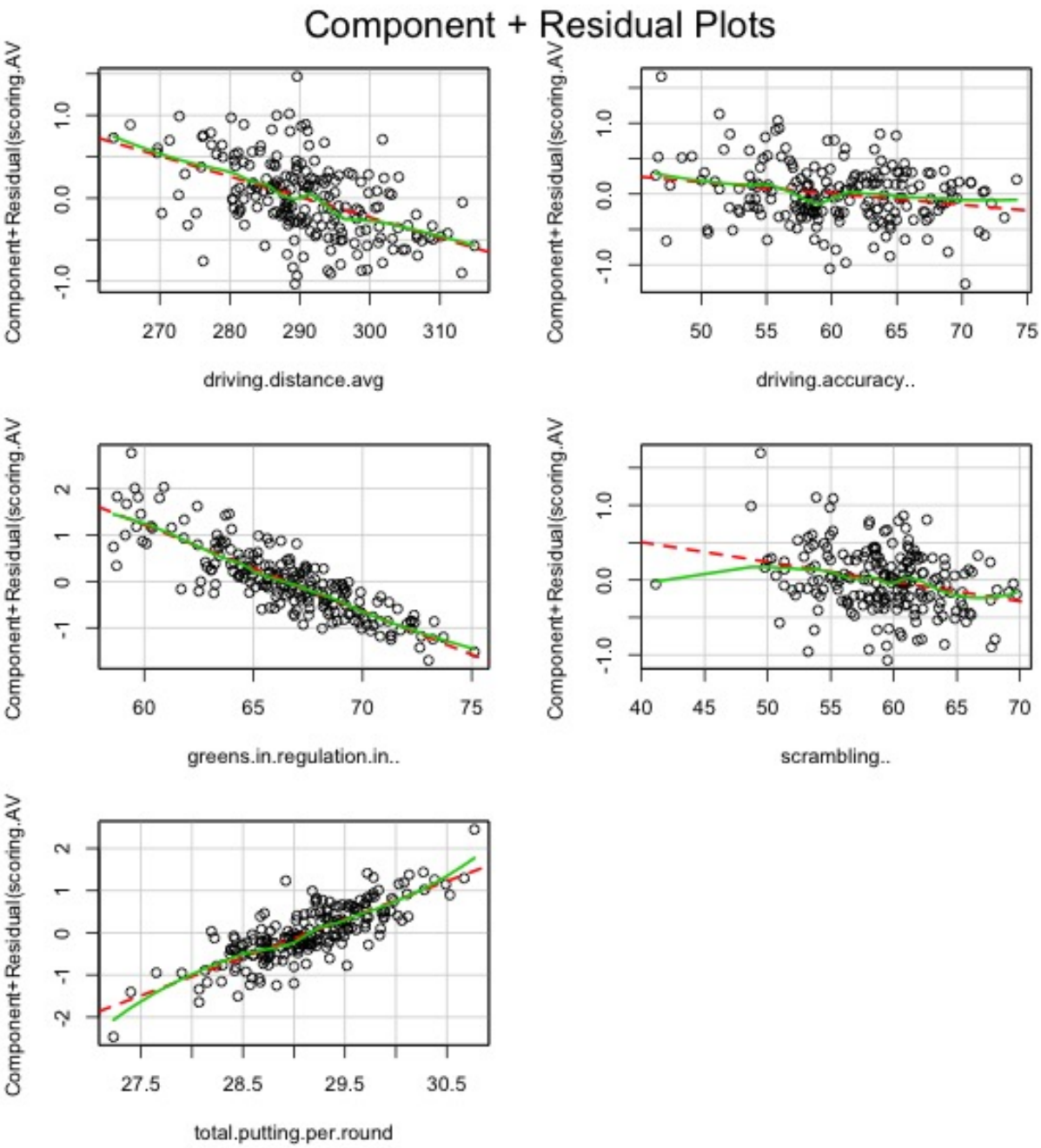
3. Greens in regulation,
4. Scrambling,
5. Total Putting per Round.

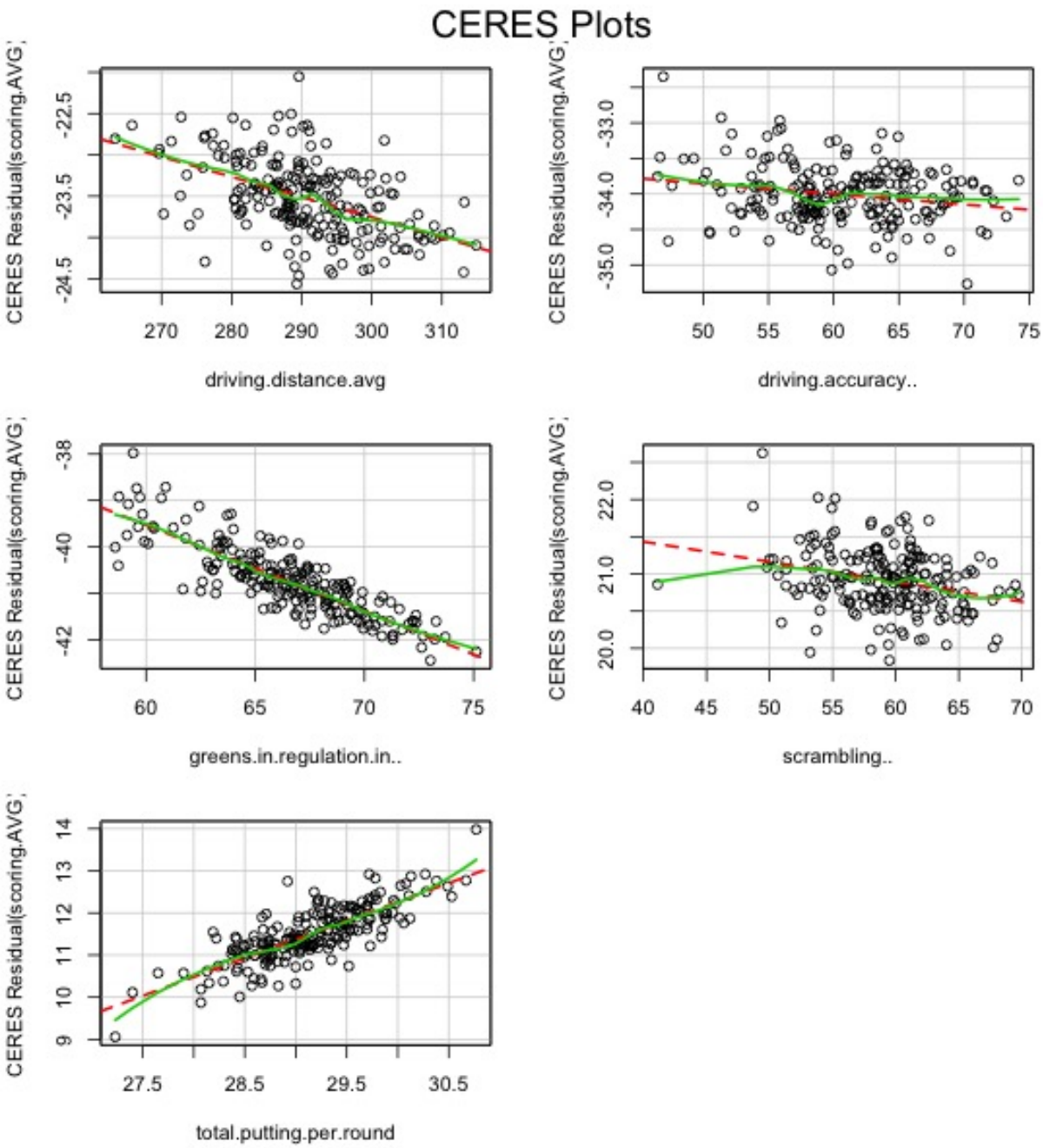
These predictor variables significantly affect the variability in terms of explaining scoring averages for players. Note that when examining the PCA factors responsible for justifying maximum variability we note that the coefficients for the above mentioned variables are higher,

Table 12: Loadings for the 4 components of PCA.

X	Comp.1	Comp.2	Comp.3	Comp.4
driving distance avg	0.33	0.86	0.16	-0.26
driving accuracy	-0.14	-0.34	0.16	-0.79
greens in regulation in	0.04	0.08	0.14	-0.42
scrambling	-0.21	0.02	0.19	-0.23
total putting per round	0.02	0.00	0.00	-0.02
avg.putting per hole	0.00	-0.00	-0.00	0.00
scramble from (10yards)	-0.19	-0.03	0.92	0.29
scramble from (20-30 yards)	-0.89	0.37	-0.20	0.00

where the first two factors justify the projected plots that are considered in the analysis. This concludes the analysis successfully showing consistency in relationship for the significant predictors located above.


Component + Residual Plots
Component+Residual(scoring.AV
driving.distance.avg
driving.accuracy..
greens.in.regulation.in..
scrambling..
total.putting.per.round


CERES Plots
CERES Residual(scoring.AVG)
-22.5
-23.5
-24.5
270
280
290
300
310
driving.distance.avg
CERES Residual(scoring.AVG)
-33.0
-34.0
-35.0
50
55
60
65
70
75
driving.accuracy..
CERES Residual(scoring.AVG)
-38
-40
-42
60
65
70
75
greens.in.regulation.in..
CERES Residual(scoring.AVG)
22.0
21.0
20.0
40
45
50
55
60
65
70
scrambling..
CERES Residual(scoring.AVG)
14
13
12
11
10
9
27.5
28.5
29.5
30.5
total.putting.per.round

References

1. *"Modern Techniques of Exploratory Data Analysis"*, Tukey, Wiley-Interscience.
2. *"Applied Multivariate Data Analysis"*, Johnson and Wichern. Prentice Hall (2007).
3. *"Elements of Statistical Learning"*, Hastie, Trevor and Tibshirani. Springer.

References from Online

Wikipedia NCSS