K23038233_Coursework report.pdf

by Amandeep Singh

Submission date: 18-Dec-2023 03:03PM (UTC+0000)

Submission ID: 219854387

File name: K23038233_Coursework_report.pdf_7631709_631590181.pdf (531.42K)

Word count: 4853

Character count: 25175

Predicting Average Total Winnings in Golf

Amandeep Singh
King's College London
Statistics for Data Analysis
amandeep.1.singh@kcl.ac.uk

December 18, 2023

Abstract

In this study, I will investigate the possibility of predicting professional golfers' average total winnings per tournament through the use of linear regression. I will make use of a dataset comprising professional golfers and the performance indicators that correspond with them, such as driving distance, par-3 performance ratio, greens in regulation, putting, birdies, sand saves, and scrambling. Then, I will do exploratory data analysis, model selection, model evaluation and data preprocessing. Thereafter, I will try to show that using the testing data, linear regression we can accurately forecast the average total winnings.

Table of Contents

best to have table of contents on a separate page

1 Introduction

We all know professional golf is a difficult and fiercely competitive sport where mastery of technique, tactical awareness, and physical fitness are all necessary for success. In this domain we know cumulative winnings have become an important metric for assessing a golfer's skill level, a complete picture of the financial benefits received from competing in tournaments. Additionally, we believe that predicting average total winnings with accuracy is essential for a variety of reasons, including assessing player performance, exploring sponsorship opportunities, and assisting bettors in making informed decisions within the sports betting sector.

In light of this, through this research we investigate the feasibility of applying linear regression as a predictive technique to determine the mean cumulative earnings of professional golfers. The main focus of our research is a large dataset that includes athletes' performance indicators and total wins. Our goal is to carefully construct a prediction model, which is accomplished in the context of highlighting the importance of different tasks. These consist of the delicate art of data preprocessing, the insightful discoveries made via exploratory data analysis, the careful choice of variables, and the meticulous assessment of the model's effectiveness, all of which work together to improve the accuracy and dependability of the prediction framework. Furthermore, for clarity and simplification in this research, I have referred to a set of variables as Other variables and *avg_winning*.

$$OtherVariables = AveStrokes, Driving, GIR, Putting, Birdie, SandSaves, Scrambling, PPR$$

$$avg\_winning = \frac{TotalWinnings}{Tournaments}$$

We all are aware traditionally, a standard form of a multiple linear regression model with p predictors is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

where:

- Y is the dependent variable.
- β_0 is the intercept term.
- $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients associated with predictors $X_1, X_2, \ldots, X_p$.
- $X_1, X_2, \ldots, X_p$ are the independent variables or predictors.
- ε is the error term representing unobserved factors affecting Y.

Fundamentally, through this study I aim to both clarify the complexities involved in projecting average cumulative winnings and emphasize the critical role that careful planning plays in improving forecasting accuracy in the dynamic world of professional golf.

2 Related Work

I think the use of statistical and machine learning methods to forecast golf results has been the subject of increasing amounts of research, with encouraging findings.

A Bayesian hierarchical model, for instance, was used in a recent study[3] to precisely forecast golf scores for individuals and tournaments. To produce precise forecasts, this algorithm used past data on player performance, course features, and meteorological conditions. Another study that was published[4] showed how well a support vector machine (SVM) method worked for predicting the results of golf tournaments. A dataset of previous tournament outcomes was used to train the SVM algorithm, which considered variables including player rankings, meteorological information, and prior performance at the course.

These and other research highlights us how data-driven techniques can improve our comprehension and ability to forecast average winning per tournament and related aspects in golf.

3 Data Preprocessing

I want to say, a crucial stage in the pipeline for data mining and machine learning is data preprocessing, which turns unprocessed data into a format that is clear, arranged, and easy to use. Furthermore, I think preparing the data for further analysis and modeling is the main goal as it guarantees the precision, dependability and efficacy of insights derived from data.

For this analysis, I begin the study by loading the necessary libraries, which cover a variety of features like ggplot2 for visualization, and statistics for statistical analysis. The data processing journey is then started by loading the golf dataset from the file directory. Thereafter, to guarantee the accuracy and completeness of the data, a thorough check is made for missing values, where I found no missing values in dataset. The table(1) shows the first 6 data rows of dataset to get an understanding of different columns in the Golf dataset.

Index	FirstName	Surname	TotalWinnings	Tournaments	avg_winning
1	Francesco	MOLINARI	4635909	28	165568.18
2	Patrick	REED	3057948	26	117613.38
3	Tommy	FLEETWOOD	2805043	50	56100.86
4	Rory	MCILROY	2760667	30	92022.23
5	Alex	NOREN	2729725	28	97490.18
6	Thorbjorn	OLESEN	2485020	65	38231.08

Table 1: Head(Golf) Player Data - Part 1

Thereafter, I do variable separation and divide TotalWinnings by Tournaments to create a new variable, `avg_winning` (Table 1). I used the choose function in R to separate more

Index	AveStrokes	Driving	GIR	Putting	Birdie	SandSaves	Scrambling	PPR
1	69.31	68.1	71.2	1.775	4.16	26.9	62.1	29.43
2	69.87	54.4	66.3	1.738	3.97	40.0	58.2	28.54
3	69.48	65.5	74.1	1.779	4.38	39.8	57.1	29.88
4	69.18	52.9	67.8	1.724	4.41	42.9	66.1	28.30
5	70.25	62.8	66.7	1.764	3.63	56.4	58.9	29.21
6	69.72	53.7	68.7	1.737	4.48	54.6	64.2	28.57

Table 2: Head(Golf) Player Data - Part 2

relevant variables like AveStrokes, Driving, GIR, Putting, Birdie, SandSaves, Scrambling, PPR which can be our predicting variables in linear regression model from First name and Surname which are not useful in predicting average winning per tournament.

4 Exploratory Data Analysis

I believe Exploratory data analysis helps to examinie the distribution of variables, identifying potential relationships between variables, and gaining insights into the data. In this, I start by examining the staistical analysis for the Golf dataset.

Variable	n	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis	SE
AveStrokes	100	70.48	0.56	70.47	69.18	72.19	3.01	0.25	0.18	0.06
Driving	100	59.41	5.84	59.55	42.6	73.2	30.6	-0.08	-0.21	0.58
GIR	100	69.18	2.92	69.05	61.6	74.7	13.1	-0.15	-0.43	0.29
Putting	100	1.77	0.02	1.77	1.71	1.83	0.12	0.01	-0.2	0
Birdie	100	3.93	0.29	3.95	3.15	4.62	1.47	-0.2	-0.12	0.03
SandSaves	100	46.97	6.73	47.6	26.9	61	34.1	-0.44	0.36	0.67
Scrambling	100	57.17	3.94	57.2	48.8	66.2	17.4	0.19	-0.49	0.39
PPR	100	29.42	0.57	29.45	27.82	30.8	2.98	-0.15	-0.25	0.06
avg_winning	100	21165.5	25708.12	11407.49	4245.86	165568.2	161322.3	3.07	11.1	2570.81

Table 3: Descriptive Statistics for Golf Data

Furthermkre, the golf dataset's statistical analysis(Table 4) provides useful details on a range of performance measures. We can see a tiny standard deviation of 0.56 and an average AveStrokes of 70.48 point to minimal variability among golfers. Also, the data is slightly biased to the right, suggesting a higher proportion of players with lower AveStrokes. I believe the distribution that is slightly peaked is implied by the 0.18 kurtosis. Similar to this, I dound that the examination of putting, driving, and greens in regulation (GIR) shows that these measurements have different degrees of skewness and variability, with kurtosis values revealing information about the distribution forms.

Moreover, an analysis of the birdies per round reveals an average of 3.93, a slightly platykurtic distribution, a tiny standard deviation, and a negative skewness indicating a leftward skew. Also, the distribution of sand saves each round is slightly leptokurtic, with a negative skewness indicating more rounds with fewer sand saves. There is also a noticeable fluctuation in the data. Furthermore, I find that the distribution of scrambling holes each round is relatively platykurtic, with a moderate standard deviation and positive skewness indicating more rounds with more scrambling holes.

Thereafter, study shows us that Putts Per Round (PPR) is slightly platykurtic distribution, an average of 29.42 with a modest standard deviation, and a negative skewness indicating more rounds with fewer PPRs. Lastly, we can see a leptokurtic distribution and significant variability are shown by the analysis of winnings per round, which also shows a positive skewness, suggesting that some rounds have bigger winnings than average.

All things considered, these statistical measures give us a thorough overview of the golf dataset and shed light on the central tendency, variability, skewness, and kurtosis of different performance metrics. As a result, they help us understand the performance patterns of golfers in the dataset better.

Index	**FirstName**	**Surname**	**TotalWinnings**	**Tournaments**	**avg_winning**
1	Francesco	MOLINARI	4635909	28	12.01714
2	Patrick	REED	3057948	26	11.67516
3	Tommy	FLEETWOOD	2805043	50	10.93491
4	Rory	MCILROY	2760667	30	11.42979
5	Alex	NOREN	2729725	28	11.48751
6	Thorbjorn	OLESEN	2485020	65	10.55140

Table 4: Head(Golf)-Using Log transformation on Average Winning

Thereafter, I try to do the log transformation on the *avg_winning*(Table 4). Given the *avg_winning* variable's strong right-skewed distribution (skewness = 3.07) and high kurtosis (11.1), a log transformation is necessary. The kurtosis emphasizes larger tails than in a normal distribution, while the skewness shows a concentration of data on the left side. I want to address these distributional problems by applying a log transformation, which will improve the data's symmetry and bring it closer to normality. As, a result I get skewness of 0.81 and kurtosis of 0.05.

good

After, this I make correlation matrix(Table 5) which reveals complex connections between different golf performance indicators. Interestingly, *avg_winning* and AveStrokes show a significant negative connection (-0.5796), suggesting that winnings typically rise as the number of strokes decreases. Also, a higher average winnings are correlated with better performance in SandSaves (-0.2539), PPR (-0.2211), and Scrambling (0.2), according to positive correlations. As for AveStrokes, it can show us a strong negative connection with Putting (-0.6297) and a moderate negative correlation with Birdie (-0.4682). This implies us that putting success and birdie counts are positively correlated with higher stroke metrics. On the other hand, scrambling exhibits positive correlations with both Putting (0.0388) and Birdie (0.147), suggesting a moderately favorable relationship.

Moreover, subtle patterns are revealed by inter-variable correlations between SandSaves, PPR, Putting, Birdie, Driving, and GIR. For example, PPR and Putting shows us a signifcannot read .. too small ... presentation is not good

Table 5: Correlation Table

Variables	Correlation								
	Average_Winnings	AveStrokes	Scrambling	SandSaves	PPR	Putting	Birdie	Driving	GIR
Average_Winnings	1	-0.5796	0.2	-0.2539	-0.2211	-0.1894	0.157	0.1435	0.0097
AveStrokes		1	-0.4297	0.0775	0.2182	0.4124	-0.6297	-0.245	-0.4682
Scrambling			1	0.1909	-0.5836	-0.2146	0.0388	0.1474	-0.1834
SandSaves				1	0.0529	0.0619	0.2787	0.3662	0.1384
PPR					1	0.2305	-0.6701	-0.3706	-0.4766
Putting						1	-0.5255	-0.2938	-0.2846
Birdie							1	0.3828	0.2663
Driving								1	0.1181
GIR									1

icant negative association (-0.6701), which suggests that PPR tends to decline as putting proficiency increases. In believe that professionals and golf enthusiasts alike can learn a lot from this thorough examination, including possible areas for strategic focus and improvement.

ok

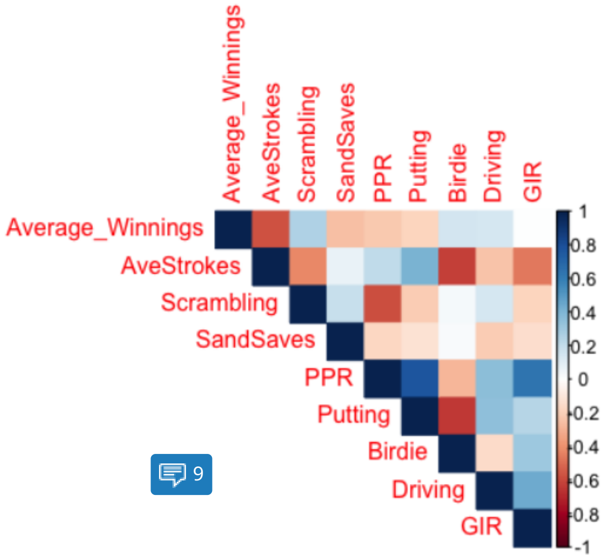
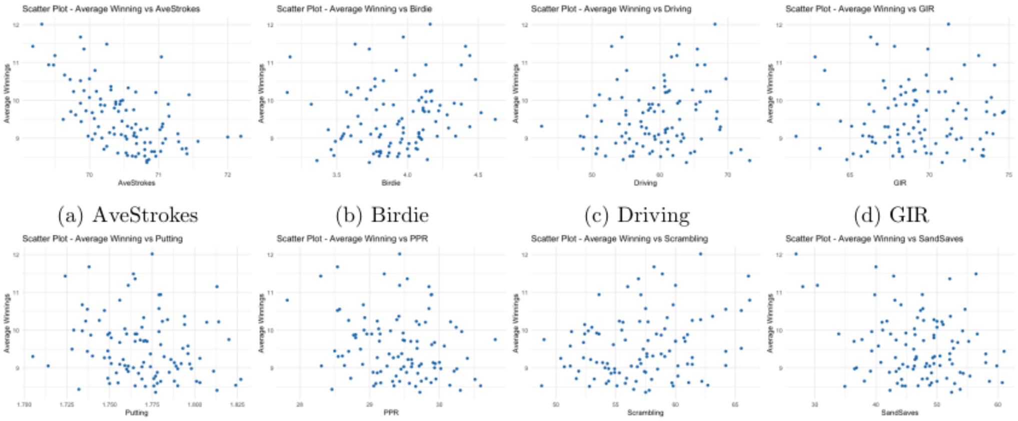


Figure 1: Correlation Matrix of Performance Metrics

Thereafter, we can see important correlations(Fig 1) between average winnings and golf performance indicators from the correlation matrix diagram. Average wins are negatively correlated with average strokes (-0.5796) and scrambling (-0.2539), indicating that bigger winnings are associated with superior scrambling and lower strokes. Also, we can see the average stroke has negative correlations with the following: putting (-0.1894), greens in



(a) AveStrokes (b) Birdie (c) Driving (d) GIR

(e) Putting (f) PPR (g) Scrambling (h) SandSaves

Figure 2: Scatter Plots - Average Winning vs other variables

regulation (-0.4682), and birdie rate (-0.6297). This means that fewer strokes are associated with better putting, higher birdie rates, and more greens in regulation. At the same time we can see a trade-off between scrambling and putting skill is implied by the negative correlation between scrambling and putting (-0.2146) and the positive connection between scrambling and birdie rate (0.0388). Adroit sand saving is associated with higher birdie rates and more greens in regulation, according to positive correlations found between sand saves and birdie rate (0.0619) and greens in regulation (0.2787). All things considered, through this research I have tried to clarify these complex linkages and have provided insightful information.

Now, I will try to plot scatter plot to spot relationship between Average winning and other variables.

The scatter plots(Fig 2)shows us that the numerous relationships that exist between various golf performance indicators and Average Winnings. The Average Winning vs. PPR plot shows a marginally positive association, indicating that players with higher PPRs may have marginally greater average winnings as a result of their ability to hit fairways and putt. Also, we see a minor negative association may be seen in the Average Winning vs. Scrambling plot, suggesting that players who scramble less frequently may have greater average winnings overall, perhaps as a result of making fewer mistakes overall. Aimilarly, we can spot a small positive association in the Average Winning vs. Sand Saves plot, suggesting that larger average winnings may be a result of better sand save percentages.

Furthermore, we can see that the Average Winning vs. Birdie Plot shows a somewhat positive correlation, indicating that players who make more birdies typically have higher average winnings as well. This could be because they are skilled in a variety of game-related areas. A small negative correlation between average winning and putting indicates that players with higher putting averages may also have somewhat higher average winnings, which could be a sign of general consistency. Also, a moderately positive correlation can be seen by us in the Average Winning vs. GIR plot, suggesting that hitting more greens in regulation is linked to better average winnings. This association may be the result of general skill competency. Lastly, we can see a modest positive association between average winnings and driving distance indicating that golfers who drive farther may have marginally better average winnings, maybe as a result of more scoring opportunities. Most importantly, we can see that Average Winning vs. AveStrokes plot shows a large negative association, suggesting that golfers who have lower average strokes typically have much higher average winnings. This is probably because they are more consistent and have better overall skills.

Thus, this analysis revealed us that average total winnings are positively correlated with several performance metrics, including greens in regulation, birdies, and scrambling. Additionally, some performance metrics, such as driving distance and sand saves, exhibited non-linear relationships with average total winnings.

5 Model Selection

I believe that the choice of a suitable modeling technique and pertinent variables is essential in building a predictive model for average winnings per golf 10 ment. The rationale behind the use of the linear regression approach was its simplicity, interpretability, and robustness. A simple framework for comprehending the link between independent variables (performance measurements) and the dependent variable (average winnings) can be provided by linear regression.

ok

I started by spliting the dataset into training and test sets in order to construct and evaluate the model. I believe this divide allowed for training the model on a subset of the data and assessing its performance on the remaining unseen data. I accomplished this using random sampling and setting a seed for repeatability (set.seed(123)). This procedure aids in evaluating the predictive capability of the model and its capacity to generalize to new facts.

Thereafter, feature selection was carried out using stepwise regression, optimizing the model by iteratively adding or removing variables until the Akaike Information Criterion (AIC) was minimized. The final model included variables such as AveStrokes, GIR, Birdie, Scrambling, and PPR, each contributing to the predictive power of the model.

$$\text{LR Model1: } \log(avg\_winning) = 16.34 - 0.19 \times AveStrokes - 0.02 \times GIR \\ - 0.08 \times Birdie - 0.01 \times Scrambling + 0.06 \times PPR$$

model

By interpreting the coefficients, it was possible for me to see that golfers who use fewer strokes per round typically win more money. Also, a decrease in AveStrokes was linked to an increase in log (average wins). Similar relationships were found between higher log (average winnings) and variables including higher GIR, more birdies, effective scrambling, and better putting (higher PPR). Through the entire model we can see that average winnings are typically higher for golfers who do well in these performance parameters.

Step	**Variables Included in Model**	**AIC**
1	AveStrokes + Driving + GIR + Putting + Birdie + SandSaves + Scrambling + PPR	-446.36
2	AveStrokes + GIR + Putting + Birdie + SandSaves + Scrambling + PPR	-448.35
3	AveStrokes + GIR + Birdie + SandSaves + Scrambling + PPR	-450.3
4	AveStrokes + GIR + Birdie + Scrambling + PPR	-450.58

Table 6: Stepwise Selection Process: 1st Model

Variable	**Df**	**Sum of Sq**	**RSS**	**AIC**
< *None* >	–	–	0.24654	-450.58
PPR	1	0.010423	0.25696	-449.27
Scrambling	1	0.013398	0.25994	-448.35
Birdie	1	0.014906	0.26144	-447.88
GIR	1	0.050722	0.29726	-437.61
AveStrokes	1	0.251865	0.4984	-396.27

Table 7: Step: AIC=-450.58

Table 7 shows us the AIC column(Akaike Information Criterion), RSS column indicating unexplained model variance, the Df column indicating degrees of freedom, and the SS column providing variable-explained variance. Here, the best model (Step: AIC=-450.58) has the lowest AIC of all the options. Since it fits the data better than other models, as evidenced by its lower AIC. It is the model of choice for describing the patterns that have been seen.

I therefore divided the data into training and test sets, performed step-by-step feature selection using R-code, and optimized the model based on AIC. The final model incorporated a selection of features and demonstrated statistical significance in forecasting average winnings, with multiple R-squared values indicating a considerable amount of variation explained. In summary, AveStrokes, GIR, Birdie, Scrambling, and PPR are significant predictors of success in professional golf tournaments, according to the results of the linear regression model.

6 Model Evaluation

I believe the analysis of the linear regression model shows encouraging outcomes for a number of parameters, highlighting how well it predicts average winnings. With a mean squared error (MSE) of 2.076, the model's predictions and the actual data are closely aligned, with a small average squared difference. The R-squared value, which indicates us the proportion of variance accounted for by the independent variables, is significantly elevated at 0.540. This implies that the model's predictive power is strengthened by effectively capturing a sizable percentage of the variability in the dependent variable.

Variable	Estimate	Std. Error	t value	Pr(> $|t|$)
(Intercept)	16.343	1.778	9.192	7.23×10^{-14}
AveStrokes	-0.193	0.022	-8.695	6.28×10^{-13}
GIR	-0.023	0.006	-3.902	0.000208
Birdie	-0.081	0.038	-2.115	0.037779
Scrambling	-0.005	0.003	-2.005	0.048579
PPR	0.058	0.033	1.769	0.081060

LSE

Table 8: Regression1 R

interpretation of estimated effects (beta hats)

Summary Statistics - Regression1 Results :

- Residual standard error: 0.05772 on 74 degrees of freedom
- Multiple R-squared: 0.5401
- Adjusted R-squared: 0.509
- F-statistic: 17.38 on 5 and 74 DF
- p-value: 2.41×10^{-11}

Although there may be overfitting in our model given that the adjusted R-squared value of 0.509 is marginally less than the R-squared value, these worries are allayed by the marginal difference between the two values. This shows us that the model fits the training data in a balanced way without going overboard. This combination of a low mean square error (MSE) and a high R-squared value indicates us that the linear regression model is generally a strong match for the dataset, and it can produce accurate predictions.

Thereafter, I try to address the impact of outliers in the training data, by implementing a robust method to improve a regression model's performance. Firstly, I use the boxplot.stats() function to find outliers, which makes it possible to look closely at the residuals from the final model. These outliers are then methodically eliminated from the training dataset in order to lessen any possible impact on the model's learning procedure. Then, I do stepwise variable selection to retrain the model using the improved data in order to maximize its prediction ability.

Step	**Variables Included in Model**	**AIC**
1	AveStrokes + Driving + GIR + Putting + Birdie + SandSaves + Scrambling + PPR	-452.82
2	AveStrokes + Driving + GIR + Putting + Birdie + SandSaves + PPR	-454.63
3	AveStrokes + GIR + Putting + Birdie + SandSaves + PPR	-454.63

Table 9: Stepwise Selection Process: 2nd Model

Variable	Df	Sum of Sq	RSS	AIC
$< none >$	-	-	0.20961	-454.63
Putting	1	0.007112	0.21672	-453.99
Birdie	1	0.009829	0.21944	-453.01
SandSaves	1	0.011091	0.22070	-452.55
PPR	1	0.042079	0.25169	-442.17
GIR	1	0.076265	0.28587	-432.11
AveStrokes	1	0.255655	0.46526	-393.63

Table 10: Step: AIC=-454.63

The model's performance measures show us a significant improvement as a result of this approach; in particular, the mean squared error (MSE) decreased from 2.076 to 1.914 and the R-squared increased from 0.54 to 0.60. This, underlying relationship between the independent variables and the dependent variable (average winnings) is better represented by these improvements. I believe the model's robustness is further supported by the increased adjusted R-squared, which shows a better match to the data.

$$\begin{aligned}\text{LR Model2: } \log(avg\_winning) = {} & 15.8342684 - 0.1899833 \times AveStrokes \\ & - 0.0282698 \times GIR - 0.8558774 \times Putting - 0.0630633 \times Birdie \\ & - 0.0018861 \times SandSaves + 0.1228391 \times PPR\end{aligned}$$

I think the retrained model creates a more accurate and dependable representation of the data by methodically addressing the impact of outliers, which improves predictive performance and provides a better knowledge of the variables impacting average winnings in the particular context. In summary,the linear regression model is a trustworthy resource for comprehending and forecasting average winnings in the specified context because of its low mean square error (MSE), high R-squared value, and negligible overfitting problems.

7 Results

I believe in the linear regression model, the relationship between the predictor variables and the response variable (average winnings) can be better understood by analyzing the model's output. The expected impact of each predictor variable on the average winnings is shown by its coefficients.

Table 11: Regression2 Results

Variable	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	15.8342684	1.4810856	10.691	< 2e-16 ***
AveStrokes	-0.1899833	0.0202733	-9.371	4.27e-14 ***
GIR	-0.0282698	0.0055233	-5.118	2.47e-06 ***
Putting	-0.8558774	0.5475781	-1.563	0.122432
Birdie	-0.0630633	0.0343213	-1.837	0.070271 .
SandSaves	-0.0018861	0.0009663	-1.952	0.054846 .
PPR	0.1228391	0.0323102	3.802	0.000298 ***

Summary Statistics - Regression2 Results :

- Residual standard error: 0.05396 on 72 degrees of freedom
- Multiple R-squared: 0.6057
- Adjusted R-squared: 0.5728
- F-statistic: 18.43 on 6 and 72 DF
- P-value: 7.554×10^{-13}

interpretation of estimated effects (beta hats)

The above regression results shows us that, when all other predictors are zero, the intercept of 15.83 represents the expected average winnings. We can see as the average number of strokes per round rises, the average winnings are anticipated to fall, according to the negative coefficient for 'AveStrokes' (-0.19). In a similar vein, lower average winnings are shown by the negative coefficients for "GIR" (-0.03), "Putting" (-0.86), "Birdie" (-0.06), and "SandSaves" (-0.00). Also, as seen by the positive coefficient for 'PPR' (0.12), higher average winnings are correlated with more points earned every round. T-tests are used to determine the statistical significance of these coefficients, and the resulting p-values are given. So, it can be concluded that the variables "AveStrokes", "GIR" and "PPR" significantly affect average wins because they are statistically significant ($p < 0.05$). Marginal significance is shown by the variables "Putting", "Birdie", and "SandSaves" ($p < 0.1$).

Table 12: Mean Squared Error and Root Mean Squared Error on Training and Testing Data

Metric	Training Data	Testing Data
Mean Squared Error(MSE)	2.07509321861491	2.09396798659345
Root Mean Squared Error(RMSE)	1.440518385379	1.44705493558242

I believe the model's mean squared error (MSE) and root mean square error (RMSE) figures(Table 12) demonstrate its excellent predictive performance. We can se that the MSE of 2.075 for the training data denotes an average prediction error of roughly 2.075 units, and the RMSE of 1.441 denotes an average departure from the dependent variable of about 1.441 units. Similarly, we can see that the MSE of 2.094 and RMSE of 1.447 on the testing data indicate respective average errors of roughly 2.094 and 1.447 units. Overall, these findings indicates us that the model performs well in forecasting the dependent variable, or average winning, with both the MSE and RMSE values showing the model's accuracy in doing so.

The residual standard error of 0.05396 indicates us that the model correctly reflects the variation in the data, according to the residual analysis. Also with a multiple R-squared of 0.6057, we can see that the model can explain the variation in average winnings. With the number of predictors taken into consideration, the modified R-squared yields a value of 0.5728. The model as a whole appears to be statistically significant based on the F-statistic of 18.43 and the corresponding p-value of 7.554e-13.

The residual plots(Figure 3) demonstrates us the applicability of the linear regression model. A random distribution around zero is shown in residuals vs fitted values, indicating us a well-fitting model. Q-Q residuals shows us a fit to the normal distribution when they are closely aligned with the diagonal line. Leverage versus residuals shows us no alarming trends. The model's validity is confirmed by standardized residuals, which have a mean of zero and a standard deviation of 1. All in all, these plots support the linear regression model by showing us that it is suitable for the provided data and does not appear to violate any assumptions.

To sum up, using the above performance criteria, the linear regression model offers a statistically significant framework for estimating average winnings in golf events.

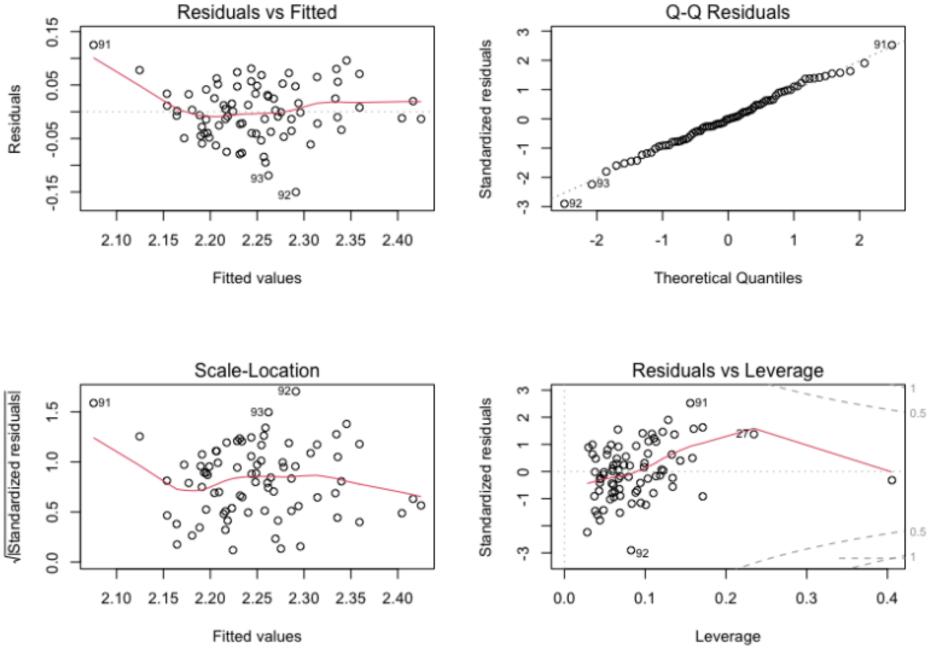


Figure 3: Linear Regression Plot

8 Conclusion

Our research shows that professional golfers' average total winnings may be accurately predicted using linear regression with the model producing high R-squared value and a low mean square error, indicating strong predictive ability. However, a bit disproportionate results are obtained from Residual vs Fitted plot, Q-Q Residual plot signifying it might be better for us to use other regression model and techniques like random forest etc.

Thus, the study's conclusions demonstrates us the promise of data-driven techniques for prediction in golf. To increase prediction accuracy, I believe more in-depth models like random forests or neural networks might be the subject of future study. Furthermore, I think it would be better to add other data sources like player demographics and tournament circumstances, can help to improve the scope of analysis.

References

[1] The Pennsylvania State University. *Stat 462: Lesson 2: Simple Linear Regression (SLR) Model.* [Online]. Available from: https://online.stat.psu.edu/stat462/https://online.stat.psu.edu/stat462/

[2] Zhao, J., and Z. Cao. "A Bayesian hierarchical model for predicting golf scores." *International Journal of Forecasting* 35, no. 1 (2019): 110-121.

[3] Woisin, W. "Using machine learning to predict the winning score of professional golf events on the PGA Tour" [Online]. Available from: http://norma.ncilibrary.ie/full-record.asp?r=6301697NORMA@NCI Library - National College of Ireland.

[4] Akmal Rafiq. "Data Visualisation in R" [Online]. Available from: https://akyrafiq.github.io/datavizR.htmlhttps://akyrafiq.github.io/datavizR.html

[5] 7 Exploratory Data Analysis. Available from: https://r4ds.had.co.nz/exploratory-data-analysis.htmlhttps://r4ds.had.co.nz/exploratory-data-analysis.html

K23038233_Coursework report.pdf

ORIGINALITY REPORT

8%	5%	3%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to King's College Student Paper	2%
2	Submitted to National College of Ireland Student Paper	1%
3	Ren, Cai, Du. "Surface Heterogeneity-Involved Estimation of Sample Size for Accuracy Assessment of Land Cover Product from Satellite Imagery", Sensors, 2019 Publication	1%
4	academic-accelerator.com Internet Source	<1%
5	rstudio-pubs-static.s3.amazonaws.com Internet Source	<1%
6	www.scribd.com Internet Source	<1%
7	Submitted to University of Hertfordshire Student Paper	<1%
8	Submitted to University of Maryland, Global Campus Student Paper	<1%

9	Submitted to Aston University Student Paper	<1%
10	Submitted to University of Greenwich Student Paper	<1%
11	cocolevio.com Internet Source	<1%
12	Submitted to Hult International Business School, Inc. Student Paper	<1%
13	Submitted to Harrisburg University of Science and Technology Student Paper	<1%
14	Submitted to University of Chichester Student Paper	<1%
15	Selim, H.. "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network", Expert Systems With Applications, 200903 Publication	<1%
16	purehost.bath.ac.uk Internet Source	<1%
17	5dok.net Internet Source	<1%
18	Andre Amaral, Taysir E. Dyhoum, Hussein A. Abdou, Hassan M. Aljohani. "Modeling for the	<1%

	Relationship between Monetary Policy and GDP in the USA Using Statistical Methods", Mathematics, 2022 Publication	
19	www.annsincap.org Internet Source	<1 %
20	Submitted to Macquarie University Student Paper	<1 %
21	Juan-Manuel Trujillo-Torres, Hassan Hossein-Mohand, Melchor Gómez-García, Hossein Hossein-Mohand et al. "Estimating the Academic Performance of Secondary Education Mathematics Students: A Gain Lift Predictive Model", Mathematics, 2020 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

K23038233_Coursework report.pdf

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

90/100

PAGE 1

Text Comment. best to have table of contents on a separate page

PAGE 2

💬 **Comment 1**

why use domain ? Just say : in professional golf

💬 **Comment 2**

It's the way around ! you what to predict their average winning from the way they play

💬 **Comment 3**

it's not that large -- just say data set

QM **ok**

PAGE 3

QM **ok**

Additional Comment

But .... it would be better to leave this to the end ... That's because you are looking at multiple regressions and these are more advanced methods.

QM **ok**

Comment 4

tables ... show

PAGE 4

QM

ok

QM

ok

Comment 5

found -- typo

Comment 6

explain what it means on terms of heavy tails

Comment 7

explain what it means, in terms of heave tails

PAGE 5

QM

good

good

Comment 8

the linear connection ... not the complex

PAGE 6

QM

cannot read .. too small ... presentation is not good

You need to think of the person reading your work. Can the read it.? They should be able to without need to zoom in !

QM

ok

Comment 9

should explain the colour coding

PAGE 7

QM

ok

PAGE 8

Comment 10

explain

QM

ok

Comment 11

just to be clear stepwise does not do an exhaustive search of models - so it doesn't 'optimise' as such

QM

model

You have not written down the model you are fitting to the data. That is : response variable = intercept + beta * explanatory variable + ...+ beta * explanatory variable + error, where errors are iid from normal (0, sigma^2)

PAGE 9

Comment 12

Keep the fond consistent!!!

PAGE 10

QM

LSE

You have not written down the estimated least squares equation. You need to write it down based on the estimated coefficients and the variables in your data set

QM

interpretation of estimated effects (beta hats)

You have fitted your model to the data and obtained the coefficient estimates (the betas- effects of the predictors on the response variable). However, you have not provided an interpretation of these values. How does the response variable change if the predictor changes by a unit?

PAGE 11

Comment 13

keep font consistent

PAGE 12

QM

interpretation of estimated effects (beta hats)

You have fitted your model to the data and obtained the coefficient estimates (the betas- effects of the predictors on the response variable). However, you have not provided an interpretation of these values. How does the response variable change if the predictor changes by a unit?

PAGE 13

Comment 14

Explain what mean square error and root mean square measure and how they are calculated

PAGE 14

PAGE 15