Stats 101A Homework #5

Damien Ha

2023-03-03

Problem 1

6.7.3

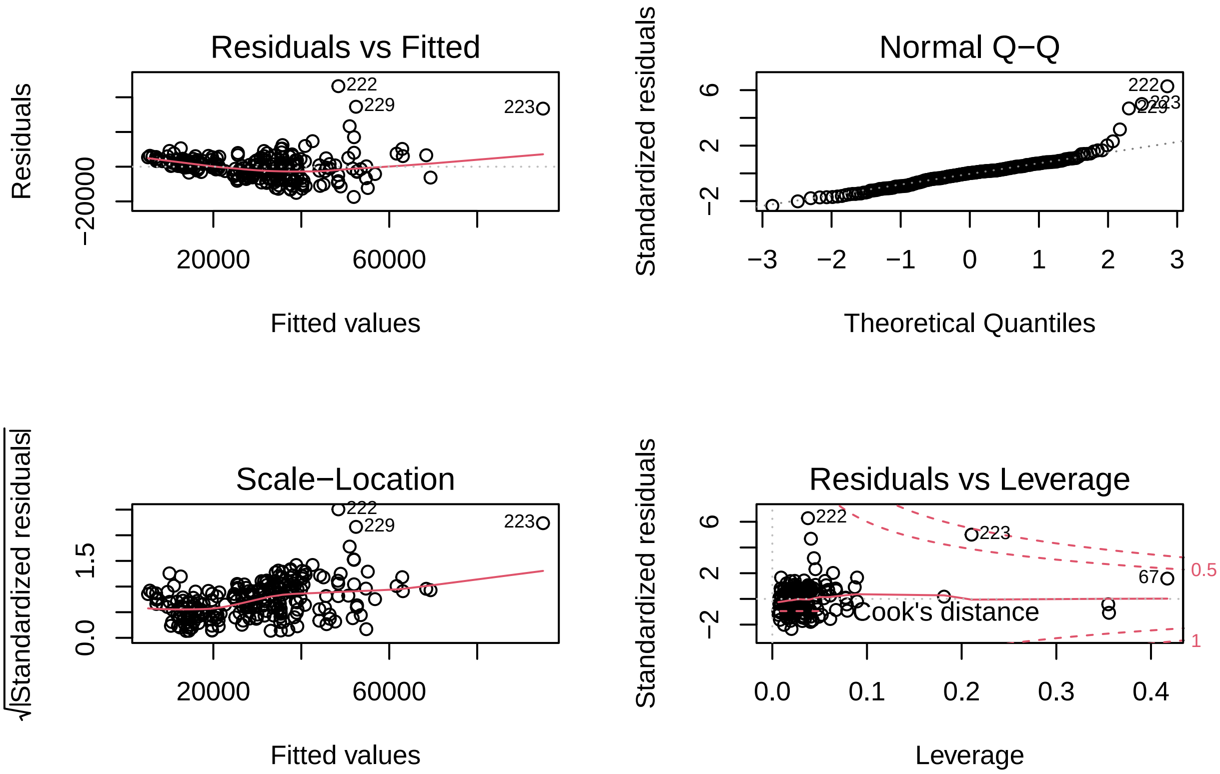
```r
cars <- read.csv("cars04.csv")
```

Part A

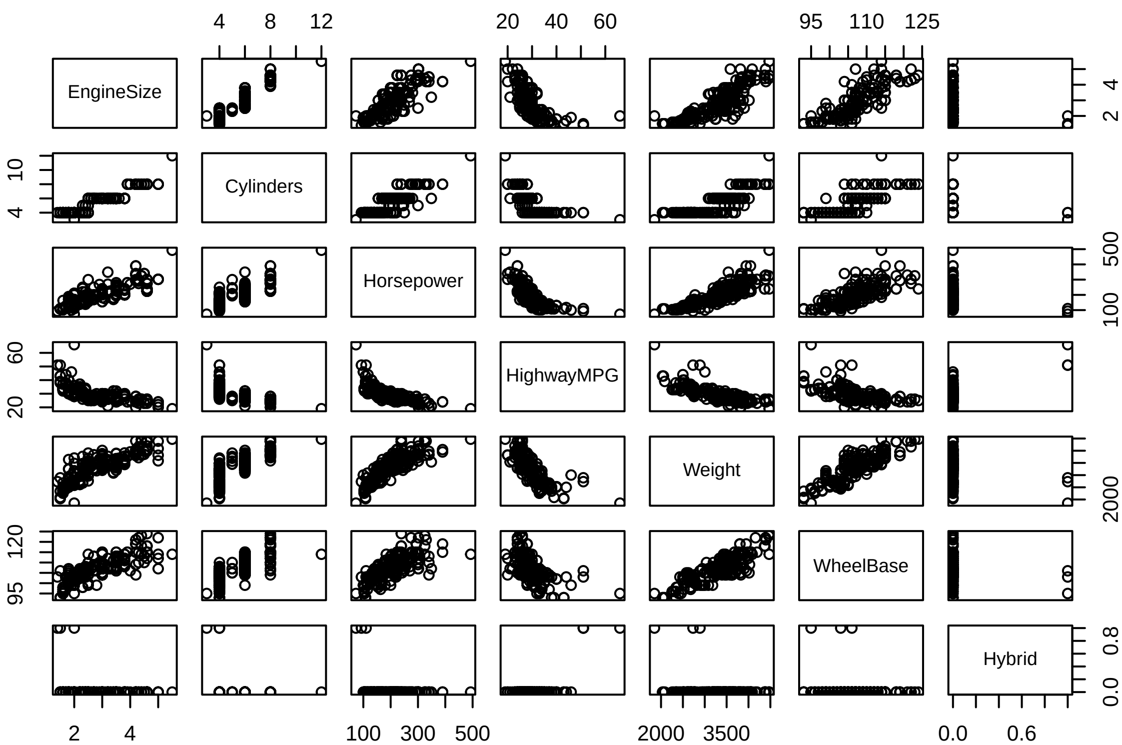
```r
m1 <- lm(SuggestedRetailPrice ~ EngineSize + Cylinders + Horsepower + HighwayMPG +
     Weight + WheelBase + Hybrid, data = cars)

summary(m1)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ EngineSize + Cylinders +
##     Horsepower + HighwayMPG + Weight + WheelBase + Hybrid, data = cars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17436  -4134    173   3561  46392
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -68965.793  16180.381  -4.262 2.97e-05 ***
## EngineSize   -6957.457   1600.137  -4.348 2.08e-05 ***
## Cylinders     3564.755    969.633   3.676 0.000296 ***
## Horsepower     179.702     16.411  10.950  < 2e-16 ***
## HighwayMPG     637.939    202.724   3.147 0.001873 **
## Weight          11.911      2.658   4.481 1.18e-05 ***
## WheelBase       47.607    178.070   0.267 0.789444
## Hybrid         431.759   6092.087   0.071 0.943562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7533 on 226 degrees of freedom
## Multiple R-squared:  0.7819, Adjusted R-squared:  0.7751
## F-statistic: 115.7 on 7 and 226 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(m1)
```




```r
plot(cars[, c(5:7, 9:11, 2)])
```



From the adjusted R^2 value, it may appear at a glance that the model could be reasonable. However, there do appear to be some model violations. The data does not look to be completely normal and some leverage points might be influencing the model.

Part B

If the plot of the residuals against fitted values produces a curved pattern, the model may not be a good fit to the data. It may suggest that the relationship is not strictly linear.

Part C

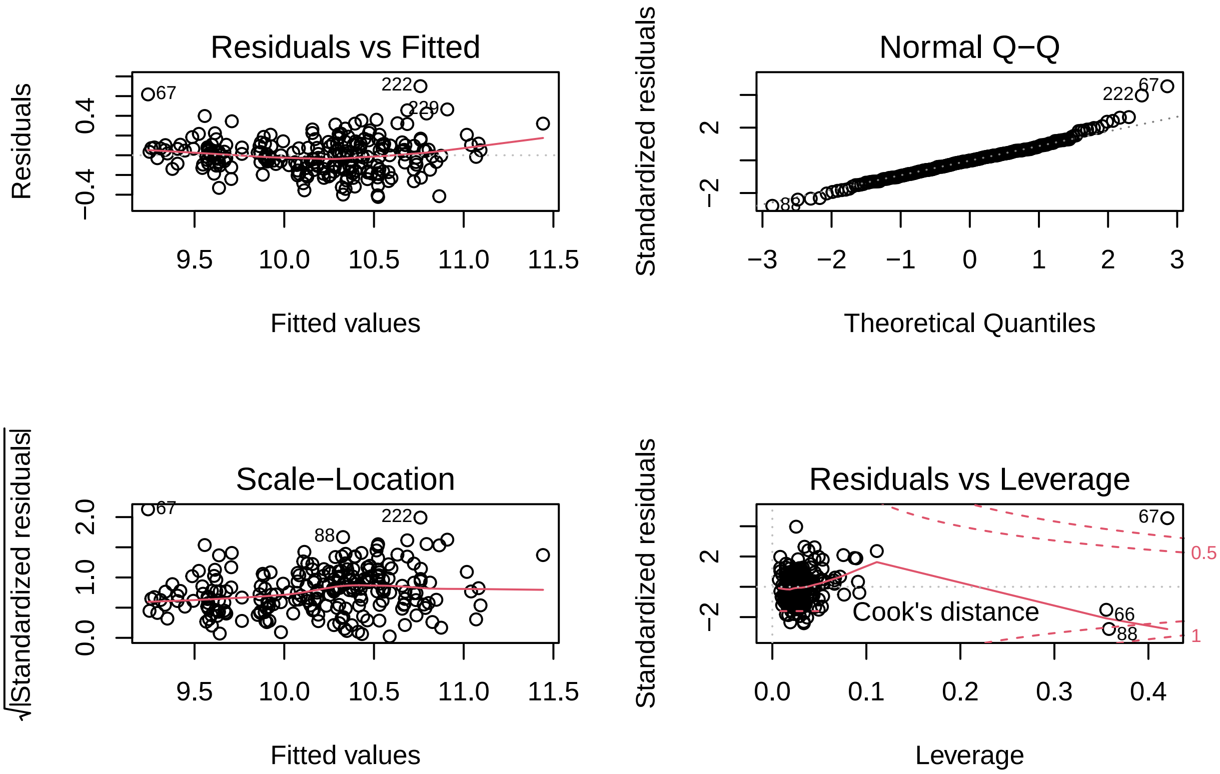
Based on Cook's distance from our diagnostic plot, point 223 looks like a bad leverage point.

Part D

```
m2 <- lm(log(SuggestedRetailPrice) ~ I(EngineSize^(0.25)) + I(log(Cylinders)) + I(log(Horsepower)) + I(
summary(m2)
```

```
##
## Call:
## lm(formula = log(SuggestedRetailPrice) ~ I(EngineSize^(0.25)) +
##     I(log(Cylinders)) + I(log(Horsepower)) + I(1/HighwayMPG) +
##     Weight + I(log(WheelBase)) + Hybrid, data = cars)
##
## 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(m2)
```



This transformed model looks to be an improvement. The residual plots now don't have much of a pattern and the greater linearity in the qq plot suggests a more normal distribution. The adjusted R^2 value is at a good level. There still seem to be some noticeable leverage points, but the model is overall better.

Part E

```
m3 <- update(m2, . ~ . - I(1/HighwayMPG) - I(log(WheelBase)))
summary(m3)
```

```
##
## Call:
## lm(formula = log(SuggestedRetailPrice) ~ I(EngineSize^(0.25)) +
##     I(log(Cylinders)) + I(log(Horsepower)) + Weight + Hybrid,
##     data = cars)
##
## 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
```

The F-statistic implies a greater significance in this case

Part F

Manufacturer is a quantitative variable, so in order to observe its effect on prices you'd have to create a dummy variable where it is represented as a numerical value.

Problem 2

6.7.5

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.1.2
```

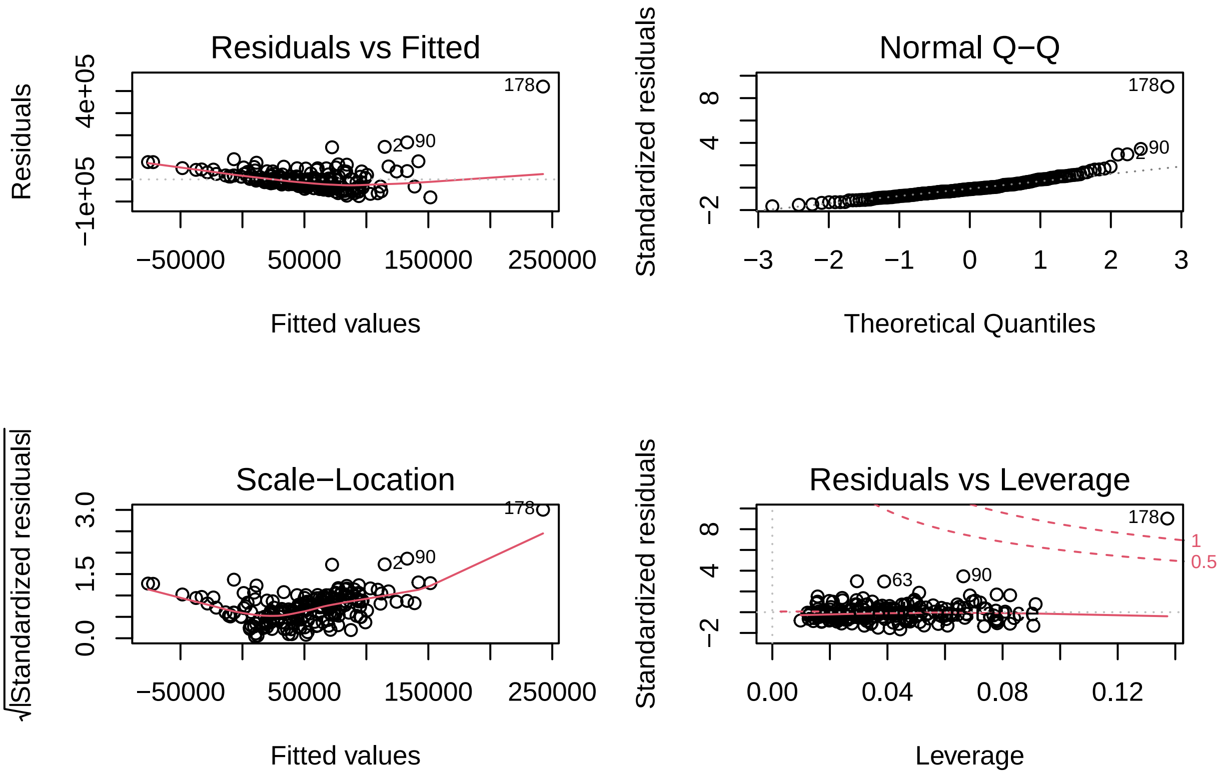
```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.2
```

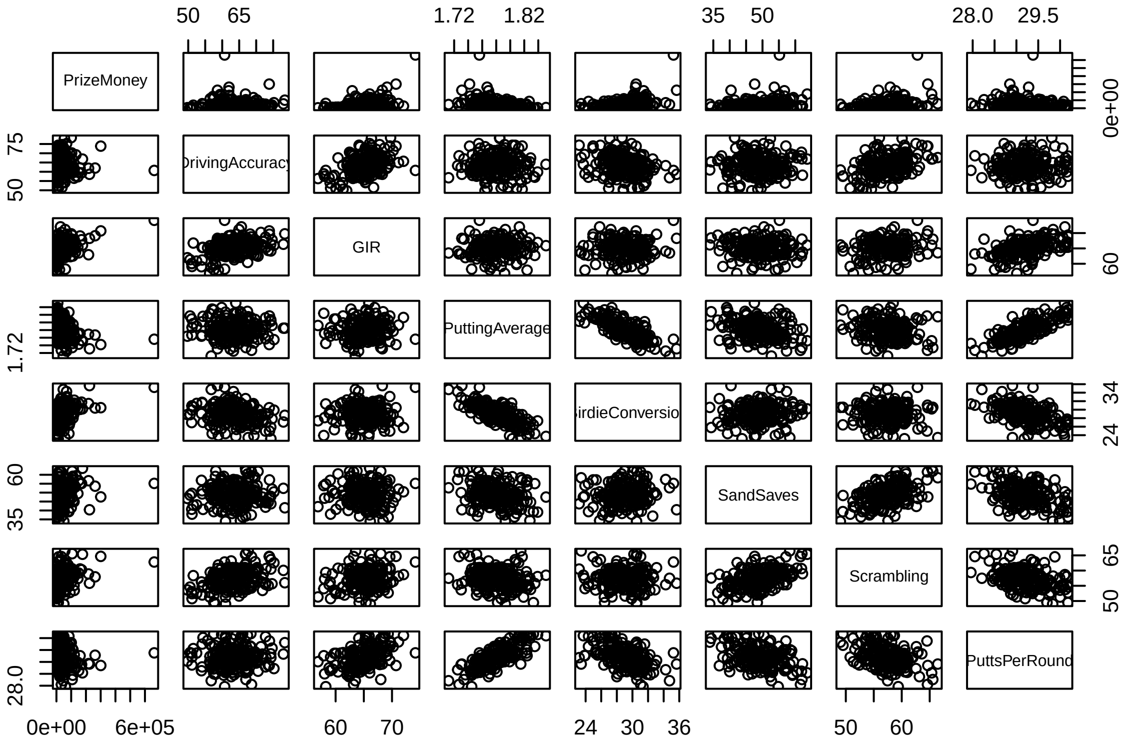
```
pgatour <- read.csv("pgatour2006.csv")

m4 <- lm(PrizeMoney ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion + SandSaves + Scramblin
summary(m4)
```

```
##
## Call:
## lm(formula = PrizeMoney ~ DrivingAccuracy + GIR + PuttingAverage +
##     BirdieConversion + SandSaves + Scrambling + PuttsPerRound,
##     data = pgatour)
##
## 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(m4)
```




```
pgatour2 <- subset(pgatour, select=c('PrizeMoney','DrivingAccuracy','GIR','PuttingAverage','BirdieConve
                                     'SandSaves', 'Scrambling', 'PuttsPerRound'))
plot(pgatour2)
```

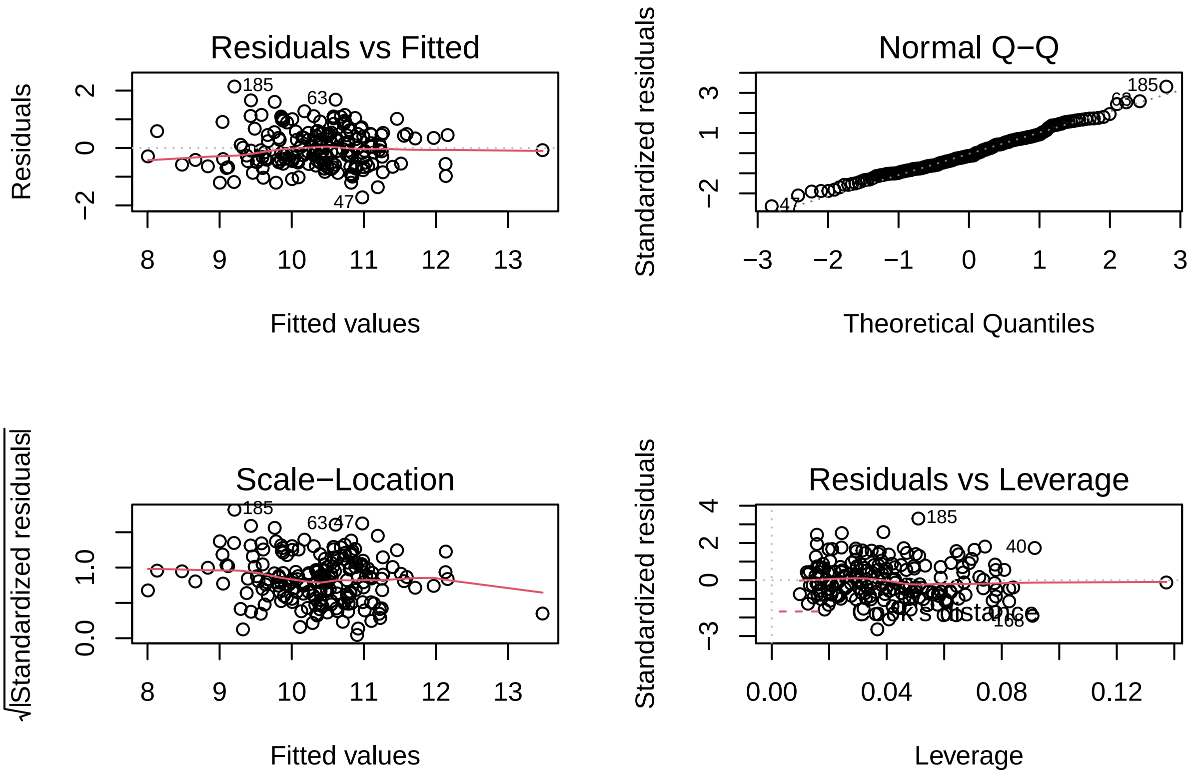


Part A

```r
m4_log <- lm(log(PrizeMoney) ~ ., pgatour2)
summary(m4_log)
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ ., data = pgatour2)
##
## 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(m4_log)
```



I agree. This transformation does appear to make the data better fitted to a linear regression model. The qq plot becomes more linear suggesting more normality and the residuals vs fitted values is not as much of a curve but more of a random scatter. This shows the transformed model better aligns with the assumptions of linear regression.

Part B

This can be seen above in part A. The better choice of full model is the second one where Y is log transformed for the reasons listed in part A. A scatterplot and 4 diagnostic plots can be seen there as well.

Part C

Based on our diagnostic plots, we may want to investigate point 185. It may be an outlier or leverage point.

Part D

There may be outlying values, like point 185 listed above, that need further investigation as they may skew our data. Also while the qqplot is fairly linear, it has some non-linear sections suggesting that the normality may not be perfect.

Part E

Changing or removing one variable can affect the statistical significance of other variable(s). We should not remove any variables because we might suddenly make one of the other insignificant variables significant.

Problem 3

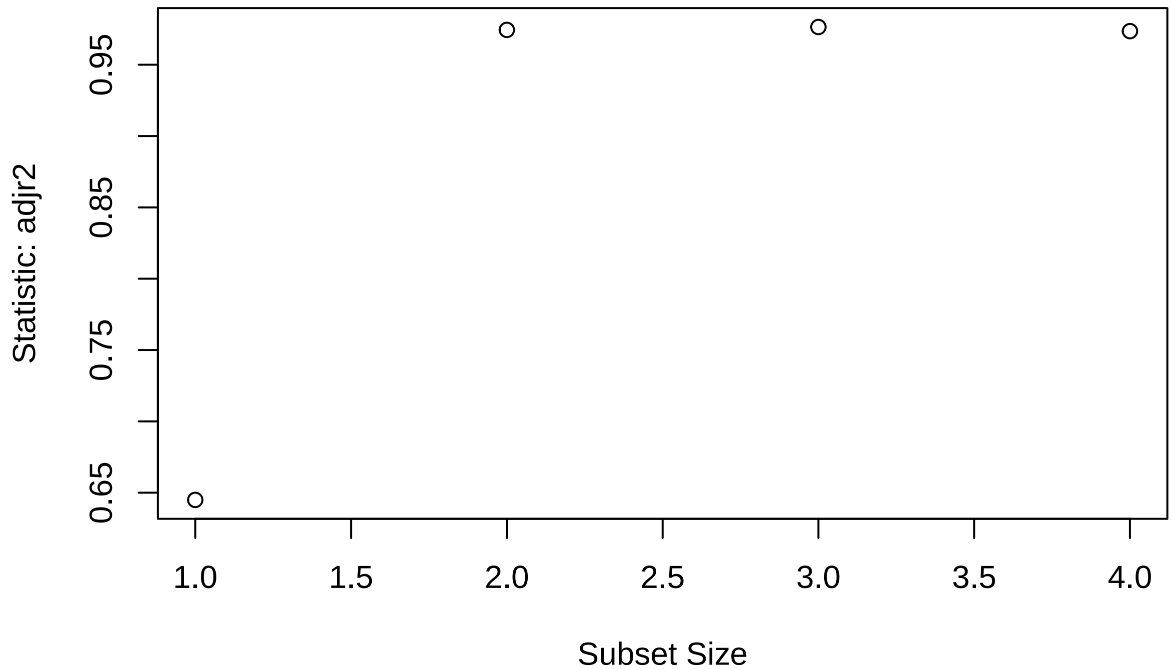
7.5.2

```
haldcement <- read.table("Haldcement.txt", header = T)
head(haldcement)
```

```
##       Y x1 x2 x3 x4
## 1  78.5  7 26  6 60
## 2  74.3  1 29 15 52
## 3 104.3 11 56  8 20
## 4  87.6 11 31  8 47
## 5  95.9  7 52  6 33
## 6 109.2 11 55  9 22
```

Part A

```
m5 <- lm(Y ~ x4, haldcement)
m6 <- lm(Y ~ x1 + x2, haldcement)
m7 <- lm(Y ~ x1 + x2 + x4, haldcement)
m8 <- lm(Y ~ x1 + x2 + x3 + x4, haldcement)
sum5 <- summary(m5)
sum6 <- summary(m6)
sum7 <- summary(m7)
sum8 <- summary(m8)

adjr2 <- c(sum5$adj.r.squared, sum6$adj.r.squared, sum7$adj.r.squared, sum8$adj.r.squared)
x <- seq(1,4,by=1)
plot(adjr2 ~ x, xlab = "Subset Size", ylab = "Statistic: adjr2")
```




```
Predictors <- c("X4", "X1, X2", "X1, X2, X4", "X1, X2, X3, X4")
AIC_col <- c(AIC(m5, k=2), AIC(m6, k=2), AIC(m7, k=2), AIC(m8, k=2))
BIC_col <- c(AIC(m5, k=log(nrow(haldcement))), AIC(m6, k=log(nrow(haldcement))), AIC(m7, k=log(nrow(hal
```

```r
allsubsets <- cbind(x, Predictors, adjr2, AIC_col, BIC_col)
allsubsets
```

```
##      x   Predictors       adjr2               AIC_col
## [1,] "1" "X4"             "0.644954869961756" "97.7440447788562"
## [2,] "2" "X1, X2"         "0.974414049442758" "64.3123927621906"
## [3,] "3" "X1, X2, X4"     "0.976447268267236" "63.8662854718626"
## [4,] "4" "X1, X2, X3, X4" "0.97356343061152"  "65.8366897916517"
##      BIC_col
## [1,] "99.4388928512408"
## [2,] "66.5721901920368"
## [3,] "66.6910322591703"
## [4,] "69.2263859364209"
```

Subset of 2 or 3 would be reasonable

Part B

```r
attach(haldcement)

m9 <- step(lm(Y ~ 1), Y ~ x1 + x2 + x3 + x4, direction="forward")
```

```
## Start:  AIC=71.44
## Y ~ 1
##
##        Df Sum of Sq     RSS    AIC
## + x4    1   1831.90  883.87 58.852
## + x2    1   1809.43  906.34 59.178
## + x1    1   1450.08 1265.69 63.519
## + x3    1    776.36 1939.40 69.067
## <none>              2715.76 71.444
##
## Step:  AIC=58.85
## Y ~ x4
##
##        Df Sum of Sq    RSS    AIC
## + x1    1    809.10  74.76 28.742
## + x3    1    708.13 175.74 39.853
## <none>              883.87 58.852
## + x2    1     14.99 868.88 60.629
##
## Step:  AIC=28.74
## Y ~ x4 + x1
##
##        Df Sum of Sq    RSS    AIC
## + x2    1    26.789 47.973 24.974
## + x3    1    23.926 50.836 25.728
## <none>              74.762 28.742
##
## Step:  AIC=24.97
## Y ~ x4 + x1 + x2
##
##        Df Sum of Sq    RSS    AIC
## <none>              47.973 24.974
```

```
## + x3     1   0.10909 47.864 26.944
```

```r
m9
```

```
##
## Call:
## lm(formula = Y ~ x4 + x1 + x2)
##
## Coefficients:
## (Intercept)           x4           x1           x2
##     71.6483      -0.2365       1.4519       0.4161
```

```r
m10 <- step(lm(Y ~ 1), Y ~ x1 + x2 + x3 + x4, direction="forward", k = log(nrow(haldcement)))
```

```
## Start:  AIC=72.01
## Y ~ 1
##
##        Df Sum of Sq     RSS    AIC
## + x4    1   1831.90  883.87 59.982
## + x2    1   1809.43  906.34 60.308
## + x1    1   1450.08 1265.69 64.649
## + x3    1    776.36 1939.40 70.197
## <none>              2715.76 72.009
##
## Step:  AIC=59.98
## Y ~ x4
##
##        Df Sum of Sq    RSS    AIC
## + x1    1    809.10  74.76 30.437
## + x3    1    708.13 175.74 41.547
## <none>              883.87 59.982
## + x2    1     14.99 868.88 62.324
##
## Step:  AIC=30.44
## Y ~ x4 + x1
##
##        Df Sum of Sq    RSS    AIC
## + x2    1    26.789 47.973 27.234
## + x3    1    23.926 50.836 27.987
## <none>              74.762 30.437
##
## Step:  AIC=27.23
## Y ~ x4 + x1 + x2
##
##        Df Sum of Sq    RSS    AIC
## <none>              47.973 27.234
## + x3    1   0.10909 47.864 29.769
```

```r
m10
```

```
##
## Call:
## lm(formula = Y ~ x4 + x1 + x2)
##
## Coefficients:
## (Intercept)           x4           x1           x2
##     71.6483      -0.2365       1.4519       0.4161
```

The model with 3 predictors seems to work best

Part C

```r
m11 <- step(lm(Y ~ x1 + x2 + x3 + x4), Y ~ x1 + x2 + x3 + x4, direction="backward")
```

```
## Start:  AIC=26.94
## Y ~ x1 + x2 + x3 + x4
##
##        Df Sum of Sq    RSS    AIC
## - x3    1    0.1091 47.973 24.974
## - x4    1    0.2470 48.111 25.011
## - x2    1    2.9725 50.836 25.728
## <none>              47.864 26.944
## - x1    1   25.9509 73.815 30.576
##
## Step:  AIC=24.97
## Y ~ x1 + x2 + x4
##
##        Df Sum of Sq    RSS    AIC
## <none>               47.97 24.974
## - x4    1      9.93  57.90 25.420
## - x2    1     26.79  74.76 28.742
## - x1    1    820.91 868.88 60.629
```

```r
m11
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x4)
##
## Coefficients:
## (Intercept)           x1           x2           x4
##     71.6483       1.4519       0.4161      -0.2365
```

```r
m12 <- step(lm(Y ~ x1 + x2 + x3 + x4), Y ~ x1 + x2 + x3 + x4, direction="backward", k = log(nrow(haldcem
```

```
## Start:  AIC=29.77
## Y ~ x1 + x2 + x3 + x4
##
##        Df Sum of Sq    RSS    AIC
## - x3    1    0.1091 47.973 27.234
## - x4    1    0.2470 48.111 27.271
## - x2    1    2.9725 50.836 27.987
## <none>              47.864 29.769
## - x1    1   25.9509 73.815 32.836
##
## Step:  AIC=27.23
## Y ~ x1 + x2 + x4
##
##        Df Sum of Sq    RSS    AIC
## - x4    1      9.93  57.90 27.115
## <none>               47.97 27.234
## - x2    1     26.79  74.76 30.437
## - x1    1    820.91 868.88 62.324
##
```

```
## Step:  AIC=27.11
## Y ~ x1 + x2
##
##        Df Sum of Sq     RSS    AIC
## <none>                57.90 27.115
## - x1    1    848.43  906.34 60.308
## - x2    1   1207.78 1265.69 64.649
```

```
m12
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##     52.5773       1.4683       0.6623
```

Part D

These models all select variables differently. The first one from part A involves fitting all possible combinations of predictor variables, from one variable to all variables, and selecting the model that has the lowest AIC or BIC. The second one from part B involves starting with a null model and then adding one predictor variable at a time until the addition of another variable no longer improves the AIC or BIC. Finally, the final one from part C is the reverse of the forward selection approach. It starts with a model that includes all predictor variables and then removes one variable at a time until removing another variable does not improve the AIC or BIC.

Part E

The 2 or 3 predictor model would be best