Stats_101A_hw9_anna_piskun

Anna Piskun

3/3/2020

From the textbook, do these problems: ### Chapter 6: 3, 4, 5

```
setwd("~/Desktop")
cars <- read.csv("cars04.csv")
View(cars)
```

3) The analyst was so impressed with your answers to Exercise 5 in Section 3.4 that your advice has been sought regarding the next stage in the data analysis, namely an analysis of the effects of different aspects of a car on its suggested retail price. Data are available for all 234 cars on the following variables: Y = Suggested Retail Price, X1 = Engine Size, X2 = Cylinders, X3 = Horse Power, X4 = Highway MPG, X5 = Weight, X6 = Wheel Base, and X7 = Hybrid, a dummy variable which is 1 for so-called hybrid cars. The first model considered for these data was: Y = B0 + B1X1 +B2X2 + B3X3 + B4X4 + B5X5 + B6X6 + B7X7 + e

a) Decide whether that is a valid model. Give reasons to support your answer.

In order to determine model validity we look to the diagnostic plots. First analyzing the residual plot, we see a slight curved pattern indicating non-linearity. Likewise, the plot has a small fanshape trend indicating non-constant variance. While the normal Q-Q plot has little deviation from a straight line, showing that the errors are normally distributed, the scale-location plot shows an increasing trend once again confirming that the constant variance condition is not met. Looking at the residuals vs. leverage plot there are some influential points. Thus, since the model does not satisfy the constant variance or linear conditions, it is not valid.

b) The plot of residuals against fitted values produces a curved pattern. Describe what, if anything can be learned about the model above from this plot.

Since the plot of residuals against fitted values produces a curved pattern, we learn that the data is not best described/fit by a linear model and that we should try another model and transform our variables.

c) identify any bad leverage points for the above model

Point 223 is potentially a bad leverage point.

d) Decide whether the new model is a valid model

Looking at the residual plot for the new model, there is no clear pattern or fan-shape trend, therefore both the linearity and constant-variance conditions are satisfied. The normal QQ plot follows a straight line, thus satisfying the normality condition. The scale-location plot has no incresing or decreasing trend, again confirming that the constant-variance condition is met. There are no high leverage or high influence points using this model. Therefore, the new model is a valid model. Additionally, the adjusted R-squared increased to 0.859 from 0.7751, meaning that more of the variability is able to be explained by our new model. Additionally, looking at the marginal model plots, all the variables are fit by the model pretty well with the loess lines being almost the same as the regression lines. However, a potential weakness in the model is that there is evidence of multiple co-linearity amongst tEngineSize, tCylinders, tHorsepower, and Weight as seen by their VIF values, which were all greater than 5.

f) The analyst's boss has compained about the new model saying that it fails to take account of the manufacturer of the vehicle (e.g., BMW vs. Toyota). Describe how the new model could be expanded in order to estimate the effect of manufacturer on suggested retail price.

The model can be expanded to add a new manufacturer variable and then test to see whether it is significant or not. We can do this by performing a partial F-test and determining whether the manufacturer of the vehicle is a statistically significant variable and helps create a better fitting model. Likewise we can take this process one step further and use the AIC to determine whether a model with this variable is better. Since this dataset has a finite sample size, using the AIC will allow us to avoid oversimplifying. By comparing a model with the new variable and one without, the model with the smaller AIC will be the better one.

4) A book on robust statistical methods published in June 2006 considers regression models for a data set taken from Jalali-Heravi and Knouz (2002). The aim of this modeling is to predict a physical property of chemical compounds called the Krafft point based on four potential predictor variables using a data set of size n=32. According to Maronna, Martin and Yohai (2006, p.380) - "The Krafft point is an important physical characteristic of the compounds called surfactants, establishing the minimum temperature at which a surface can be used."

Variables: Y = Krafft Point (KPOINT), X1 = Randic Index (RA), X2 = Heat of Formation (HEAT), X3 = Reciprocal of Volume of the tail of the molecule (VTINV), x4 = Reciprocal of dipole moment (DIPINV).

First model to be considered: Y = B0 + B1X1 + B2X2 + B3X3 + B4X4 + e

a) Decide whether the above model is a valid model

Looking at the diagnostic plots for the above model we can determine whether or not its valid. There is no trend or pattern in the residual plot indicating that the linearity condition is satisfied. Likewise, there is no fanshape in the residual plot, meaning that the constant variance condition is satisfied. However, looking at the Normal QQ plot there is some slight deviation from a straight line which may indicate non-normality of the errors. The scale-location plot shows no increasing or decreasing trend once again confirming that the constant variance condition is satisfied. The Residuals vs. Leverage plot shows only one potentially influential point. Since the normality of errors condition is not satisfied, this model is not valid.

b) The plots of standardized residuals against RA and VTINV produce curved patterns. Describe what, if anything can be learned about the above model from these plots. Give a reason to support your answer.

Curved patterns in the plots of standardized residuals against RA and VTINV indicate that a non-linear relationship may exist between the two variables and that a transformation may result in a better fitting overall model.

c) Jalali-Heravi and Knouzgive give "four criteria of correlation coefficient (r), standard deviation (s), F value for the statistical significance of the model and the ratio of the number of observations to the number of descriptors in the equation" for choosing between competing regression models. Provide a detailed critique of this suggestion.

Firstly, the correlation coefficient (r) is not useful in deciding between competing models since it only measures the strength and relationship of two variables. As such, multiple r values would have to be compared but since the variables used vary from model to model you wouldn't be able to reliably compare values. Likewise, it only measures a linear relationship and is not suitable for any sort of non-linear regression. Looking at our answer to a, we determined that a linear model is not valid so using the correlation coefficient to compare this model would be useless.

The standard deviation (s) serves as a measure of the variation in our data which could potentially be used to test for the normality of our residuals, however, diagnosting plots and other measures tend to be more useful. In fact, since we are trying to distinguish between competing models, it is better to assess the accuracy of our model through looking at the standard error of the regression. The standard error of the regression represents the average distance that the observed values fall from the regression line (so unlike standard deviation, it accounts for the error of our actual model). It shows how wrong the regression model is throught using the

response variable. In this case, smaller values are preferred because they indicate that the observations are closer to the model's fitted line.

The F-test answers the question regarding whether or not the variables used in a model are associated with the response variable. It tests the null hypothesis that all regression coefficients are equal to 0, and the alternative hypothesis that at least one is not equal to 0. If the resulting F-value is small then we fail to reject the null hypothesis, while large f-values are consistent with the alternative hypothesis and that one of the variables is non-zero. Therefore the F-test and resulting F-valus tests whether the relationship between the response variable and set of predictor variables is statistically significant (indicating whether the variables used in a model are useful) and is therefore helpful in choosing between competing regression models.

When using multiple linear regression, we assume that the variables are independent. When the number of independent variables (aka descriptors) is greater than the number of observations, we cannot apply multiple linear regression. As such, the ratio of the number of observations to number of descriptors is a useful criteria in distinguishing between competing models (because it can determine whether or not we can apply a regression model all together).

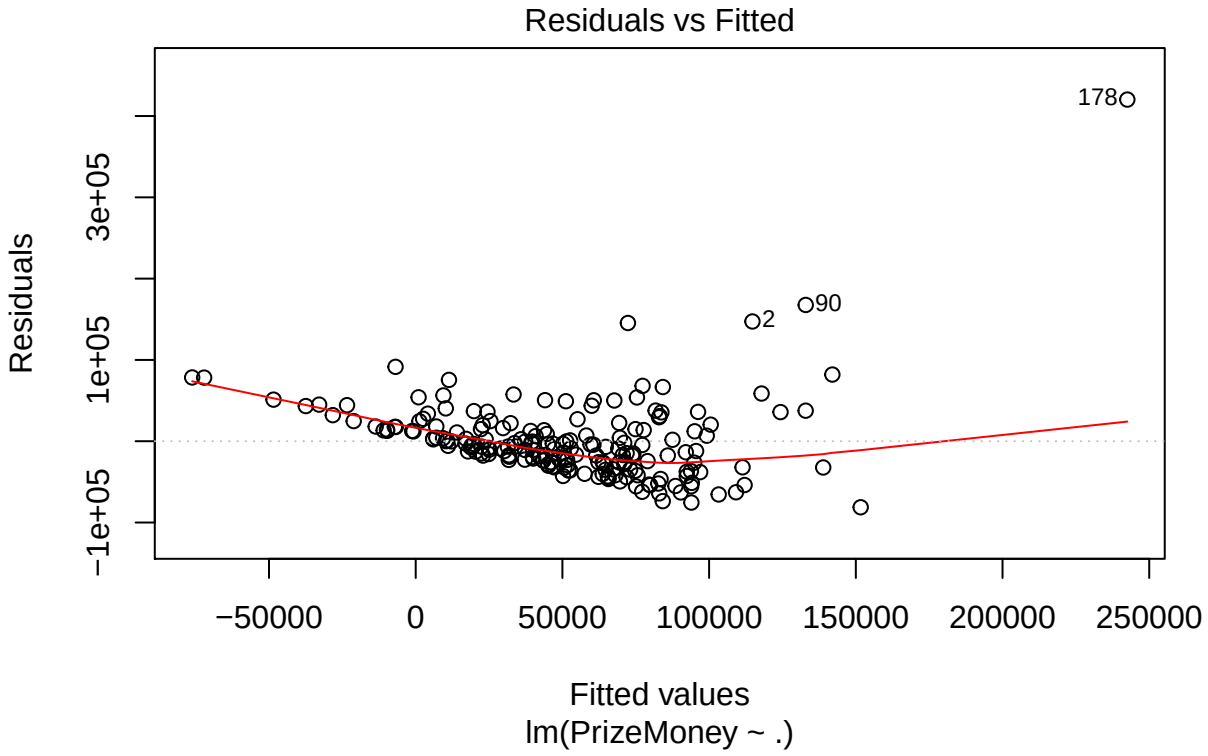
5) An avid fan of the PGA tour with limited background in statistics has sought your help in answering one of the age old questions in golf, namely, what is the relative importance of each different aspect of the game on average prize money in professional golf?

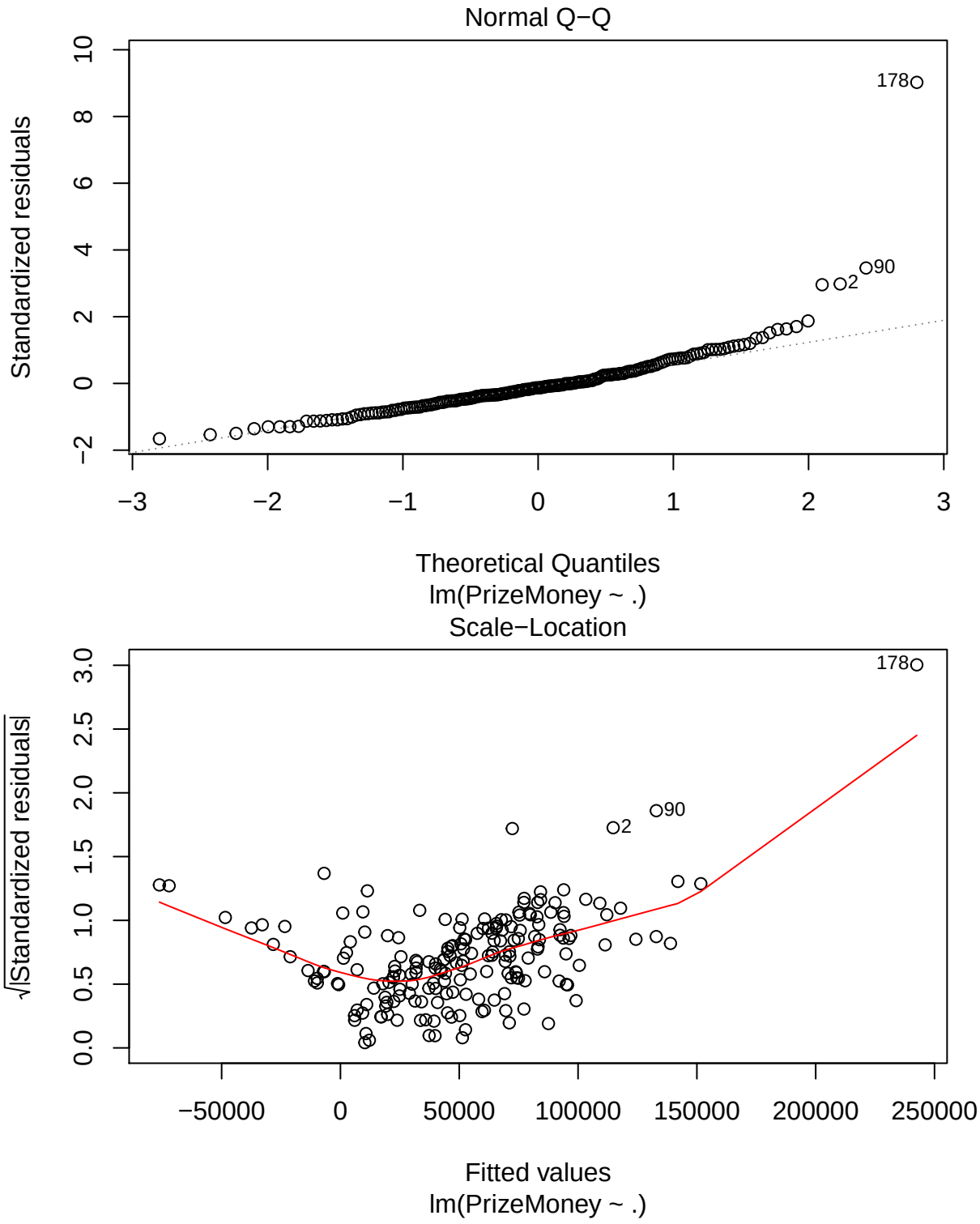
```
library(alr3)
```

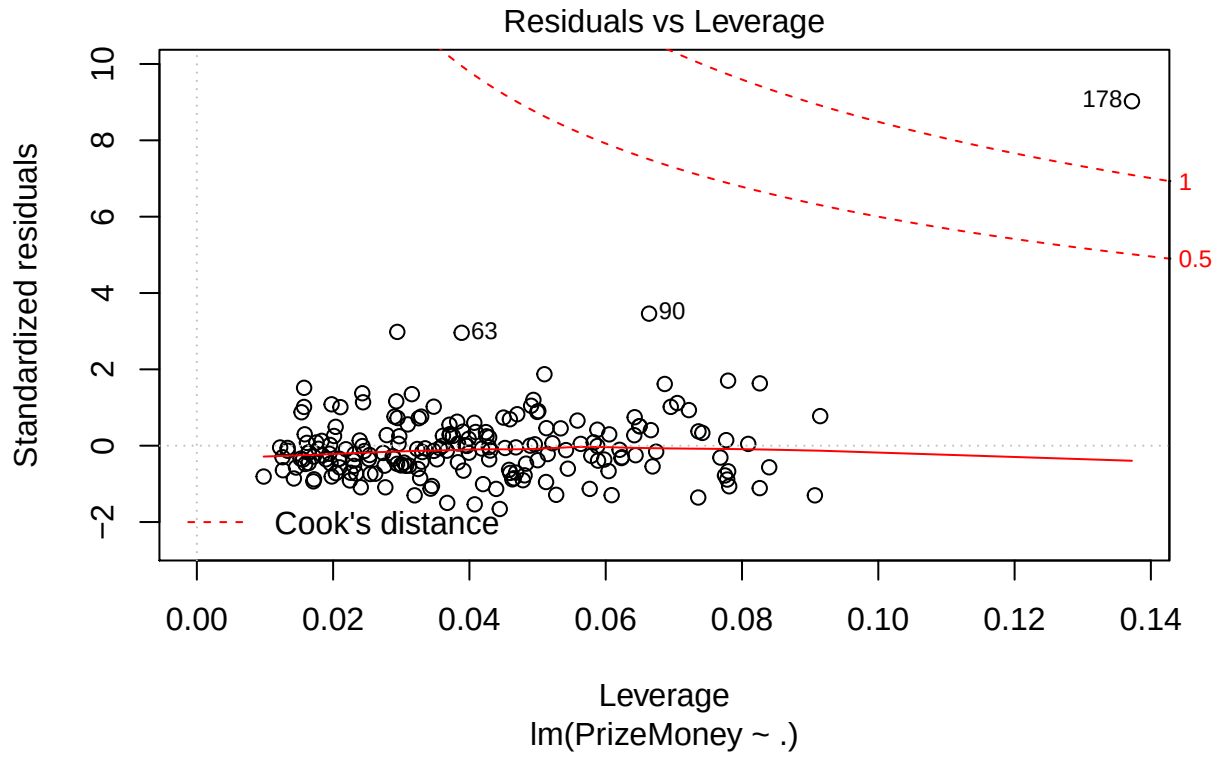
```
## Loading required package: car
```

```
## Loading required package: carData
```

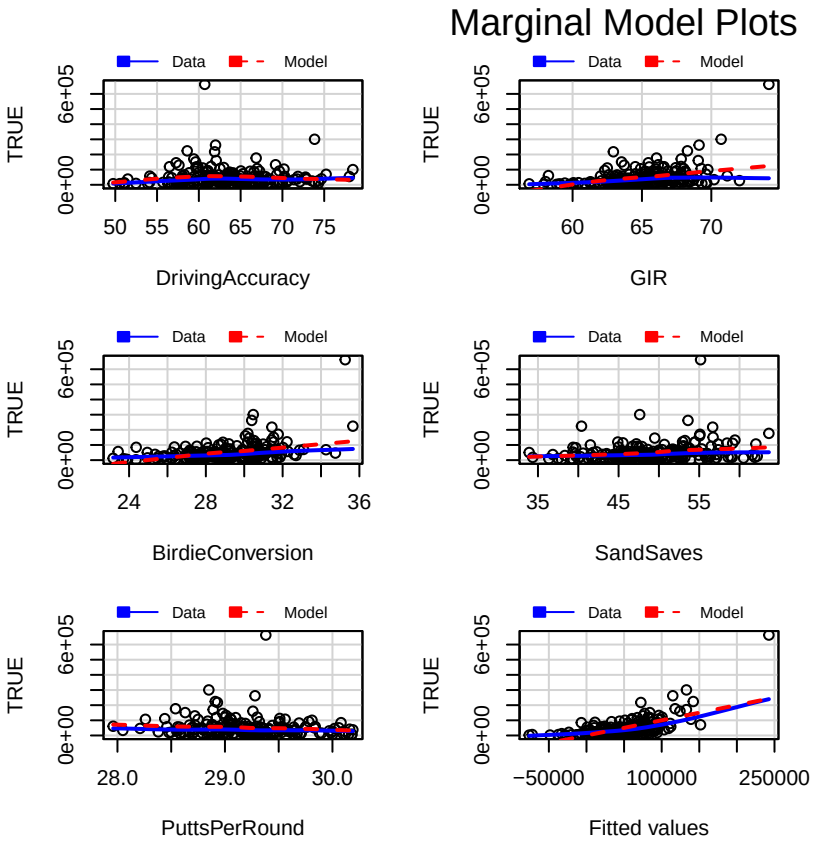
```
golf <- read.csv("pgatour2006.csv")

new.golf <- dplyr::select(golf, PrizeMoney, DrivingAccuracy, GIR, PuttingAverage, BirdieConversion, San

model <- lm(PrizeMoney~., data = new.golf)
plot(model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(PrizeMoney ~ .)
178
2
90
Scale−Location
√|Standardized residuals|
Fitted values
lm(PrizeMoney ~ .)




```
mmps(model)
```



a) A statistician from Australia has recommended to the analyst that they not transform any of the

predictor variables but that they transform Y using the log transformation. Do you agree with this recommendation? Give reasons to support your answer.

```
library(alr3)
summary(powerTransform(model))
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts argument
## ignored
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.0337           0      -0.0701       0.1376
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df    pval
## LR test, lambda = (0) 0.4054804  1 0.52427
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 335.2384  1 < 2.22e-16
```

Looking at the boxcox method, since there is a large p-value (>0.05) for lambda = 0, we fail to reject the null hypothesis that we should do a log transform. Looking at the pvalue for lambda = 1 (test that no transformation is needed) it is smaller than 0.05, therefore we reject the null hypothesis and the transformed model is best. Thus, the BoxCox method recommends a log transformation of Y and the Australian statistician's recommendation is supported.

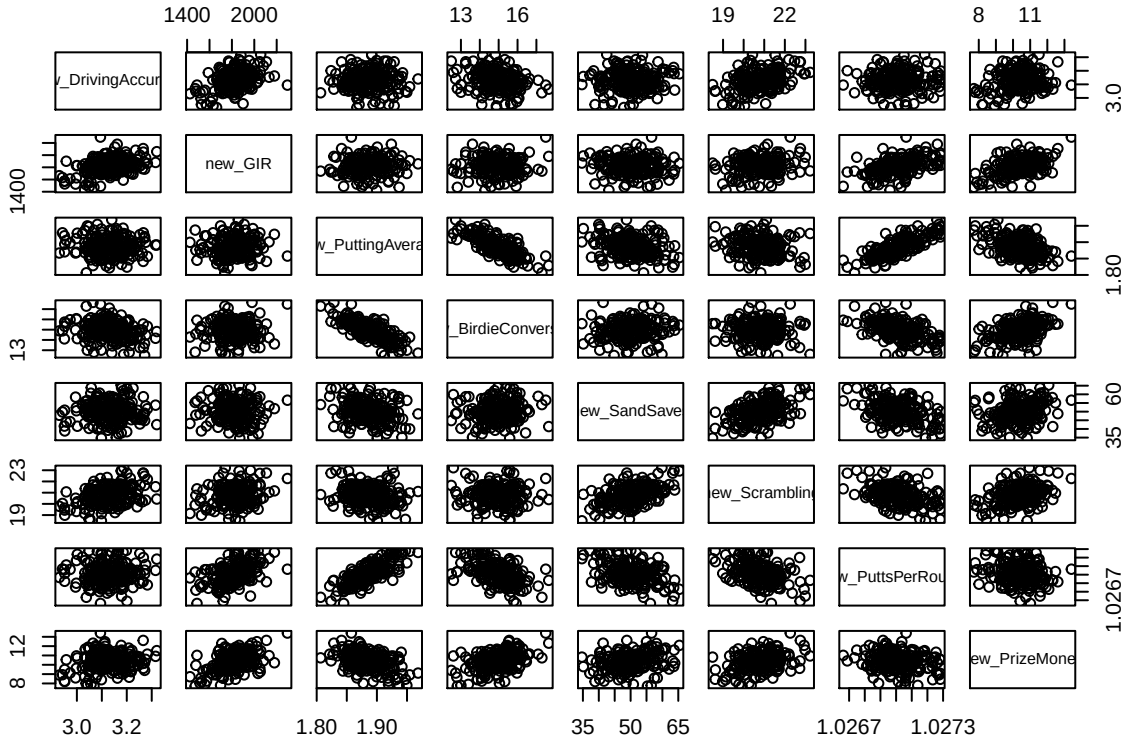
b) Develop a valid full regression model containing all seven potential predictor variables listed above. Ensure that you provide justification for your choice of full model, which includes scatter plots of the data, plots of standardized residuals, and any other relevant diagnostic plots.

```
library(alr3)
summary(powerTransform(cbind(new.golf$DrivingAccuracy, new.golf$GIR, new.golf$PuttingAverage, new.golf$
```

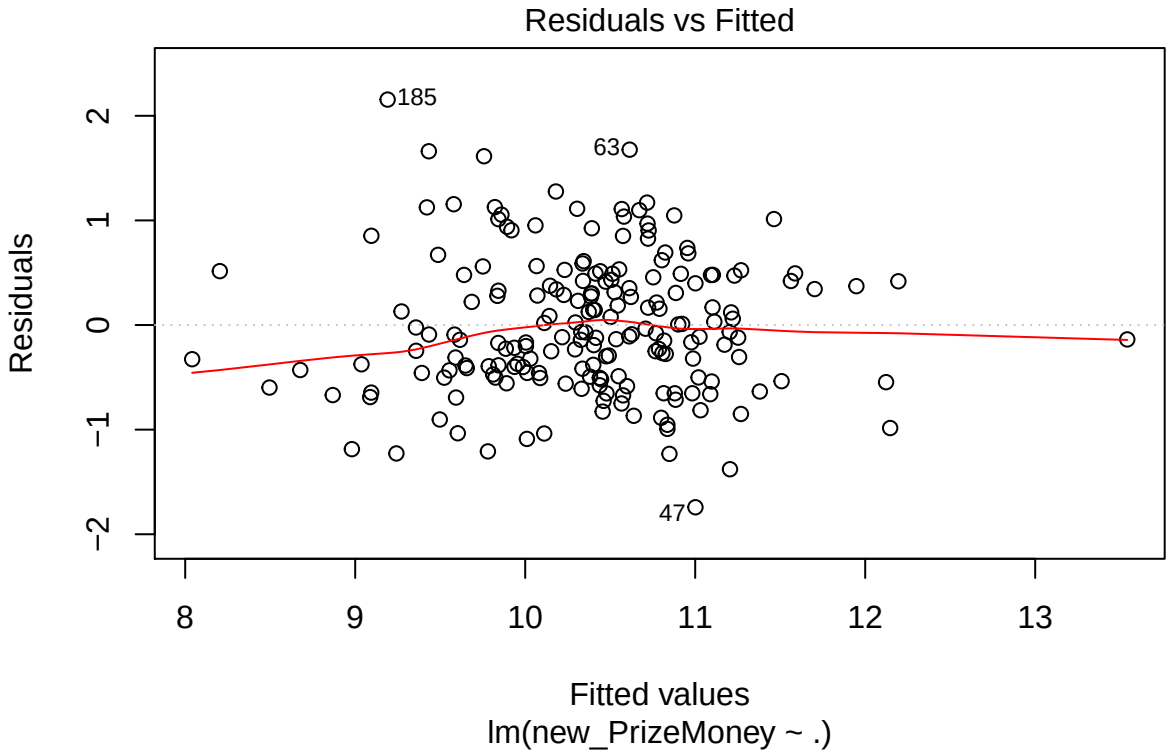
```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.2751           1      -0.8984       1.4486
## Y2    1.7972           1       0.3116       3.2827
## Y3    1.0999           1      -3.4384       5.6383
## Y4    0.8033           1      -0.2707       1.8772
## Y5    1.0064           1       0.0634       1.9493
## Y6    0.7495           1      -0.6752       2.1742
## Y7    0.0079           1      -3.2324       3.2483
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                       LRT df     pval
## LR test, lambda = (0 0 0 0 0 0 0) 13.46843  7 0.061485
##
## Likelihood ratio test that no transformations are needed
##                                       LRT df    pval
## LR test, lambda = (1 1 1 1 1 1 1) 3.687514  7 0.81498
```

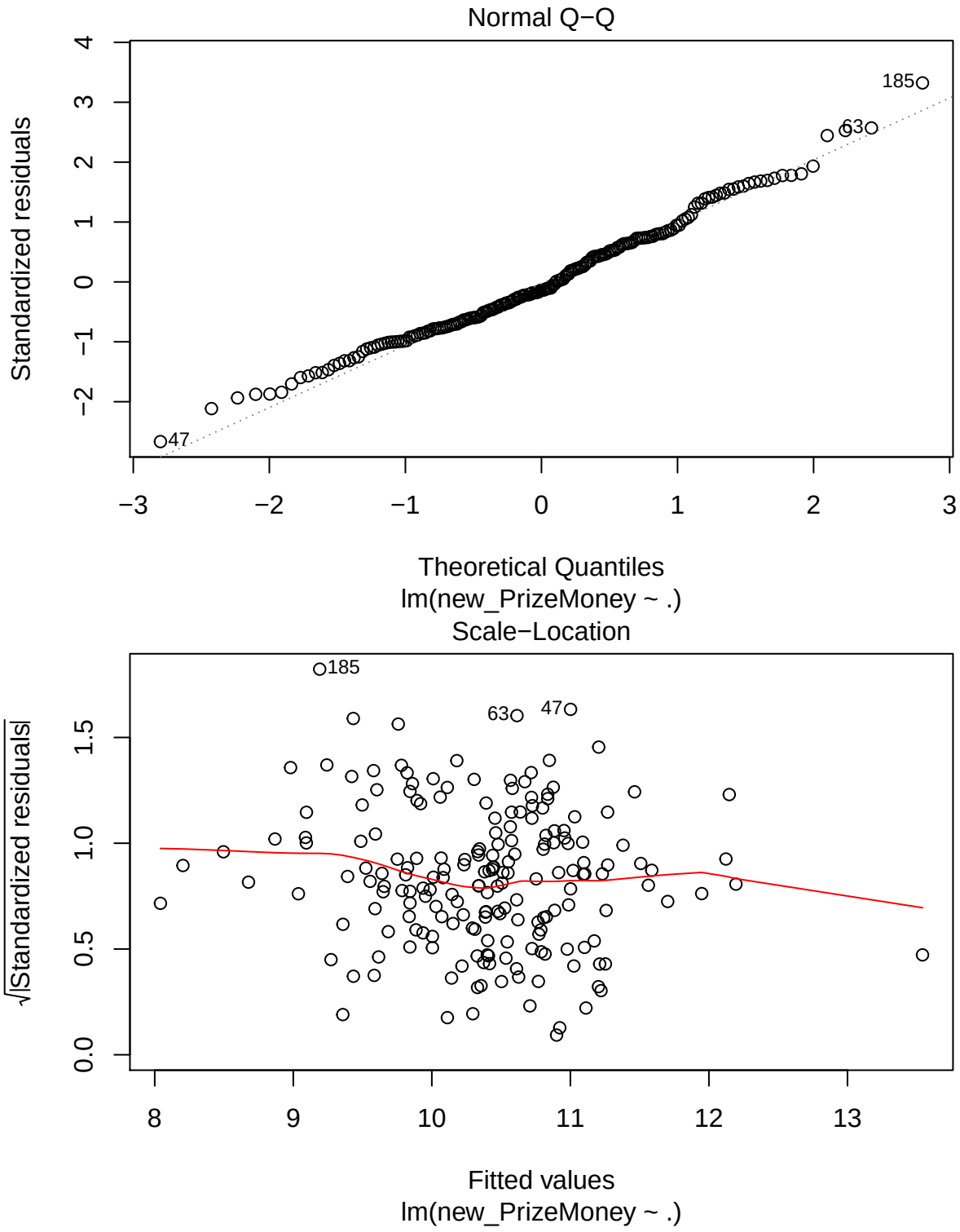
```
t.golf <- dplyr::transmute(new.golf, new_DrivingAccuracy = DrivingAccuracy^0.2751, new_GIR = GIR^1.7972

t.model <- lm(new_PrizeMoney~., data = t.golf)
```

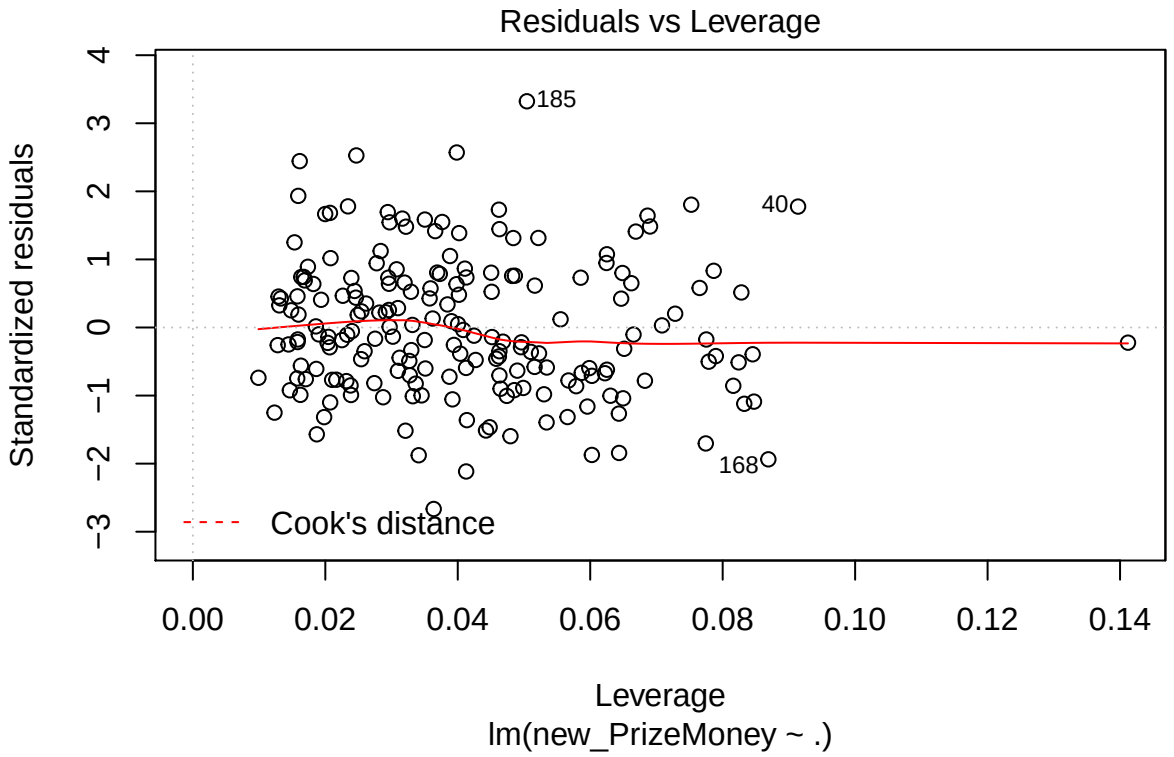
```
plot(t.golf)
```




```
plot(t.model)
```



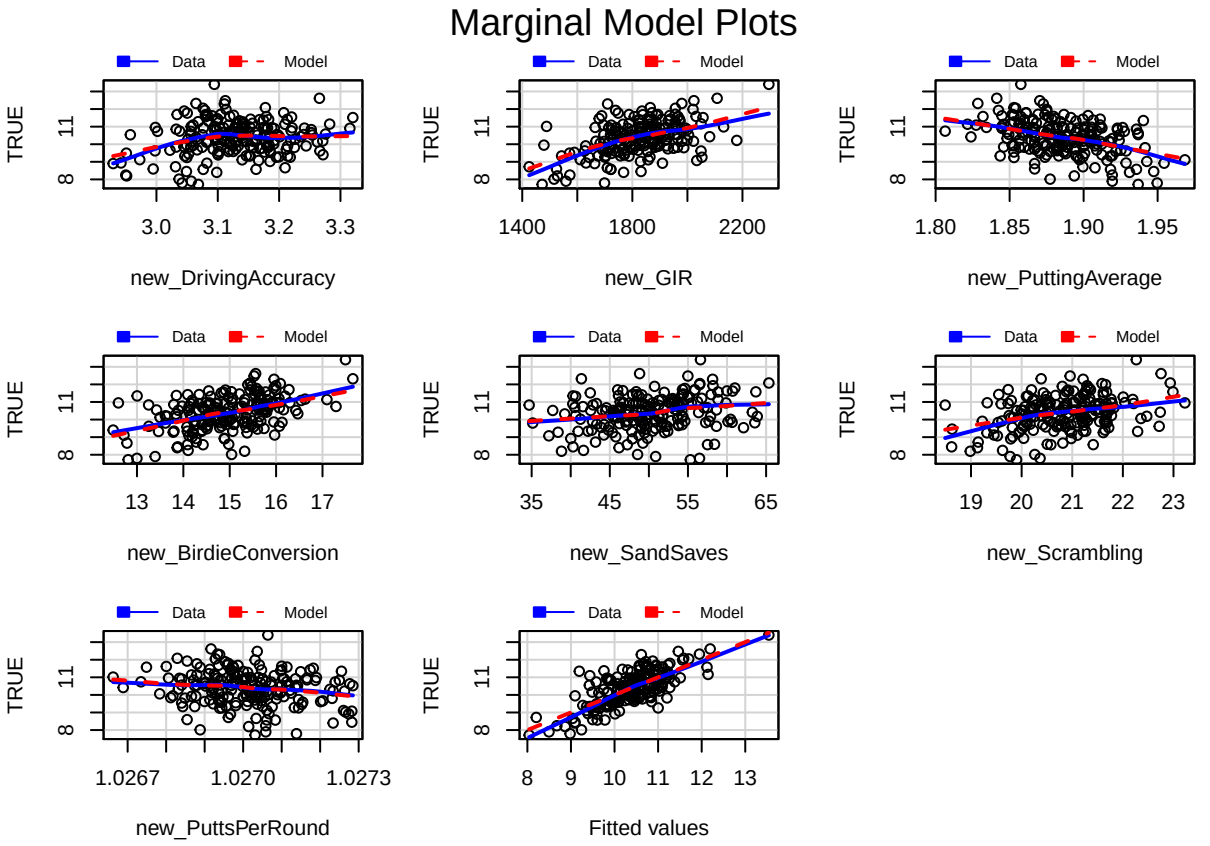

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(new_PrizeMoney ~ .)
185
63
47
Scale−Location
√|Standardized residuals|
Fitted values
lm(new_PrizeMoney ~ .)




```
summary(t.model)
```

```
##
## Call:
## lm(formula = new_PrizeMoney ~ ., data = t.golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74192 -0.47644 -0.09115  0.43785  2.15502
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.083e+03  1.753e+03   0.618 0.537540
## new_DrivingAccuracy  -1.760e-01  8.678e-01  -0.203 0.839475
## new_GIR               3.869e-03  8.809e-04   4.392 1.87e-05 ***
## new_PuttingAverage   -8.086e-01  5.953e+00  -0.136 0.892093
## new_BirdieConversion  3.838e-01  9.758e-02   3.933 0.000118 ***
## new_SandSaves         1.462e-02  9.577e-03   1.527 0.128510
## new_Scrambling        1.999e-01  1.176e-01   1.699 0.090993 .
## new_PuttsPerRound    -1.059e+03  1.716e+03  -0.617 0.537750
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6656 on 188 degrees of freedom
## Multiple R-squared:  0.5555, Adjusted R-squared:  0.539
## F-statistic: 33.57 on 7 and 188 DF,  p-value: < 2.2e-16
```

```
mmps(t.model)
```



Looking at the residual plot for the model using both the transformed predictor variables and response variable, we see that there is no clear trend or pattern indicating that the linearity condition is satisfied. There is no fan-shape in the residuals, thus indicating that the constant-variance condition is also satisfied. The normal QQ plot shows little to no deviation from a straight line, and therefore shows that the errors are normally distributed and that the normality condition is satisfied. The scale-location plot shows no increasing or decreasing trend, once again supporting the fact that the constant-variance condition is satisfied. The residual plot shows no high leverage or influential points. Therefore, our model is valid. Furthermore, analyzing the marginal model plots, we see that the regression lines for every predictor variable is essentially the same as their respective loess lines indicating that our model fits the data well.

c) Identify any points that should be investigated. Give one or more reasons to support each point chosen.

Looking at the Residuals vs. Leverage plot, it may be useful to investigate 185, 40, and 168 as they are outside of the [-2, 2] residual range and are therefore outliers which may the result of error and skew our model. Likewise, 185, 63, and 47 are much further away from the center line in comparison to the rest of the points that are relatively equidistant from the center line. Since these points have very large residuals, we see that they are outliers (with potential to be influential/have leverage) and may indicate either a sample oddity or error that alters the fit of our model.

d) Describe any weaknesses in your model.

While our model is valid and fits the data relatively well, a weakness in it is the presence of multiple outliers that require further investigation to see whether or not they should be removed. Likewise, the normal QQ plot still slightly deviates from a straight line which potentially indicates non-normality of errors. Looking at the matrix plots for the transformed data frame, there are clear trends in the plots between multiple variables indicating multiple collinearity. There is a clear trend between *DrivingAccuracy* and *GIR*, *PuttingAverage* and *BirdieConversion*, *PuttingAverage* and *PuttsPerRound* and many more. Multicollinearity is a problem because it undermines the statistical significance of our predictor variables and as such makes our estimators biased. Thus, this indicates a weakness in our model.

e) The golf fan wants to remove all predictors with insignificant t-values from the full model in a single step. Explain why you would not recommend this approach.

I would not recommend this approach since the t-statistics are interpreted given that all other variables are controlled for (or in the model). In multiple regression, the t-statistic tests whether the variable is significant given the other variables. Therefore it can be dangerous to remove all of the predictors with insignificant t-values in one step, since if predictors are associated with each other, removing one could change the significance of others. If we remove all of the variables at once we risk losing important variables, and should instead proceed with a more automated method like forward or backward stepwise regression that allows us to test the significance of variables on a step by step basis. Thus, I would not recommend this approach for fear of leaving out useful variables and creating an oversimplified model.

Chapter 7: 3 (use only AIC and BIC)

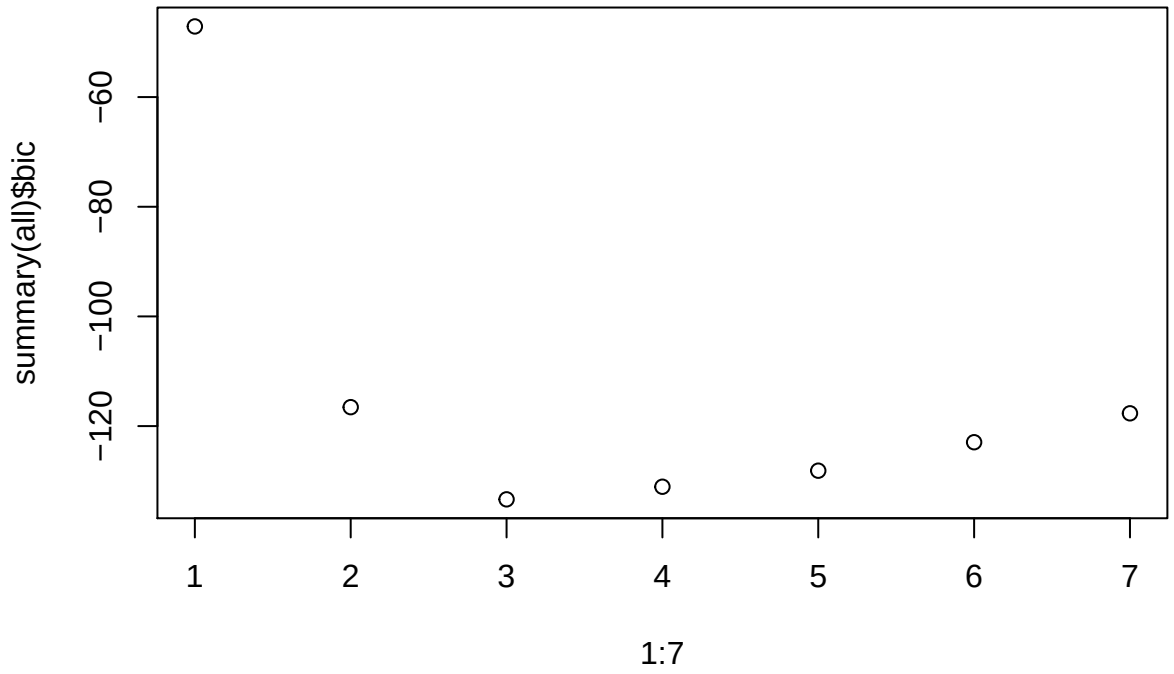
3) This is a continuation of Exercise 5 in Chapter 6. The golf fan was so impressed with your answers to part 1 that your advice has been sought re the next stage in the data analysis, namely using model selection to remove the redundancy in the full model developed in part 1.

$\log(Y) = B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5 + B6X6 + B7X7 + e$

Interest centers on using variable selection to choose a subset of the predictors to model the transformed version of Y. Throughout this question we shall assume that the above model is a valid model for the data.

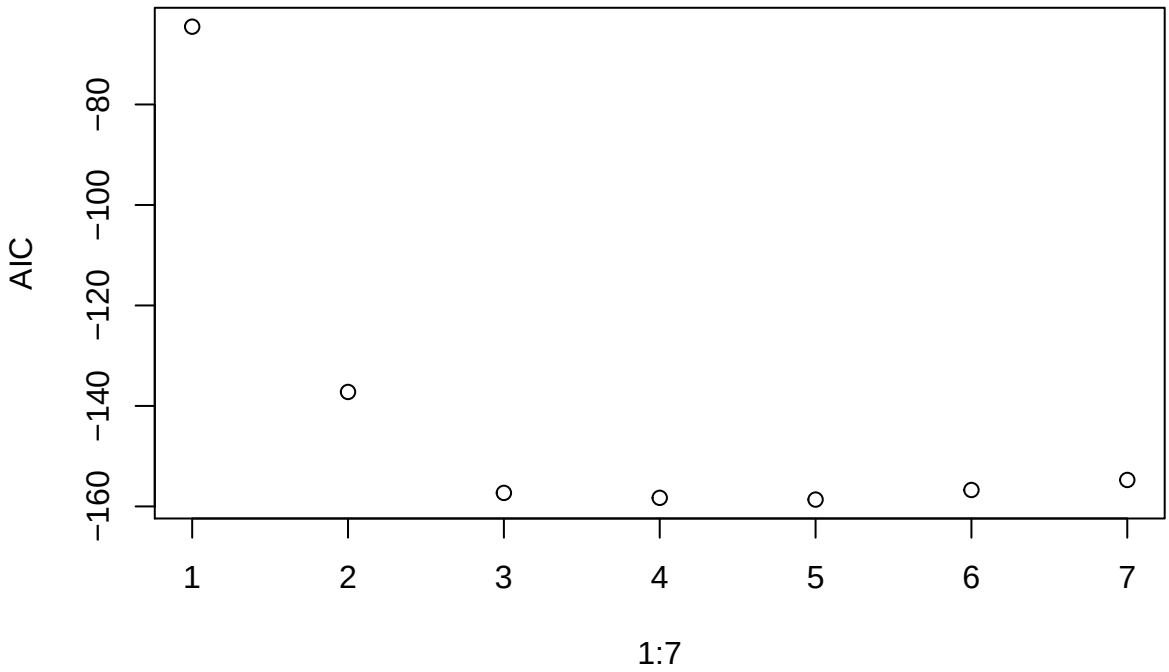
a) Identify the optimal model or models based on AIC and BIC from the approach based on all possible subsets.

```
library(leaps)

y.golf <- dplyr::transmute(new.golf, new_DrivingAccuracy = DrivingAccuracy, new_GIR = GIR, new_PuttingA

all <- regsubsets(new_PrizeMoney~new_DrivingAccuracy+new_GIR+new_PuttingAverage+new_BirdieConversion+ne
plot(1:7, summary(all)$bic)
```




```
#3 variable model has lowest BIC

Rss <- summary(all)$rss
```

```r
n <- nrow(y.golf)
p <- 1:7
AIC <- n * log(Rss/n) + 2 * p
plot(1:7, AIC)
```




```r
#5 variable model has lowest AIC

summary(all)
```

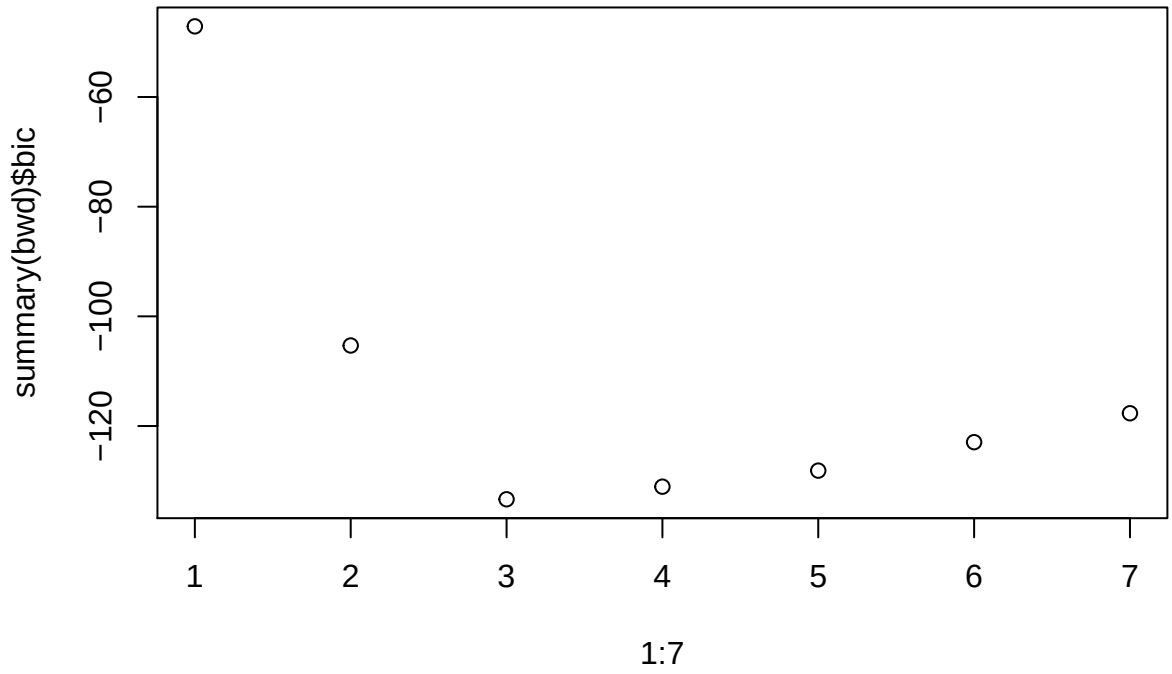
```
## Subset selection object
## Call: regsubsets.formula(new_PrizeMoney ~ new_DrivingAccuracy + new_GIR +
##     new_PuttingAverage + new_BirdieConversion + new_SandSaves +
##     new_Scrambling + new_PuttsPerRound, data = y.golf, method = "exhaustive")
## 7 Variables  (and intercept)
##                       Forced in Forced out
## new_DrivingAccuracy       FALSE      FALSE
## new_GIR                   FALSE      FALSE
## new_PuttingAverage        FALSE      FALSE
## new_BirdieConversion      FALSE      FALSE
## new_SandSaves             FALSE      FALSE
## new_Scrambling            FALSE      FALSE
## new_PuttsPerRound         FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          new_DrivingAccuracy new_GIR new_PuttingAverage new_BirdieConversion
## 1  ( 1 ) " "                 "*"     " "                 " "
## 2  ( 1 ) " "                 "*"     " "                 " "
## 3  ( 1 ) " "                 "*"     " "                 "*"
## 4  ( 1 ) " "                 "*"     " "                 "*"
## 5  ( 1 ) " "                 "*"     " "                 "*"
## 6  ( 1 ) "*"                 "*"     " "                 "*"
## 7  ( 1 ) "*"                 "*"     "*"                 "*"
##          new_SandSaves new_Scrambling new_PuttsPerRound
```

```
## 1  ( 1 ) " "           " "           " "
## 2  ( 1 ) " "           " "           "*"
## 3  ( 1 ) " "           "*"           " "
## 4  ( 1 ) "*"           "*"           " "
## 5  ( 1 ) "*"           "*"           "*"
## 6  ( 1 ) "*"           "*"           "*"
## 7  ( 1 ) "*"           "*"           "*"
```

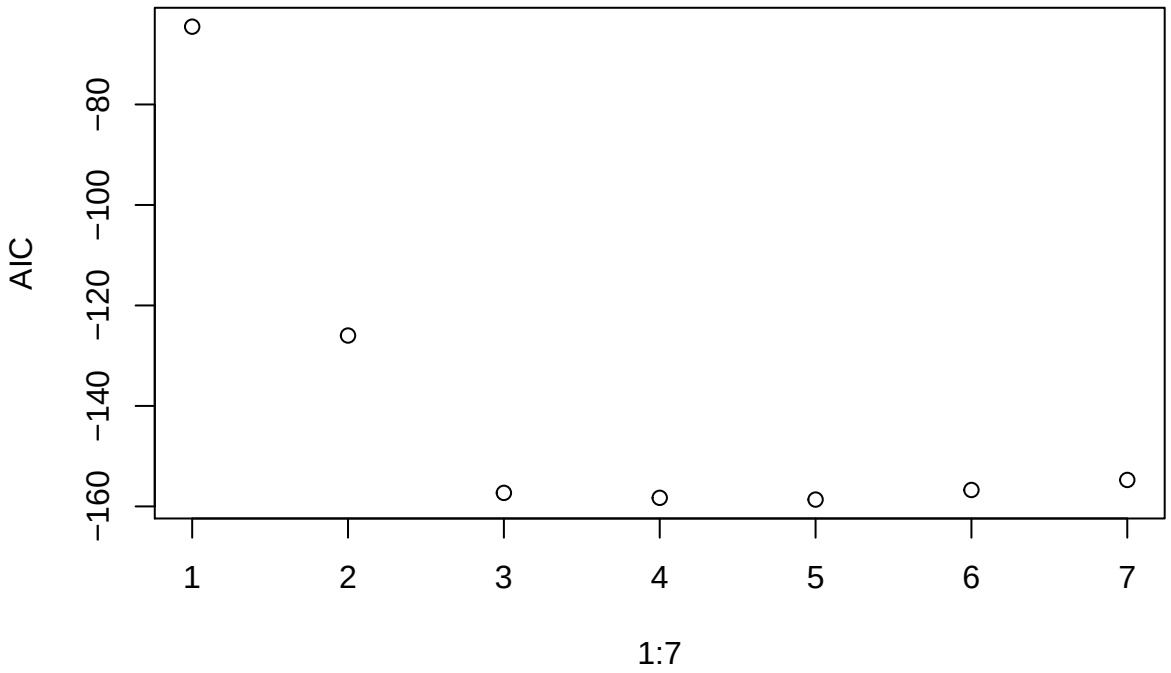
Based on all possible subsets, the optimal model according to BIC is a 3 variable model with *GIR*, *BirdieConversion*, and *Scrambling*. While the optimal model according to AIC is a 5 variable model with *GIR*, *BirdieConversion*, *SandSaves*, *Scrambling*, and *PuttsPerRound*.

b) Identify the optimal model or models based on AIC and BIC from the approach based on backward selection.

```
bwd <-regsubsets(new_PrizeMoney~new_DrivingAccuracy+new_GIR+new_PuttingAverage+new_BirdieConversion+new_
plot(1:7, summary(bwd)$bic)
```




```
#3 variable model has the lowest BIC
RSs <- summary(bwd)$rss
n <- nrow(y.golf)
p <- 1:7
AIC <- n * log(RSs/n) + 2 * p
plot(1:7, AIC)
```




```
# 5 variable model has the lowest AIC
summary(bwd)
```

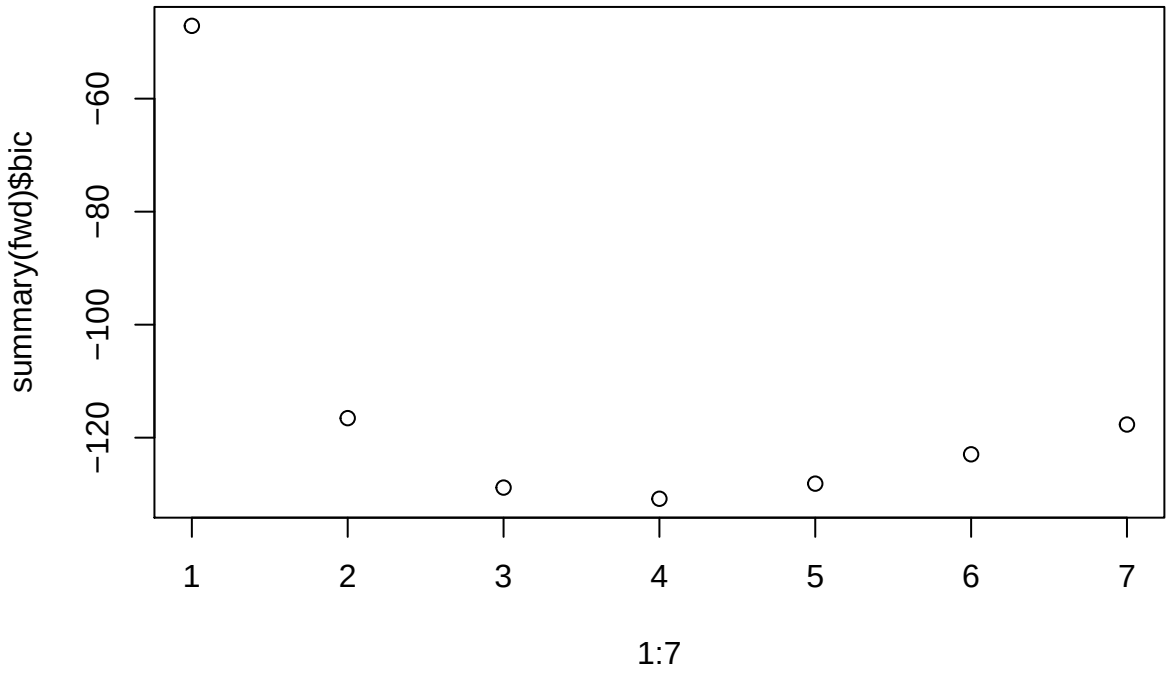
```
## Subset selection object
## Call: regsubsets.formula(new_PrizeMoney ~ new_DrivingAccuracy + new_GIR +
##     new_PuttingAverage + new_BirdieConversion + new_SandSaves +
##     new_Scrambling + new_PuttsPerRound, data = y.golf, method = "backward")
## 7 Variables  (and intercept)
##                       Forced in Forced out
## new_DrivingAccuracy       FALSE      FALSE
## new_GIR                   FALSE      FALSE
## new_PuttingAverage        FALSE      FALSE
## new_BirdieConversion      FALSE      FALSE
## new_SandSaves             FALSE      FALSE
## new_Scrambling            FALSE      FALSE
## new_PuttsPerRound         FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: backward
##          new_DrivingAccuracy new_GIR new_PuttingAverage new_BirdieConversion
## 1  ( 1 ) " "                 "*"     " "                 " "
## 2  ( 1 ) " "                 "*"     " "                 "*"
## 3  ( 1 ) " "                 "*"     " "                 "*"
## 4  ( 1 ) " "                 "*"     " "                 "*"
## 5  ( 1 ) " "                 "*"     " "                 "*"
## 6  ( 1 ) "*"                 "*"     " "                 "*"
## 7  ( 1 ) "*"                 "*"     "*"                 "*"
##          new_SandSaves new_Scrambling new_PuttsPerRound
## 1  ( 1 ) " "           " "           " "
## 2  ( 1 ) " "           " "           " "
## 3  ( 1 ) " "           "*"           " "
## 4  ( 1 ) "*"           "*"           " "
## 5  ( 1 ) "*"           "*"           "*"
## 6  ( 1 ) "*"           "*"           "*"
```

```
## 7  ( 1 ) "*"           "*"           "*"
```

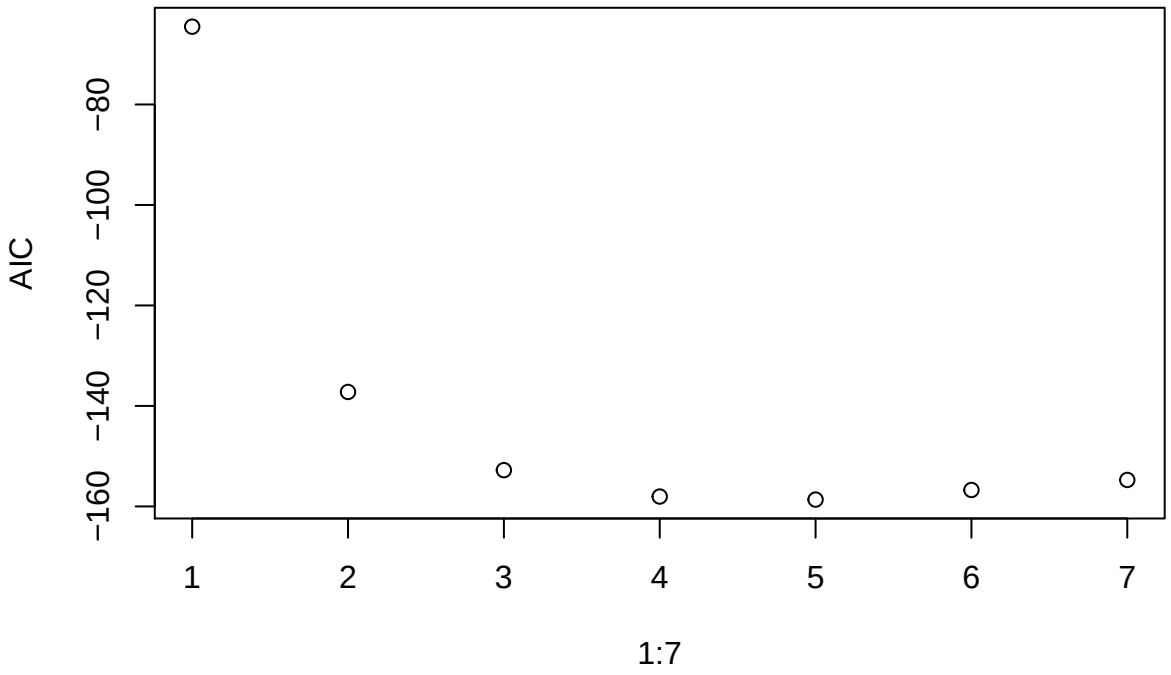
Based on backward selection, the optimal model according to BIC is a 3 variable model with *GIR*, *BirdieConversion*, and *Scrambling*. However, the optimal model according to AIC is a 5 variable model with *GIR*, *BirdieConverison*, *Scrambling*, *SandSaves*, and *PuttsPerRound*.

c) Identify the optimal model or models based on AIC and BIC from the approach based on forward selection.

```
library(leaps)

fwd <- regsubsets(new_PrizeMoney~new_DrivingAccuracy+new_GIR+new_PuttingAverage+new_BirdieConversion+new
plot(1:7, summary(fwd)$bic)
```




```
# 4 variable model has lowest BIC

rss <- summary(fwd)$rss
n <- nrow(y.golf)
p <- 1:7
AIC <- n * log(rss/n) + 2 * p
plot(1:7, AIC)
```




```
summary(fwd)
```

```
## Subset selection object
## Call: regsubsets.formula(new_PrizeMoney ~ new_DrivingAccuracy + new_GIR +
##     new_PuttingAverage + new_BirdieConversion + new_SandSaves +
##     new_Scrambling + new_PuttsPerRound, data = y.golf, method = "forward")
## 7 Variables  (and intercept)
##                       Forced in Forced out
## new_DrivingAccuracy       FALSE      FALSE
## new_GIR                   FALSE      FALSE
## new_PuttingAverage        FALSE      FALSE
## new_BirdieConversion      FALSE      FALSE
## new_SandSaves             FALSE      FALSE
## new_Scrambling            FALSE      FALSE
## new_PuttsPerRound         FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: forward
##          new_DrivingAccuracy new_GIR new_PuttingAverage new_BirdieConversion
## 1  ( 1 ) " "                 "*"     " "                 " "
## 2  ( 1 ) " "                 "*"     " "                 " "
## 3  ( 1 ) " "                 "*"     " "                 "*"
## 4  ( 1 ) " "                 "*"     " "                 "*"
## 5  ( 1 ) " "                 "*"     " "                 "*"
## 6  ( 1 ) "*"                 "*"     " "                 "*"
## 7  ( 1 ) "*"                 "*"     "*"                 "*"
##          new_SandSaves new_Scrambling new_PuttsPerRound
## 1  ( 1 ) " "           " "            " "
## 2  ( 1 ) " "           " "            "*"
## 3  ( 1 ) " "           " "            "*"
## 4  ( 1 ) " "           "*"            "*"
## 5  ( 1 ) "*"           "*"            "*"
## 6  ( 1 ) "*"           "*"            "*"
## 7  ( 1 ) "*"           "*"            "*"
```

```
# 5 variable model has the lowest AIC
```

Based on forward selection, the model with the lowest BIC has 4 variables consisting of GIR, $PuttsPerRound$, $BirdieConversion$ and $Scrambling$. The model with the lowest AIC includes 5 variables including GIR, $PuttsPerRound$, $BirdieConversion$, $SandSaves$, and $Scrambling$.

d) Carefully explain why the models chosen in a) and c) are not the same while those in a) and b) are the same.

One of the reasons why the models chosen in a and c are not the same while those in a and b are the same are due to the fact that the three different methods that were applied in the arlier questions approach the final model from different directions. In a) best subsets is applied so the algorithm goes through all possible variations of the model including the full model which explains why it results in the same model as b, which goes through backwards selection. This means it starts the algorithm with all potential predictors (aka the full model) and step by step deletes predictors with large p-values. Part c uses forward stepwise regression which starts with the singular predictor that has the highest correlation (r-squared) with the log transform of PrizeMoney. Since forward selection starts with only one variable, it requires fewer checks but also lends itself to potentially miss predictors that are checked in both the backward and best subsets methods (since both check for the full model). These missing predictors may be significant and thus result in an oversimplified model. Therefore, the models in a and c are not the same while a and b are.

Another potential reason for getting different results in a and c is potentially due to co-linearity in the predictors. Collinearity is a condition in which some of the independent variables are highly correlated and one predictor variable (in a multiple regression model) can be linearly predicted from the others. If there was no co-linearity at all, then adding and removing predictors would not change any of the p-values, so the three methods (best subsets, forward stepwise, and backward stepwise) would produce the same results. Since best subsets and forward stepwise regression did not produce the same results this means that there is a chance that co-linearity is present amongst some of the variables.

e) Recommend a final model. Give detailed reasons to support your choice.

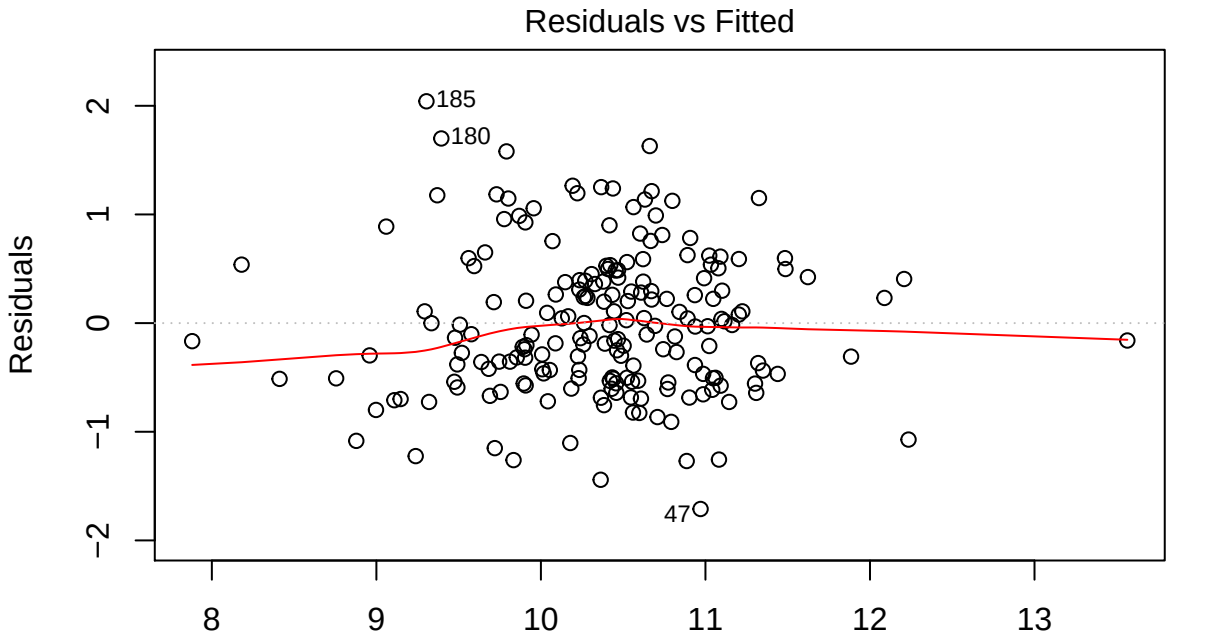
```
library(car)
five.model <- lm(new_PrizeMoney~new_GIR + new_PuttsPerRound + new_BirdieConversion + new_SandSaves + ne
vif(five.model)
```

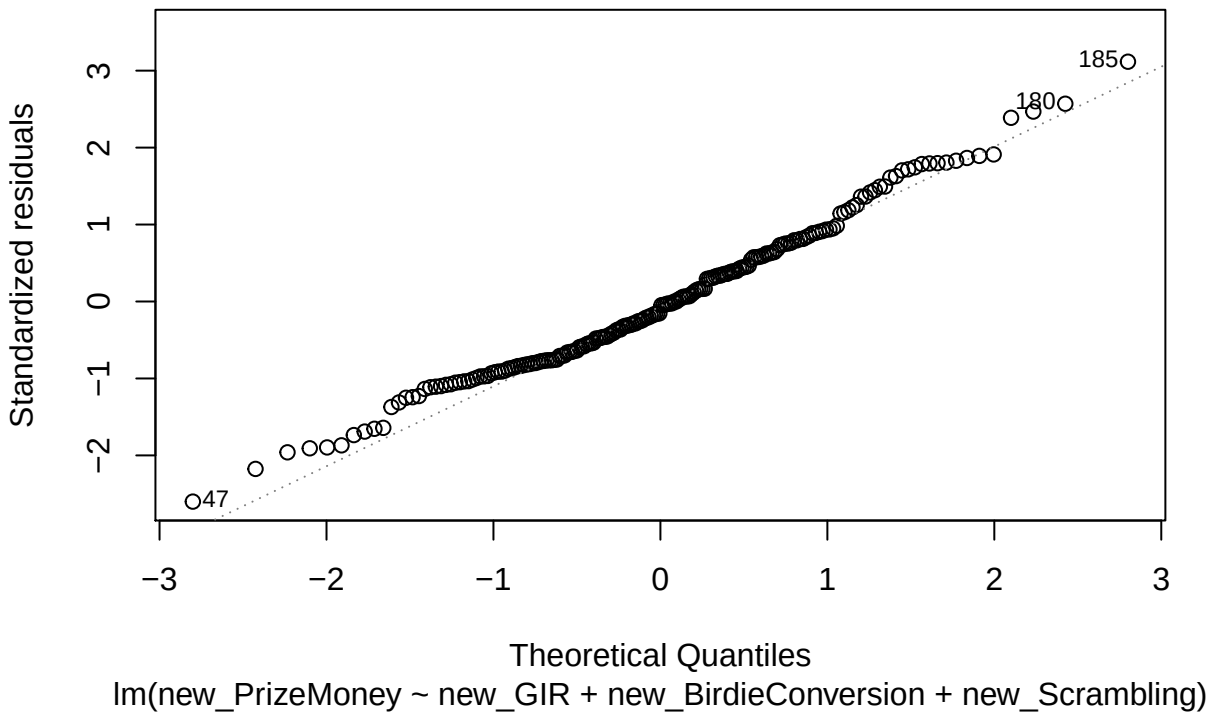
```
##              new_GIR    new_PuttsPerRound new_BirdieConversion
##             2.730165             4.652336             2.322693
##        new_SandSaves       new_Scrambling
##             1.441054             2.734765
```

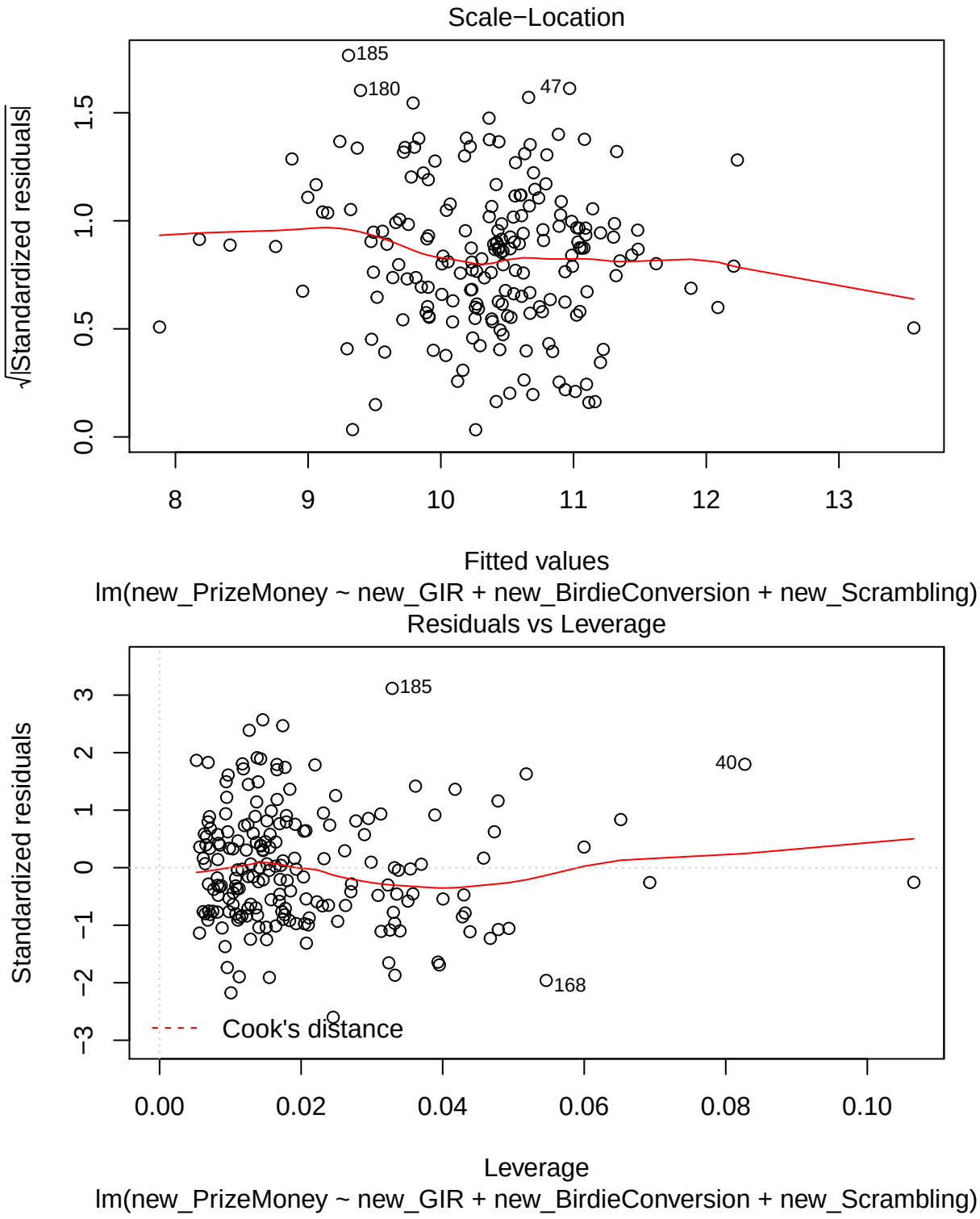
```
three.model <- lm(new_PrizeMoney~new_GIR + new_BirdieConversion + new_Scrambling, data = y.golf)
vif(three.model)
```

```
##              new_GIR new_BirdieConversion       new_Scrambling
##             1.040396             1.001936             1.040516
```

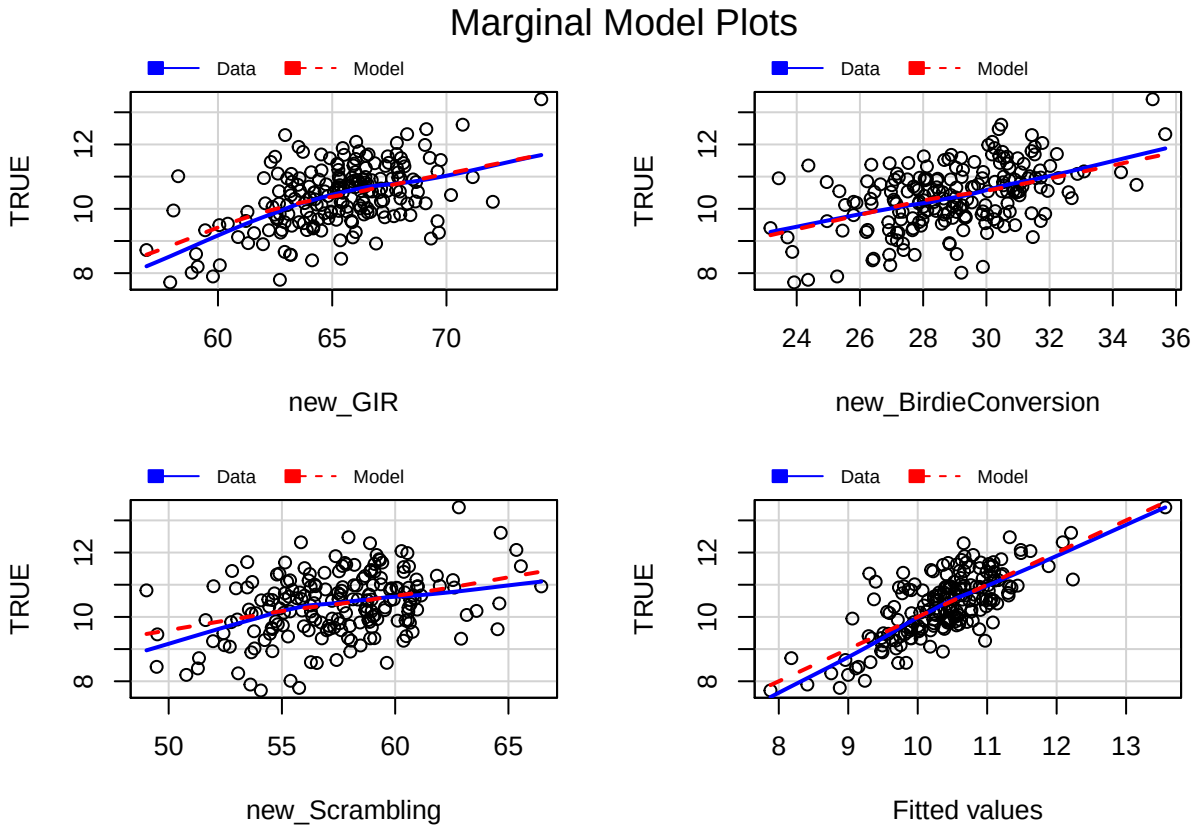
```
plot(three.model)
```


Residuals vs Fitted
Residuals
185
180
47
2
1
0
−1
−2
8
9
10
11
12
13
Fitted values
lm(new_PrizeMoney ~ new_GIR + new_BirdieConversion + new_Scrambling)


Normal Q−Q
Standardized residuals
185
180
47
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(new_PrizeMoney ~ new_GIR + new_BirdieConversion + new_Scrambling)




```
mmps(three.model)
```



For the final model I recommend the 3 variable model found using the best subsets approach and including GIR, BirdieConversion, and Scrambling. Since this approach tests all the best possible model combinations, we know that either the 3 variable or 5 variable model would be the best one. Furthermore, since the 3 variable model is simpler, it is preferred to the 5 variable one. Likewise, using the variance inflation factor to test for collinearity in the 5 variable model, while none had values greater than 5, PuttsPerRound had a value of 4.652 which may indicate slight potential correlation with another variable in the model. The model satisfies the linearity condition (no pattern in the residual plot), normality condition (little deviation from a straight line), and constant variance condition (no fanshape in the residual plot or increasing or decreasing trend in scale location plot). Looking at the marginal model plots, the loess lines are almost the same as the regression lines indicating good fit. Therefore, all together, I would recommend the 3 variable model.

f) Interpret the regression coefficients in the final model. Is it necessary to be cautious about taking these results too literally?

```
summary(three.model)
```

```
##
## Call:
## lm(formula = new_PrizeMoney ~ new_GIR + new_BirdieConversion +
##     new_Scrambling, data = y.golf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71081 -0.50717 -0.06683  0.41975  2.04147
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -11.08314    1.45712  -7.606 1.23e-12 ***
## new_GIR                 0.15658    0.01787   8.761 1.01e-15 ***
```

```
## new_BirdieConversion   0.20625    0.02164   9.531  < 2e-16 ***
## new_Scrambling         0.09178    0.01539   5.965 1.16e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6661 on 192 degrees of freedom
## Multiple R-squared:  0.5453, Adjusted R-squared:  0.5382
## F-statistic: 76.75 on 3 and 192 DF,  p-value: < 2.2e-16
```

Controlling for all other variables in the model, with an additional percentage point in GIR (green in regulation), on average, the amount of prize money per tournament increases by a factor of 0.15658.

Given that the model accounts for all other variables, for an additional percentage point in Birdie Conversion, on average, the amount of prize money per tournament increases by a factor of 0.20625.

Finally, given that the model accounts for all other variables, for a one percent difference in Scrambling, on average, the amount of prize money per tournament increases by a factor of 0.09178.

Since no model is ever perfect nor can it exactly predict the future, it is necessary to be cautious about taking these results too literally. Likewise, since we did not account for multi-colinearity the actual predictive power of this model may be inflated and thus should not be taken too literally.