Q-3 Consider again the auto.csv dataset from Q-1.

i. Perform linear regression on mpg as the response with the following predictors: **cylinders, displacement, weight, acceleration, year, origin.**

```
> autodata.mod2 = lm(mpg ~ cylinders+displacement+weight+acceleration+year+origin,
+              data= Auto_data_csv)
```

ii. Provide the summary report.

```
> autodata.mod2 = lm(mpg ~ cylinders+displacement+weight+acceleration+year+orig$
+              data= Auto_data_csv)
> summary(autodata.mod2)

Call:
lm(formula = mpg ~ cylinders + displacement + weight + acceleration +
    year + origin, data = Auto_data_csv)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5640 -2.1692 -0.0382  1.8196 13.0720

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.974e+01  4.168e+00  -4.737 3.06e-06 ***
cylinders    -4.447e-01  3.211e-01  -1.385   0.1668
displacement  1.719e-02  7.189e-03   2.390   0.0173 *
weight       -6.838e-03  5.812e-04 -11.767  < 2e-16 ***
acceleration  1.557e-01  7.777e-02   2.002   0.0460 *
year          7.647e-01  4.973e-02  15.378  < 2e-16 ***
origin        1.346e+00  2.706e-01   4.975 9.87e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.33 on 385 degrees of freedom
Multiple R-squared:  0.8208,    Adjusted R-squared:  0.818
F-statistic: 293.9 on 6 and 385 DF,  p-value: < 2.2e-16
```

iii. Which predictors do not have influence on **mpg** (in statistical sense) and why?

Analyzing the p-values associated with each predictor's t-statistic. All predictors are statistically significant except cylinders. We can also observe that displacement and acceleration also have relatively less p-value, so it's better to ignore these predictors.

```
> autodata.mod3 = lm(mpg ~ displacement+weight+acceleration+year+origin,
+              data= Auto_data_csv)
> summary(autodata.mod3)

Call:
lm(formula = mpg ~ displacement + weight + acceleration + year +
    origin, data = Auto_data_csv)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3110 -2.1671 -0.0526  1.8293 13.0061

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.054e+01  4.133e+00  -4.970 1.01e-06 ***
displacement  1.060e-02  5.398e-03   1.963   0.0503 .
weight       -6.904e-03  5.799e-04 -11.904  < 2e-16 ***
acceleration  1.522e-01  7.782e-02   1.956   0.0512 .
year          7.639e-01  4.978e-02  15.344  < 2e-16 ***
origin        1.319e+00  2.702e-01   4.881 1.55e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.334 on 386 degrees of freedom
Multiple R-squared:  0.8199,    Adjusted R-squared:  0.8175
F-statistic: 351.4 on 5 and 386 DF,  p-value: < 2.2e-16
```

iv. Re-run the model with the remaining subset of predictors that have influence on **mpg**. Provide the summary report and comment on how this differs from part-iii in terms p-value, R^2 etc.

The higher the p value the least is the influence on response. So, from the above summary we must exclude cylinders, displacement, acceleration.

- HERE from the summary we can observe increase in F- Statistics.
- R^2 is decreased from 0.8208 to 0.8175.

```
> autodata.mod4 = lm(mpg ~ weight+year+origin,
+                    data= Auto_data_csv)
> summary(autodata.mod4)

Call:
lm(formula = mpg ~ weight + year + origin, data = Auto_data_csv)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9440 -2.0948 -0.0389  1.7255 13.2722

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.805e+01  4.001e+00  -4.510 8.60e-06 ***
weight      -5.994e-03  2.541e-04 -23.588  < 2e-16 ***
year         7.571e-01  4.832e-02  15.668  < 2e-16 ***
origin       1.150e+00  2.591e-01   4.439 1.18e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.348 on 388 degrees of freedom
Multiple R-squared:  0.8175,    Adjusted R-squared:  0.816
F-statistic: 579.2 on 3 and 388 DF,  p-value: < 2.2e-16

>
```

```
> anova(autodata.mod2,autodata.mod4)
Analysis of Variance Table

Model 1: mpg ~ cylinders + displacement + weight + acceleration + year +
    origin
Model 2: mpg ~ weight + year + origin
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    385 4269.0
2    388 4348.1 -3   -79.153 2.3795 0.06932 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Graduate

v. Analyze by considering *all* the predictors in the dataset and how it influences **mpg** as response.

```
> autodata.mod5= lm(mpg ~ cylinders+displacement+horsepower+weight+accele
+              data= Auto_data_csv)
> summary(autodata.mod5)

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin + name, data = Auto_data_csv)

Residuals:
   Min     1Q Median     3Q    Max
-5.646  0.000  0.000  0.000  5.646

Coefficients: (1 not defined because of singularities)
                                      Estimate Std. Error t value
(Intercept)                           0.187305  12.773943   0.015
cylinders                            -0.918096   0.616653  -1.489
displacement                          0.003041   0.015621   0.195
horsepower                           -0.042342   0.029070  -1.457
weight                               -0.004193   0.001209  -3.467
acceleration                         -0.481449   0.171537  -2.807
year                                  0.636498   0.112195   5.673
origin                                1.324264   4.243221   0.312
nameamc ambassador dpl                3.371358   3.245610   1.039
nameamc ambassador sst                3.364264   3.270924   1.029
nameamc concord                      -0.122500   3.275058  -0.037
nameamc concord d/l                  -1.686548   3.422955  -0.493
nameamc concord dl 6                 -0.529687   3.514452  -0.151
nameamc gremlin                      -0.426606   2.958293  -0.144
nameamc hornet                        0.269026   2.902972   0.093
nameamc hornet sportabout (sw)       -0.403111   3.527112  -0.114
```

namechevrolet vega 2300	5.309133	3.978383	1.334
namechevrolet woody	0.242401	4.035002	0.060
namechevy c10	-1.686820	3.292205	-0.512
namechevy c20	4.381943	3.755983	1.167
namechevy s-10	4.822281	4.016031	1.201
namechrysler cordoba	3.158331	3.288955	0.960
namechrysler lebaron medallion	-2.791925	4.039754	-0.691
namechrysler lebaron salon	-4.524894	3.582800	-1.263
namechrysler lebaron town @ country (sw)	2.084281	3.302393	0.631
namechrysler new yorker brougham	5.282481	3.390324	1.558
namechrysler newport royal	4.391942	3.291715	1.334
namedatsun 1200	7.452422	6.449055	1.156
namedatsun 200-sx	-5.232326	6.424023	-0.814
namedatsun 200sx	2.817519	6.441163	0.437
namedatsun 210	5.606103	6.195917	0.905
namedatsun 210 mpg	5.070308	6.390048	0.793
namedatsun 280-zx	5.896050	6.830585	0.863
namedatsun 310	4.643895	6.277737	0.740
namedatsun 310 gx	4.043465	6.266014	0.645
namedatsun 510	-2.468844	6.400190	-0.386
namedatsun 510 (sw)	3.062356	6.557580	0.467
namedatsun 510 hatchback	6.552699	6.395446	1.025
namedatsun 610	-3.382103	6.545652	-0.517
namedatsun 710	1.331250	6.306882	0.211
namedatsun 810	-3.215340	6.816828	-0.472
namedatsun 810 maxima	-2.442319	6.860787	-0.356
namedatsun b-210	2.371920	6.357360	0.373
namedatsun b210	3.331328	6.462449	0.515
namedatsun b210 gx	9.604655	6.404554	1.500
namedatsun f-10 hatchback	2.950464	6.329051	0.466
namedatsun pl510	0.981674	6.254240	0.157
namedodge aries se	0.403815	4.041675	0.100
namedodge aries wagon (sw)	-2.256829	4.019898	-0.561

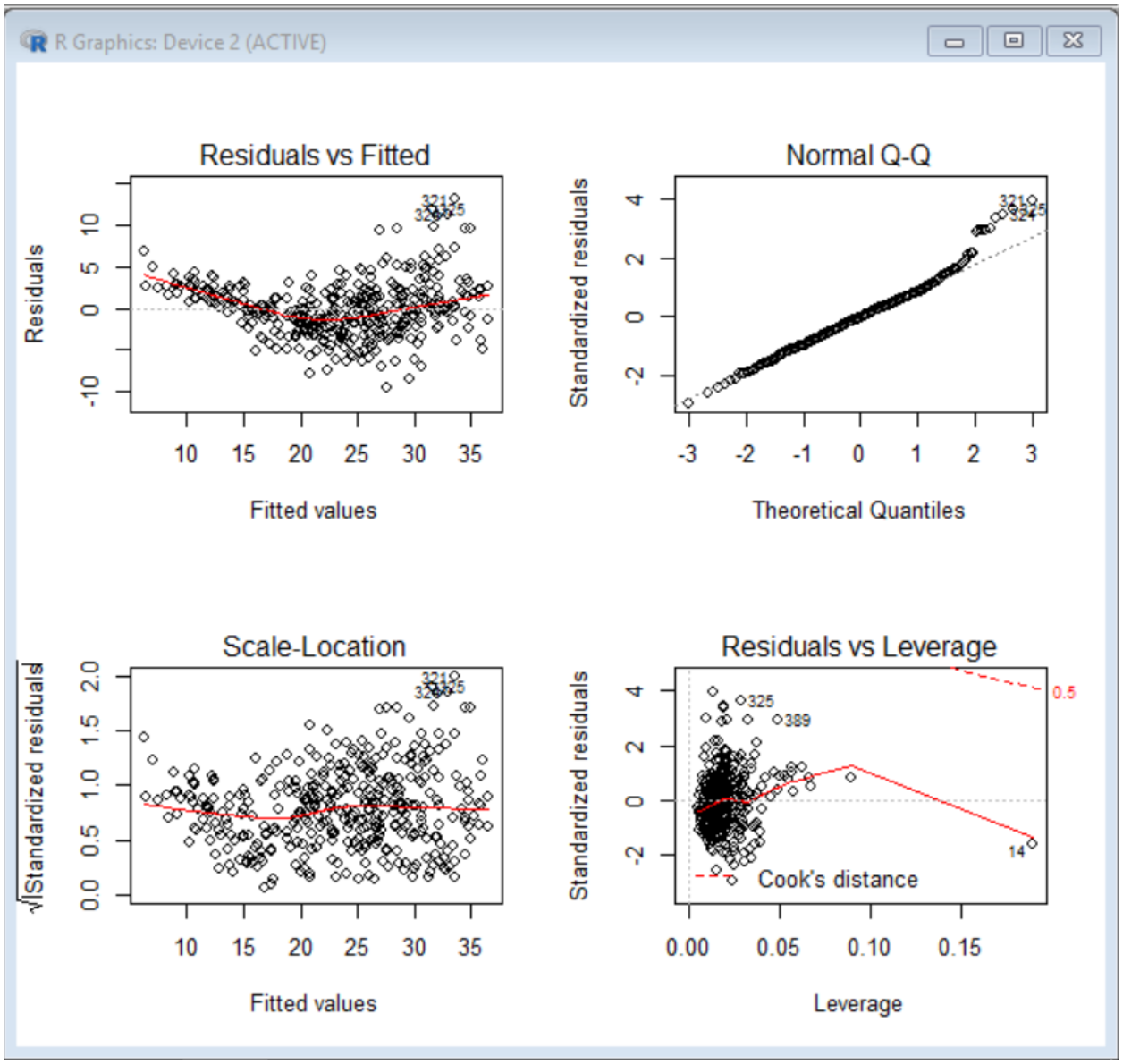
```
nametriumph tr7 coupe                        0.069741 .
namevokswagen rabbit                         0.383382
namevolkswagen 1131 deluxe sedan             0.676597
namevolkswagen 411 (sw)                      0.784470
namevolkswagen dasher                        0.749866
namevolkswagen jetta                         0.731190
namevolkswagen model 111                     0.617833
namevolkswagen rabbit                        0.703474
namevolkswagen rabbit custom                 0.745228
namevolkswagen rabbit custom diesel          4.24e-05 ***
namevolkswagen rabbit l                      0.338159
namevolkswagen scirocco                      0.770470
namevolkswagen super beetle                  0.929370
namevolkswagen type 3                        0.834581
namevolvo 144ea                              0.441829
namevolvo 145e (sw)                          0.357530
namevolvo 244dl                              0.607987
namevolvo 245                                0.425122
namevolvo 264gl                              0.135809
namevolvo diesel                             0.049758 *
namevw dasher (diesel)                       1.23e-05 ***
namevw pickup                                5.00e-05 ***
namevw rabbit                                0.091774 .
namevw rabbit c (diesel)                     2.49e-05 ***
namevw rabbit custom                               NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.272 on 85 degrees of freedom
Multiple R-squared:  0.9816,    Adjusted R-squared:  0.9153
F-statistic:  14.8 on 306 and 85 DF,  p-value: < 2.2e-16
```

From above statistics, we can identify that name is not the valid predictor.

```
> autodata.mod6 = lm(mpg~.-name, data=Auto_data_csv)
> summary(autodata.mod6)

Call:
lm(formula = mpg ~ . - name, data = Auto_data_csv)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

- **Relationship**: There is a relationship between the predictors and the response mpg by testing the null hypothesis of whether all the regression coefficients are zero. The F -statistic is far from 1 (with a small p-value), indicating evidence against the null hypothesis.
- Looking at the p-values associated with each predictor's t-statistic, we see that displacement, weight, year, and origin have a statistically significant relationship, while cylinders, horsepower, and acceleration do not.
- The relation between cylinders, horsepower, weight is negative. So, the relationship between mpg and cylinders, horsepower, weight predictors are negative. For the other predictors, its positive.

```
> par(mfrow=c(2,2))
> plot(autodata.mod6)
```



- The fit does not appear to be accurate because there is a discernible curve pattern to the residuals plots. From the leverage plot, point 14 appears to have high leverage, although not a high magnitude residual.