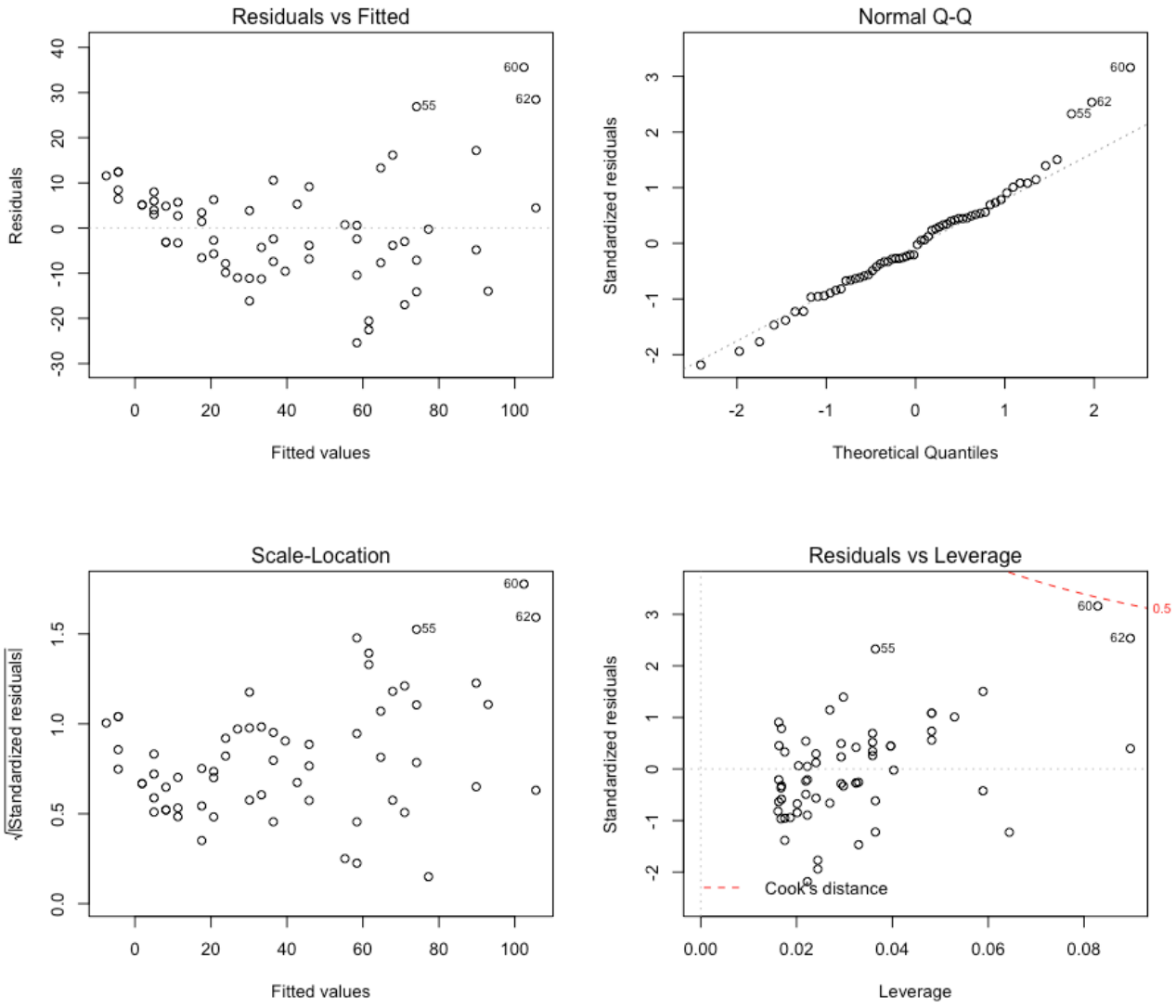
Homework 2

Name: Zixin Ouyang

1.

install.packages("alr4")
library("alr4")
stopmod<-lm(Distance~Speed,data=stopping)
(a)

```
> par(mfrow=c(2,2))
> plot(stopmod, add.smooth=FALSE)
```



(b) We use the residuals-versus-fitted plot to check the assumed mean function(linearity). The trend is not roughly flat and the vertical spread is not equal, so the assumed mean function is not very solid.

(c) Both the residuals-versus-fitted plot and the scale-location plot show that the variance is not constant, the residuals increase when the fitted values get larger.

(d) The largest (most positive) least-squares residual value is 35.60783, and the smallest (most negative) least-squares residual value is -25.40952.6.

```
> max(residuals(stopmod))
[1] 35.60783
> min(residuals(stopmod))
[1] -25.40952
```

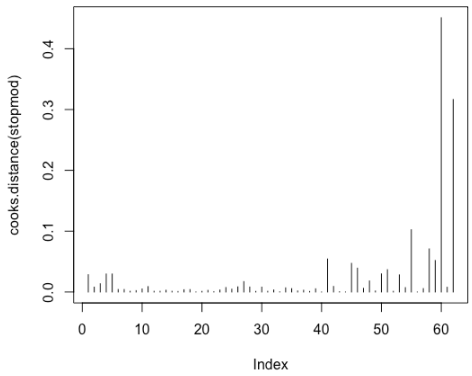
(e) Observation 61 has the largest leverage value, and the leverage value is 0.0896.

```
> which.max(hatvalues(stopmod))
61
61
> hatvalues(stopmod)[61]
        61
0.08967251
```

(f) According to the normal Q-Q plot, there are several outliers in the data set. Some points deviate from the straight line.

(g) According to the residuals versus leverage plot, all Cook's Distances are less than 0.5, so there is no highly influential point. We could also use R to calculate the Cook's Distance and verify the conclusion.

```
> cooks.distance(stopmod)
          1           2           3           4           5           6           7           8
2.851169e-02 7.922295e-03 1.362428e-02 2.963726e-02 2.963726e-02 4.098688e-03 4.098688e-03 1.251745e-03
          9          10          11          12          13          14          15          16
2.226069e-03 5.010332e-03 8.908747e-03 1.235103e-03 1.235103e-03 2.947582e-03 1.211326e-03 8.271565e-04
         17          18          19          20          21          22          23          24
3.665400e-03 3.933200e-03 1.868146e-04 1.073468e-03 2.698044e-03 6.077598e-04 3.273780e-03 7.347350e-03
         25          26          27          28          29          30          31          32
4.666982e-03 8.458898e-03 1.707784e-02 8.133440e-03 9.799569e-04 7.947843e-03 1.143106e-03 3.339743e-03
         33          34          35          36          37          38          39          40
3.548934e-04 6.795745e-03 5.497583e-03 1.709641e-03 2.946114e-03 9.291838e-04 5.275372e-03 4.113440e-05
         41          42          43          44          45          46          47          48
5.424219e-02 9.103441e-03 4.877581e-04 2.929265e-05 4.711739e-02 3.913056e-02 6.077298e-03 1.818542e-02
         49          50          51          52          53          54          55          56
1.678841e-03 2.984051e-02 3.665101e-02 1.126364e-03 2.824621e-02 7.179696e-03 1.024174e-01 1.060929e-05
         57          58          59          60          61          62
5.593290e-03 7.084428e-02 5.182358e-02 4.509000e-01 7.792085e-03 3.165446e-01
> which(cooks.distance(stopmod) >= 1)
named integer(0)
> plot(cooks.distance(stopmod), type="h")
```

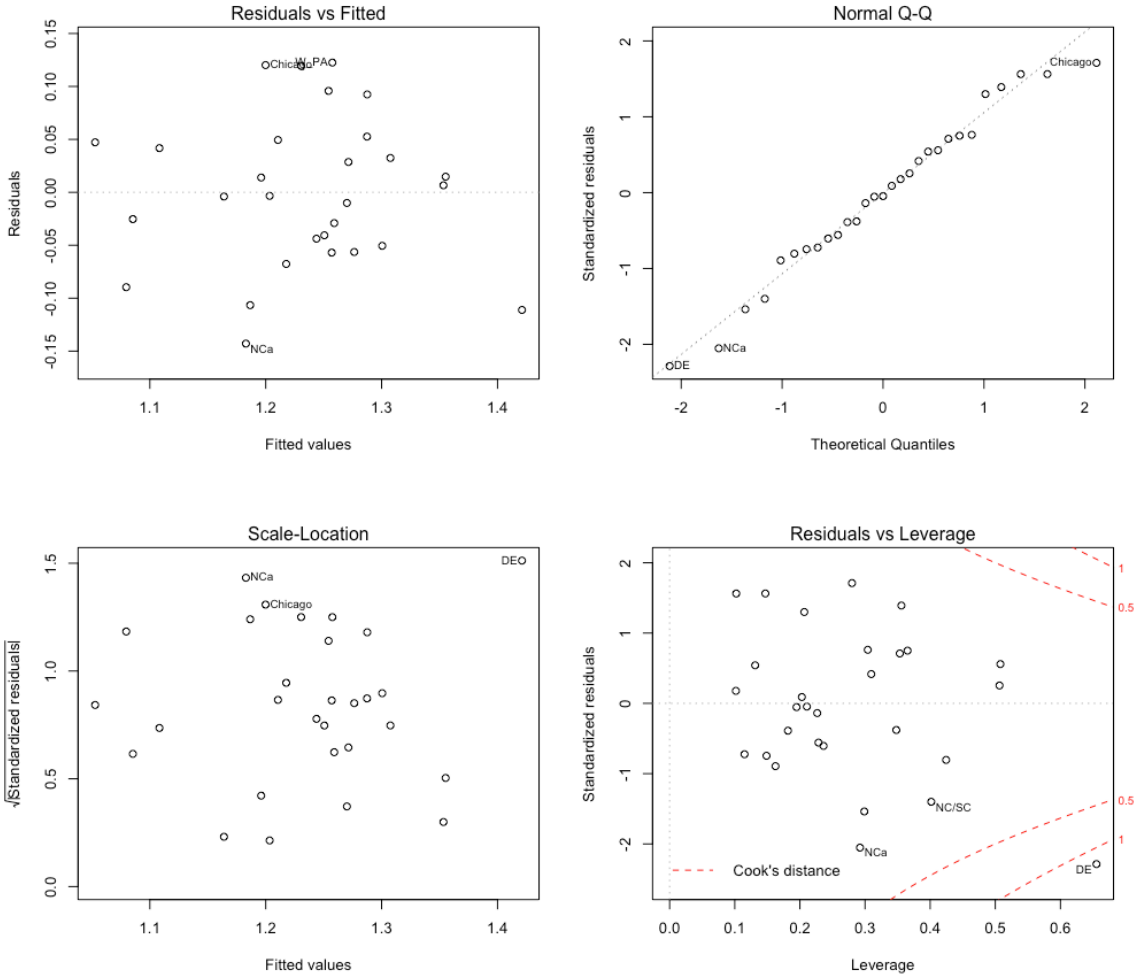


2.

```
> drugmod<-lm(COST~RXPM+GS+RI+COPAY+AGE+F+MM,data=drugcost)
```

(a)

```
> par(mfrow=c(2,2))
> plot(drugmod, add.smooth=FALSE)
```



(b)We use the residuals-versus-fitted plot to check the assumed mean function(linearity). The residuals are between -0.15 to 0.15. The trend is roughly flat and the vertical spread is almost equal, so the assumed mean function can hold.

(c) The scale-location plot show that there is not clear pattern of residuals , so the variance is constant.

(d) The largest (most positive) least-squares residual value is 0.1225225, and the smallest (most negative) least-squares residual value is -0.142888.

```
> max(residuals(drugmod))
[1] 0.1225225
> min(residuals(drugmod))
[1] -0.142888
```

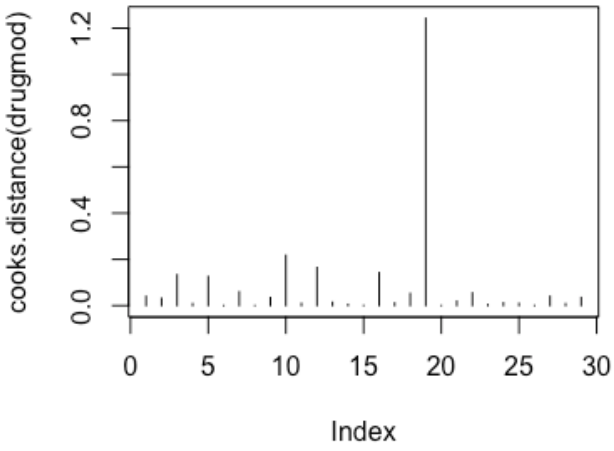
(e) Observation DE has the largest leverage value, and the leverage value is 0.6553194.

```
> which.max(hatvalues(drugmod))
DE
19
> hatvalues(drugmod)[19]
       DE
0.6553194
```

(f) According to the normal Q-Q plot, DE, Chicago and Nca are outliers in the data set.

(g) According to the residuals versus leverage plot, most Cook's Distances are less than 0.5. But DE has Cook's Distances greater than 1, so the observation DE is highly influential. We could also use R to calculate the Cook's Distance and verify the conclusion.

```
> hatvalues(drugmod)[19]
       DE
0.6553194
> cooks.distance(drugmod)
         MN1          MN2          MN3           GA          GA2          AZ1          AZ2           TN
4.026002e-02 3.173883e-02 1.337041e-01 8.508650e-03 1.260317e-01 8.617347e-05 5.969978e-02 6.999681e-05
   San_Diego          NCa         SoCA        NC/SC           LA           FL       Dallas      Chicago
3.441171e-02 2.174260e-01 9.583553e-03 1.644654e-01 1.417312e-02 5.537856e-03 6.998861e-04 1.422969e-01
     Houston           NJ           DE Mid-Atlantic     Richmond           NY       C/E_PA         S_NE
1.153971e-02 5.260164e-02 1.242687e+00 2.565483e-04 1.936044e-02 5.501197e-02 4.189496e-03 1.214343e-02
   St._Louis           OH    Cincinnati     Columbus         W_PA
9.722857e-03 4.480640e-04 4.046480e-02 8.264996e-03 3.476433e-02
> which(cooks.distance(drugmod) >= 1)
DE
19
> plot(cooks.distance(drugmod), type="h")
```

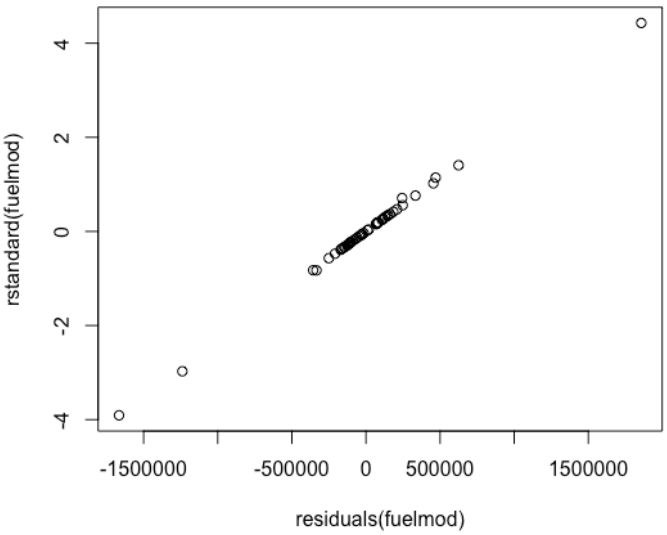


3.

```
> fuelmod<-lm(FuelC~Tax+Drivers+Income,data=fuel2001)
```

(a)

```
> plot(rstandard(fuelmod)~residuals(fuelmod))
```



(b) The standardized residuals are obtained by dividing the ordinary residuals by their standard errors. When x_i is far from the mean, it will tend to have small variance. When the variance is small, it will tend to have large standardized residuals.

(c)

```
> rstudent(fuelmod)
         AL          AK          AZ          AR          CA          CO          CT          DE          DC
-0.46986746 -0.82491434 -0.22552020 -0.26084510  0.70545119 -0.05630035  0.19227116  0.24295695  0.30391038
         FL          GA          HI          ID          IL          IN          IA          KS          KY
-3.26347861  1.14716565 -0.56496830 -0.06721422 -0.37602240  0.41402139  0.15750910 -0.10053959  0.04140851
         LA          ME          MD          MA          MI          MN          MS          MO          MT
 0.36507551 -0.23761917  0.75825045 -0.29709900  0.46684060  1.42180335 -0.07658889  0.55145654 -0.06313120
         NE          NV          NH          NJ          NM          NY          NC          ND          OH
 0.02320195  0.26378418  0.25899886  0.16981675 -0.38521666 -4.70659975  0.32209073 -0.16259656 -0.31755855
         OK          OR          PA          RI          SC          SD          TN          TX          UT
 0.15315873 -0.35581807 -0.82241372  0.34623857  0.17893242 -0.03417366 -0.12936141  5.74349084 -0.25316822
         VT          VA          WA          WV          WI          WY
-0.19927711  1.02258646 -0.22841001 -0.33722937  0.18520584 -0.28225433
```

(d)States FL, NY, and TX are identified as outliers

```
> critval1<-qt(0.05/2,df=df.residual(fuelmod)-1, lower=FALSE)
> which(abs(rstudent(fuelmod)) > critval1)
FL NY TX
10 33 44
```

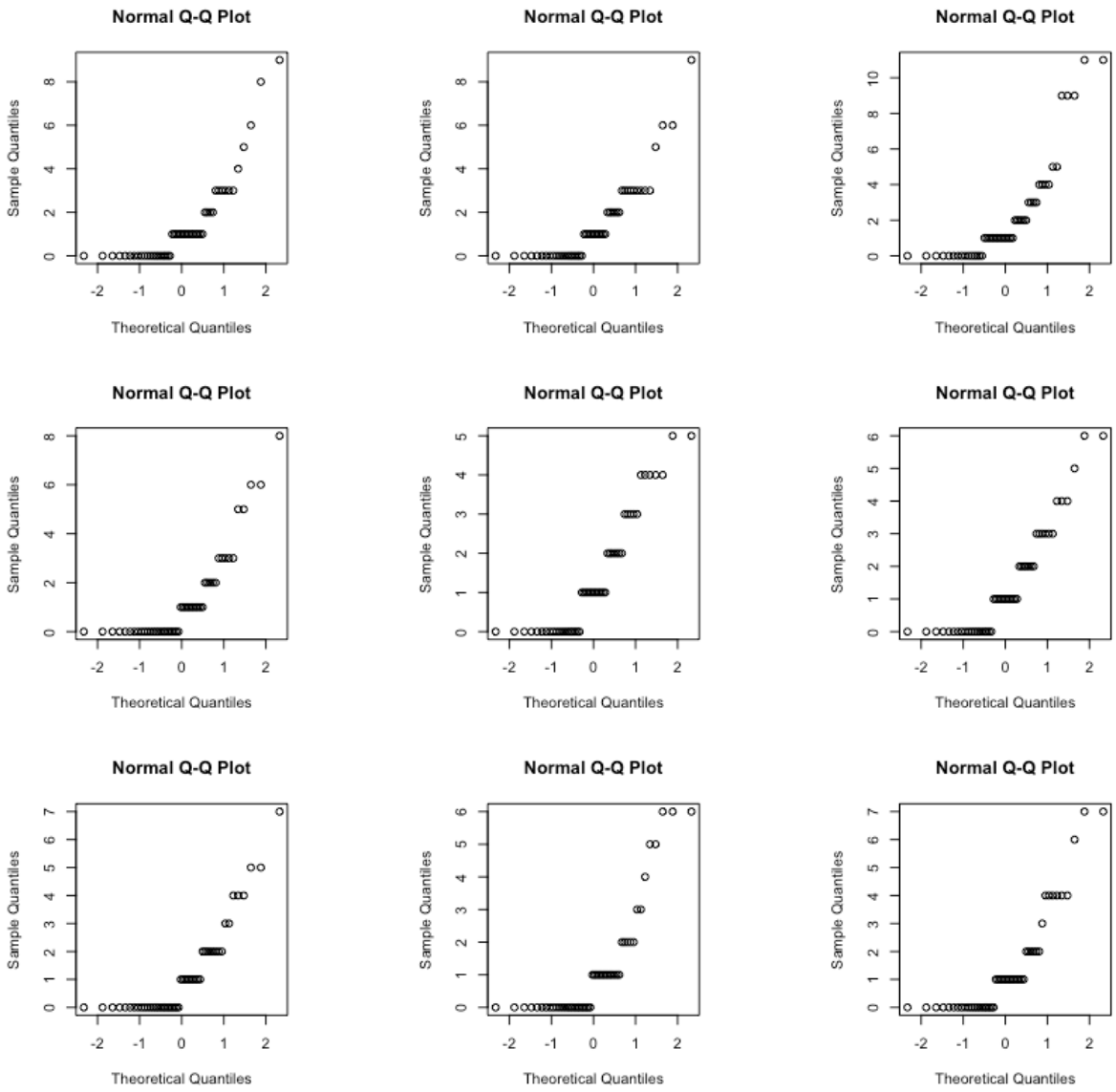
(e)State NY and TX are identified as outliers

```
> critval2<-qt(0.05/(2*nobs(fuelmod)),df=df.residual(fuelmod)-1, lower=FALSE)
> which(abs(rstudent(fuelmod)) > critval2)
NY TX
33 44
```

4.

(a)

```
> par(pty="s")
> par(mfrow=c(3,3))
> for(i in 1:9){
+ qqnorm(rgeom(50,0.4))
+ }
```



(b)

The points in normal Q-Q plot cannot form a straight line.
There are more points clustered at the lower quantiles.