Stats 101A Homework 4 (Lecture 1B)

Charles Liu (304804942)

2/28/2020

Loading Necessary Packages:

```
library(readr)
library(car)
```

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

Problem 1:

```
getwd()
```

```
## [1] "C:/Users/cliuk/Documents/UCLA Works/UCLA Winter 2020/Stats 101A/Homeworks/HW 4"
```

```
Prob1 <- read_csv("C:/Users/cliuk/Documents/UCLA Works/UCLA Winter 2020/Stats 101A/Homeworks/HW 4/overdu
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   LATE = col_double(),
##   BILL = col_double(),
##   TYPE = col_character()
## )
```

```
names(Prob1)
```

```
## [1] "X1"   "LATE" "BILL" "TYPE"
```

```
attach(Prob1)

m1 <- lm(LATE ~ BILL)
summary(m1)
```

```
##
## Call:
## lm(formula = LATE ~ BILL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.846 -17.212  -0.793  19.007  47.774
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.98390    5.96405   8.716 9.84e-14 ***
## BILL        -0.01264    0.03128  -0.404    0.687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.72 on 94 degrees of freedom
## Multiple R-squared:  0.001734,   Adjusted R-squared:  -0.008885
## F-statistic: 0.1633 on 1 and 94 DF,  p-value: 0.687
```

Problem 2:

```
Prob2 <- read.table("C:/Users/cliuk/Documents/UCLA Works/UCLA Winter 2020/Stats 101A/Homeworks/HW 4/Lat
attach(Prob2)
```

2a)

```
m2 <- lm(Quality ~ EndofHarvest * Rain)
summary(m2)
```

```
##
## Call:
## lm(formula = Quality ~ EndofHarvest * Rain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6833 -0.5703  0.1265  0.4385  1.6354
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.16122    0.68917   7.489 3.95e-09 ***
## EndofHarvest      -0.03145    0.01760  -1.787   0.0816 .
## Rain               1.78670    1.31740   1.356   0.1826
## EndofHarvest:Rain -0.08314    0.03160  -2.631   0.0120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7578 on 40 degrees of freedom
## Multiple R-squared:  0.6848, Adjusted R-squared:  0.6612
## F-statistic: 28.97 on 3 and 40 DF,  p-value: 4.017e-10
```

The p-value = 0.0120 < 0.05 –> we can say it is significant.

2b)

The Linear Equation Estimated: Quality = 5.16122 + (-0.03145)*(EndofHarvest)* + *(1.78670)*(Rain) + (-0.08314)*(EndofHarvest)*(Rain)

Case (i): No unwanted rain at the harvest; Quality = 5.16122 + (-0.03145)*(EndofHarvest)* + *(1.78670)*(0) + (-0.08314)*(EndofHarvest)*(0) for Rain = 0 –> Quality = 5.16122 + (-0.03145)*(EndofHarvest) –> slope = -0.03145. Thus to decrease by 1 unit of quality –> -1/-0.03145 –> 31.7965 days OR approximately 32 days.

Case (ii): Some unwanted rain at the harvest; Quality = 5.16122 + (-0.03145)*(EndofHarvest)* + *(1.78670)*(1) + (-0.08314)*(EndofHarvest)*(1) for Rain = 1 –> Quality = 6.94792 + (-0.11459)*(EndofHarvest) –> slope =

-0.11459. Thus to decrease by 1 unit of quality –> -1/-0.11459 –> 8.726765 days OR approximately 9 days.

Problem 3:

3a)

My 2 Concerns are: (1) multicollinearity –> important & (2) linearity assumptions –> normality.

Problem 4:

```
Prob4 <- read_csv("C:/Users/cliuk/Documents/UCLA Works/UCLA Winter 2020/Stats 101A/Homeworks/HW 4/cars
```

```
## Parsed with column specification:
## cols(
##   `Vehicle Name` = col_character(),
##   Hybrid = col_double(),
##   SuggestedRetailPrice = col_double(),
##   DealerCost = col_double(),
##   EngineSize = col_double(),
##   Cylinders = col_double(),
##   Horsepower = col_double(),
##   CityMPG = col_double(),
##   HighwayMPG = col_double(),
##   Weight = col_double(),
##   WheelBase = col_double(),
##   Length = col_double(),
##   Width = col_double()
## )
```

```
attach(Prob4)
```

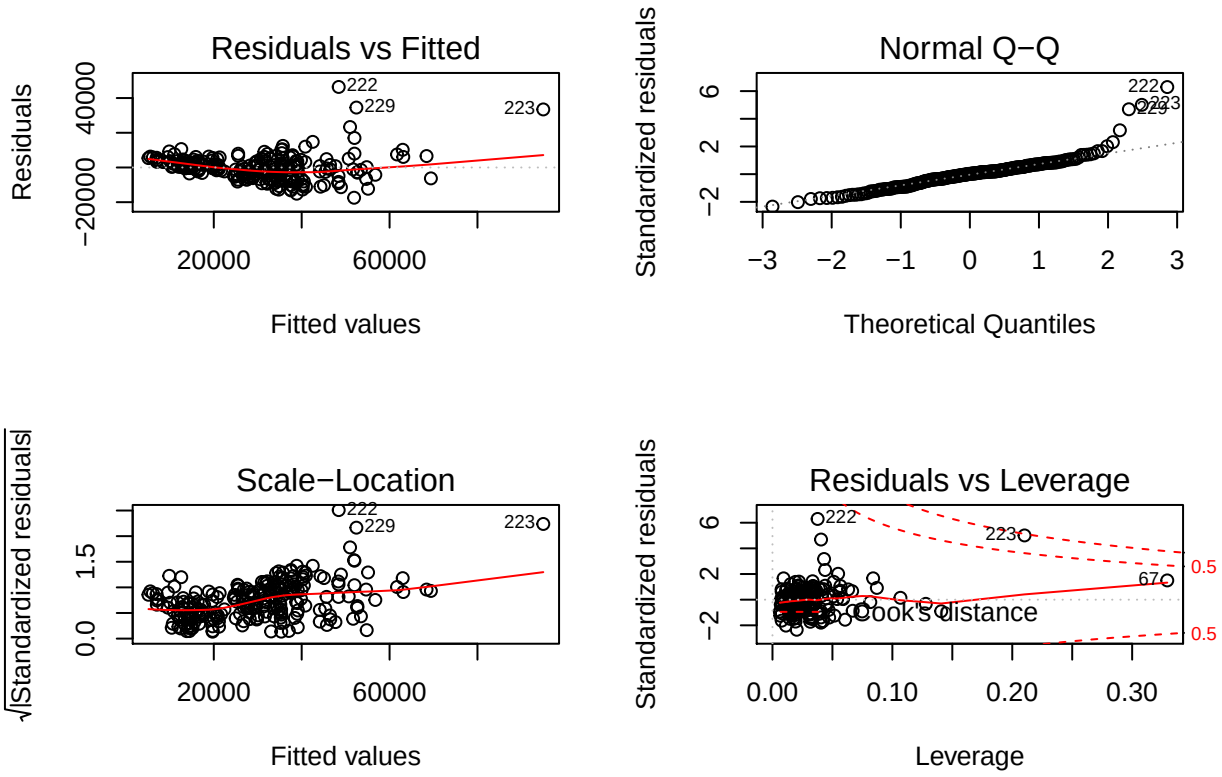
4a)

```
m3 <- lm(SuggestedRetailPrice ~ EngineSize + Cylinders + Horsepower + HighwayMPG +
           Weight + WheelBase)
summary(m3)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ EngineSize + Cylinders +
##     Horsepower + HighwayMPG + Weight + WheelBase)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17433  -4124    156   3573  46392
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -69349.733  15213.086  -4.559 8.42e-06 ***
## EngineSize   -6962.501   1595.046  -4.365 1.93e-05 ***
## Cylinders     3569.471    965.224   3.698 0.000272 ***
## Horsepower     179.805     16.311  11.024  < 2e-16 ***
## HighwayMPG     647.663    148.915   4.349 2.06e-05 ***
```

```
## Weight          11.965      2.538   4.714 4.24e-06 ***
## WheelBase       46.585    177.095   0.263 0.792751
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7517 on 227 degrees of freedom
## Multiple R-squared:  0.7819, Adjusted R-squared:  0.7761
## F-statistic: 135.6 on 6 and 227 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(m3)
```



Not a valid model because two predictors are not significant, the plot violates the normality assumption, violates outliers and leverages, and violates linearity.

4b)

The residuals plot has a curved pattern, so we should do some transformation. It violated the linearity assumption.

4c)

```r
leverage <- hatvalues(m3)

### h_ii > 2 * (p + 1)/n for p = "number of predictors"
nrow(Prob4) # n = 234 & there are 7 predictors (p = 7)
```

```
## [1] 234
```

```
which(leverage >= 2 * (7 + 1)/234 & abs(rstandard(m3)) >= 2)
```

```
## 223
## 223
```

```
### 223 & 223 --> bad leverage

### h_ii > 2 * average(h_ii)
which(leverage >= 2 * mean(leverage) & abs(rstandard(m3)) >= 2)
```

```
## 223
## 223
```

```
### 223 & 223 --> bad leverage
```

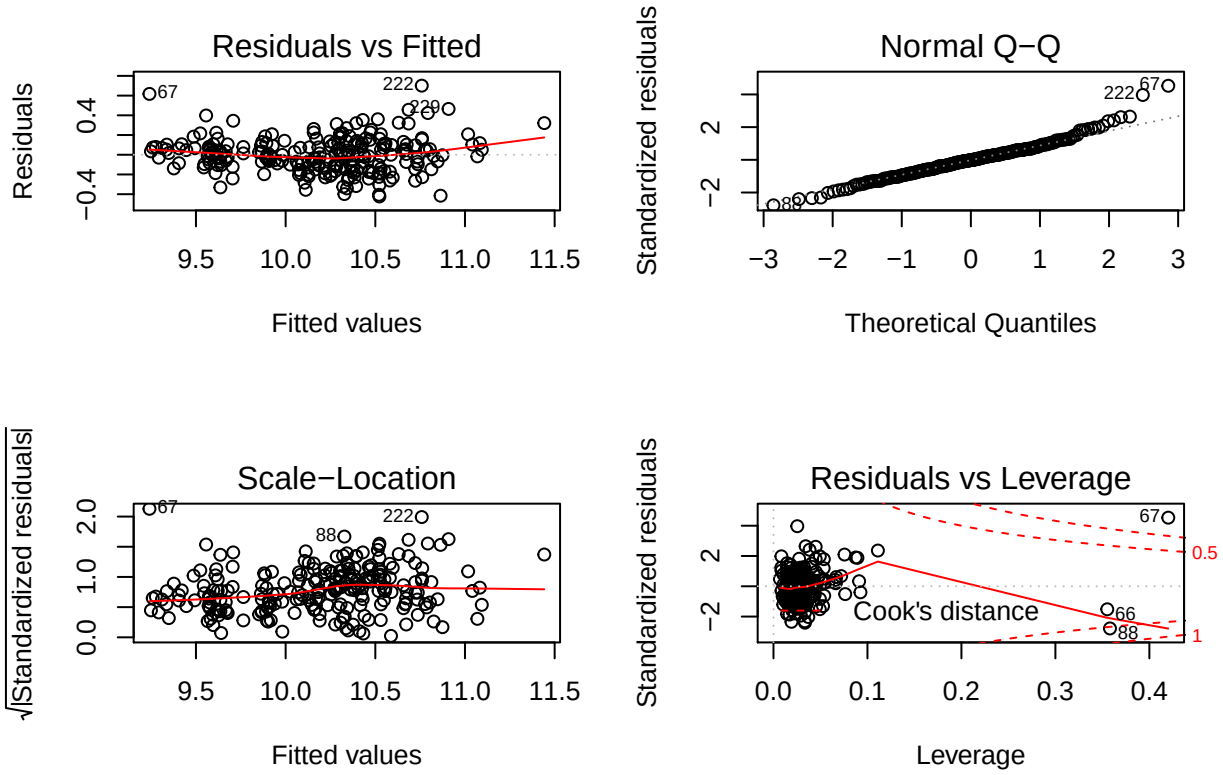
We can see that 223 and 223 are bad leverages.

4d)

```
m4 <- lm(log(SuggestedRetailPrice) ~ I(EngineSize^(0.25)) + I(log(Cylinders)) +
                  I(log(Horsepower)) + I(1/HighwayMPG) + Weight + I(log(WheelBase)) +
                  Hybrid)
summary(m4)
```

```
##
## Call:
## lm(formula = log(SuggestedRetailPrice) ~ I(EngineSize^(0.25)) +
##     I(log(Cylinders)) + I(log(Horsepower)) + I(1/HighwayMPG) +
##     Weight + I(log(WheelBase)) + Hybrid)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42288 -0.10983 -0.00203  0.10279  0.70068
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.703e+00  2.010e+00   2.838  0.00496 **
## I(EngineSize^(0.25)) -1.575e+00  3.332e-01  -4.727 4.01e-06 ***
## I(log(Cylinders))       2.335e-01  1.204e-01   1.940  0.05359 .
## I(log(Horsepower))      8.992e-01  8.876e-02  10.130  < 2e-16 ***
## I(1/HighwayMPG)         8.029e-01  4.758e+00   0.169  0.86614
## Weight                  5.043e-04  6.367e-05   7.920 1.07e-13 ***
## I(log(WheelBase))      -6.385e-02  4.715e-01  -0.135  0.89240
## Hybrid                  6.422e-01  1.150e-01   5.582 6.78e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1789 on 226 degrees of freedom
## Multiple R-squared:  0.8621, Adjusted R-squared:  0.8578
## F-statistic: 201.8 on 7 and 226 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(m4)
```



It is an improvement of the old model (m3), but it is still an invalid model. The new model (m4) has 2 predictors that are not significant, the plot the violates outliers and leverages, and violates linearity assumption.

4e)

```
m5 <- lm(log(SuggestedRetailPrice) ~ I(EngineSize^(0.25)) + I(log(Cylinders)) +
           I(log(Horsepower)) + Weight + Hybrid)
summary(m5)
```

```
##
## Call:
## lm(formula = log(SuggestedRetailPrice) ~ I(EngineSize^(0.25)) +
##     I(log(Cylinders)) + I(log(Horsepower)) + Weight + Hybrid)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42224 -0.11001 -0.00099  0.10191  0.70205
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.422e+00  3.291e-01  16.474  < 2e-16 ***
## I(EngineSize^(0.25)) -1.591e+00  3.157e-01  -5.041 9.45e-07 ***
## I(log(Cylinders))       2.375e-01  1.186e-01   2.003   0.0463 *
## I(log(Horsepower))      9.049e-01  8.305e-02  10.896  < 2e-16 ***
## Weight                  5.029e-04  5.203e-05   9.666  < 2e-16 ***
```

```
## Hybrid               6.340e-01  1.080e-01   5.870 1.53e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1781 on 228 degrees of freedom
## Multiple R-squared:  0.862,  Adjusted R-squared:  0.859
## F-statistic: 284.9 on 5 and 228 DF,  p-value: < 2.2e-16
```

```
anova(m4, m5)
```

```
## Analysis of Variance Table
##
## Model 1: log(SuggestedRetailPrice) ~ I(EngineSize^(0.25)) + I(log(Cylinders)) +
##     I(log(Horsepower)) + I(1/HighwayMPG) + Weight + I(log(WheelBase)) +
##     Hybrid
## Model 2: log(SuggestedRetailPrice) ~ I(EngineSize^(0.25)) + I(log(Cylinders)) +
##     I(log(Horsepower)) + Weight + Hybrid
##   Res.Df    RSS Df  Sum of Sq     F Pr(>F)
## 1    226 7.2337
## 2    228 7.2358 -2 -0.0021769 0.034 0.9666
```

Since the p-value for the F-test is large, we can remove the 2 predictors.

4f)

A new categorical variable with Manufacturer, then add it to the regression model.

Problem 5:

```
Prob5 <- read_csv("C:/Users/cliuk/Documents/UCLA Works/UCLA Winter 2020/Stats 101A/Homeworks/HW 4/pgatou
```

```
## Parsed with column specification:
## cols(
##   Name = col_character(),
##   TigerWoods = col_double(),
##   PrizeMoney = col_double(),
##   AveDrivingDistance = col_double(),
##   DrivingAccuracy = col_double(),
##   GIR = col_double(),
##   PuttingAverage = col_double(),
##   BirdieConversion = col_double(),
##   SandSaves = col_double(),
##   Scrambling = col_double(),
##   BounceBack = col_double(),
##   PuttsPerRound = col_double()
## )
```

```
attach(Prob5)
```

5a)

```
summary(powerTransform(cbind(PrizeMoney, DrivingAccuracy, GIR, PuttingAverage,
                             BirdieConversion, SandSaves, Scrambling, PuttsPerRound) ~ 1))
```

```
## bcPower Transformations to Multinormality
```

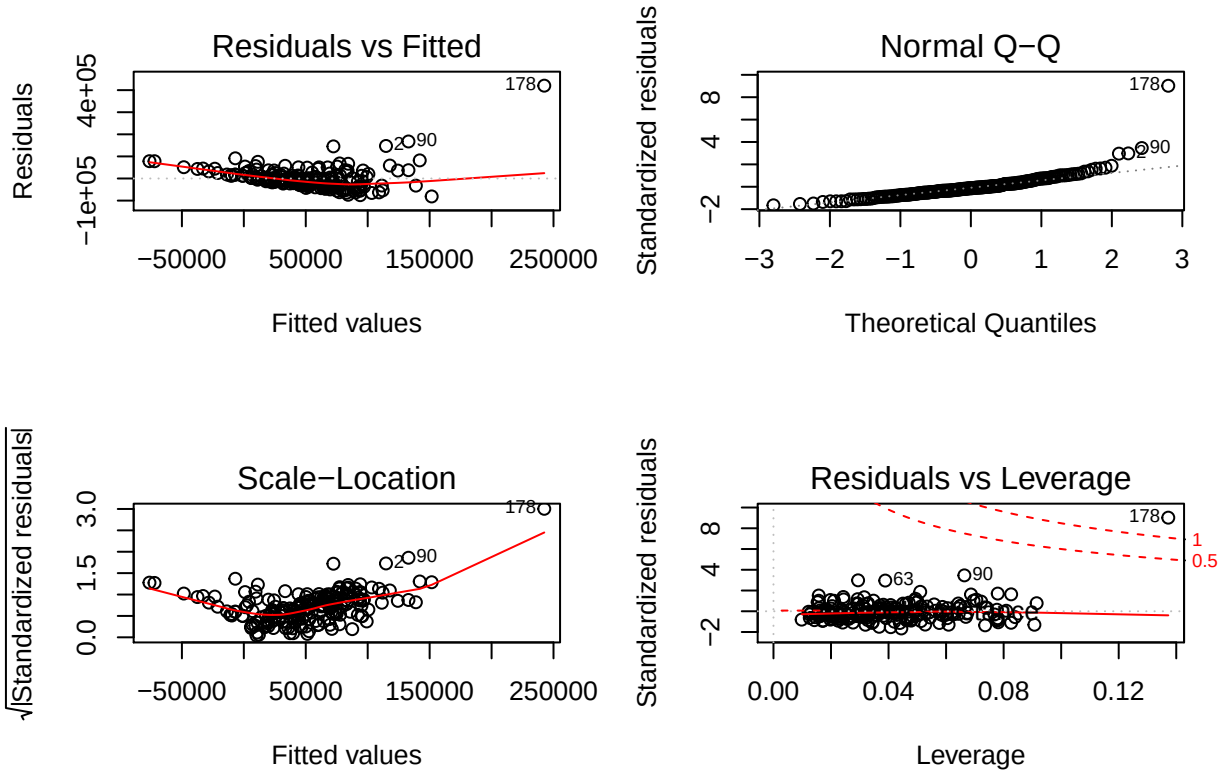
```
##                   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## PrizeMoney          0.0364           0      -0.0684       0.1413
## DrivingAccuracy     0.2859           1      -0.8679       1.4397
## GIR                 1.5181           1       0.1076       2.9285
## PuttingAverage      0.9967           1      -0.9977       2.9911
## BirdieConversion    0.8962           1      -0.1107       1.9031
## SandSaves           0.9921           1       0.0733       1.9109
## Scrambling          0.6827           1      -0.7106       2.0759
## PuttsPerRound      -0.0085           1      -3.1342       3.1172
##
## 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 0) 14.69767  8 0.065298
##
## Likelihood ratio test that no transformations are needed
##                                           LRT df       pval
## LR test, lambda = (1 1 1 1 1 1 1 1) 338.243  8 < 2.22e-16
```

Yes, I agree with this PowerTransformation because the LRT test shows that all parameters equal to zero.

```
m6a_1 <- lm(PrizeMoney ~ DrivingAccuracy + GIR + PuttingAverage +
           BirdieConversion + SandSaves + Scrambling + PuttsPerRound)
summary(m6a_1)
```

```
##
## Call:
## lm(formula = PrizeMoney ~ DrivingAccuracy + GIR + PuttingAverage +
##     BirdieConversion + SandSaves + Scrambling + PuttsPerRound)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -81239 -26260  -6521  17539 420230
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1165233.1   587382.9  -1.984 0.048737 *
## DrivingAccuracy     -1835.8      889.2  -2.065 0.040326 *
## GIR                  9671.3     3309.4   2.922 0.003899 **
## PuttingAverage     -47435.3   521566.4  -0.091 0.927631
## BirdieConversion    10426.0     3049.6   3.419 0.000771 ***
## SandSaves            1182.1      744.8   1.587 0.114184
## Scrambling           4741.3     2400.8   1.975 0.049749 *
## PuttsPerRound        5267.5    35765.7   0.147 0.883070
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50140 on 188 degrees of freedom
## Multiple R-squared:  0.4064, Adjusted R-squared:  0.3843
## F-statistic: 18.39 on 7 and 188 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(m6a_1)
```

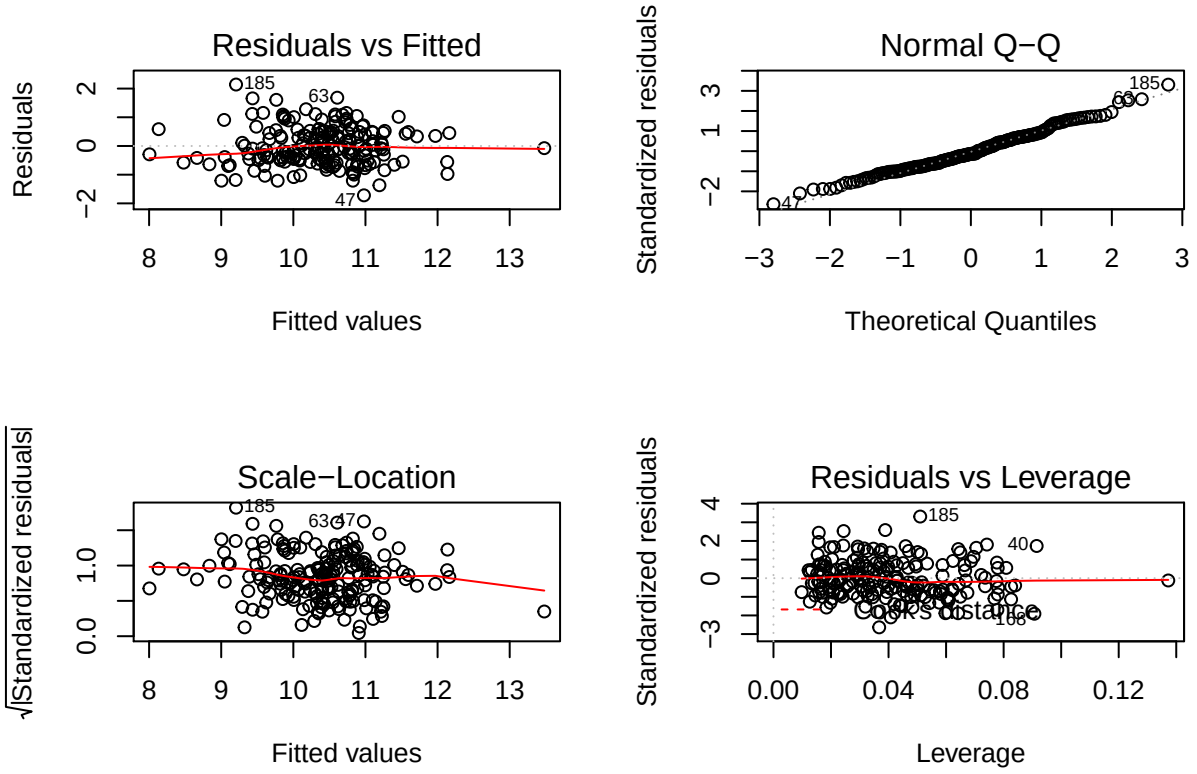



```
m6a_2 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage +
            BirdieConversion + SandSaves + Scrambling + PuttsPerRound)
summary(m6a_2)
```

```
##
## 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
```

```
par(mfrow = c(2, 2))
plot(m6a_2)
```



I agree with the log Transformation because the R-squared is greater with the log(y) and the diagnostic plots are better. For instance, the Residuals vs. Fitted is more equally spread and is a straight horizontal line. The normality assumption is fulfilled, Scale-Location plot is satisfied, and lastly the outliers and leverages are okay.

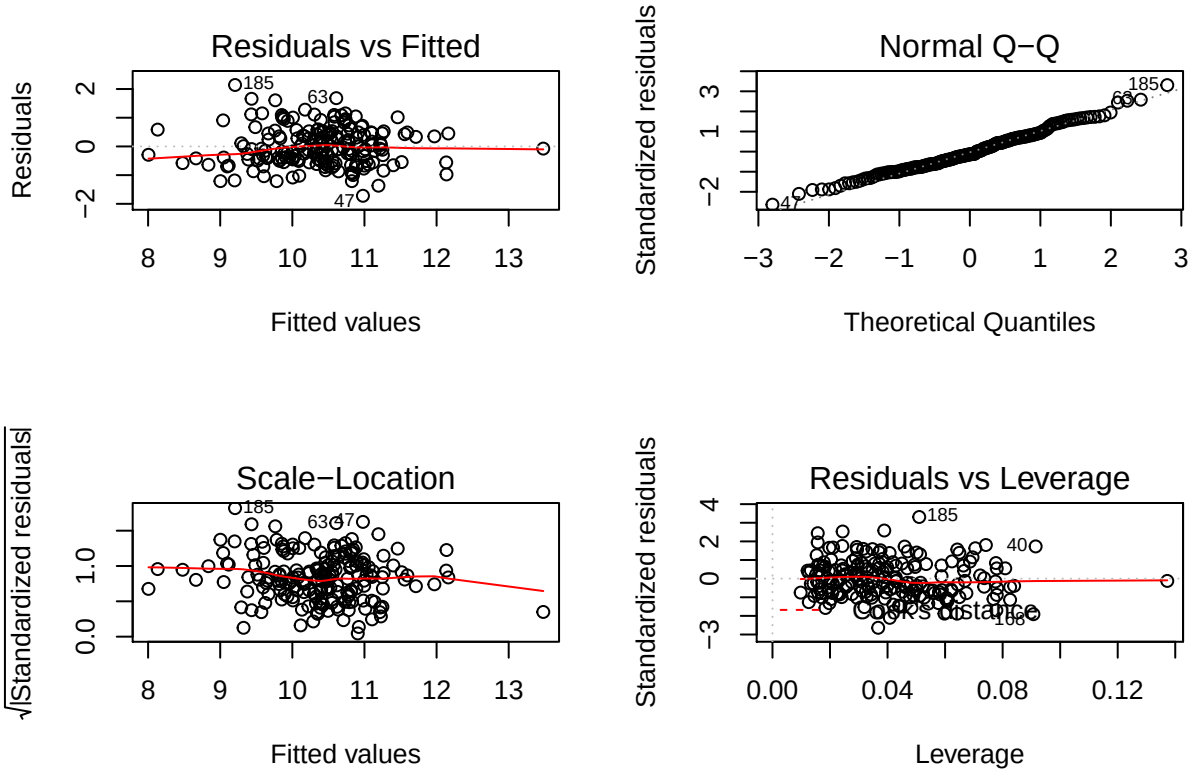
5b)

```
m6b_1 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage +
              BirdieConversion + SandSaves + Scrambling + PuttsPerRound)
summary(m6b_1)
```

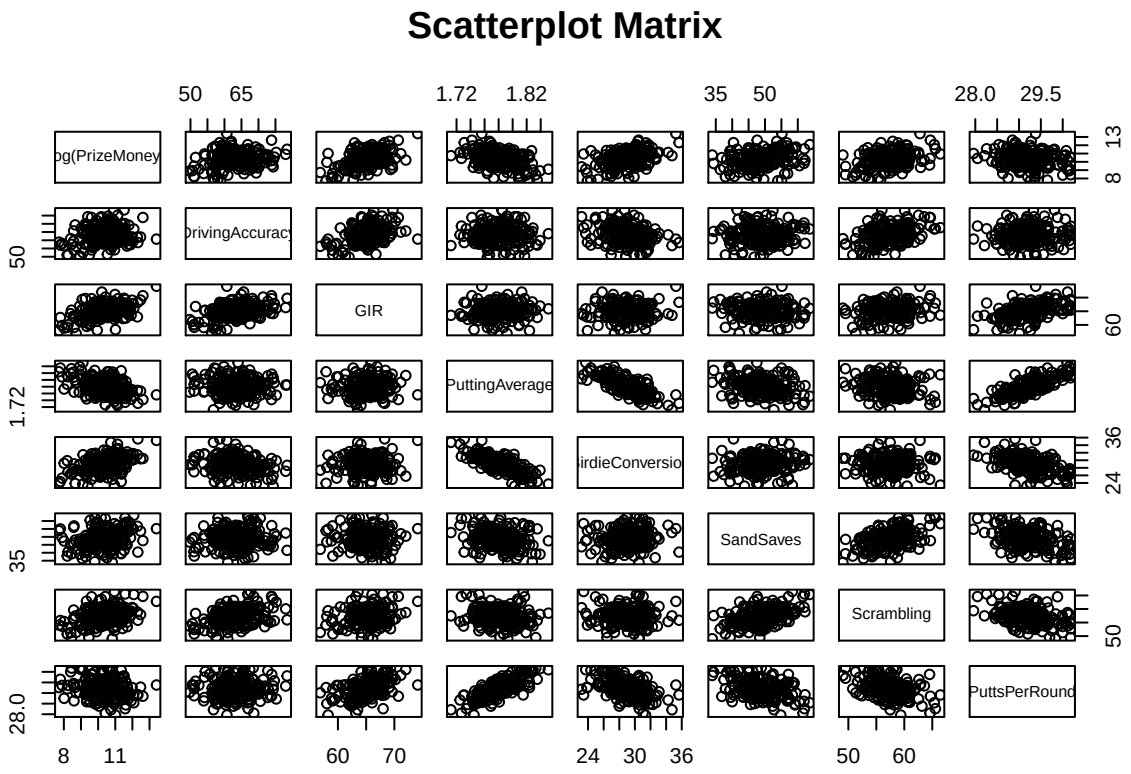
```
##
## Call:
## lm(formula = log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage +
##     BirdieConversion + SandSaves + Scrambling + PuttsPerRound)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71949 -0.48608 -0.09172  0.44561  2.14013
##
```

```
## 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
```

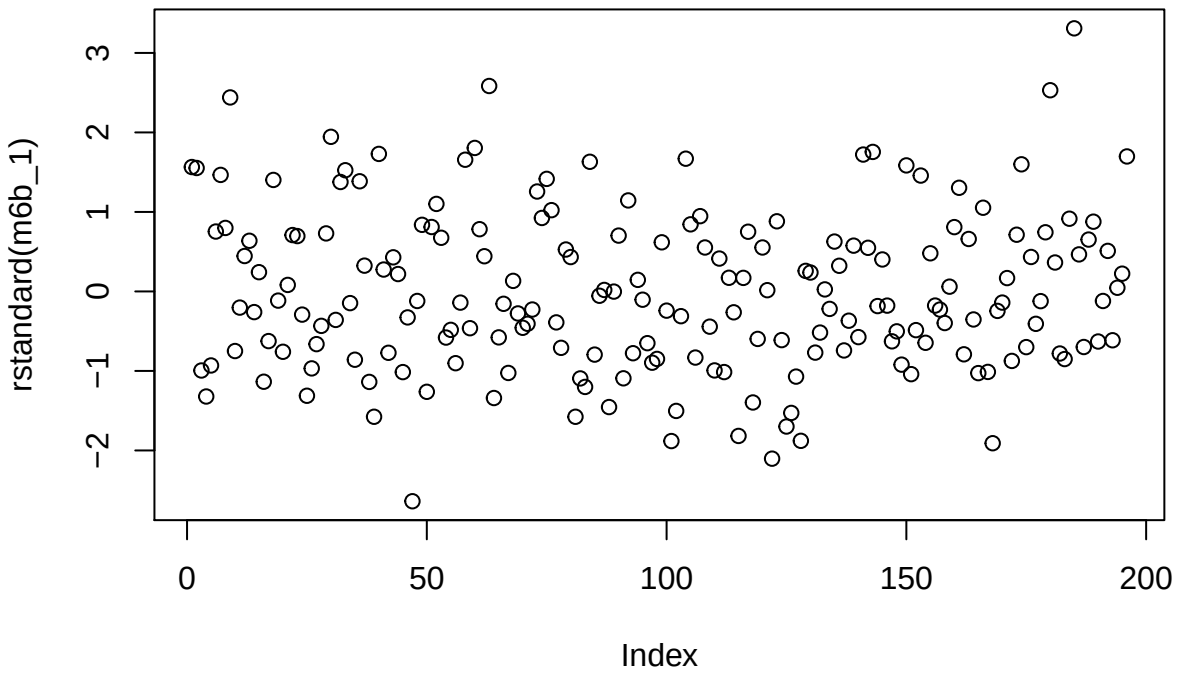
```r
par(mfrow = c(2, 2))
plot(m6b_1)
```




```r
pairs(~ log(PrizeMoney) + DrivingAccuracy + GIR + PuttingAverage +
        BirdieConversion + SandSaves + Scrambling + PuttsPerRound, main = "Scatterplot Matrix")
```




```
par(mfrow = c(1, 1))
plot(rstandard(m6b_1))
```



After looking at the diagnostic plots, we can see that the log(y) is a good choice as they satisfy the assumptions for linearity and normality.

5c)

```
leverage_5 <- hatvalues(m6b_1)
which(leverage_5 >= 2 * mean(leverage_5))
```

```
##  16  40  70  77 168 178
##  16  40  70  77 168 178
```

```
## Leverages are 16, 40, 70, 77, 168, 178

which(abs(rstandard(m6b_1)) >= 2)
```

```
##   9  47  63 122 180 185
##   9  47  63 122 180 185
```

```
## Outliers are 9, 47, 63, 122, 180, 185
```

The leverages are: 16, 40, 70, 77, 168, 178. The outliers are: 9, 47, 63, 122, 180, 185.

5d)

```
vif(m6b_1)
```

```
##  DrivingAccuracy              GIR   PuttingAverage BirdieConversion
##         1.796616         6.294969        12.900789         3.511898
```

```
##        SandSaves       Scrambling    PuttsPerRound
##         1.461506         4.470203        19.355667
```

It has multicollinearity. For the variables with ViF > 5 are: GIR, PuttingAverage, and PuttsPerRound.

5e)

No, because removing one predictor may influence the whole model. At least, only p-value should not determine this.