STATS 413 Hw2

Shu Zhou

2020/9/28

This is the Assignment 1 of STATS 413 Author: Shu Zhou UMID: 19342932

Exercise 1

We cannot reject the hypothesis that "TV" and "radio" has significant impact on "Sales". But we can reject the hypothesis that "newspaper" is significant according to its P-value (<0.8599)

Exercise 3

(a.)

$Y = 50 + 20 \times GPA + 0.07 \times IQ + 35 \times Gender + 0.01 \times (GPA \times IQ) - 10 \times (GPA \times Gender)$
Point (iii) is correct, Since when GPA is high enough(which is greater than 3.5), males would earn more than females.

(b.)

For a female with IQ of 110 and a GPA of 4.0,
$Y = 50 + 20 \times 4 + 0.07 \times 110 + 35 \times 1 + 0.01 \times 4 \times 110 - 10 \times 4 \times 1 = 137.1 k dollars$

(c.)

False, since the coefficient β_4 is not zero, it reflects some interactions between GPA and experience. Since both GPA and IQ are great in great scales (GPA from 0-4 Experience can go to the hundreds), the impact of this interaction might cause a great impact on the salary.

Exercise 7

Given: The simple linear regression is calculated by

$$TSS = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2$$

$$RSS = \sum_i (y_i - \hat{y_i})^2 = \sum_i (y_i - \frac{\sum_j x_j y_j}{\sum_j x_j^2 x_i})^2$$

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{\sum_j y_j^2 - (\sum_i y_i^2 - 2\sum_i y_i(\frac{\sum_j x_j y_j}{\sum_j x_j^2})x_i + \sum_i(\frac{\sum_j x_j y_j}{\sum_j x_j^2})^2 x_i^2)}{\sum_j y_j^2} = \frac{2\frac{(\sum_i x_i y_i)^2}{\sum_j x_j^2} - \frac{(\sum_i x_i y_i)^2}{\sum_j x_j^2}}{\sum_j y_j^2} = (\frac{\sum_i x_i y_i}{\sum_j x_j \sum_j y_j})^2$$

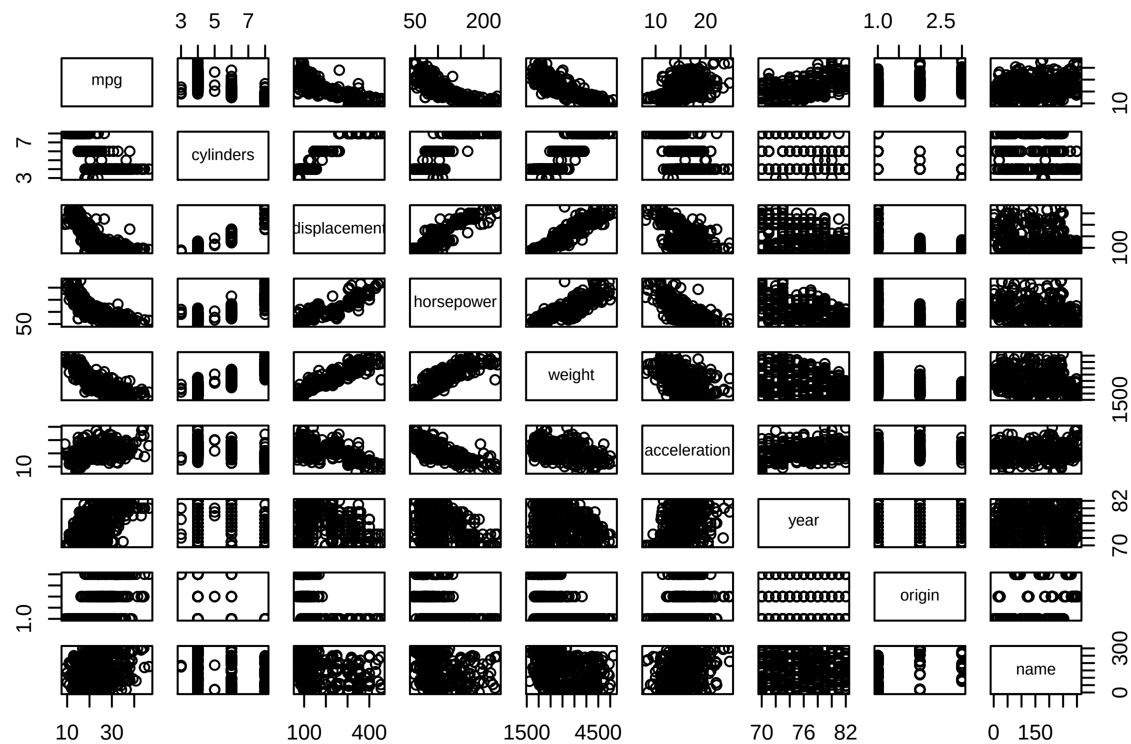
$$Cor(X, Y) = \frac{\sum_i x_i y_i}{\sum_j x_j \sum_j y_j}$$

Hence, Q.E.D.

Exercise 9

(a.)

```r
data("Auto")
pairs(Auto)
```



(b.)

```r
cor(subset(Auto, select=-name))
```

```
##                     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.)

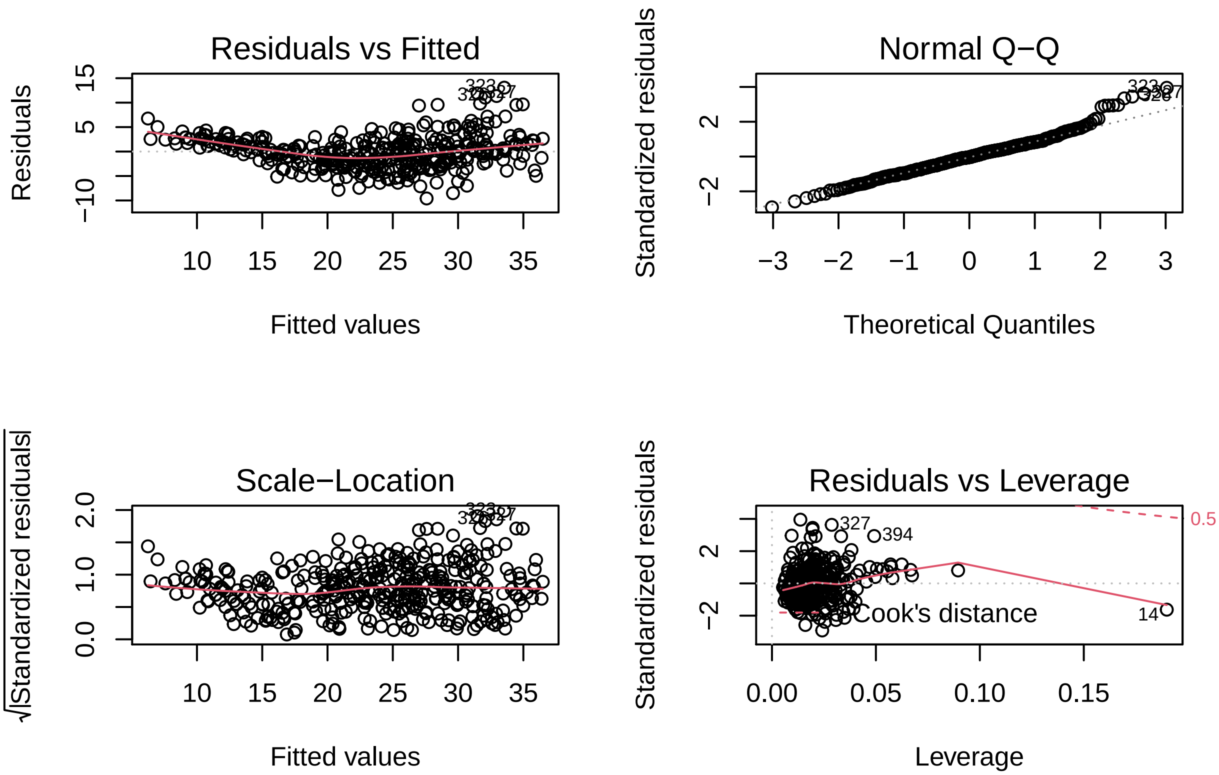
```r
fit <- lm(mpg ~ . - name, data = Auto)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## 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
```

- There is a relationship between predictors and response.
- The variables "displacement", " weight", "acceleration","year" and "origin" have significant impact on mpg.
- The coefficient ot the “year” variable suggests that the an increase of 1 year would cause an increase of 0.7507727 in “mpg” .

(d.)

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



The plot of residuals-fitted values indicates that there is no relationship between the residuals and fitted values, which reflects non-linearity.

The plot of residual vs. Leverages indicates observation 14 has high leverage.

(e.) From the correlation data, we can observe that 1.cylinders and displacement 2.displacement and weight 3. horsepower and weight have great correlation (>0.9)

```r
fit2 <- lm(mpg ~ cylinders * displacement+displacement * weight+horsepower*weight, data = Auto[, 1:8])
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight + horsepower * weight, data = Auto[, 1:8])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9295  -2.1066  -0.3601   1.8641  15.7110
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.529e+01  3.185e+00  20.500  < 2e-16 ***
## cylinders              -1.368e+00  8.075e-01  -1.694  0.09111 .
## displacement           -4.133e-02  2.352e-02  -1.757  0.07972 .
## weight                 -8.126e-03  1.338e-03  -6.076 2.97e-09 ***
## horsepower             -2.323e-01  5.624e-02  -4.130 4.46e-05 ***
## cylinders:displacement  7.378e-03  3.666e-03   2.013  0.04486 *
```

```
## displacement:weight    -4.032e-06  8.604e-06  -0.469  0.63958
## weight:horsepower       4.630e-05  1.606e-05   2.883  0.00416 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.916 on 384 degrees of freedom
## Multiple R-squared:  0.7527, Adjusted R-squared:  0.7482
## F-statistic:   167 on 7 and 384 DF,  p-value: < 2.2e-16
```

According to the p-values, we can see that the interaction between 1. displacement and weight 2. horsepower and weight are statistically significant. While the interaction between cylinders and displacement can be rejected and not significant.

(f.)

```r
fit0 <- lm(mpg~displacement+weight+year+origin, Auto[, 1:8])
fit3 <- lm(mpg~displacement+I(sqrt(weight))+year+origin, Auto[, 1:8])
fit4 <- lm(mpg~displacement+I(log(weight))+year+origin, Auto[, 1:8])
fit5 <- lm(mpg~displacement+I(weight^2)+year+origin, Auto[, 1:8])
summary(fit0)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + weight + year + origin, data = Auto[,
##     1:8])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8102 -2.1129 -0.0388  1.7725 13.2085
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.861e+01  4.028e+00  -4.620 5.25e-06 ***
## displacement  5.588e-03  4.768e-03   1.172    0.242
## weight       -6.575e-03  5.571e-04 -11.802  < 2e-16 ***
## year          7.714e-01  4.981e-02  15.486  < 2e-16 ***
## origin        1.226e+00  2.670e-01   4.593 5.92e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.346 on 387 degrees of freedom
## Multiple R-squared:  0.8181, Adjusted R-squared:  0.8162
## F-statistic: 435.1 on 4 and 387 DF,  p-value: < 2.2e-16
```

```r
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + I(sqrt(weight)) + year + origin,
##     data = Auto[, 1:8])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7348 -2.0154  0.0539  1.6762 13.0776
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.514591   4.305605   0.584   0.5595
## displacement      0.008465   0.004422   1.915   0.0563 .
## I(sqrt(weight)) -0.784635   0.057758 -13.585  < 2e-16 ***
## year              0.790391   0.047908  16.498  < 2e-16 ***
## origin            1.030154   0.257008   4.008 7.34e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.211 on 387 degrees of freedom
## Multiple R-squared:  0.8325, Adjusted R-squared:  0.8308
## F-statistic: 480.9 on 4 and 387 DF,  p-value: < 2.2e-16
```

```
summary(fit4)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + I(log(weight)) + year + origin,
##     data = Auto[, 1:8])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7136 -1.9214  0.0447  1.5790 12.9864
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    131.274483  11.082986  11.845  < 2e-16 ***
## displacement     0.007711   0.004052   1.903 0.057810 .
## I(log(weight)) -21.584745   1.451851 -14.867  < 2e-16 ***
## year             0.804835   0.046532  17.296  < 2e-16 ***
## origin           0.836143   0.250485   3.338 0.000925 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.113 on 387 degrees of freedom
## Multiple R-squared:  0.8425, Adjusted R-squared:  0.8409
## F-statistic: 517.7 on 4 and 387 DF,  p-value: < 2.2e-16
```

```
summary(fit5)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + I(weight^2) + year + origin,
##     data = Auto[, 1:8])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0988  -2.2549  -0.1057   1.8704  13.4702
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.609e+01  4.349e+00  -5.999 4.56e-09 ***
## displacement -9.114e-03  5.118e-03  -1.781   0.0757 .
## I(weight^2)  -7.068e-07  9.075e-08  -7.789 6.28e-14 ***
## year          7.336e-01  5.380e-02  13.635  < 2e-16 ***
```

```
## origin        1.488e+00  2.900e-01   5.132 4.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.628 on 387 degrees of freedom
## Multiple R-squared:  0.7861, Adjusted R-squared:  0.7839
## F-statistic: 355.7 on 4 and 387 DF,  p-value: < 2.2e-16
```

From the result, we can see that $\log(weight)$ and $\sqrt(weight)$ have greater coefficient than weight however $weight^2$have less coefficient.

Exercise 15

(a.)

```
lmp <- function (modelobject) {
    if (class(modelobject) != "lm") stop("Not an object of class 'lm' ")
    f <- summary(modelobject)$fstatistic
    p <- pf(f[1],f[2],f[3],lower.tail=F)
    attributes(p) <- NULL
    return(p)
}
data("Boston")
fit.zn <- lm(crim ~ zn,data = Boston)
fit.indus <- lm(crim ~ indus,data = Boston)
fit.chas <- lm(crim ~ chas,data = Boston)
fit.nox <- lm(crim ~ nox,data = Boston)
fit.rm <- lm(crim ~ rm,data = Boston)
fit.rad <- lm(crim ~ rad,data = Boston)
fit.tax <- lm(crim ~ tax,data = Boston)
fit.ptratio <- lm(crim ~ ptratio,data = Boston)
fit.black <- lm(crim ~ black,data = Boston)
fit.lstat <- lm(crim~lstat,data = Boston)
fit.medv <- lm (crim~medv,data = Boston)
lmp(fit.zn)
```

```
## [1] 5.506472e-06
```

```
lmp(fit.indus)
```

```
## [1] 1.450349e-21
```

```
lmp(fit.chas)
```

```
## [1] 0.2094345
```

```
lmp(fit.nox)
```

```
## [1] 3.751739e-23
```

```
lmp(fit.rm)
```

```
## [1] 6.346703e-07
```

```
lmp(fit.rad)
```

```
## [1] 2.693844e-56
```

```r
lmp(fit.tax)
```

```
## [1] 2.357127e-47
```

```r
lmp(fit.ptratio)
```

```
## [1] 2.942922e-11
```

```r
lmp(fit.black)
```

```
## [1] 2.487274e-19
```

```r
lmp(fit.lstat)
```

```
## [1] 2.654277e-27
```

```r
lmp(fit.medv)
```

```
## [1] 1.173987e-19
```

We can see that only the variable "chas" is non-significant.

(b.)

```r
data("Boston")
fit.lm <- lm(crim~., data=Boston)
summary(fit.lm)
```

```
##
## 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
```

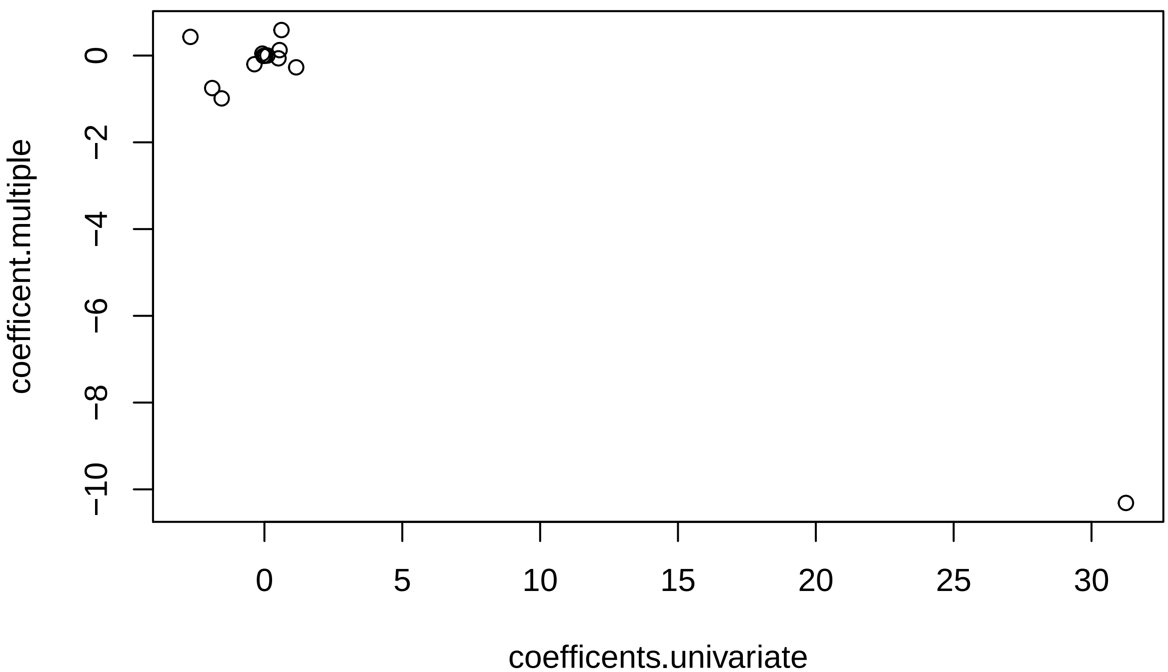
Hence, We can reject the null hypothesis for "zn", "dis", "rad", "black" and "medv".

(c.)

```r
function(x){coefficients(lm(Boston[, x]))}
```

```
## function(x){coefficients(lm(Boston[, x]))}
```

```r
results <- combn(names(Boston), 2, function(x) { coefficients(lm(Boston[, x])) })
coefficents.univariate <- unlist(results)[seq(2,26,2)]
coefficent.multiple <- coefficients(fit.lm)[-1]
plot(coefficents.univariate, coefficent.multiple)
```



Only the variable "chas" is non-significant when performing univariate regression, however more variables become non-significant in mutiple regression due to the impact of other variables.

(d.)

```r
data("Boston")

lm(crim ~ poly(zn,3),data = Boston)
```

```
##
## Call:
## lm(formula = crim ~ poly(zn, 3), data = Boston)
##
## Coefficients:
##  (Intercept)  poly(zn, 3)1  poly(zn, 3)2  poly(zn, 3)3
##        3.614       -38.750        23.940       -10.072
```

```r
lm(crim ~ poly(indus,3),data = Boston)
```

```
##
## Call:
## lm(formula = crim ~ poly(indus, 3), data = Boston)
##
## Coefficients:
##     (Intercept)  poly(indus, 3)1  poly(indus, 3)2  poly(indus, 3)3
##           3.614           78.591          -24.395          -54.130
```

```r
lm(crim ~ poly(nox,3),data = Boston)
```

```
##
## Call:
## lm(formula = crim ~ poly(nox, 3), data = Boston)
##
## Coefficients:
##   (Intercept)  poly(nox, 3)1  poly(nox, 3)2  poly(nox, 3)3
##         3.614         81.372        -28.829        -60.362
```

```r
lm(crim ~ poly(rm,3),data = Boston)
```

```
##
## Call:
## lm(formula = crim ~ poly(rm, 3), data = Boston)
##
## Coefficients:
##  (Intercept)  poly(rm, 3)1  poly(rm, 3)2  poly(rm, 3)3
##        3.614       -42.379        26.577        -5.510
```

```r
lm(crim ~ poly(rad,3),data = Boston)
```

```
##
## Call:
## lm(formula = crim ~ poly(rad, 3), data = Boston)
##
## Coefficients:
##   (Intercept)  poly(rad, 3)1  poly(rad, 3)2  poly(rad, 3)3
##         3.614        120.907         17.492          4.698
```

```r
lm(crim ~ poly(tax,3),data = Boston)
```

```
##
## Call:
## lm(formula = crim ~ poly(tax, 3), data = Boston)
##
## Coefficients:
##   (Intercept)  poly(tax, 3)1  poly(tax, 3)2  poly(tax, 3)3
##         3.614        112.646         32.087         -7.997
```

```r
lm(crim ~ poly(ptratio,3),data = Boston)
```

```
##
## Call:
## lm(formula = crim ~ poly(ptratio, 3), data = Boston)
##
## Coefficients:
```

```
##       (Intercept)  poly(ptratio, 3)1  poly(ptratio, 3)2  poly(ptratio, 3)3
##             3.614             56.045             24.775            -22.280
```

```r
lm(crim~poly(lstat,3),data = Boston)
```

```
##
## Call:
## lm(formula = crim ~ poly(lstat, 3), data = Boston)
##
## Coefficients:
##     (Intercept)  poly(lstat, 3)1  poly(lstat, 3)2  poly(lstat, 3)3
##           3.614           88.070           15.888          -11.574
```

```r
lm(crim~poly(medv,3),data = Boston)
```

```
##
## Call:
## lm(formula = crim ~ poly(medv, 3), data = Boston)
##
## Coefficients:
##    (Intercept)  poly(medv, 3)1  poly(medv, 3)2  poly(medv, 3)3
##          3.614         -75.058          88.086         -48.033
```

From the result, we can see that there is evidence of non-linear association for all of the predictors except "chas".