Statistics for Data Analysis - Assessment B

Yongxin Cai

December, 2022

Introduction

We are dealing with golfer's data set. This data contains top 100 European tour players in 2018 (Source: http://www.europeantour.com). For each golfer, we have their first name, last name, and 11 variables which indicate their ability in golf. The ultimate goal of this report is to predict the average winnings per tournament of each professional golfers on the European Tour, and in particular to explore the factors that may influence this, so we are going to find out the relationship between their average winnings and their ability in different aspects of game.

Methods

1. Using numeric summary and graphical summary to explore data. Since all variables except names are numeric and have no missing values, basicStats function is used for the descriptive statistics, boxplot is used to observe the variation and extreme value, and both histograms and density plot are used to find out the distribution of each variables.
2. Calculating the correlation coefficient of each pair of two variables, combining this coefficients with scatter plot, so that we could see the relationship between each variables, and make transformation for non-linear related variables. This could also help with exploring any collinearity between predictor variables.
3. Clarifying the response variable and predictor variables, with two assumptions (relationship between response variable and predictor variables are linear, the error terms are random variables with $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$), fitting all predictor variables into a linear model.
4. Checking the collinearity in the full model, removing any predictors with high statistical evidence of collinearity.
5. Variable selection, this involves backward elimination, forward selection, stepwise regression, exhaustive search. Through compare the p-value of each variable and criterion of each model to work out a best fitted model.
6. Working with the selected model, checking its assumption. Drawing the residual-fitted plot and qqplot to check the linearity of the data and the assumptions about error variables.
7. Predicting. Randomly pick a golfer and construct 95% confidence interval of the response variable to predict his/her average winning per tournament.
8. Conclusion. Summarising key findings, reflextions and limitations.

1. Exploratory Analysis

1.1 numeric summary and graphical summary

	TotalWinnings	Tournaments	AveStrokes	Driving	GIR
Minimum	342715.00	20.00	69.18	42.60	61.60
Maximum	4635909.00	90.00	72.19	73.20	74.70
Range	4293194.00	70.00	3.01	30.60	13.10
1. Quartile	453026.50	44.00	70.08	55.08	67.35
3. Quartile	1095418.25	72.00	70.83	62.95	71.33
IQR	642391.75	28.00	0.75	7.88	3.98
Mean	926390.80	57.32	70.48	59.41	69.18
Median	624680.50	62.50	70.47	59.55	69.05
Stdev	722693.16	18.71	0.56	5.84	2.92
Skewness	2.27	-0.45	0.25	-0.08	-0.15
Kurtosis	6.54	-1.01	0.18	-0.21	-0.43

Table 1: Numerical summary table for golfer's data 1-2

There are 10 numeric variables and 2 character variables in the golfer data. The following summary is about the numeric variables, based on the statistical (Table 1 & 2) and graphical representations (Appendix Figure 2 & 3 & 4) of the data.

- **TotalWinnings**: The golfer's total winnings in 2018 on the European Tour. Total winnings has a large range 4293194, compared to its range it has a relatively narrow IQR, the middle 50% of total winning values has a range 642391.75. It can also be observed from its box plot, the box has a narrow shape and 7 extreme values. These 7 people has winned too much compared to the others, their winnings exceed more than 1.5 times than the middle 50% of total winning values. They are Francesco M, Patrick R, Tommy F, Rory M, Alex N, Thorbjorn O, Jon R. We will pay more attention to these people when fit in the model. Due to the large range and IQR, total winning also have a large value of standard deviation, it spreads widely. Compare to its mean and median, the skewness, also from the density plot, we could see a right skewed distribution. From its excess kurtosis, it has a large value compared the kurtosis of normal distribution, which means it has a heavier tail than normal distribution.
- **Tournaments**: The number of tournaments in which the golfer competed in 2018 on the European Tour. Tournaments is a discrete variable from minimum 20 to maximum 90. It has a IQR 28, which means 50% of golfer completed 44 to 72 tournaments on the European Tour. It has a higher median compared to its mean, and a negetive skewness value which could be observed from its boxplot, number of tournaments are more concentrated at the bottom of the box. According to its histogram and density plot, its value is quite spread out and does not have the bell shape, so not normal distributed.
- **AveStrokes**: The average number of strokes taken per round. AveStrokes has a narrow range (72.19, 60.18), the middle 50% data concentrates between 70.08 and 70.83. It has a equal mean and median, and a small value of skewness (0.25, nearly 0). By looking at its boxplot and denstiy plot, it can be observed that it is likely to follow a normal distribution, it has a bell shape and it is symmetric, especially its excess kurtosis is just 0.18 higher than a normal distribution. However, from its boxplot, there are two extreme values, which come from Yusaku M and Alvaro Q, they have more than 1.5 times higher average number of strokes compared to the middle 50% average number of strokes.
- **Driving**: Driving Accuracy is the percentage of time a player is able to hit the fairway with his tee shot. Driving accuracy has a range between 42.6 and 73.2, and its middle 50% data concentrates in (55.08, 62.59). It has a equal mean and median, from its boxplot, we can see the symmetry of the box. It also has a extreme value, which appears in Paul D, who got 42.6 in driving accuracy, which is less than 1.5 times of middle 50% of the driving accuracy values. From its density plot, it can be obserced that it is symmetric (-0.08 in skewness) and in a bell shape, it has a fat tail (-0.21 in excess kurtosis) and with standard deviation 5.84, a little bit spread out.

- **GIR**: Greens in Regulation is the percentage of time a player was able to hit the green in regulation. GIR has a range between 61.6 and 74.7, its middle 50% data concentrates in (67.35, 71.33). It has a equal mean and median, a symmetric box in the boxplot. And a low value of standard deviation 2.92, indicating a narrow spread of its data. Its density plot has a peak in the middle and it has a low skewness -0.15, more data are concentrated above 70, so it has a relatively fatter tail on RHS, it can also be observed from its excess kurtosis, it has value of -0.43, which shows a fat tail compared to a normal distribution.

	Putting	Birdie	SandSaves	Scrambling	PPR
Minimum	1.71	3.15	26.90	48.80	27.82
Maximum	1.83	4.62	61.00	66.20	30.80
Range	0.12	1.47	34.10	17.40	2.98
1. Quartile	1.75	3.76	43.17	54.27	29.06
3. Quartile	1.79	4.12	51.73	59.85	29.82
IQR	0.03	0.37	8.55	5.58	0.76
Mean	1.77	3.93	46.97	57.17	29.42
Median	1.77	3.94	47.60	57.20	29.45
Stdev	0.02	0.29	6.73	3.94	0.57
Skewness	0.01	-0.20	-0.44	0.19	-0.15
Kurtosis	-0.20	-0.12	0.36	-0.49	-0.25

Table 2: Numerical summary table for golfer's data 2-2

- **Putting**: Putting Average measures putting performance on those holes where the green is hit in regulation (GIR). Putting average has a narrow range (1.71, 1.83), also a narrow IQR (1.75, 1.79), which cause a low standard deviation 0.02. It has a equal mean and median, and a low value in skewness. From its boxplot, we can see a symmetric box lie on the middle. And it also have a symmetric bell shape in its density plot. Its excess kurtosis shows it is very close to a normal distribution.
- **Birdie**: The number of birdies per round. Birdie has maximum 4.62 and minimum 3.15, its middle 50% data is in the interval (3.76, 4.12). Its has a equal mean and median, which can be also observed from its symmetric boxplot, however it has a negative skewness, refers to a fatter tail on the left side of the distribution. This fatter tail on LHS may come from its two extreme points, 3.15, 3.17 from Matt K and Xander S. They got a lower value in number of birdies, 1.5 times than the middle 50% birdie data.
- **SandSaves**: the percentage of time a player was able to finish the hole in two shots from a green-side sand bunker. SandSaves a range of 34.10, its middle 50% data has a lower IQR, they are concentrated in (43.17, 51.73). Its mean is slightly lower than its median which causes a negaetive skewness -0.44, we can also see a slightly left skewed distribution in the density plot. Its excess kurtosis indicates it has a fatter tail than a normal distribution. Its boxplot shows there are two extreme values, 26.9 from Francesco M and 28.0 from Xander S, which means they got a lower sand save percentage, 1.5 times lower than the middle 50% SandSaves data.
- **Scrambling**: the percentage of time that a player misses the green in regulation. Scrambling has a range 17.40 and IQR 5.58. Its mean and median are equal, a symmetric box can be seen from its boxplot. It has a slightly positive (0.19) skewed distribution, and a thinner tail (excess kurtosis -0.49) compared to a normal distribution.
- **PPR**: Putts Per Round is the average total number of putts per round. PPR has a large value (27.82, 30.80) but a small range 2.98 and also a small IQR 0.76. It has a equal mean and median, a symmetric box can be observed from its boxplot. There is a extreme point, which is the minimum 27.82 from Cameron S, who has a 1.5 times less than the middle 50% PPR data. Its density plot indicates a sightly left skewed distribution (skewness -0.15) and a thinner tail than normal distribution (excess kurtosis -0.25).

Since we are interested in the average winnings per tournament of each golfer. We add a new column call avg_win_per_tour to the dataset, this column uses TotalWinnings divided by Tournaments, i.e. the mount

of total winnings divided by the number of tournaments for each golfer. We could also make a numeric and graphical summary for this continuous variable.

Min	Max	Q1	Q3	Mean	Median	sd	Skewness	Kurtosis
4245.86	165568.2	7408.03	22233.27	21165.5	11407.49	25708.12	3.07	11.1

Table 3: Numerical summary table for average winnings per tournament

- **avg_win_per_tour**: it has a wide range from 4245 to 165568, and a wide IQR 14825, which causes a high value of standard deviation. From its boxplot in *Figure 3*, it can be observed there are 9 extreme value which are all 1.5 times higher than the middle part of data. The box is not symmetric and it also reflects on the density plot, the distribution is clearly not normal. It has a higher value in median than mean, and this causes a significant right skewness.

1.2 Association between variables

Now, calculate the correlation coefficient between continuous variables, combine the coefficient with the scatterplot (Appendix Figure 5). We can observe that there are only 4 pairs which have linear relationship: Birdie has a negative linear relationship with AveStrokes (-0.630) and also with Putting (-0.649). PPR has a positive linear relationship with GIR (0.649) and also with Putting (0.771). So we should be aware of these pairs when fit them in the model to avoid collinearity. There is no other relationship for rest of variables, particularly, there is no linear relationship between avg_win_per_tour and other vairables.

2. Fit in model

2.1 Variables and assumptions

Since we are interested in predicting the average winnings per tournament of professional golfers, we let avg_win_per_tour be the response variable. By looking at its numeric summary and graphical summary, we notice that avg_win_per_tour has a wide range and spread out, and it is right skewed and not normally distributed. With these reason, I make a log transformation [1] to the response variable to help with its skewness. For the predictor variables, golfer's name does not affect the response variable theoretically, so we let the rest continuous variables be the possible predictor variables. Since the number of predictor variables is bigger than 1, we are going to fit these variable into a multiple linear model. Before doing this, we make two assumptions:

- The relationship between response variable and predictor variables is linear.
- There is an error vairable for each case and the error terms are random variables with $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

2.2 Variable sections

2.2.1 Test of all the predictors

Now, we fit all numeric variables into a multiple linear model with response variable log(avg_win_per_tour) and rest of others as predictor variables, i.e. $\log(avg\_win\_per\_tour) = \beta_0 + \beta_1 AveStrokes + \beta_2 Driving + \beta_3 GIR + \beta_4 Putting + \beta_5 Birdie + \beta_6 SandSaves + \beta_7 Scrambling + \beta_8 PPR + \epsilon$.

This full-model summary shows that there are 4 predictor variables which are significant to the response variable, they are AveStrokes, GIR, Birdie, SandSaves, and the intercept is also significant.

Then make a null model, there is no predictor variable in this linear model, i.e. $\log(avg\_win\_per\_tour) = \beta_0 + \epsilon$, and compare with the full model. We use anova to compare two model to see whether any of the predictors are useful in predicting the response variable. In terms of hypothsis tests about the β's:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_8 = 0$$

vs

H_α : At least one β_j for j = 1,...,8 is not equal to 0

From its extremely small p-value (4.961e-15) which is definitely less than the level of significance $\alpha = 0.05$, we have the statistical evidence to say that, we reject H_0, which means that at least one predictor is useful in predicting the response variable.

2.2.2 Collinearity

Before the variable selection, we should check the collinearity in the model, since it may affect the estimation of the regression parameters which make the estimated values of parameters not accurate.

From the scatter plot in previous exploratory analysis, we have noticed that there are 4 pairs predictor variables which have moderate linear relationship between each other, they are Birdie with AveStrokes and with Putting, PPR with GIR, Putting. Based on these 4 pairs, we are going to consider remove Birdie and PPR to avoid collinearity. But before removing any predictor variables, we calculate the variance inflation factor first for each predictor variables to measure the inflation of the standard error. We have the following result:

AveStrokes	Driving	GIR	Putting	Birdie	SandSaves	Scrambling	PPR
4.028299	1.720515	9.174723	9.490852	3.546780	1.193683	5.549496	24.621189

Table 4: variance inflation factor

For PPR, we have its VIF = 24.62 >10, which means that the estimated value of PPR is biased, PPR has high correlation with other predictor variables. As for Birdie, although it has moderate linear relationship with AveStrokes and with Putting, its correlation coefficient is not very big (below 0.65), and its VIF is not above 10, so we are going to only drop PPR from the full model. After dropping, we have a new model with 7 predictor variables, and let us check the VIF again:

AveStrokes	Driving	GIR	Putting	Birdie	SandSaves	Scrambling
3.995198	1.709126	3.203286	3.181857	3.420827	1.170422	2.085015

Table 5: variance inflation factor after dropping

After moving the biased parameter PPR, all VIF values are now below 10. The new model is $\log(avg\_win\_per\_tour) = \beta_0 + \beta_1 AveStrokes + \beta_2 Driving + \beta_3 GIR + \beta_4 Putting + \beta_5 Birdie + \beta_6 SandSaves + \beta_7 Scrambling + \epsilon$.

The adjusted R^2 has changed from 0.5603 to 0.5644, and AIC has reduced from 175.56 to 173.72. We also make the significant test by using anova to check whether there is a significant relationship between PPR in the linear regression model. The p-value we got is 0.7009, which is bigger than $\alpha = 0.05$. Thus, according to these statistical evidence, we could say that, PPR has no significant influence to the linear model and remove PPR could lead to a better fitted model.

2.2.3 Testing based procedures

We are going to find out which predictor variables can best predict the response variable by the rest based procedures. Let the new full model be $\log(avg\_win\_per\_tour) = \beta_0 + \beta_1 AveStrokes + \beta_2 Driving + \beta_3 GIR + \beta_4 Putting + \beta_5 SandSaves + \beta_6 Scrambling + \epsilon$.

We run forward selection, backward elimination and stepwise regression on the same data, and we got a same answer, the suggested best fit model is: $\log(avg\_win\_per\_tour) = \beta_0 + \beta_1 AveStrokes + \beta_2 GIR + \beta_3 Birdie + \beta_4 Scrambling + \beta_5 SandSaves + \beta_6 Putting + \epsilon$.

2.2.4 Criterion based procedures

There is also another method of variable selection. We are going to use best subsets regression and look at the different criterion value when we keep adding variables to the model. The order of adding variables to the null model is AveStrokes, GIR, Birdie, Scrambling, SandSaves, Putting, Driving. And every time we add one variable, we calculate the following criterion of the model: Mallow's C_p, BIC, and adjusted R^2. We are looking for the model with smallest value in Mallow's C_p, smallest value in BIC and the highest value in adjusted R^2. The result shows below:

	AveStrokes	GIR	Birdie	Scrambling	SandSaves	Putting	Driving
Mallow's C_p	54.899784	36.967992	23.650608	10.361709	6.663289	6.011099	8.000000
BIC	-31.73215	-41.29328	-49.12604	-58.70016	-59.94043	-58.17697	-53.58386
adjusted R^2	0.3291921	0.4117945	0.4751777	0.5397646	0.5612781	0.5689845	0.5643521

Table 6: Criterion of the model

We could observe that, when we add the sixth variable to the model, Mallow's C_p reaches the lowest and adjusted R^2 reaches the highest. BIC got its lowest value when the fifth variable is adding to the model. Based on this, we should compare the model with first 6 variables to the model with 5 variables to see if the sixth variable, i.e. Putting, is significant to the linear regression. Use anova to test the following models:

$$\log(avg\_win\_per\_tour) =$$

$$\beta_0 + \beta_1 AveStrokes + \beta_2 GIR + \beta_3 Birdie + \beta_4 Scrambling + \beta_5 SandSaves + \beta_6 Putting$$

$$vs$$

$$\log(avg\_win\_per\_tour) =$$

$$\beta_0 + \beta_1 AveStrokes + \beta_2 GIR + \beta_3 Birdie + \beta_4 Scrambling + \beta_5 SandSaves$$

We got p-value = 0.105 which is higher than the level of significant $\alpha = 0.05$. So we have the statistical evidence to say that, it does not have significant effect, we could drop Putting from the model. However, the model with Putting has a higher value of adjusted R^2, i.e. the probability of variability in response variable explained by its assumed linear relationship with predictor variables, which indicates this variable can help with building a better linear model, so we are going to keep it in the model then we have the fitted model which is same as the result from the testing based procedures. The model after variable selection is $\log(avg\_win\_per\_tour) = \beta_0 + \beta_1 AveStrokes + \beta_2 GIR + \beta_3 Birdie + \beta_4 Scrambling + \beta_5 SandSaves + \beta_6 Putting + \epsilon$.

3 Modelling

3.1 Check assumption [2]

Now, we have found a suggested regression model, we are going to check the assumption. The assumptions we made before are:

- There is a linear relationship between response variable and predictor variables.
- There is an error variable for each unit.
- The true mean of all error variables is equal to zero.
- All error variables have the same variance.
- The error variables are independent and identically distributed; and their distribution is normal distribution.

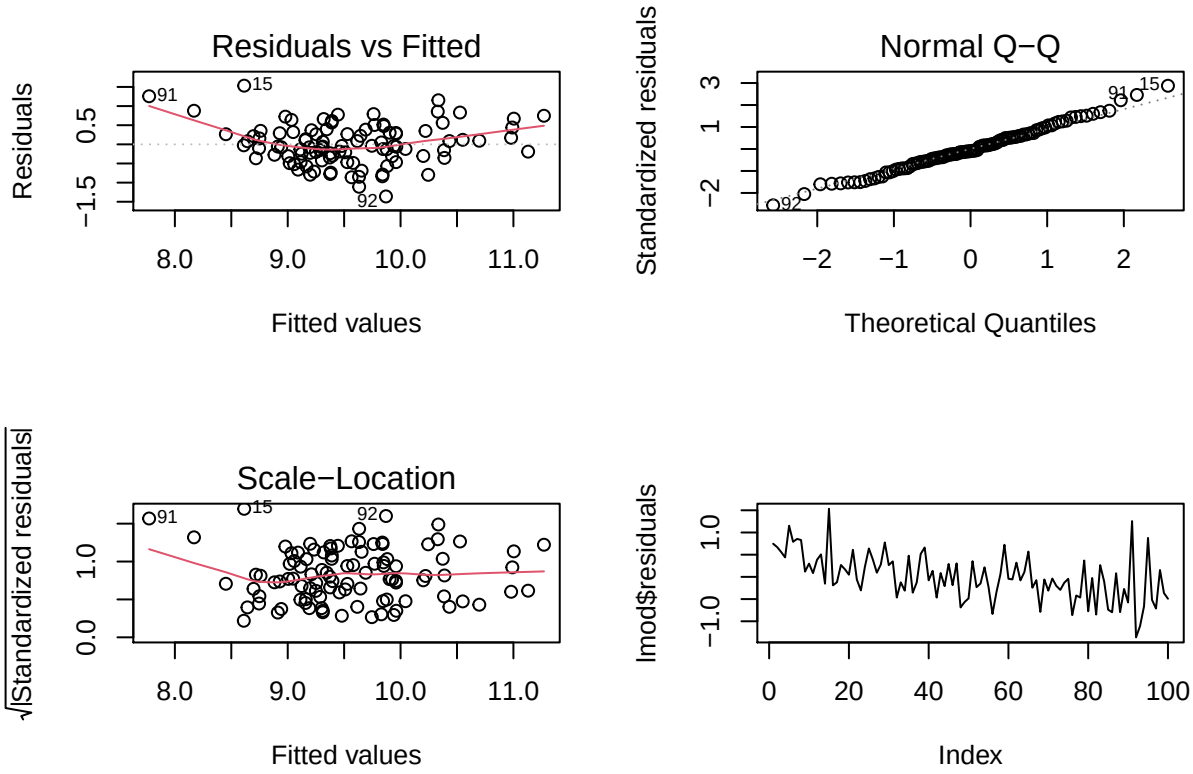

Figure 1: Diagnostic plot

To check the linearity of the data and zero mean of all error variables, we look at its residual-fitted plot. (*Figure 1*)

The residual-fitted plot shows the residual points may have patterns, they are more denser in the middle part, and the three outliers have influenced this plot a lot. The outliers are Brandon S, Alvaro Q, Jacques K, they are not fitting the model very well, especially, Alvaro Q has an extreme value in AveStrokes which may cause this outlier. They have influenced the residual-fitted plot a lot, so we are going to check whether they are influential points and then move these three points from the data. Also, we would ideally want to see the red line flat on 0, which would indicate that the residual errors have a mean value of zero. In plot, the red line is above 0 for low fitted values and high fitted values. In the middle part, the residual errors have a mean value of zero. We would then remove the outliers to see if the model fitted better.

To check the variance of the error variables, we look at its scale-location plot. (*Figure 1*)

From its scale-location plot, we could see that the residual points are not all equally spread out for low fitted values, it may be influence by its outliers. We could also use the Non-Constant Error Variance (NCV) Test. The p-value is 0.50, which is greater than the level of significance $\alpha = 0.05$, so we have a statistical evidence to say that, we fail to reject H_0 (H_0: variance of residuals is constant). So we meet the assumption of constant variance for all error variables.

Next, check whether the error variables are independent and identically distributed and their distribution is normal distribution, we look at its qqplot. (*Figure 1*)

From qqplot, we could see that the residual points are away from the black line for low theoretical quantiles and high theoretical quantiles, and the residual points are following the black line for the middle part. We could also use the Anderson-Darling normality test to check its normality. The Anderson-Darling normality test has p-value 0.9015, which is greater than the level of significance $\alpha = 0.05$, so we have a statistical evidence to say that, we fail to reject H_0 (H_0 : residuals are normally distributed). To check whether the residuals are correlated, we apply Ljung–Box test, the p-value we got is 0.068, which is greater than the level

of significance $\alpha = 0.5$, so we reject the null hypothesis, there is a correlation between the residuals and the model exhibit a lack of fit. We could also observe this from its residual-index plot (*Figure 1*), as there might be a trend to its residuals.

So we meet all the assumption except the assumption of independent and identically distributed distributed error variables. We have already done the transformation of response variable, and there is no sense to do another transformation to our predictor variables since all of them are almost normally distributed and they are not skewed. So the final model we have is $\log(avg\_win\_per\_tour) = \beta_0 + \beta_1 AveStrokes + \beta_2 GIR + \beta_3 Birdie + \beta_4 Scrambling + \beta_5 SandSaves + \beta_6 Putting + \epsilon$.

3.2 Prediction

Since there are 3 outliers which have bad fits in the linear model, also affect the diagnostic plot, we are going to check its Cook's distance and compare with the threshold 0.04 (Appendix figure 6). All 3 outliers in the diagnostic plot all have high values of Cook's distance which beyond the threshold, so we consider them as influential points, and we are going to remove them from the dataset. After dropping these 3 rows, we got a better model with higher adjusted R^2 0.6577. And all predictor variables become significant (all p-values less than level of significance $\alpha = 0.05$).

Now, we use the least square equation: $\log(avg\_win\_per\_tour_i) = \hat{\beta}_0 + \hat{\beta}_1 AveStrokes_i + \hat{\beta}_2 GIR_i + \hat{\beta}_3 Birdie_i + \hat{\beta}_4 Scrambling_i + \hat{\beta}_5 SandSaves_i + \hat{\beta}_6 Putting_i$ to make prediction where:

- i = 1, 2, 3, ... , 97
- $\hat{\beta}_0 = 161.26$, meaning that if a golfer attend a game, and all his performance is zero, i.e. he does nothing, he will get $e^{161.26} = 1.082251 * 10^{70}$. This is not making sense.
- $\hat{\beta}_1 = -2.040$, meaning that with other things being same, every one unit increase in AveStrokes will lead to a $e^{-2.04} = 0.130$ increase in average winning per tournament.
- $\hat{\beta}_2 = -0.187$, meaning that with other things being same, every one unit increase in GIR will lead to a $e^{-0.187} = 0.829$ increase in average winning per tournament.
- $\hat{\beta}_3 = -0.886$, meaning that with other things being same, every one unit increase in Birdie will lead to a $e^{-0.886} = 0.412$ increase in average winning per tournament.
- $\hat{\beta}_4 = -0.068$, meaning that with other things being same, every one unit increase in Scrambling will lead to a $e^{-0.0680} = 0.934$ increase in average winning per tournament.
- $\hat{\beta}_5 = -0.022$, meaning that with other things being same, every one unit increase in SandSaves will lead to a $e^{-0.022} = 0.978$ increase in average winning per tournament.
- $\hat{\beta}_6 = 7.580$, meaning that with other things being same, every one unit increase in Putting will lead to a $e^{7.580} = 1958.629$ increase in average winning per tournament.

Now, pick a random person (Jens D) to make the prediction. According to the fitted model, we got that Jens D will win 5328 per tournament. And we have 95% confidence interval that this person (Jens D) will win between 1942 and 14620 per tournament. Actually, in 2018 European Tour, this person won 7423.57, which is in the interval of our confidence interval.

4. Conclusion and Limitation

Conclusion

The final model is: $\log(avg\_win\_per\_tour) = \beta_0 + \beta_1 AveStrokes + \beta_2 GIR + \beta_3 Birdie + \beta_4 Scrambling + \beta_5 SandSaves + \beta_6 Putting + \epsilon$. We could make a prediction of average winning per tournament by input the golfer's AveStrokes, GIR, Birdie, Scrambling, SandSaves and Putting data. This model is a linear model, with error term in normal distribution, the error variables have mean of zero and constant variance.

Reflection

- This model has a problem with its error variables, the error variables are correlated, which suggests that there is additional information in the data that has not been exploited in the current model.
- Each predictor variable is an indicator of a golfer's ability, the higher these indicators are the better the golfer is, theoretically, a golfer with high ability will win more tournaments and prizes. Since we made a log transformation for response variable, it always causes an increase in average winnings. However, the estimate of all predictor variables except Putting are negative, which means that the golfer with less ability in AveStrokes, GIR, Birdie, Scrambling, SandSaves are going to win more in the tournament. This is unreasonable.
- After log transformation the response variable is not normally distributed, it may suggest that we could consider a generalised linear model. [3]
- The confidence interval for predicting average winning per tournament is too wide, The reasons for this may be a high variability.

Limitation

- Each row of data may be correlated, as these golfers may have taken part in many races at the same time, so that each individual's results may affect each other.

Appendix

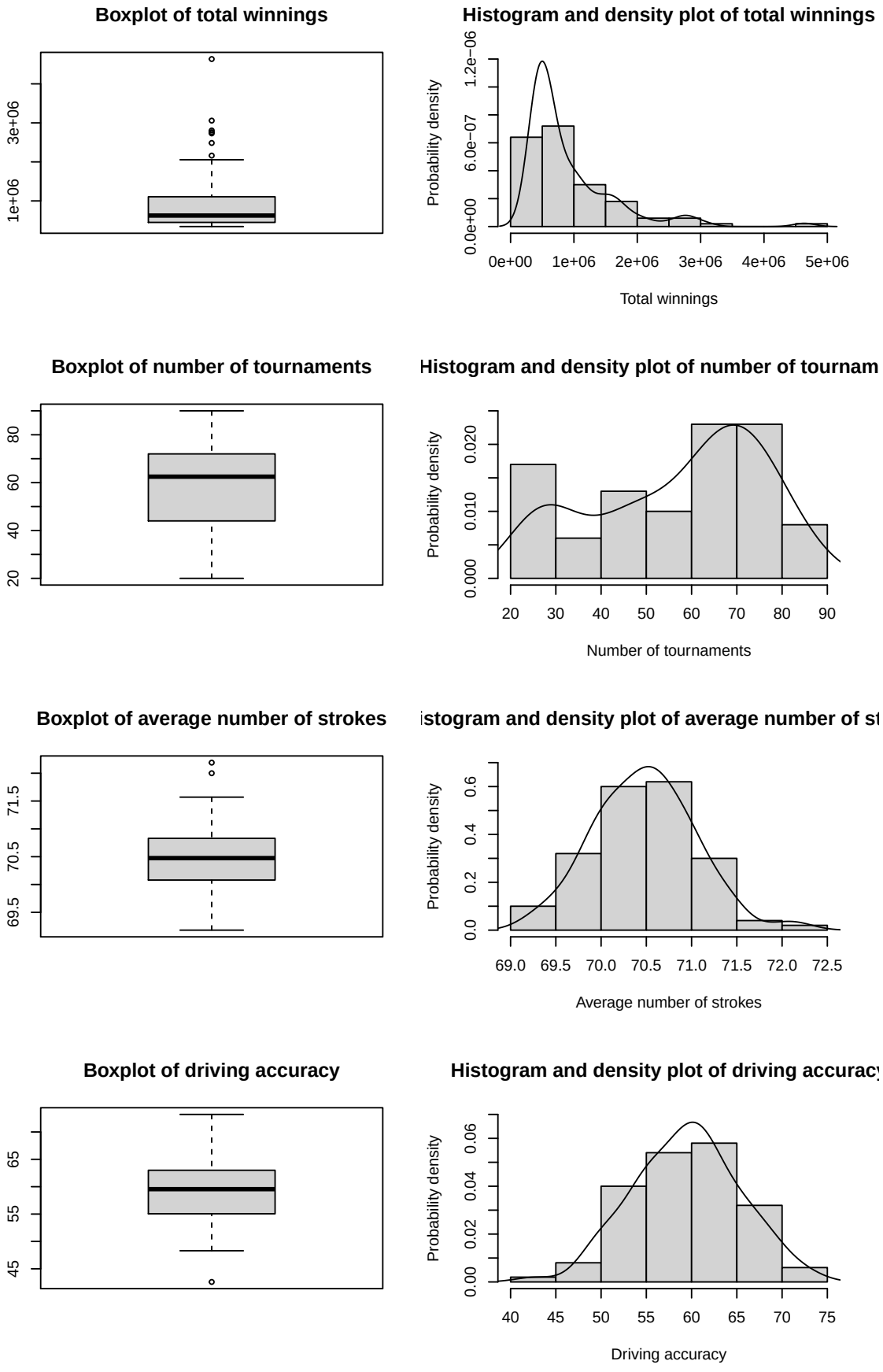

Figure 2: Boxplots and probability density plots of variables in golfer data 1-3

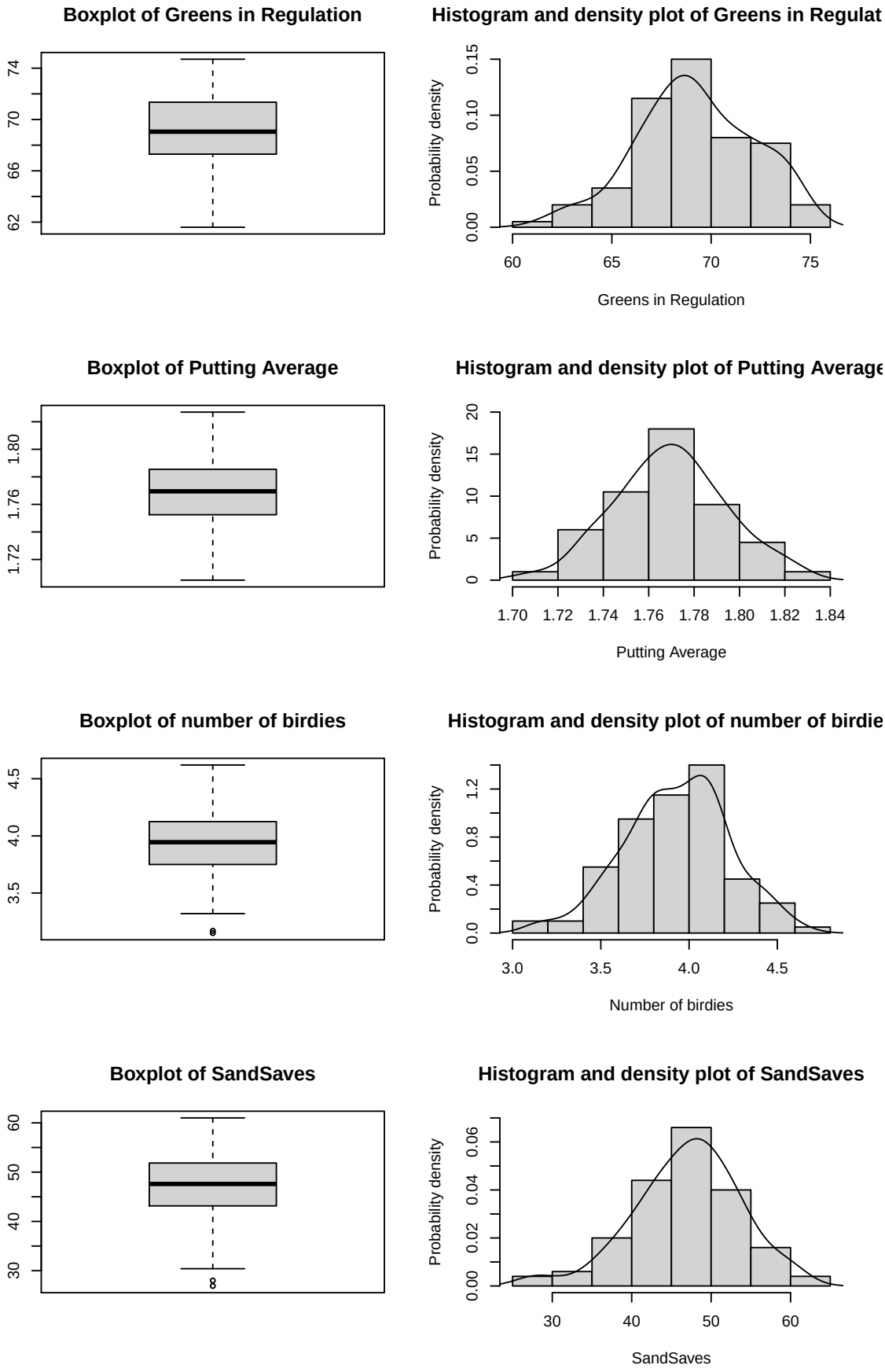

Figure 3: Boxplots and probability density plots of variables in golfer data 2-3

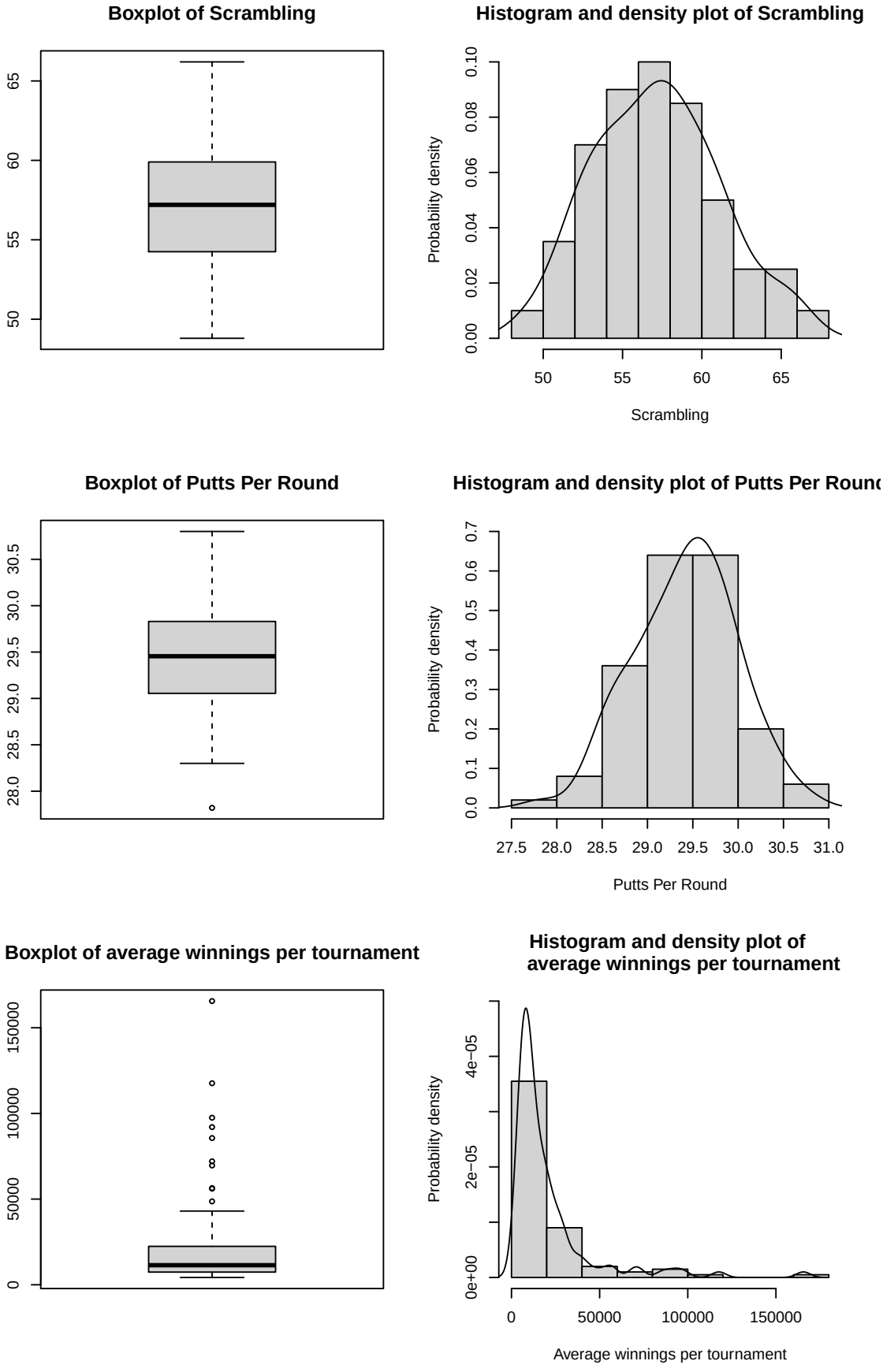

Figure 4: Boxplots and probability density plots of variables in golfer data 3-3

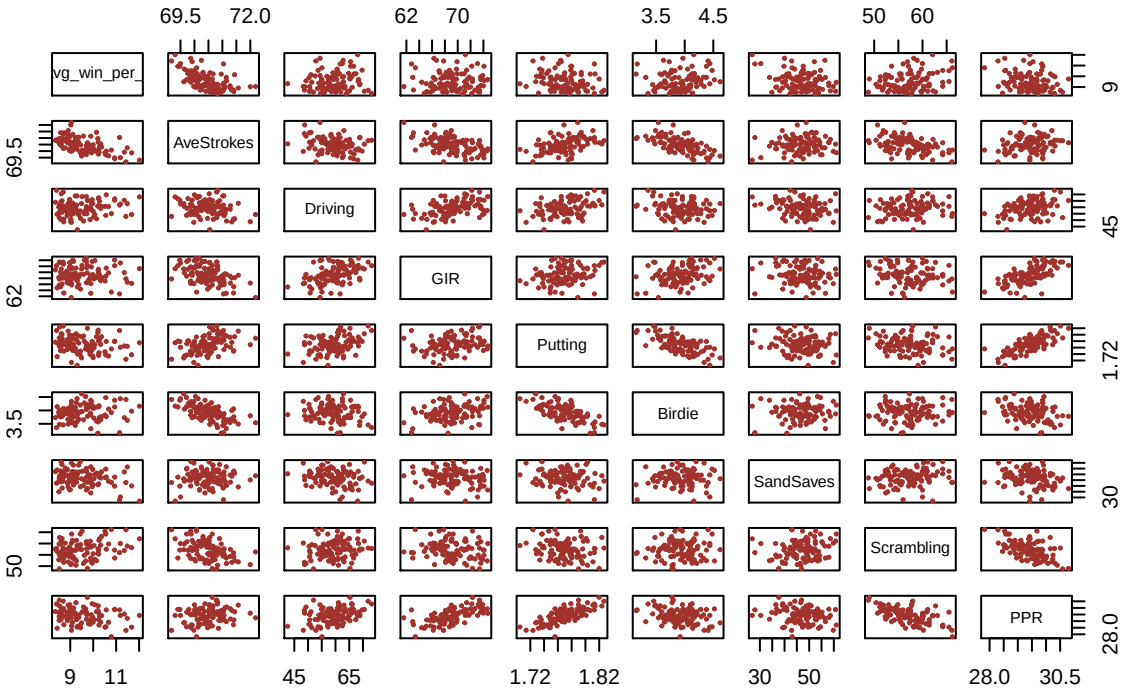

Figure 5: Scatterplot matrix for contiunous variable pairs on Golfer Data

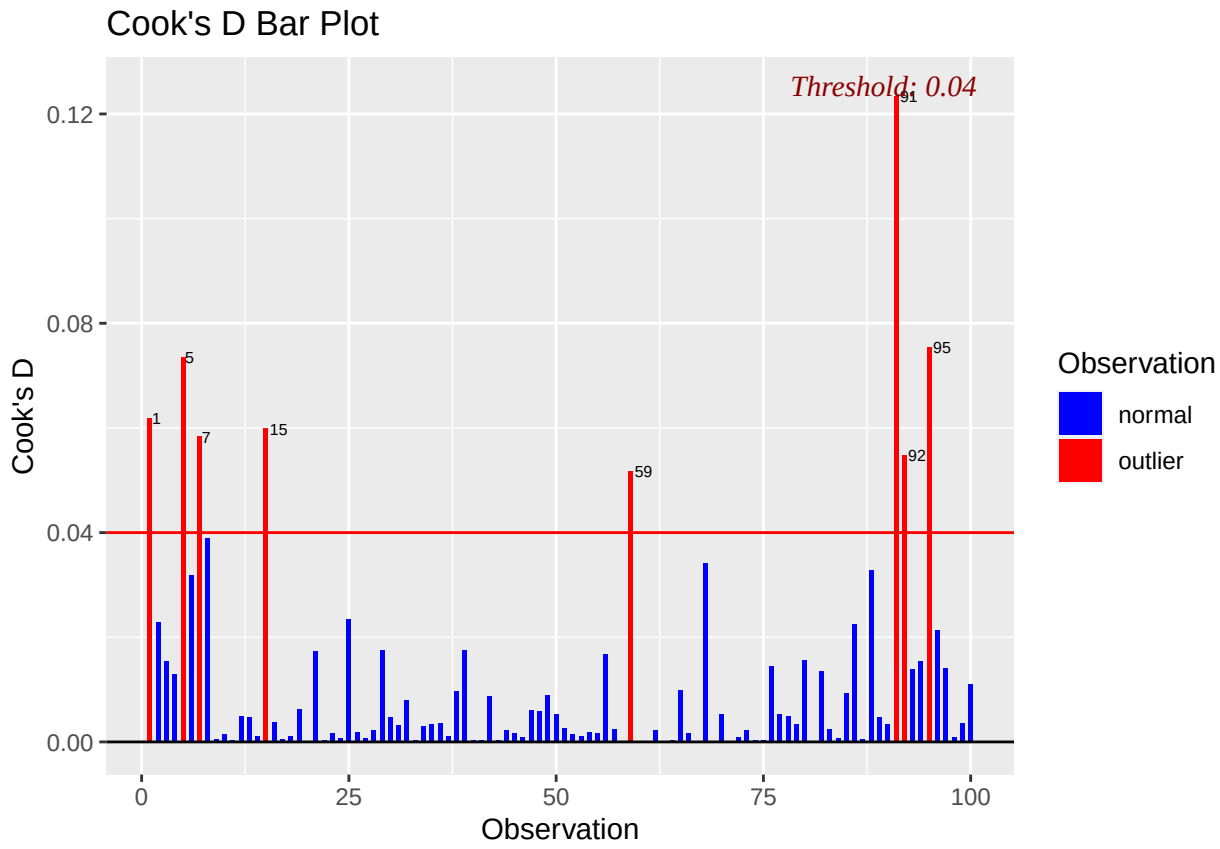

Figure 6: Cook's distance

Reference

[1] Log transformation, [Online] Avaliable: https://lmc2179.github.io/posts/multiplicative.html

[2] Assumption check, [Online] Avaliable: https://godatadrive.com/blog/basic-guide-to-test-assumptions-of-linear-regression-in-r

[3] Generalized Linear Models, [Online] Avaliable: https://vitalflux.com/generalized-linear-models-explained-with-examples/