Quiz1

Duc Le

10/22/2020

Problem 1

```
library(MASS)
library(ggplot2)
data(mpg)

head(mpg)
```

1a

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

```
mpg.var = (mpg['cty'] + mpg['hwy'])/2
mpg = cbind(mpg, mpg.var)
names(mpg)[12] = 'avg.mpg'
drop = c("cty","hwy")

df = mpg[,!(names(mpg) %in% drop)]
```

```
dim(df)
```

1b

```
## [1] 234  10
```

```
names(df)
```

```
##  [1] "manufacturer" "model"        "displ"        "year"         "cyl"
##  [6] "trans"        "drv"          "fl"           "class"        "avg.mpg"
```

```
summary(df)
```

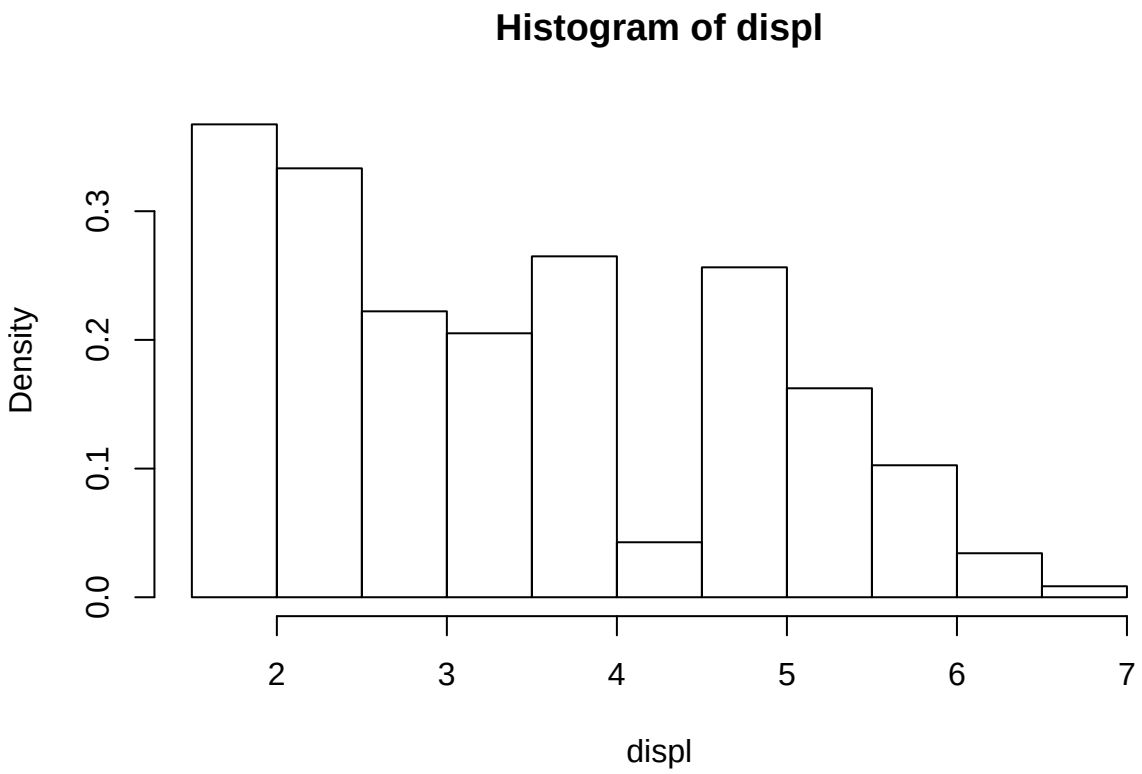
```
##  manufacturer          model               displ            year
##  Length:234         Length:234         Min.   :1.600   Min.   :1999
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999
```

```
##  Mode  :character   Mode  :character   Median :3.300   Median :2004
##                                        Mean   :3.472   Mean   :2004
##                                        3rd Qu.:4.600   3rd Qu.:2008
##                                        Max.   :7.000   Max.   :2008
##       cyl           trans               drv                fl
##  Min.   :4.000   Length:234         Length:234         Length:234
##  1st Qu.:4.000   Class :character   Class :character   Class :character
##  Median :6.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :5.889
##  3rd Qu.:8.000
##  Max.   :8.000
##     class              avg.mpg
##  Length:234         Min.   :10.50
##  Class :character   1st Qu.:15.50
##  Mode  :character   Median :20.50
##                     Mean   :20.15
##                     3rd Qu.:23.50
##                     Max.   :39.50
```

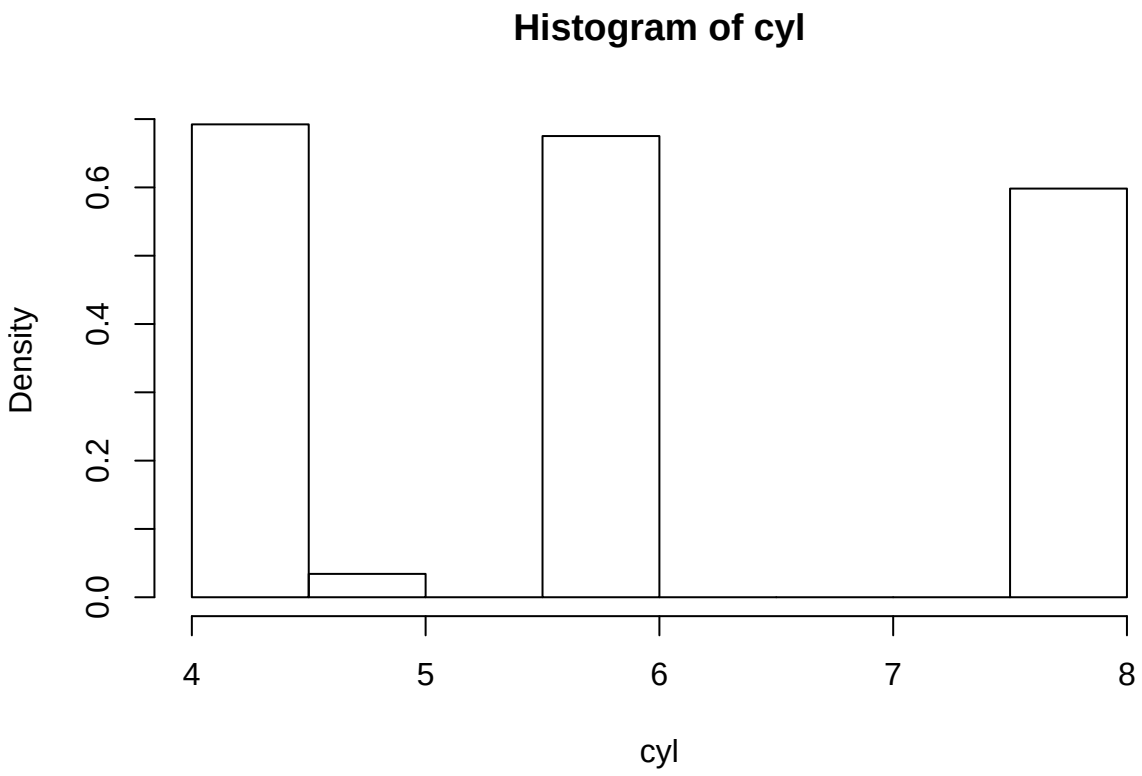
```
str(df)
```

```
## 'data.frame':    234 obs. of  10 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
##  $ avg.mpg     : num  23.5 25 25.5 25.5 21 22 22.5 22 20.5 24 ...
```

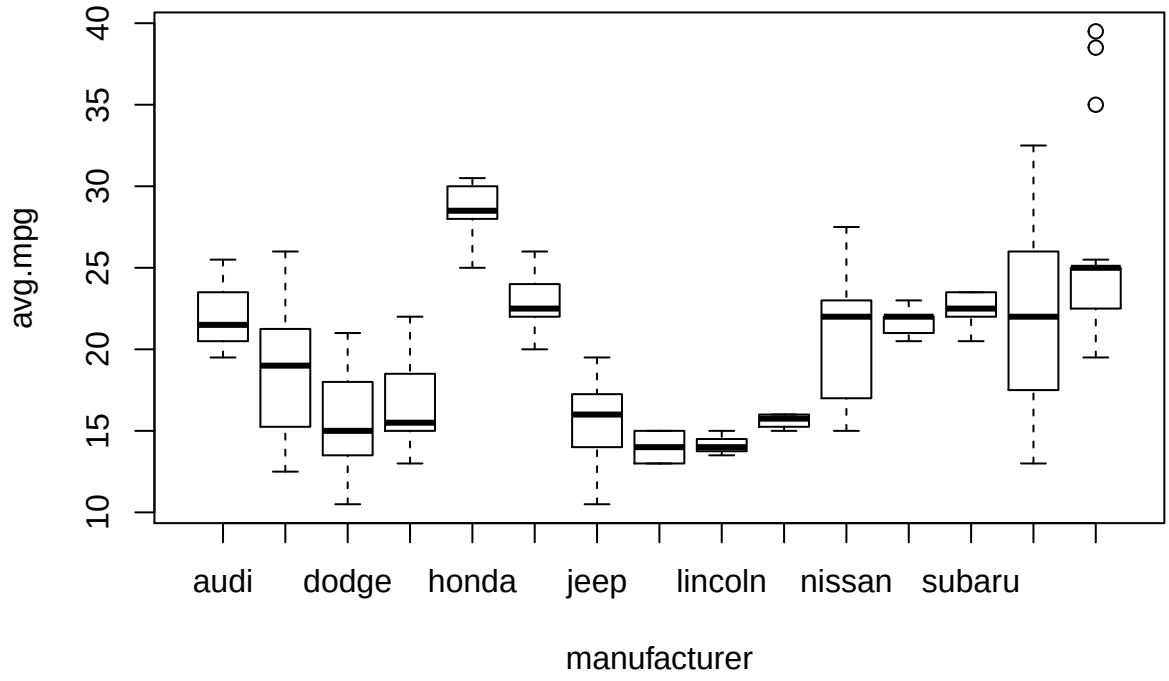
```
attach(df)

hist(displ, freq = F)
```



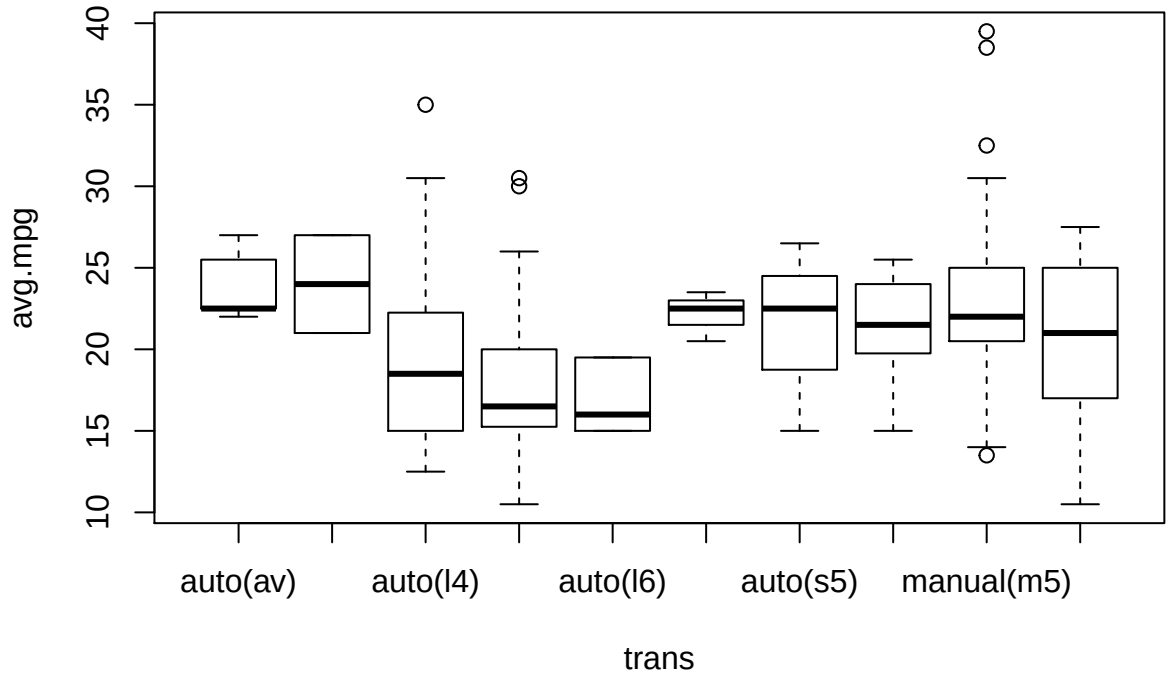

```
hist(cyl, freq = F)
```



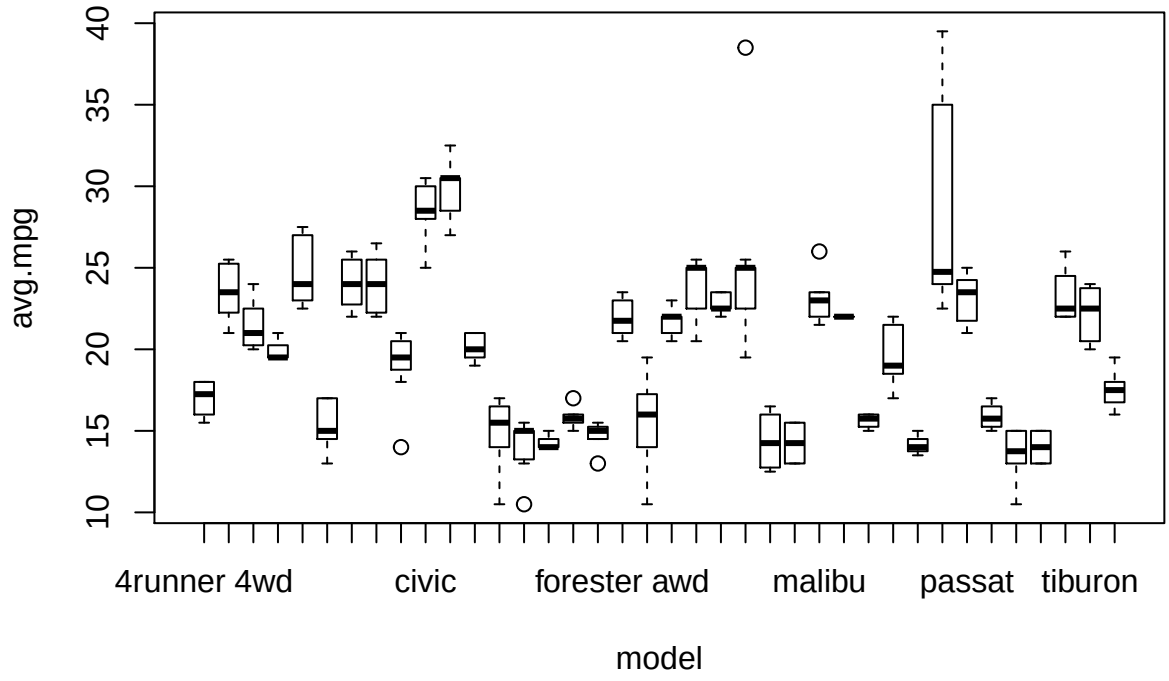

```
boxplot(avg.mpg ~ manufacturer)
```



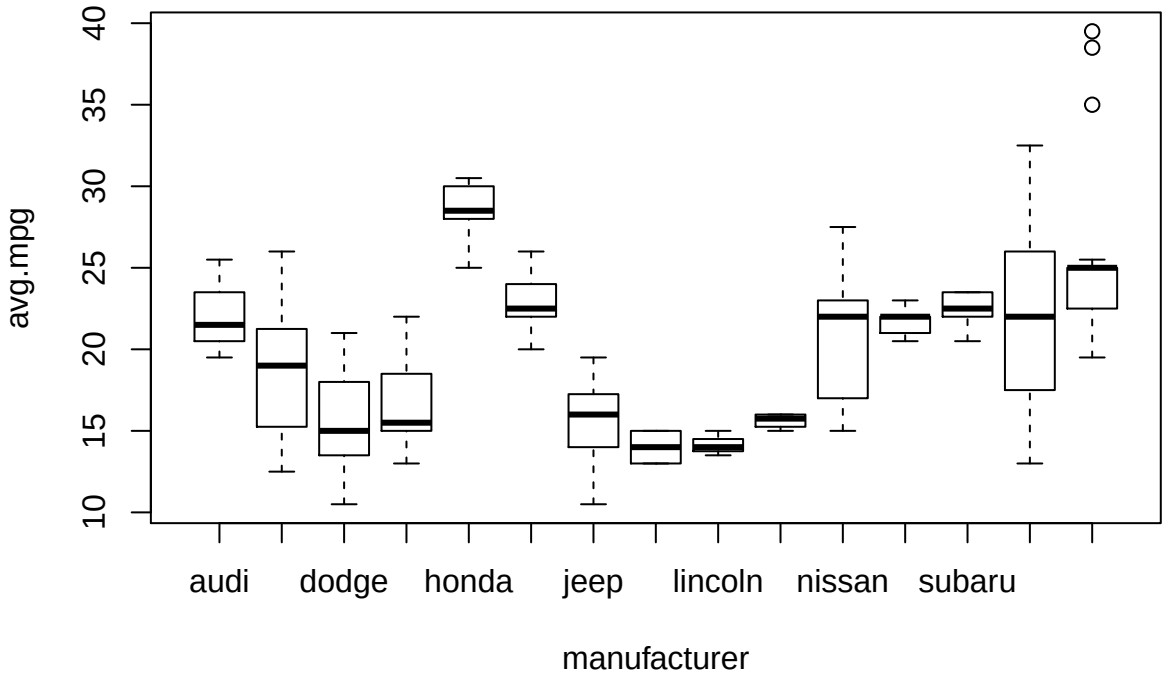

```
boxplot(avg.mpg ~ trans)
```




```
boxplot(avg.mpg ~ model)
```




```
boxplot(avg.mpg ~ manufacturer)
```



1c

From looking at the avg mpg by manufacturer in this dataset, Volkswagen seems to have a few outliers if we base it off the boxplot. Further analysis can be done using other statistical tests to examine this hypothesis.

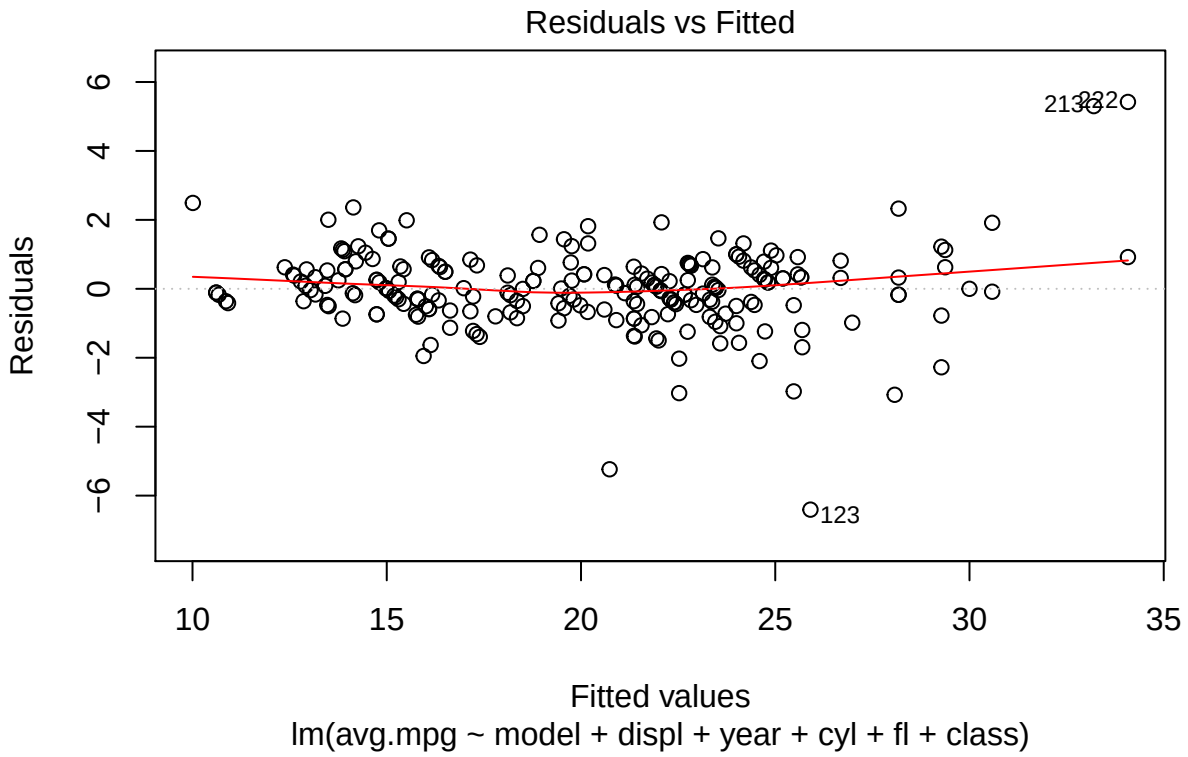
1d

```
mlm = lm(avg.mpg ~., data = df)
best.lm = lm(avg.mpg ~model + displ + year + cyl + fl + class, data = df)
```

This following model with these features give the lowest AIC score. Thus it's probably the most efficient linear model to use.

1e

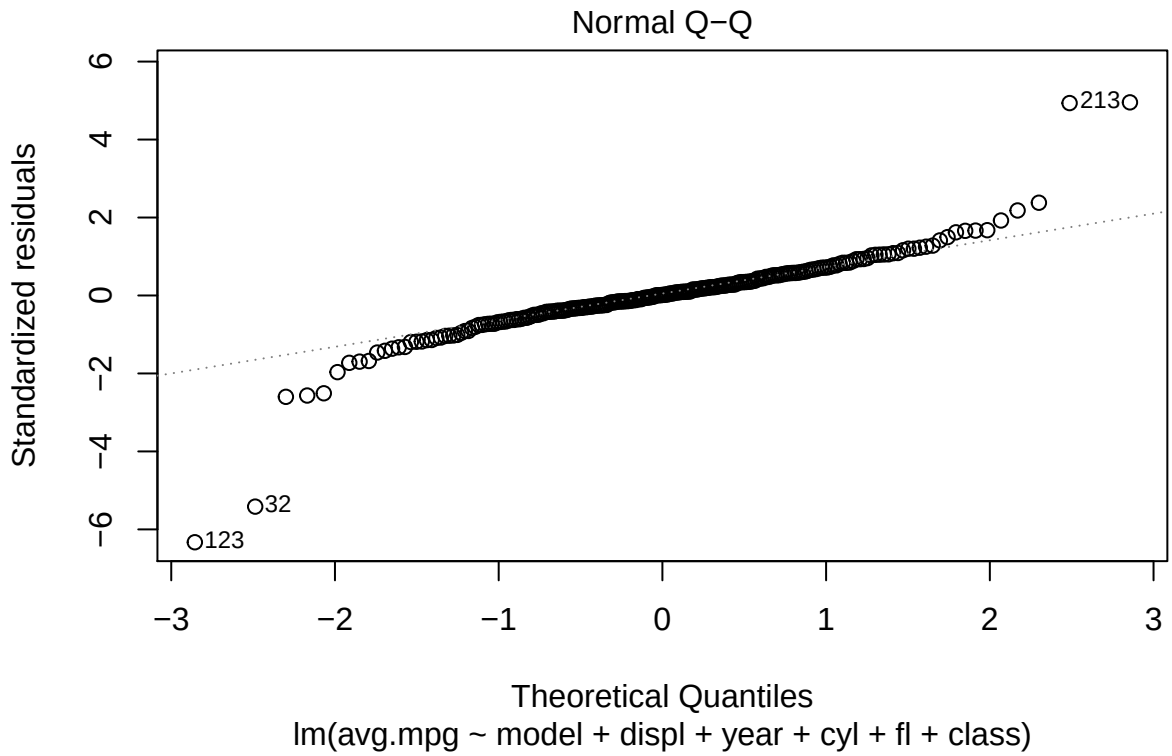
```
plot(best.lm, which = c(1))
```



Judging from the first Residuals vs. Fitted Values plot, we see that the red line representing the mean value of the residuals is hovering over 0 for the majority of the fitted values. However, it does deviate a bit from 0 towards the end. This could be caused by an value with high leverage + influence.

```
plot(best.lm, which = c(2))
```

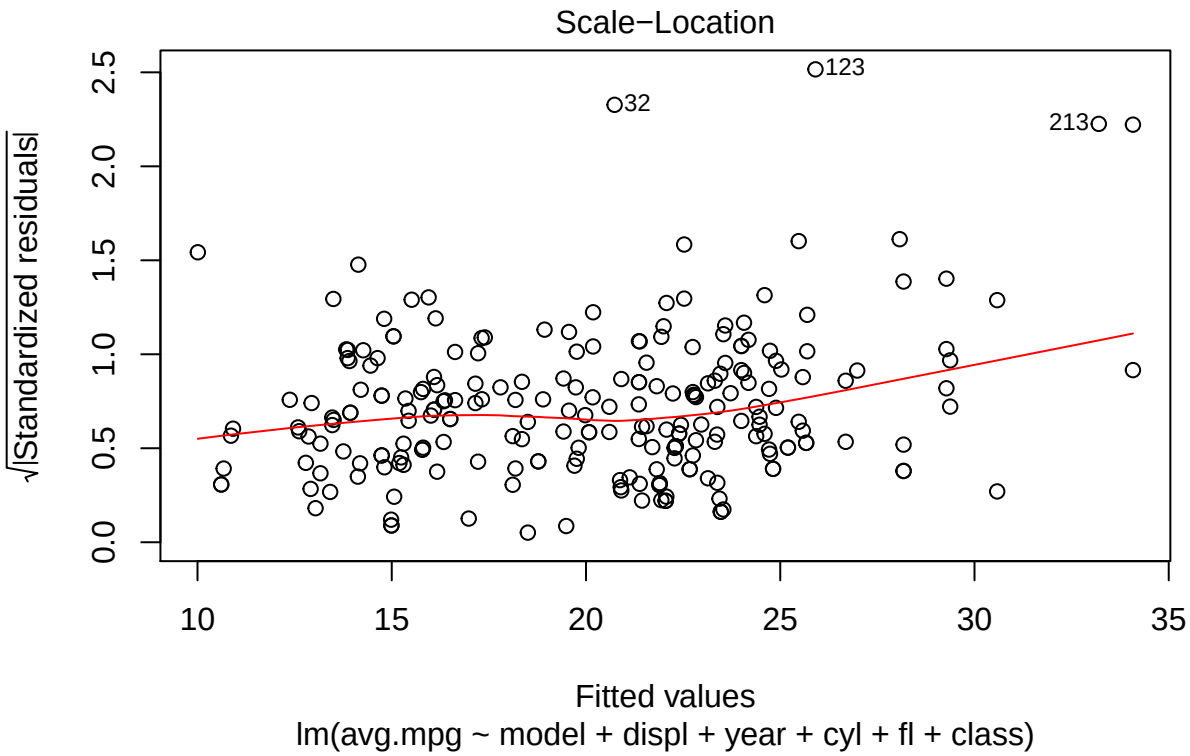
```
## Warning: not plotting observations with leverage one:
##   107
```



The QQ plot shows us good results as the stand. residuals stay on the line.

```
plot(best.lm, which = c(3))
```

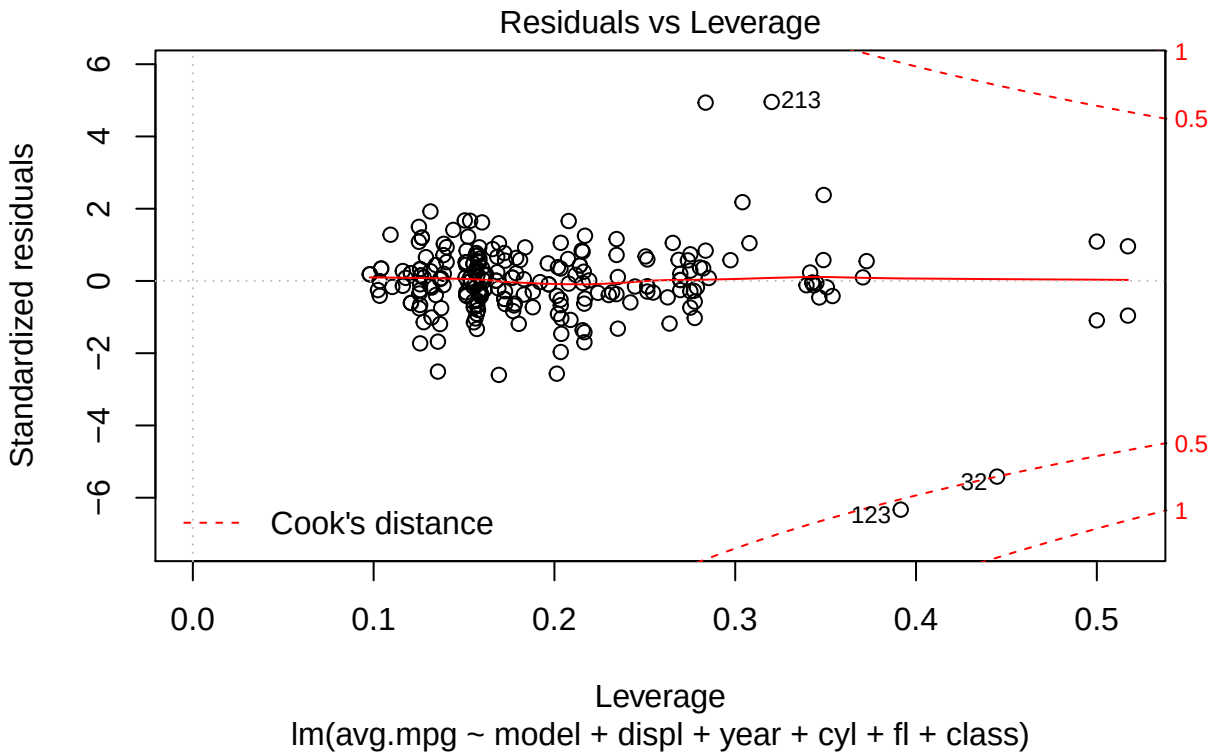
```
## Warning: not plotting observations with leverage one:
##   107
```



The Stand. Res. vs Fitted Values plot does not look too good since the red line is not near 1 where we want it to be. As we can also see at the top, we have some fitted values with extremely high residuals. We can analyze their Cook's distance to conclude our residuals analysis.

```
plot(best.lm, which = c(5))
```

```
## Warning: not plotting observations with leverage one:
##   107
```



Looking at this graph, we see that there are multiple points with high leverage, but 2 in particular with a Cook's Distance > 1, thus indicates that they are a great influence to our line of best fit. Eliminating these samples could possibly improve our linear model.

1f Looking at our result from part d), the number of cofficients involved could be overwhelming. A more efficient way to analyze the significance of these predictors could be to look at their associated p-values and compare them to our desired alpha. Which is 0.05. Any feature with an associated p-value below alpha should be considered a significant predictor to our model, thus leaving us room to eliminate any other unnecessary features.

```
stepAIC(best.lm, direction = "both")
```

1g

```
## Start:  AIC=163.43
## avg.mpg ~ model + displ + year + cyl + fl + class
##
##         Df Sum of Sq    RSS    AIC
## - class  2      3.20 318.03 161.79
## <none>               314.83 163.43
## - displ  1      4.70 319.53 164.90
## - cyl    1     30.88 345.71 183.33
## - year   1     78.05 392.88 213.26
## - model 33    350.95 665.78 272.68
## - fl     4    424.54 739.37 355.21
##
```

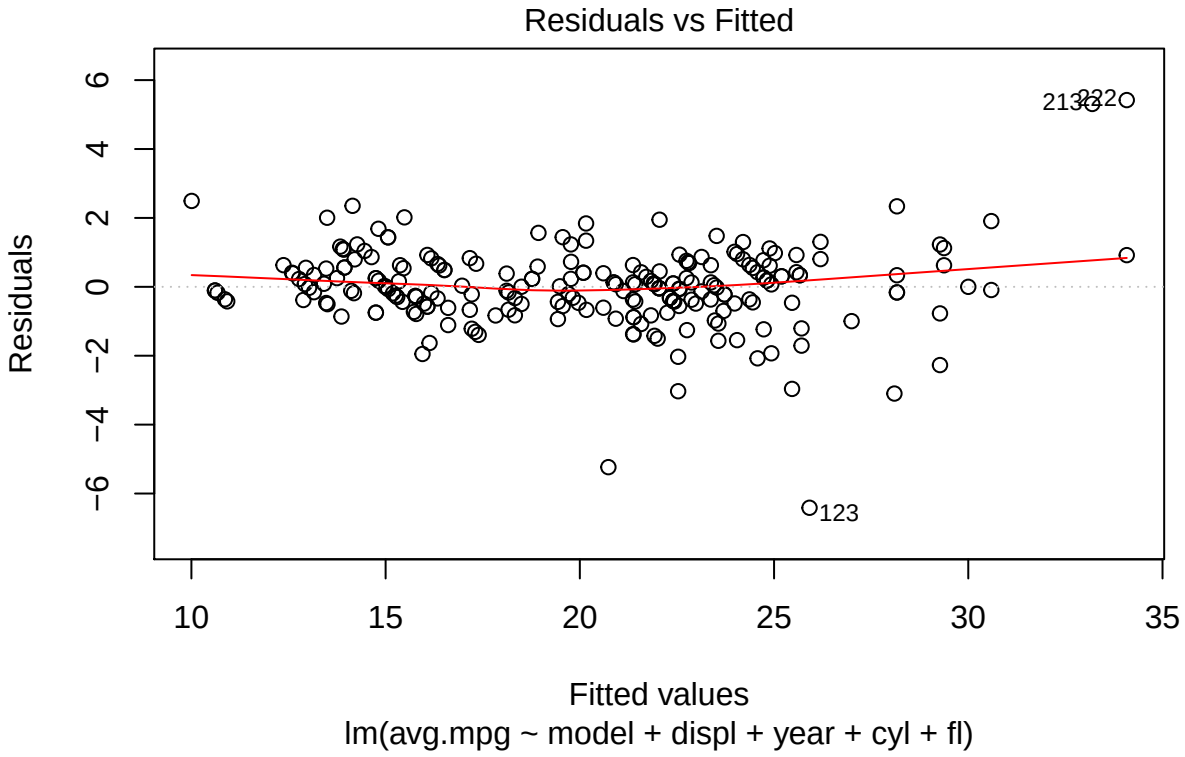
```
## Step:  AIC=161.79
## avg.mpg ~ model + displ + year + cyl + fl
##
##         Df Sum of Sq     RSS    AIC
## <none>                318.03 161.79
## - displ  1      4.87  322.89 163.35
## + class  2      3.20  314.83 163.43
## - cyl    1     29.82  347.85 180.77
## - year   1     83.40  401.43 214.29
## - fl     4    425.26  743.29 352.45
## - model 37   1162.72 1480.75 447.73
```

```
##
## Call:
## lm(formula = avg.mpg ~ model + displ + year + cyl + fl, data = df)
##
## Coefficients:
##                  (Intercept)                      modela4
##                    -269.65947                      6.33180
##              modela4 quattro              modela6 quattro
##                      4.27103                      4.46730
##                  modelaltima      modelc1500 suburban 2wd
##                      6.43872                      1.49990
##                   modelcamry            modelcamry solara
##                      5.85157                      5.76453
##             modelcaravan 2wd                   modelcivic
##                      2.37866                      9.23576
##                modelcorolla                modelcorvette
##                     10.45299                      7.25273
##      modeldakota pickup 4wd              modeldurango 4wd
##                     -0.27669                     -0.33565
##         modelexpedition 2wd            modelexplorer 4wd
##                      0.10989                      0.06491
##        modelf150 pickup 4wd            modelforester awd
##                     -0.13442                      2.98905
##     modelgrand cherokee 4wd              modelgrand prix
##                     -0.52550                      5.92871
##                     modelgti             modelimpreza awd
##                      5.33811                      3.96671
##                   modeljetta         modelk1500 tahoe 4wd
##                      5.86583                     -0.86407
## modelland cruiser wagon 4wd                   modelmalibu
##                     -0.15711                      5.20879
##                  modelmaxima         modelmountaineer 4wd
##                      5.48409                     -0.02279
##                 modelmustang           modelnavigator 2wd
##                      4.04129                      0.47806
##              modelnew beetle                   modelpassat
##                      6.75362                      5.87081
##         modelpathfinder 4wd     modelram 1500 pickup 4wd
##                      0.21794                     -0.59117
##            modelrange rover                   modelsonata
##                     -0.31497                      5.07486
##                modeltiburon       modeltoyota tacoma 4wd
```

```
##                     3.34282                    -0.16249
##                       displ                        year
##                    -0.54623                     0.14672
##                         cyl                         fld
##                    -0.80369                     7.93723
##                         fle                         flp
##                    -4.76678                    -1.79243
##                         flr
##                    -0.62492
```

```
best.lm2 = lm(avg.mpg ~ model + displ + year + cyl + fl, data = df)
```

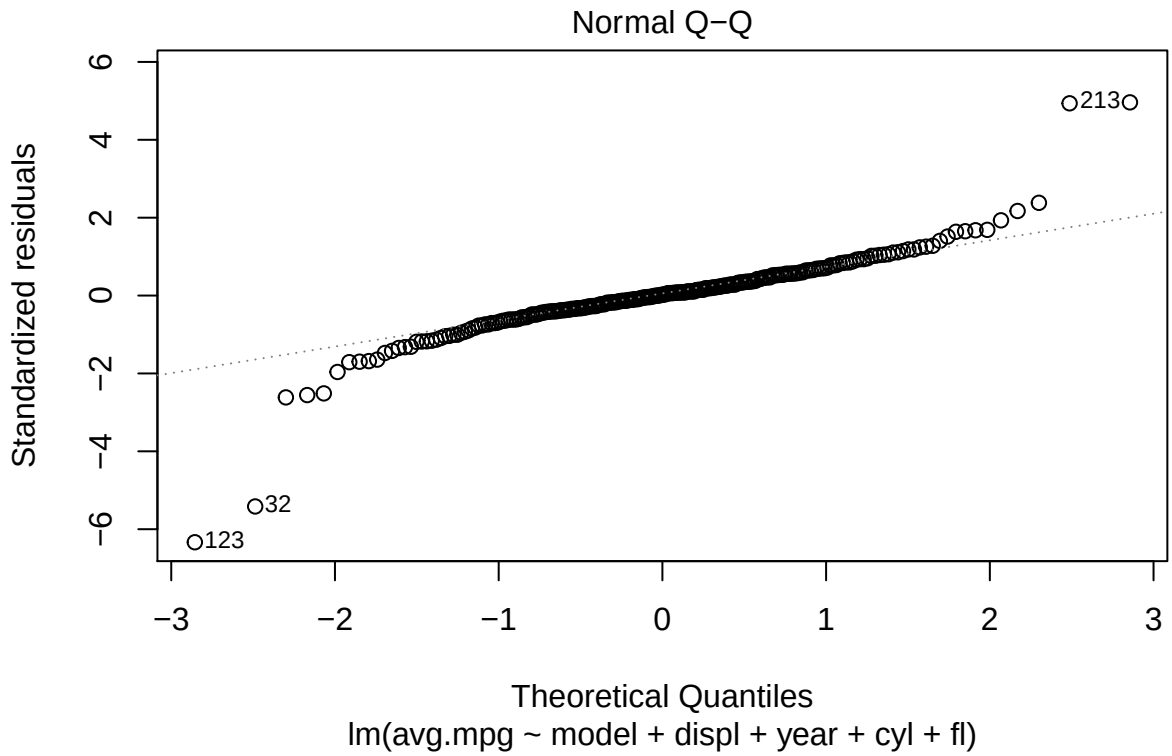
```
plot(best.lm2, which = c(1))
```



1h

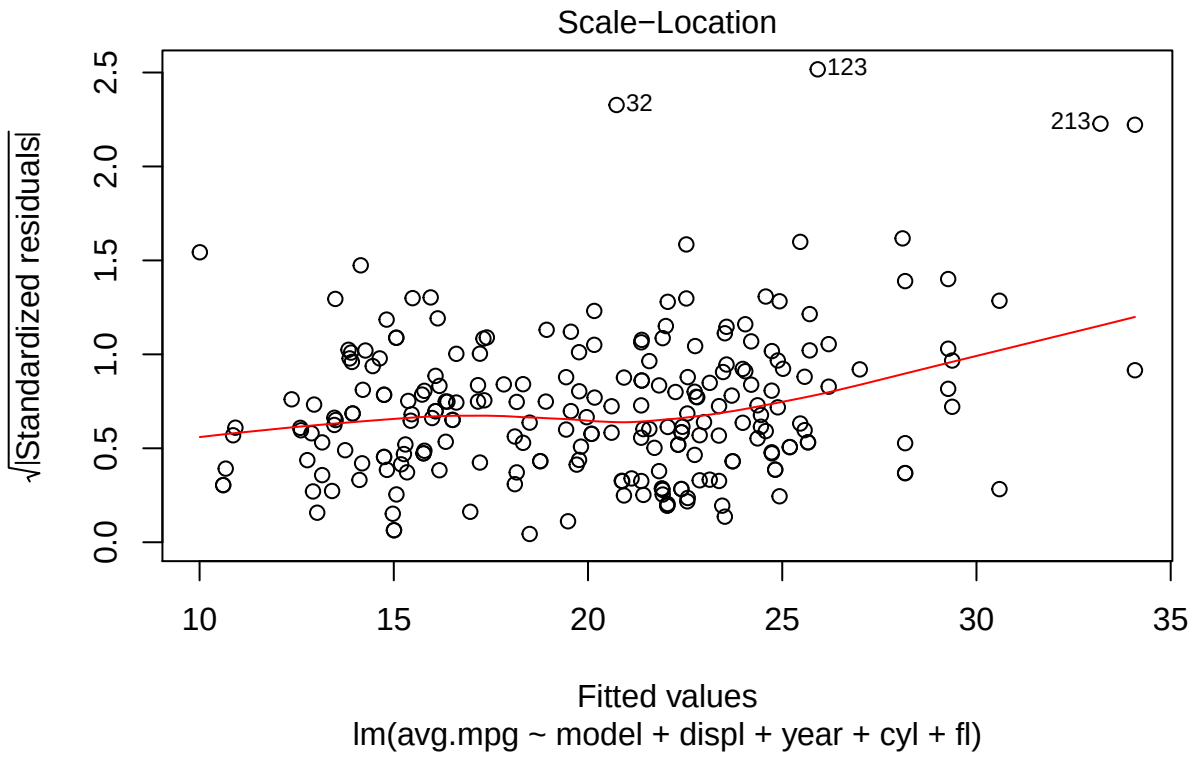
```
plot(best.lm2, which = c(2))
```

```
## Warning: not plotting observations with leverage one:
##   107
```



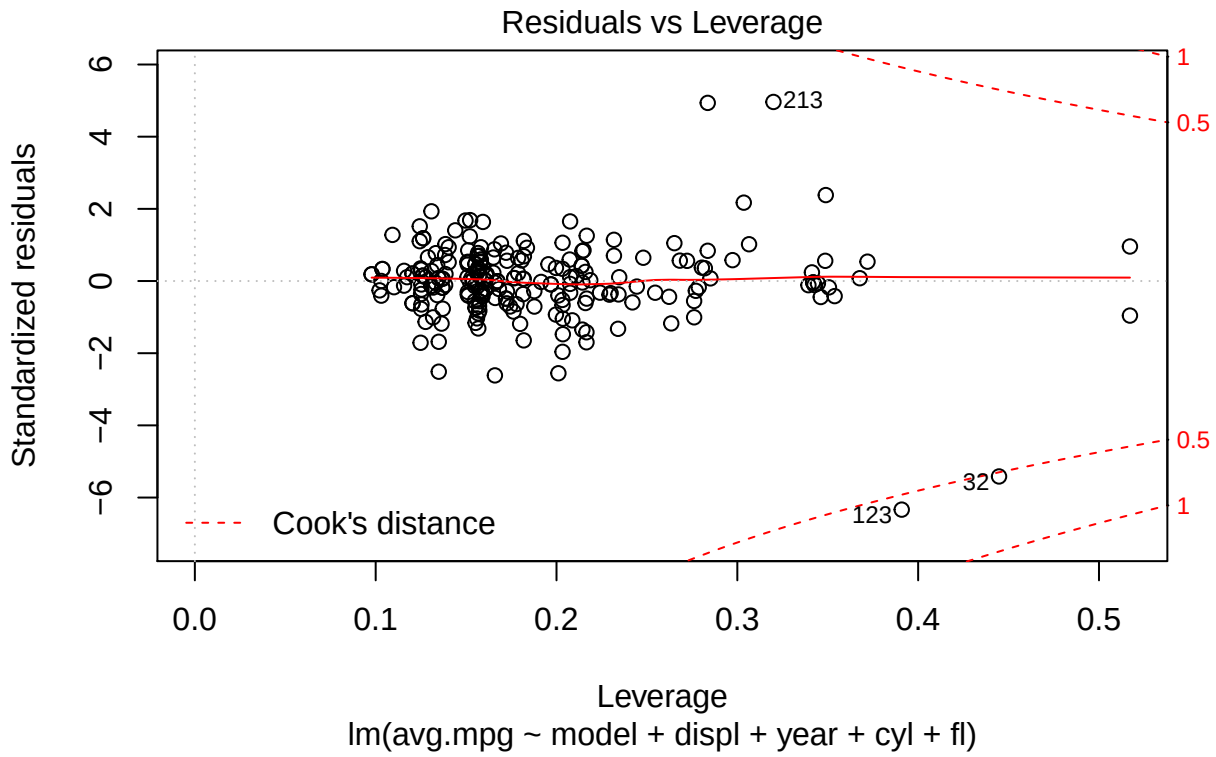

```
plot(best.lm2, which = c(3))
```

```
## Warning: not plotting observations with leverage one:
##   107
```




```
plot(best.lm2, which = c(5))
```

```
## Warning: not plotting observations with leverage one:
##   107
```




```
summary(best.lm2)
```

1i

```
##
## Call:
## lm(formula = avg.mpg ~ model + displ + year + cyl + fl, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4131 -0.4689  0.0206  0.5598  5.4201
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -269.65947   41.77359  -6.455 8.87e-10 ***
## modela4                       6.33180    0.87054   7.273 9.06e-12 ***
## modela4 quattro               4.27103    0.84819   5.035 1.11e-06 ***
## modela6 quattro               4.46730    1.02113   4.375 2.01e-05 ***
## modelaltima                   6.43872    0.77492   8.309 1.84e-14 ***
## modelc1500 suburban 2wd       1.49990    0.88302   1.699 0.091039 .
## modelcamry                    5.85157    0.74052   7.902 2.20e-13 ***
## modelcamry solara             5.76453    0.74228   7.766 4.99e-13 ***
## modelcaravan 2wd              2.37866    0.66185   3.594 0.000415 ***
## modelcivic                    9.23576    0.80333  11.497  < 2e-16 ***
## modelcorolla                 10.45299    0.85508  12.225  < 2e-16 ***
```

```
## modelcorvette                  7.25273    0.97519   7.437 3.49e-12 ***
## modeldakota pickup 4wd        -0.27669    0.70765  -0.391 0.696245
## modeldurango 4wd              -0.33565    0.77604  -0.433 0.665863
## modelexpedition 2wd            0.10989    0.96671   0.114 0.909615
## modelexplorer 4wd              0.06491    0.76295   0.085 0.932294
## modelf150 pickup 4wd          -0.13442    0.75737  -0.177 0.859318
## modelforester awd              2.98905    0.78900   3.788 0.000204 ***
## modelgrand cherokee 4wd       -0.52550    0.73914  -0.711 0.477983
## modelgrand prix                5.92871    0.80034   7.408 4.15e-12 ***
## modelgti                       5.33811    0.85211   6.265 2.47e-09 ***
## modelimpreza awd               3.96671    0.73894   5.368 2.32e-07 ***
## modeljetta                     5.86583    0.74410   7.883 2.47e-13 ***
## modelk1500 tahoe 4wd          -0.86407    0.98003  -0.882 0.379073
## modelland cruiser wagon 4wd   -0.15711    1.10440  -0.142 0.887026
## modelmalibu                    5.20879    0.79435   6.557 5.09e-10 ***
## modelmaxima                    5.48409    0.94223   5.820 2.48e-08 ***
## modelmountaineer 4wd          -0.02279    0.85385  -0.027 0.978737
## modelmustang                   4.04129    0.70357   5.744 3.64e-08 ***
## modelnavigator 2wd             0.47806    0.98352   0.486 0.627483
## modelnew beetle                6.75362    0.82062   8.230 2.99e-14 ***
## modelpassat                    5.87081    0.86223   6.809 1.27e-10 ***
## modelpathfinder 4wd            0.21794    0.85775   0.254 0.799708
## modelram 1500 pickup 4wd      -0.59117    0.73571  -0.804 0.422680
## modelrange rover              -0.31497    0.89979  -0.350 0.726694
## modelsonata                    5.07486    0.74822   6.783 1.47e-10 ***
## modeltiburon                   3.34282    0.77874   4.293 2.82e-05 ***
## modeltoyota tacoma 4wd        -0.16249    0.72440  -0.224 0.822756
## displ                         -0.54623    0.32124  -1.700 0.090703 .
## year                           0.14672    0.02084   7.040 3.45e-11 ***
## cyl                           -0.80369    0.19091  -4.210 3.95e-05 ***
## fld                            7.93723    1.54120   5.150 6.51e-07 ***
## fle                           -4.76678    1.47614  -3.229 0.001464 **
## flp                           -1.79243    1.40143  -1.279 0.202466
## flr                           -0.62492    1.38522  -0.451 0.652408
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.297 on 189 degrees of freedom
## Multiple R-squared:  0.9465, Adjusted R-squared:  0.934
## F-statistic: 75.97 on 44 and 189 DF,  p-value: < 2.2e-16
```

This model that resulted from the 2-way stepAIC function produces results very similar to our previous model in 1d. This is plausible since the AIC difference between the two models si not significant. We can once again analyze the coefficients of each predictor or their p-values to measure their level of significance.

1j For this particular case, I would probably prefer the second model (from 1g) since it does get the job done with more simplicity.