COVER PAGE

STAT 608 Homework 03, Summer 2017

Please write your name and email address clearly below, then convert to PDF and attach as the first page of your homework upload.

NAME: RAJAN KAPOOR

EMAIL: r.kapoor@tamu.edu.

45/49

(1) $H_0 : \gamma_1 = \gamma_2$

$H_A : \gamma_1 \neq \gamma_2$

or

$H_0 : Y_j = \gamma_0 + \gamma_1 z_{j1} + \gamma_1 z_{j2} + \epsilon_j$

$H_A : Y_j = \gamma_0 + \gamma_1 z_{j1} + \gamma_2 z_{j2} + \epsilon_j$

$$F = \frac{RSS_{(reduced)} - RSS_{(full)}}{df_{reduced} - df_{full}} \div \frac{RSS_{full}}{df_{full}}$$

$df_{full} = 60 - 3 = 57$

$df_{reduced} = 60 - 2 = 58$

$F \sim F_{2,57}$

If F is statistically significant, we favor the full model and reject the null hypothesis. i.e. p-value less than threshold (eg 0.001).

(2)

(2.1) Unbiased estimate of the error variance

$$S^2 = \frac{RSS}{df} = \frac{100}{30-4} = \frac{100}{26} = 38.46$$

(2.2) Possible numerical value of $\hat{\beta}_1$

$$t_c = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} \sim t\text{-distribution with } df = 26$$

p-value = 0.02

$\Rightarrow P(T > t_c) = 0.02$

For p-value of 0.02, the critical value is given by

$$\Rightarrow t_c = 2.479 \quad \text{or} \quad t_c = -2.479$$

$$\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = 2.479 \quad \text{or} \quad \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = -2.479$$

$$\Rightarrow \hat{\beta}_1 = 2.479 \times 0.5 \quad \text{or} \quad \hat{\beta}_1 = -2.479 \times 0.5$$

$$\Rightarrow \hat{\beta}_1 = 1.2395 \quad \text{or} \quad \hat{\beta}_1 = -1.2395$$

(2.3) 95% CI for $\beta_2 = \hat{\beta}_2 \pm t\left(\frac{0.05}{2}, 26\right) \times se(\hat{\beta}_2)$

$$= 1 \pm 0.3385\,(0.25)$$

$$(0.915375,\ 1.084625)$$

(2.4) $R^2 = 1 - \frac{RSS}{SST} = 0.9$

$$\Rightarrow \frac{RSS}{SST} = 0.1 \Rightarrow SST = 1000$$

$$\Rightarrow RSS + SS_{reg} = 1000$$

$$\Rightarrow SS_{reg} = 900 \qquad (\because RSS = 100)$$

(2.5) $$F = \frac{SS_{reg}/p}{RSS/n-p-1}$$

$$= \frac{900}{3} \Big/ \frac{100}{26} = 3 \times 26 = 78 \sim F_{3,26}$$

At 5% Level, $f_c = 2.975$
($\alpha = 0.05$)

F value $> f_c \Rightarrow$ Reject H_0.

(3) (3.1) $H = X(X'X)^{-1}X'$ (uppercase letters denote Matrix)

For idempotent matrix $AA' = A$

$$HH' = X(X'X)^{-1}X' \left[X(X'X)^{-1}X'\right]'$$

$$= X(X'X)^{-1}X' \; X'' \; (X'X)^{-1}X' \qquad [\because (X'X)' = X'X \;\&\; (X^{-1})' = (X')^{-1}]$$

$$= X\underbrace{(X'X)^{-1}X'X}(X'X)^{-1}X' \qquad [\because X'' = X]$$

$$= XI(X'X)^{-1}X' \qquad [\because X^{-1}X = I]$$

$$= X(X'X)^{-1}X' \qquad [\because XI = X]$$

$$= H$$

(3.2) $Var(\vec{\hat{e}}|X) = Var((I-H)\vec{y}|X)$

$$= (I-H)\,Var(\vec{y}|X)(I-H)^T$$

$[\because H = f(X)$ & $Var(A\vec{y}) = A\,Var(\vec{y})A^T]$

$$= (I-H)\Sigma(I-H)'$$

$$= (I-H)\sigma^2 I (I-H)'$$

$$= (I-H)(I-H)'\sigma^2$$

$$= (I-H)(I-H')\sigma^2$$

$$= (I - HI - IH' + HH')\sigma^2$$

$$= (I - H - H + H)\sigma^2 \qquad \because HH' = H$$

$$= (I-H)\sigma^2$$

(4) $\vec{\hat{e}} = (I - H)\vec{y}$

$\vec{\hat{e}} = \left(I - X(X'X)^{-1}X'\right)\vec{y}$ — (1)

To generate given expression, premultiply $\vec{\hat{e}}$ by X'

$$LHS = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \hat{e}_3 \\ \hat{e}_4 \end{bmatrix} = \begin{bmatrix} \hat{e}_1 + \hat{e}_3 \\ \hat{e}_1 + 2\hat{e}_2 + \hat{e}_4 \end{bmatrix}$$

$$RHS = X'\left(I - X(X'X)^{-1}X'\right)\vec{y}$$

$$= \left(X'I - X'X(X'X)^{-1}X'\right)\vec{y}$$

$$= (X' - IX')\vec{y}$$

$$= (X' - X')\vec{y} = 0_{2\times 3}$$

Comparing second row in LHS & RHS.

$\Rightarrow \hat{e}_1 + 2\hat{e}_2 + \hat{e}_4 = 0$. Yes.

(5)

$$\vec{Y} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 2 & 1 \\ 1 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} \beta_A \\ \beta_B \end{bmatrix} + \vec{\hat{e}}$$

(6) Yes. The Q-Q plots of all predictor variables follow straight line, so no transformation is needed.

(a) for response variable, the Q-Q plot is not a straight line but becomes straight after log transformation.

The Y vs each predictor variable plots show increasing variance without transformation. Also box-cox transformation plot shows $\lambda = 0$ lies within 95% Confidence interval of max. likelihood.

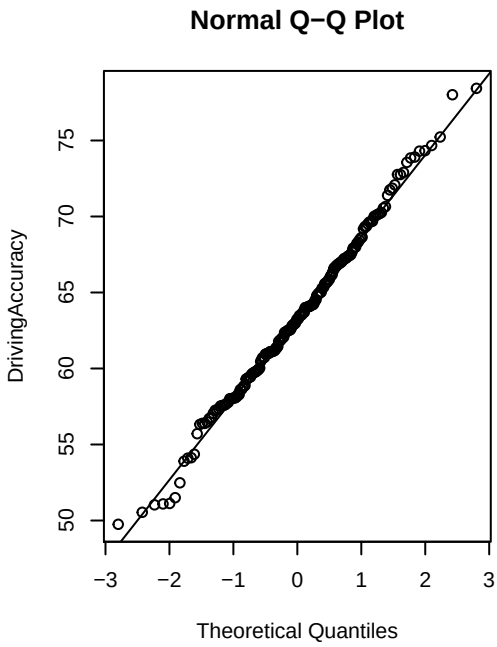
Since $\lambda = 0$ corresponds to log transformation, the recommendation is correct.

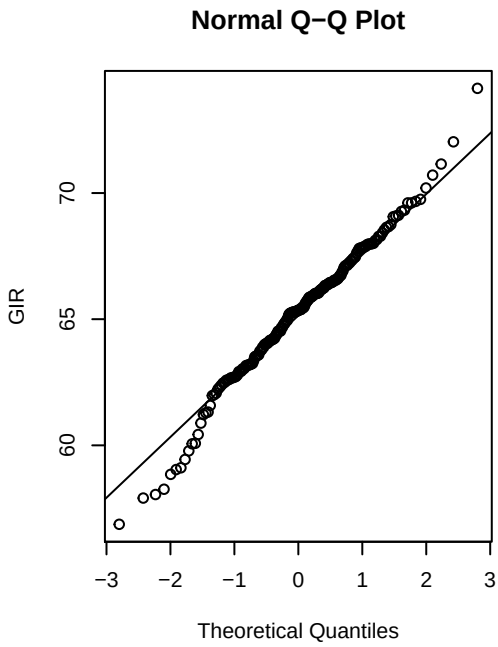
(b) See page -

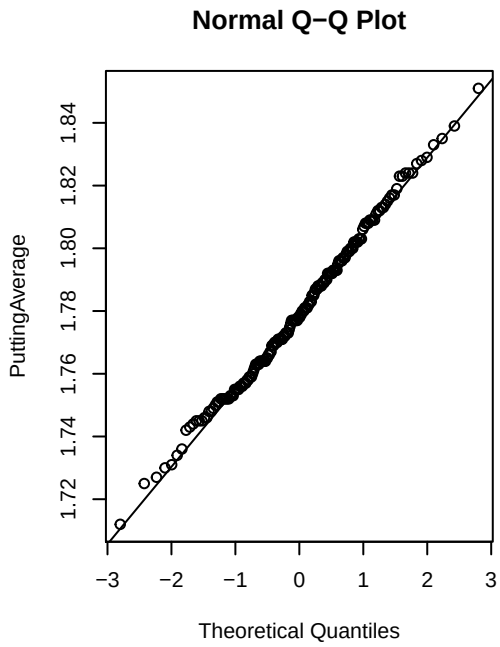
(c) From leverage plot, effects of high leverage points should be looked into before finalizing the model.

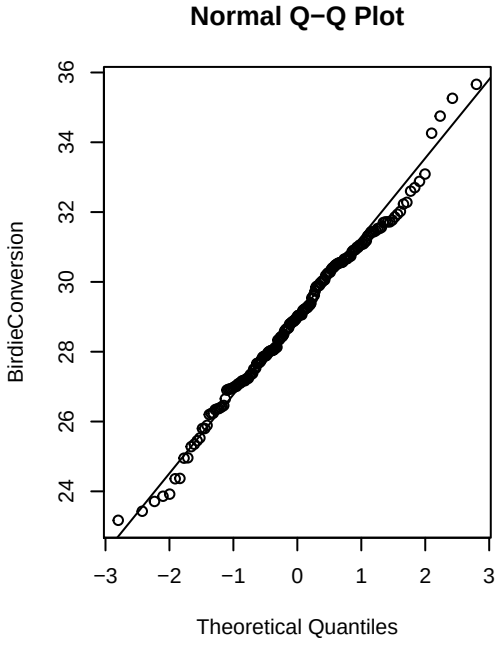
(d) Statistically insignificant predictors can be removed

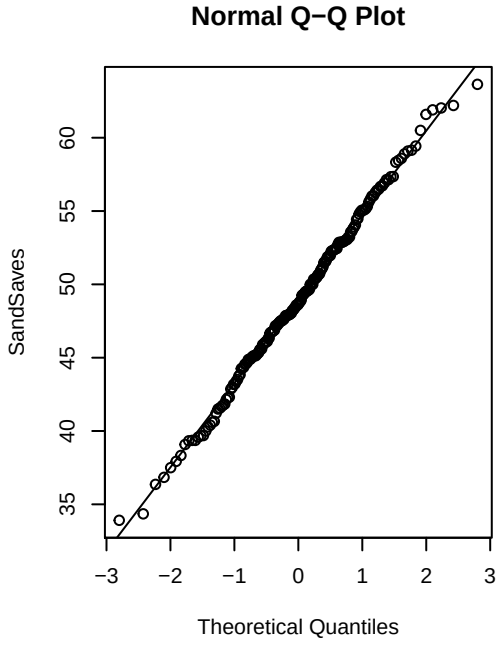
(e) Removing all predictors with insignificant p values at once is not recommended because then we'll be comparing model with subset of predictor variables with original model. p-value can only be used to remove one predictor variable at a time. Partial f-test should be used to decide if a subset of predictor variables should be removed at once.

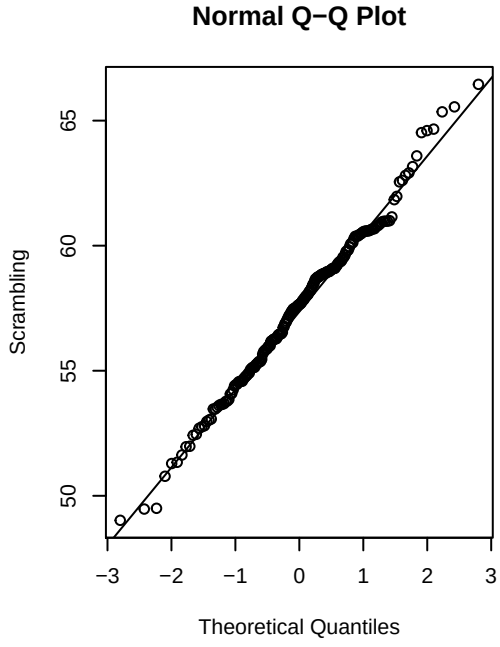

Normal Q−Q Plot
DrivingAccuracy
Theoretical Quantiles

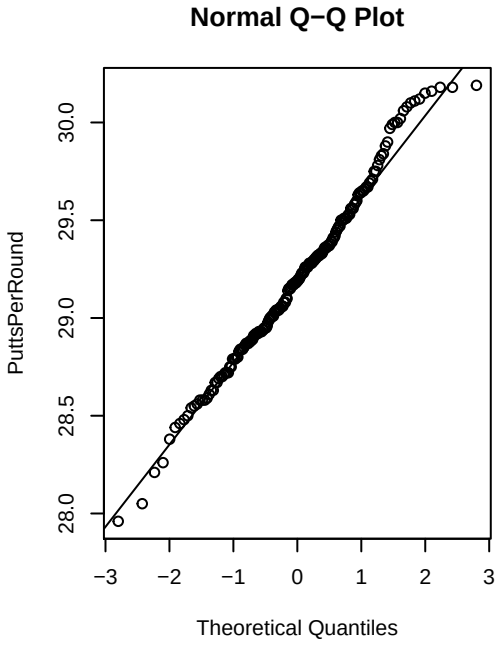

Normal Q−Q Plot
GIR
Theoretical Quantiles

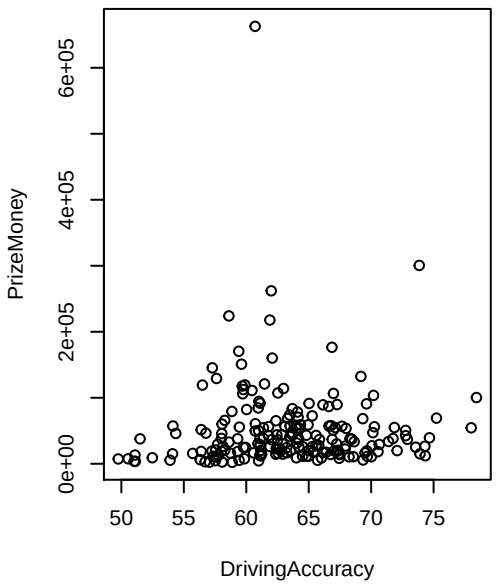

Normal Q−Q Plot
PuttingAverage
Theoretical Quantiles

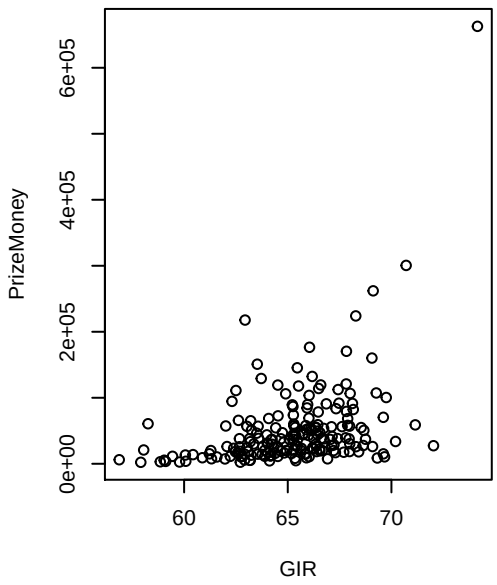

Normal Q−Q Plot
BirdieConversion
Theoretical Quantiles

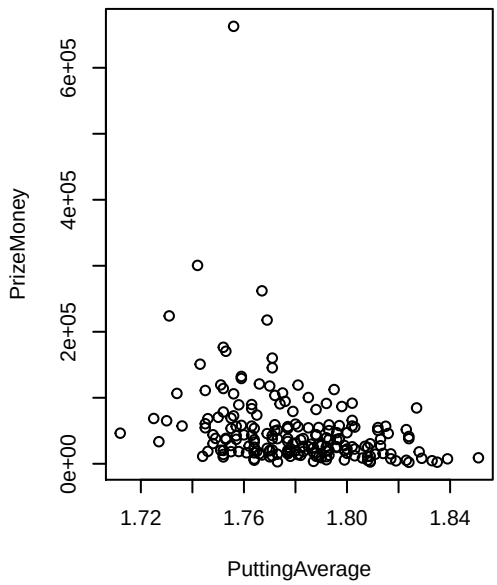

Normal Q−Q Plot
SandSaves
Theoretical Quantiles

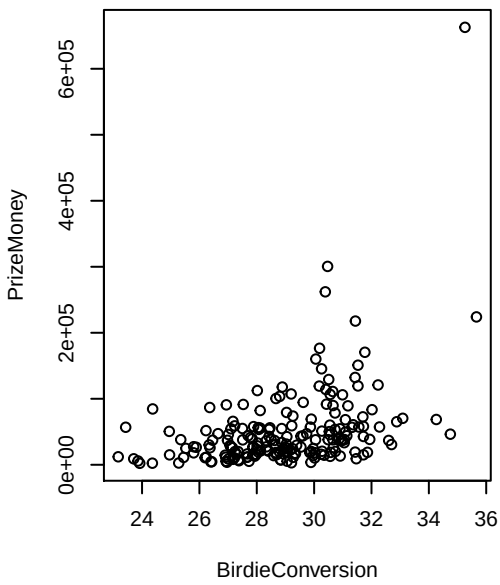

Normal Q−Q Plot
Scrambling
Theoretical Quantiles

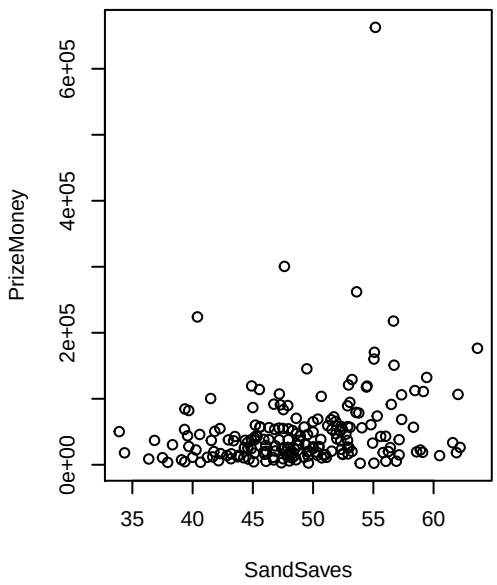

Normal Q−Q Plot
PuttsPerRound
Theoretical Quantiles

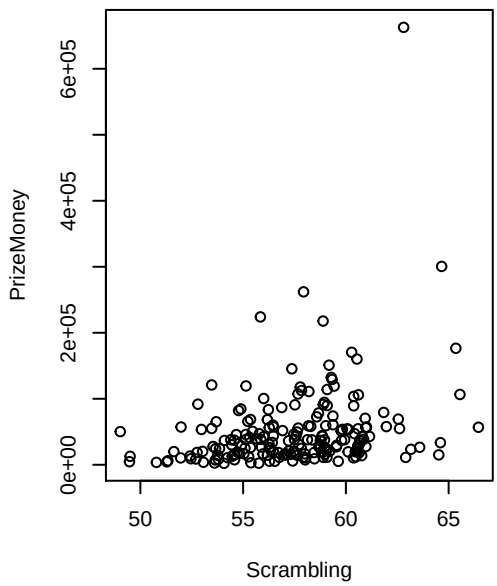

PrizeMoney
0e+00
2e+05
4e+05
6e+05
50
55
60
65
70
75
DrivingAccuracy

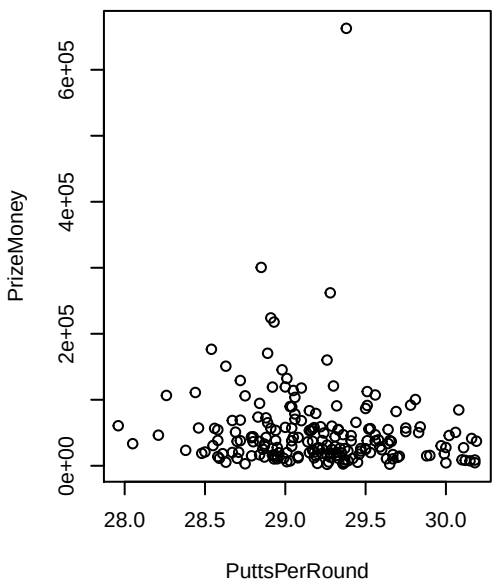

PrizeMoney
0e+00
2e+05
4e+05
6e+05
60
65
70
GIR

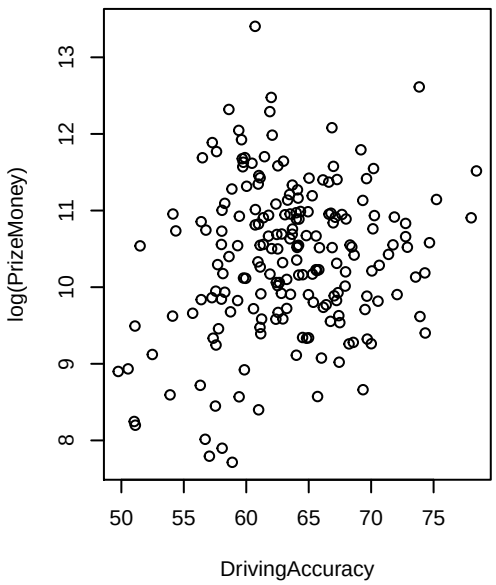

PrizeMoney
0e+00
2e+05
4e+05
6e+05
1.72
1.76
1.80
1.84
PuttingAverage

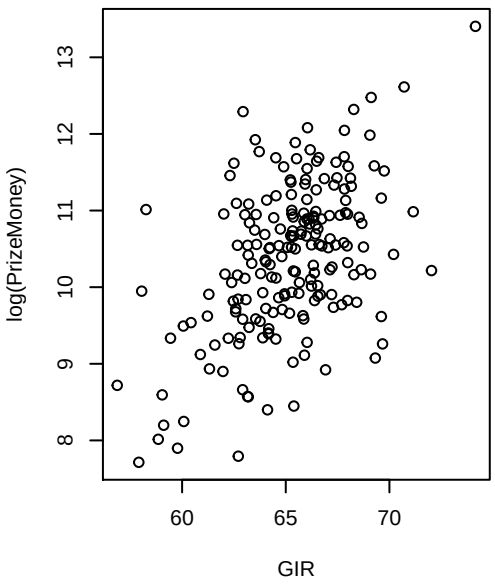

PrizeMoney
0e+00
2e+05
4e+05
6e+05
24
26
28
30
32
34
36
BirdieConversion

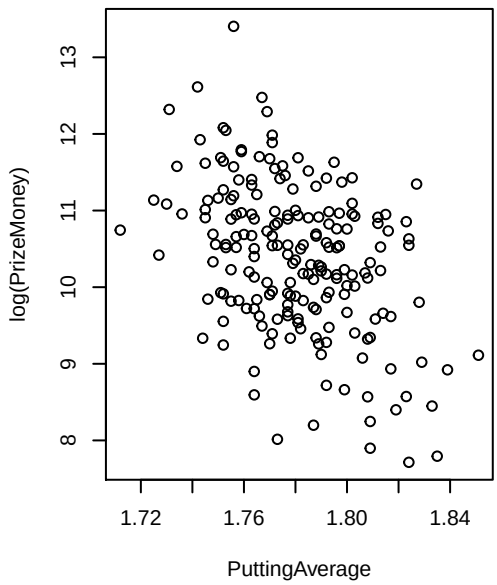

PrizeMoney
0e+00
2e+05
4e+05
6e+05
35
40
45
50
55
60
SandSaves

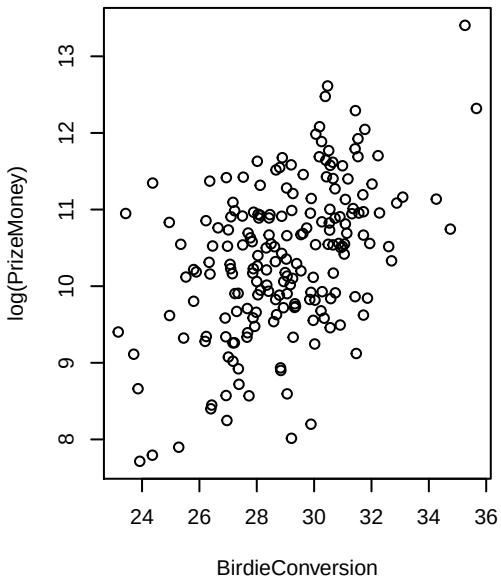

PrizeMoney
0e+00
2e+05
4e+05
6e+05
50
55
60
65
Scrambling

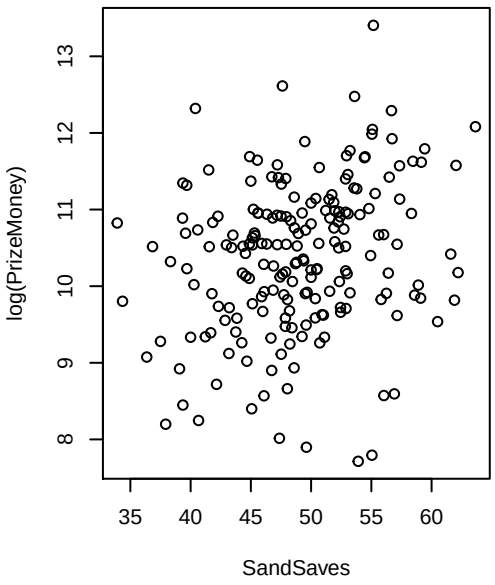

PrizeMoney
0e+00
2e+05
4e+05
6e+05
28.0
28.5
29.0
29.5
30.0
PuttsPerRound

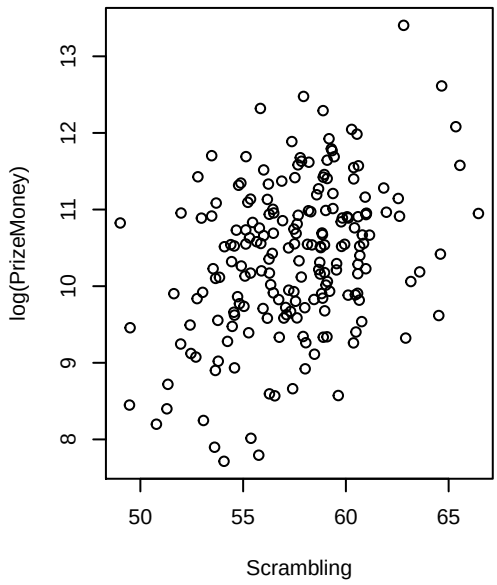

log(PrizeMoney)
8
9
10
11
12
13
50
55
60
65
70
75
DrivingAccuracy

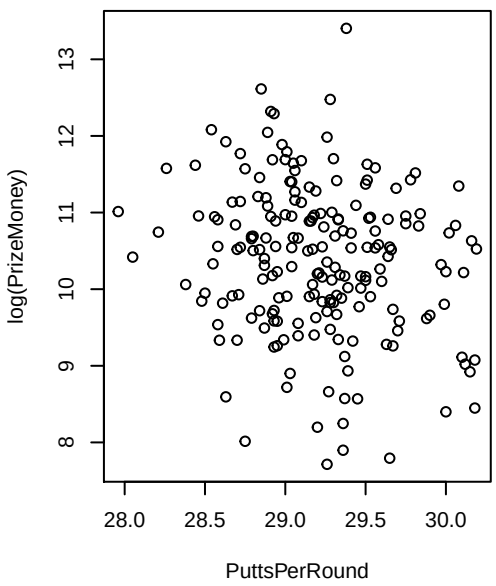

log(PrizeMoney)
8
9
10
11
12
13
60
65
70
GIR

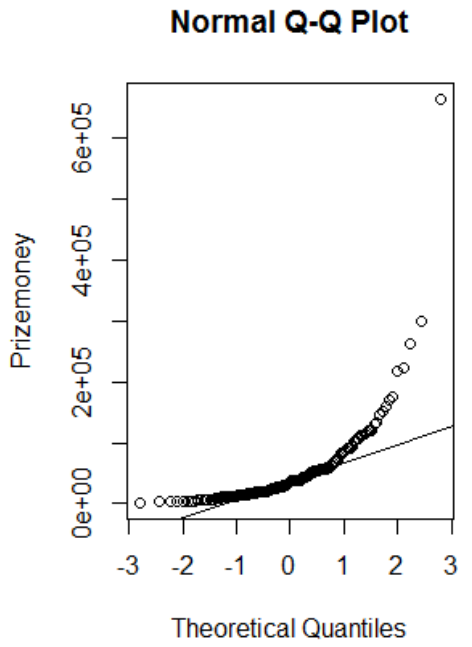

log(PrizeMoney)
8
9
10
11
12
13
1.72
1.76
1.80
1.84
PuttingAverage

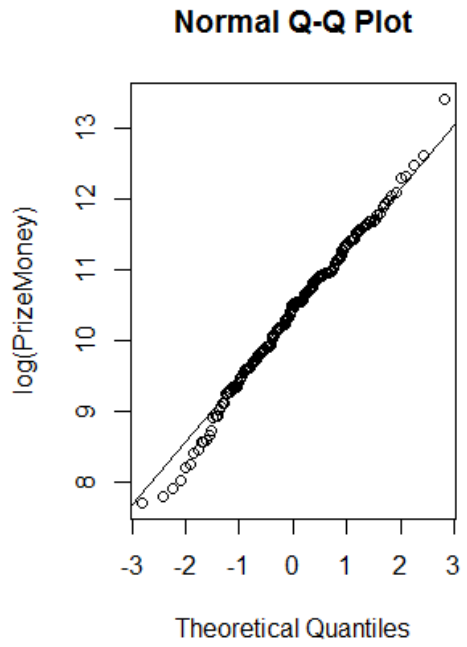

log(PrizeMoney)
8
9
10
11
12
13
24
26
28
30
32
34
36
BirdieConversion

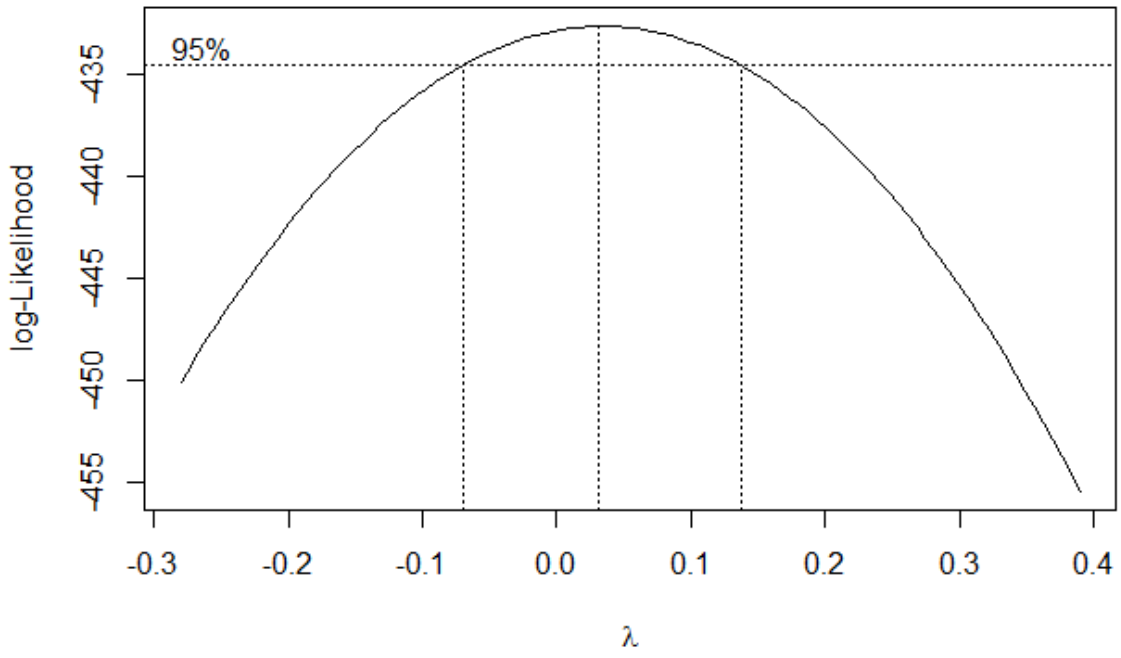

log(PrizeMoney)
8
9
10
11
12
13
35
40
45
50
55
60
SandSaves

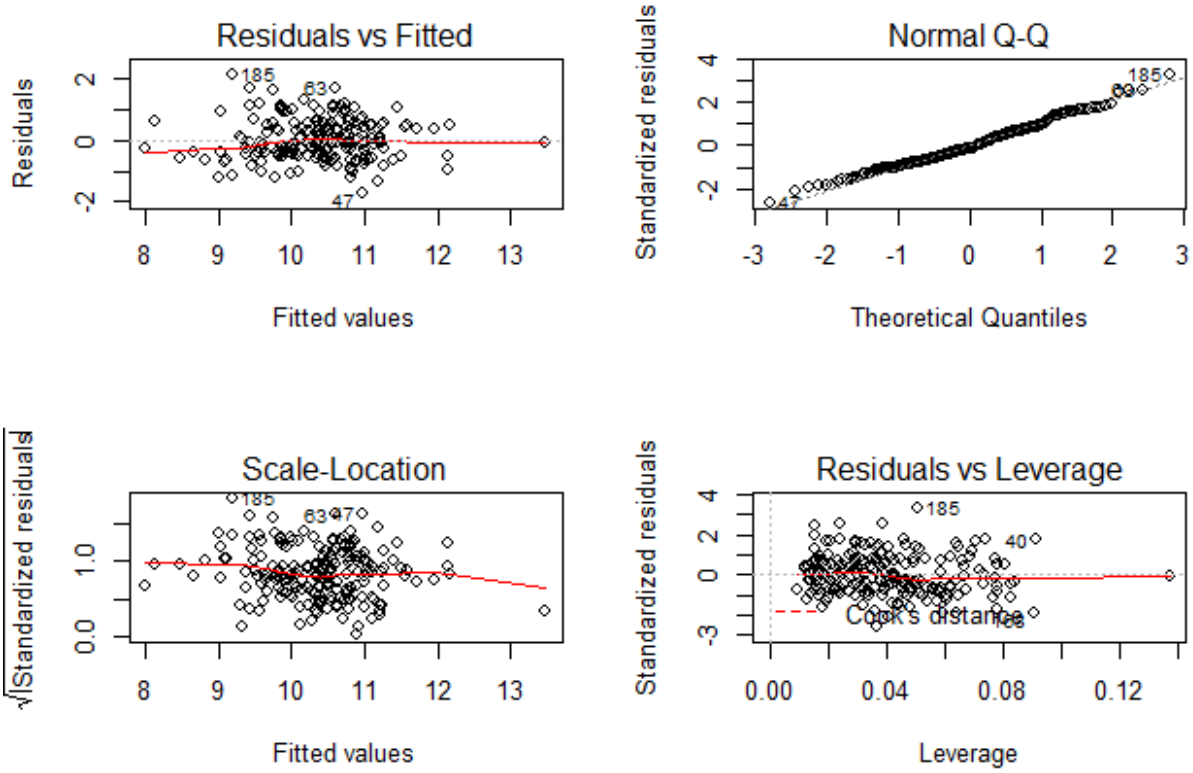

log(PrizeMoney)
8
9
10
11
12
13
50
55
60
65
Scrambling


log(PrizeMoney)
8
9
10
11
12
13
28.0
28.5
29.0
29.5
30.0
PuttsPerRound


Normal Q-Q Plot
Prizemoney
0e+00
2e+05
4e+05
6e+05
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Normal Q-Q Plot
log(PrizeMoney)
8
9
10
11
12
13
-3
-2
-1
0
1
2
3
Theoretical Quantiles


95%
log-Likelihood
-435
-440
-445
-450
-455
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
λ



Call:
lm(formula = log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage +
BirdieConversion + SandSaves + Scrambling + PuttsPerRound)

Residuals:

Min	1Q	Median	3Q	Max
-1.71949	-0.48608	-0.09172	0.44561	2.14013

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.194300	7.777129	0.025	0.980095	
DrivingAccuracy	-0.003530	0.011773	-0.300	0.764636	
GIR	0.199311	0.043817	4.549	9.66e-06	***
PuttingAverage	-0.466304	6.905698	-0.068	0.946236	
BirdieConversion	0.157341	0.040378	3.897	0.000136	***
SandSaves	0.015174	0.009862	1.539	0.125551	
Scrambling	0.051514	0.031788	1.621	0.106788	
PuttsPerRound	-0.343131	0.473549	-0.725	0.469601	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.6639 on 188 degrees of freedom
Multiple R-squared: 0.5577, Adjusted R-squared: 0.5412
F-statistic: 33.87 on 7 and 188 DF, p-value: < 2.2e-16