Cpts575 Hw4

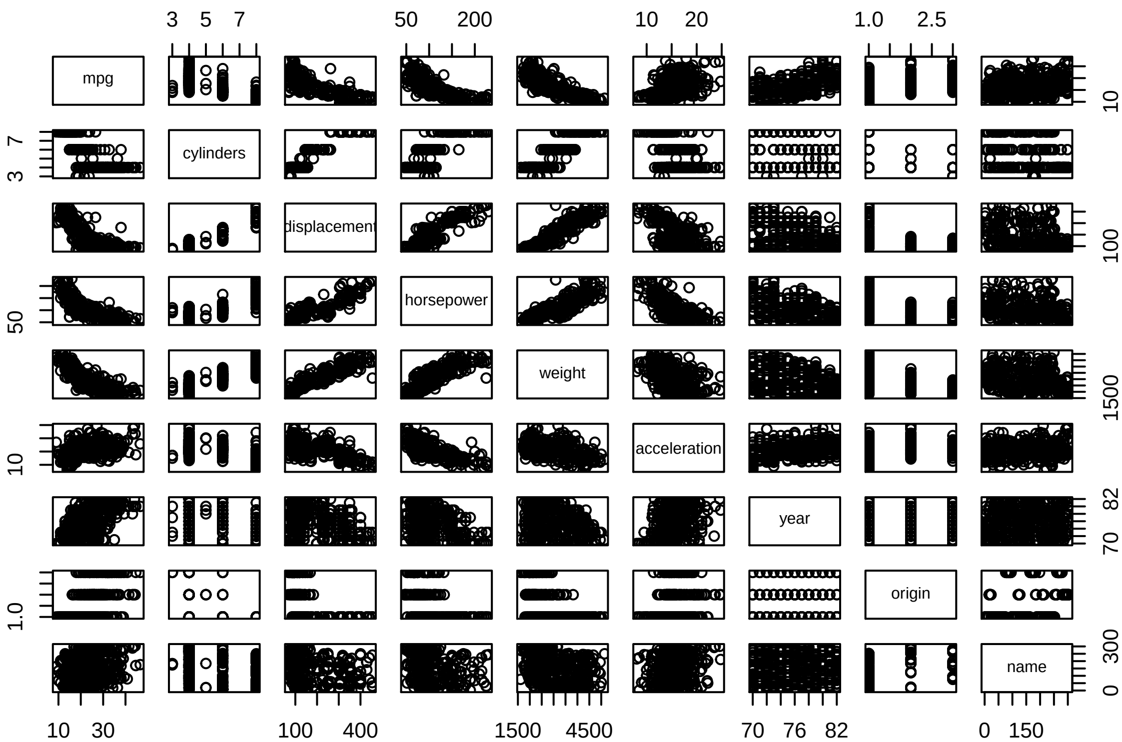
Mengxiao

Part 1

```r
library(dplyr)
library(graphics)
Auto = read.csv("https://scads.eecs.wsu.edu/wp-content/uploads/2017/09/Auto.csv", na.string = '?')
Auto = na.omit(Auto)
#Auto = Auto[Auto$horsepower != '?',] #Moving out the missing data
```

a. Produce a scatterplot matrix

```r
pairs(Auto)
```



b. Compute the matrix of correlations.

```r
Auto2 = Auto %>% dplyr::select(-name)
cor(Auto2)
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
```

```
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

c. Perform a multiple linear regression.

```r
lr = lm(mpg~., data = Auto2)
summary(lr)
```

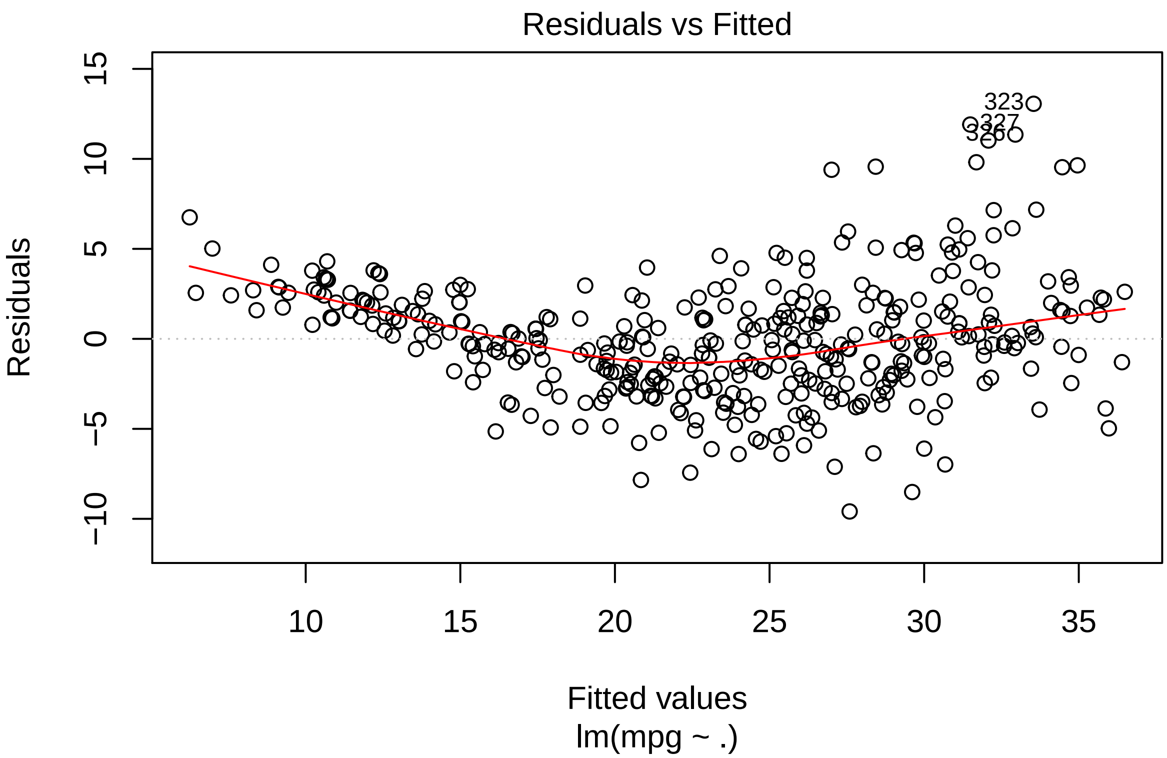
```
##
## Call:
## lm(formula = mpg ~ ., data = Auto2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -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
```

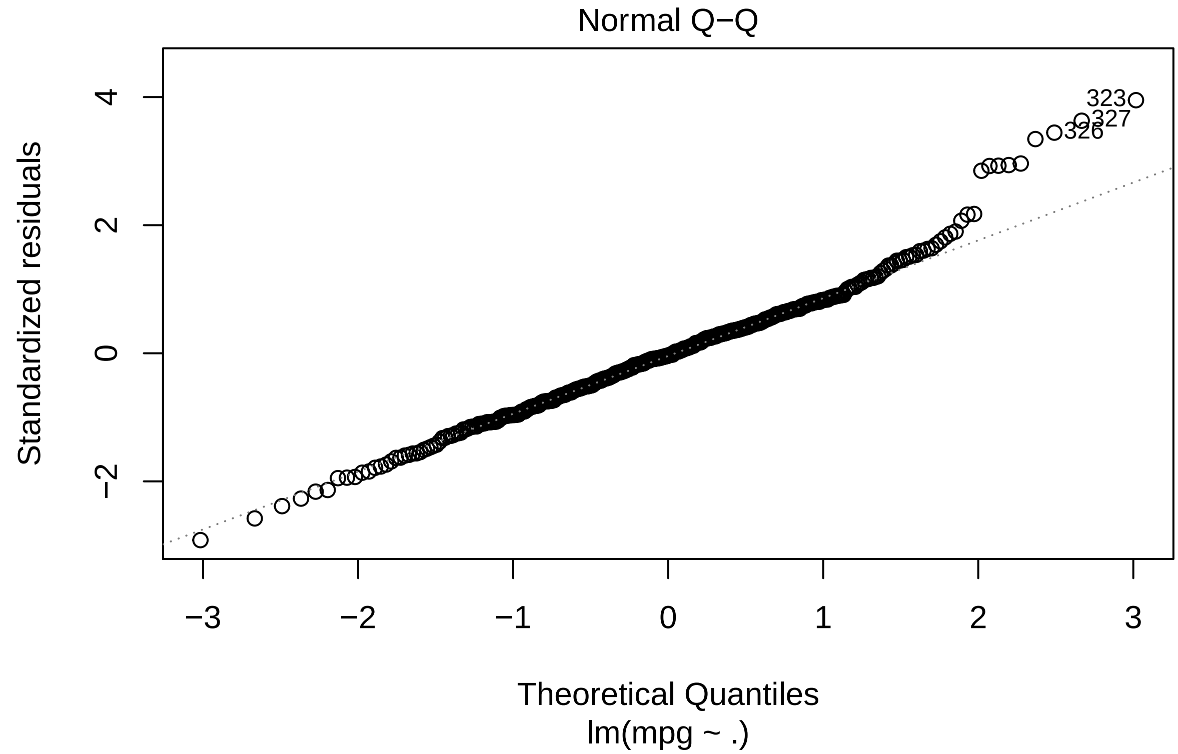
i. I think the 'displacement', 'weight', 'year' and 'origin' have the statistically significant relationship with the 'mpg', since their P-value are less than '0.05', they are significant.

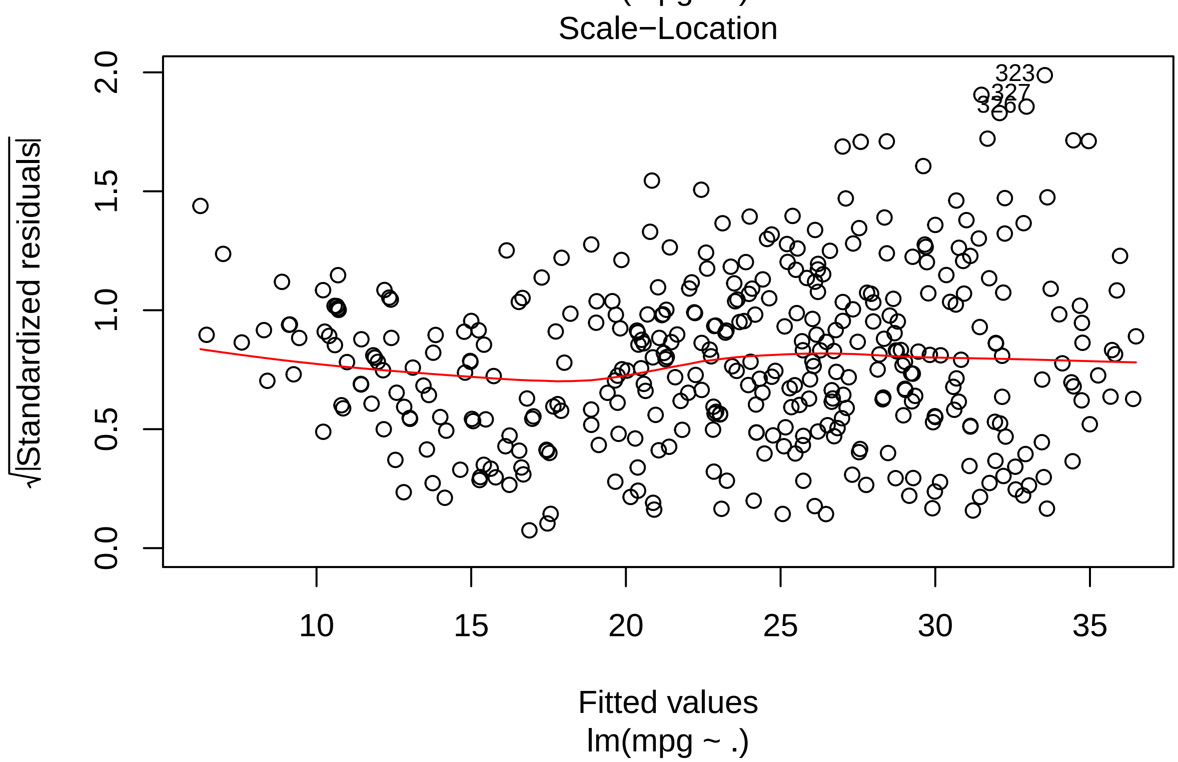
ii. Means when the value of displacement increase 1%, the mpg will increase 0.019896%.

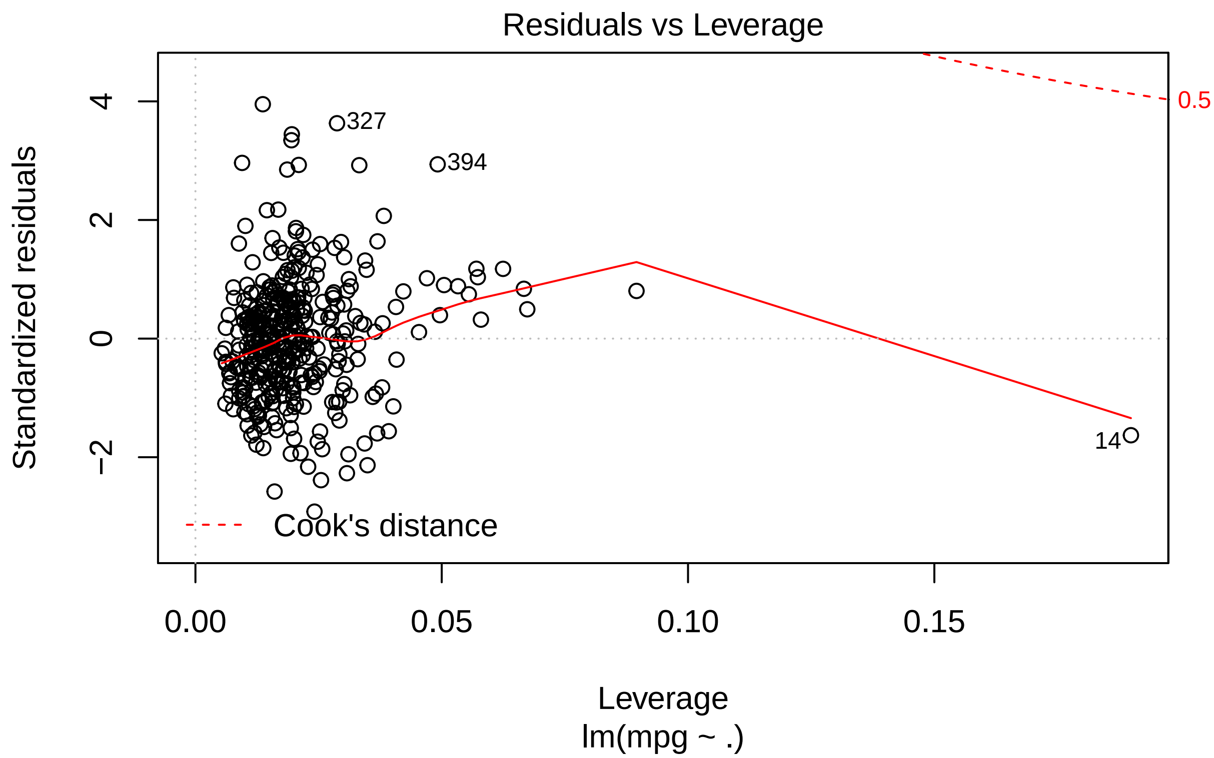
d. Produce diagnostic plots of the linear regression fit.

```r
plot(lr)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ .)
323
327
326


Scale−Location
√|Standardized residuals|
Fitted values
lm(mpg ~ .)
323
327
326



The residual plots looks good, but still have some outliers.

Yes, it identifies some unusually outliers

e. Fit linear regression models with interaction effects.

```
lm_e = lm(mpg~cylinders*displacement + weight*displacement, data=Auto2)
summary(lm_e)
```

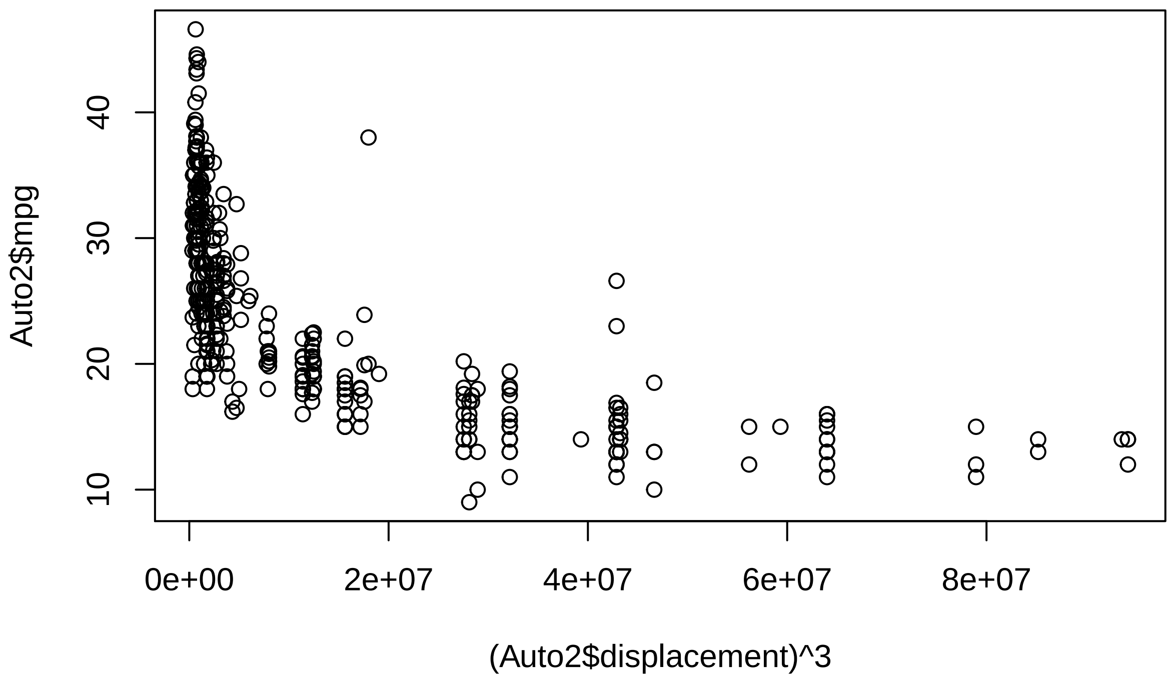
```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + weight * displacement,
##     data = Auto2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2934  -2.5184  -0.3476   1.8399  17.7723
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.262e+01  2.237e+00  23.519  < 2e-16 ***
## cylinders              7.606e-01  7.669e-01   0.992    0.322
## displacement          -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
## weight                -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
## cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
## displacement:weight    2.128e-05  5.002e-06   4.254 2.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.103 on 386 degrees of freedom
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
## F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

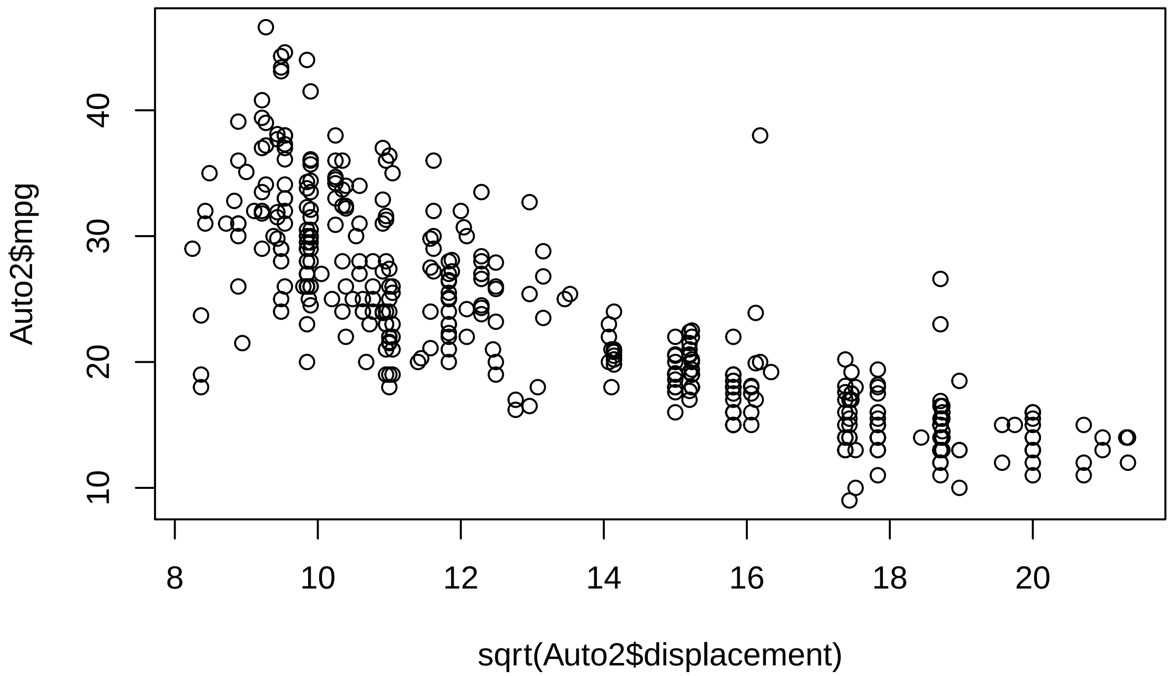
We can see from the summay that displacement and weight have statistically signifcant relationship, but the relationship between cylinders and displacement is not significant.

f. Try transformations of the variables with X^3 and log(X).

```r
plot((Auto2$displacement)^3, Auto2$mpg)
```




```r
plot(sqrt(Auto2$displacement), Auto2$mpg)
```



It looks like the distribution is more aggregated of X^3.

Part 2

```r
library(MASS)
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

a.

```r
lm_zn = lm(crim~zn, data=Boston)
lm_indus = lm(crim~indus, data=Boston)
lm_chas = lm(crim~chas, data=Boston)
lm_nox = lm(crim~nox, data=Boston)
lm_rm = lm(crim~rm, data=Boston)
```

```r
lm_age = lm(crim~age, data=Boston)
lm_dis = lm(crim~dis, data=Boston)
lm_rad = lm(crim~rad, data=Boston)
lm_tax = lm(crim~tax, data=Boston)
lm_ptratio = lm(crim~ptratio, data=Boston)
lm_black = lm(crim~black, data=Boston)
lm_lstat = lm(crim~lstat, data=Boston)
lm_medv = lm(crim~medv, data=Boston)
```

I find that only 'chas' don't have statistically significant relationship with crim, all of other variables have significant relationship.

b.

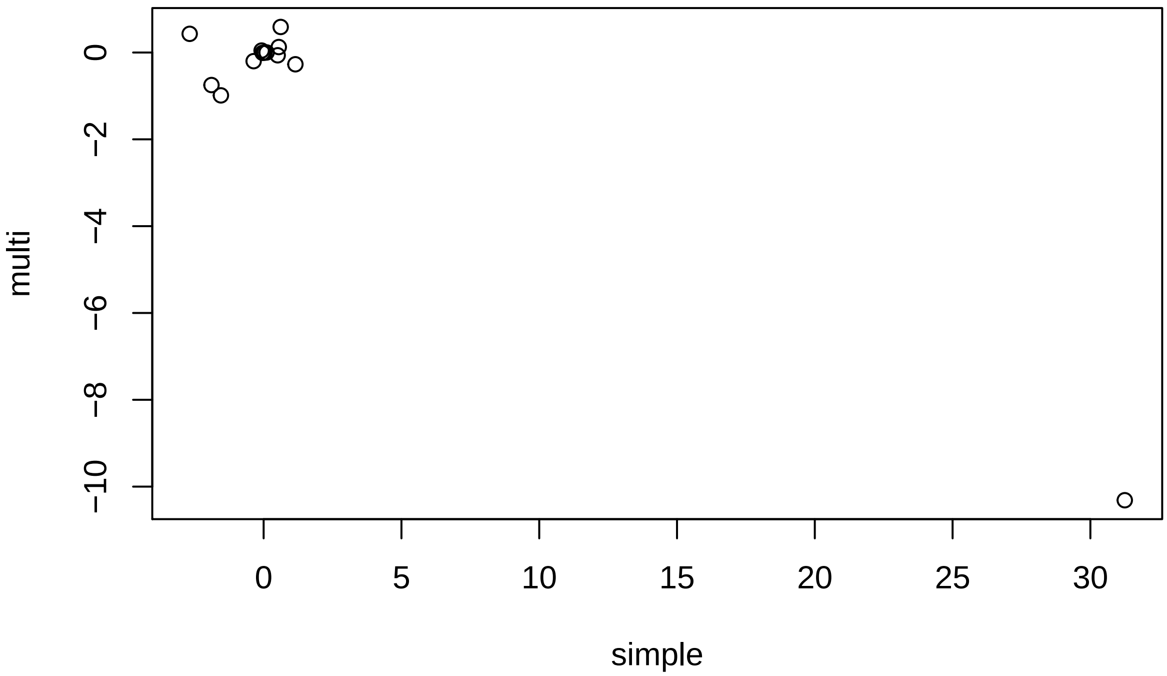
```r
lm_mul = lm(crim~., data=Boston)
summary(lm_mul)
```

```
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

In my opinion, we can reject the 'zn', 'dis', 'rad', 'black' and 'medv', since their P-value are all less than 0.05.

c.How do your results from (a) compare to your results from (b)?

```
simple = c(lm_zn$coefficients[2], lm_indus$coefficients[2], lm_chas$coefficients[2], lm_nox$coefficient
           lm_rm$coefficients[2], lm_age$coefficients[2], lm_dis$coefficients[2], lm_rad$coefficients[2
           lm_tax$coefficients[2], lm_ptratio$coefficients[2], lm_black$coefficients[2], lm_lstat$coeff
           lm_medv$coefficients[2])
multi = c(lm_mul$coefficients)
multi = multi[-1]
plot(simple, multi)
```



The coefficients of simple is much higher than it of multiple, that the arrange of simple is 0 to 30 and for multiple is from -10 to 0. In my opinion, it is because simple predict only shows whether two variables have relationship and the rate of relation, but the multiple predict shows the rate of different variables' influence.

d. Is there evidence of non-linear association between any of the predictors and the response?

```
lm_zn2 = lm(crim~poly(zn, 3), data=Boston)
lm_indus2 = lm(crim~poly(indus, 3), data=Boston)
lm_nox2 = lm(crim~poly(nox, 3), data=Boston)
lm_rm2 = lm(crim~poly(rm, 3), data=Boston)
lm_age2 = lm(crim~poly(age, 3), data=Boston)
lm_dis2 = lm(crim~poly(dis, 3), data=Boston)
lm_rad2 = lm(crim~poly(rad, 3), data=Boston)
lm_tax2 = lm(crim~poly(tax, 3), data=Boston)
lm_ptratio2 = lm(crim~poly(ptratio, 3), data=Boston)
lm_black2 = lm(crim~poly(black, 3), data=Boston)
lm_lstat2 = lm(crim~poly(lstat, 3), data=Boston)
lm_medv2 = lm(crim~poly(medv, 3), data=Boston)

summary(lm_zn2)
```

```r
summary(lm_indus2)
summary(lm_nox2)
summary(lm_rm2)
summary(lm_age2)
summary(lm_dis2)
summary(lm_rad2)
summary(lm_tax2)
summary(lm_ptratio2)
summary(lm_black2)
summary(lm_lstat2)
summary(lm_medv2)
```

I have found that only the 'black' don't have non-linear association, since the P-value of quandratic and cubic coefficients are all higher than 0.05. The other variables all have non-linear association, but some of them only have quandratic association and the other have cubic.

Part 3

a.

i.The prediction would be not impartial and not exact.

ii. It means the weight of each coefficients cannot be seperated exactly.

iii. The prediction will have more error since the cofidence intervals are not exact, sometimes we will accept some variables that don't significant before.

b. Use the covariates between two errors to constrain the correlation error.