Assignment 2

Yue Zhang

2025-09-22

```
setwd("/Users/yuezhang/Documents/Biostat/PH1976")
getwd()
library(ISLR2)
library(ggplot2)
library(tidyr)
library(dtplyr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(stats)
library(broom)
```

1.

Answer.
Intercept: Null hypothesis: when spending $0 on TV, radio, and newspaper advertising, the expected number of units sold is 0. As the p value is less than 0.0001, we have enough evidence to reject the null hypothesis. Even without advertising, there is a non-zero baseline level of sales.

TV: Null hypothesis: changing TV advertising while holding radio and newspaper fixed does not change expected sales. As the p value is less than 0.0001, we have enough evidence to reject the null. TV advertising is positively associated with the number of sales.

Radio Null hypothesis: changing radio advertising while holding TV and newspaper fixed does not change expected sales. As the p value is less than 0.0001, we have enough evidence to reject the null hypothesis. Radio advertising is positively associated with the number of sales.

Newspaper: Null hypothesis: changing newspaper advertising while holding TV and radio fixed does not change expected sales. As the p value is 0.8599 which is higher than 0.05, we fail to reject the null hypothesis. Newspaper advertising does not have a strong relationship with the expected number of sales.

10.

(a).

```
carseats_model1 = lm(data = Carseats, Sales ~ Price + Urban + US)
```

(b).

```r
summary(carseats_model1)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

Intercept: if price = 0, and the store is located in an rural location outside the US, the expected sales will be 13.04 units. Price: while holding urban and US constant, for each $1 increase in price, the expected sales decreased by 0.054 units. The p-value is less than 0.05, indicating that price and sales have a strong negative correlation. UrbanYes: while holding US and price constant, stores located in urban areas sell 0.022 fewer units than stores located in rural areas. The p-value is 0.936 which is larger than 0.05, we don't have enough evidence to say that urban is a significant factor. US: while holding price and urban constant, stores located in the US sell 1.201 more units than stores located outside the US. The p-value is less than 0.05, thus whether or not located in the US and sales have a strong positive relationship.

(c).

$\text{Sales} = -0.054\ \text{Price} - 0.022\ \text{UrbanYes} + 1.201\ \text{USYes} + 13.043$. UrbanYes = 1 if Urban = "Yes", UrbanYes = 0 if Urban = "No". USYes = 1 if US = "Yes", USYes = 0 if US = "No".

(d).

Based on the p-values, we have enough evidence to reject the null hypothesis for price and US as their p-values are less than 0.05.

(e).

```r
carseats_model2 = lm(data = Carseats, Sales ~ Price + US)
```

(f).

```r
summary(carseats_model2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

```
anova(carseats_model1, carseats_model2)
```

```
## Analysis of Variance Table
##
## Model 1: Sales ~ Price + Urban + US
## Model 2: Sales ~ Price + US
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    396 2420.8
## 2    397 2420.9 -1  -0.03979 0.0065 0.9357
```

For the first model, the R squared is 0.2393. For the second model, the R squared is also 0.2393. The ANOVA test shows that the p-value is 0.9357, indicating the model doesn't improve when adding Urban. Therefore, the second model fits as well as the first model.

(g)

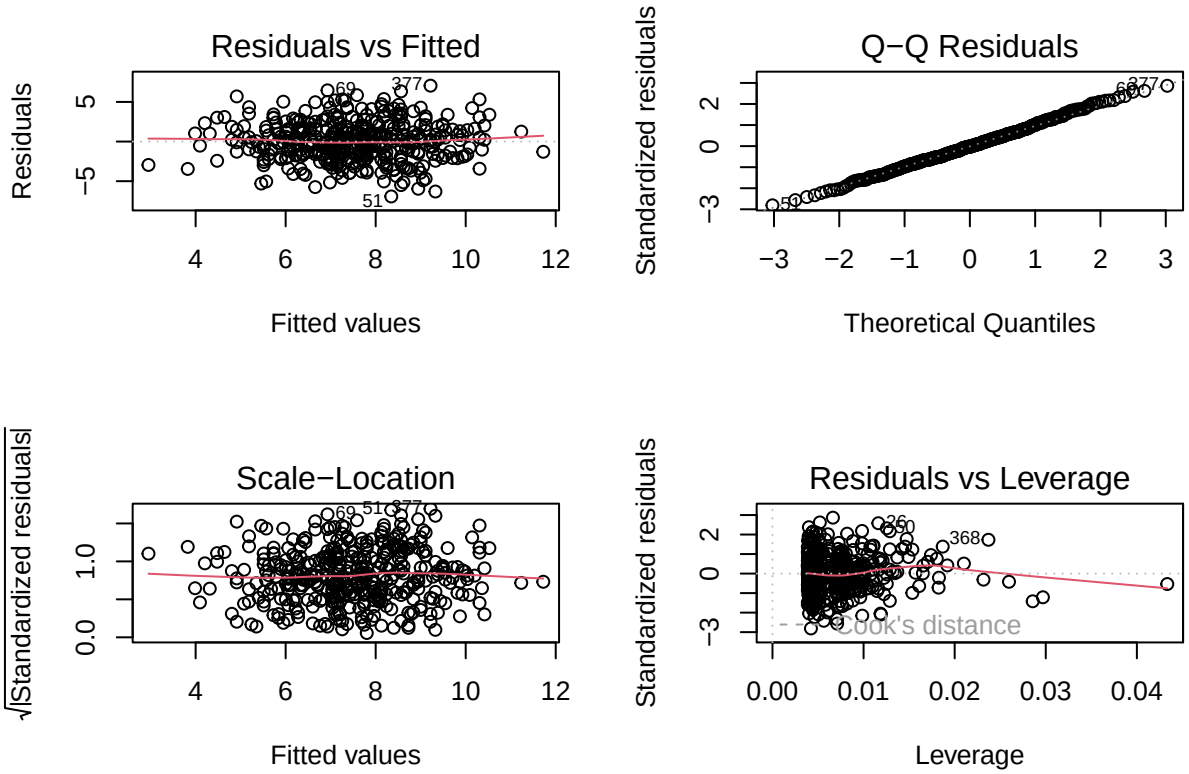
```
confint(carseats_model2, level = 0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

95% CI for intercept: [11.790, 14.271]
95% CI for price: [-0.065, -0.044]
95% CI for USYes: [0.692, 1.708]

(h).

```
par(mfrow = c(2, 2))
plot(carseats_model2)
```



From the residuals vs. fitted plot, we can see that observation 69, 377 and 51 have larger residuals. The residuals vs. leverage plot shows that observation 26, 50 and 368 have the largest Cook's distance but the values are still small. Therefore, we don't have strong evidence of outliers or high leverage observations after checking the plots. We still need more precise tests like DFITTS or DFBETA to examine further.

12.

(a). $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$ $\hat{\beta}_2 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$ Therefore, if we want the two coefficients be the same: $\sum_{i=1}^n y_i^2 = \sum_{i=1}^n x_i^2$ or $(\sum_{i=1}^n x_i y_i = 0)$

(b).

```
set.seed(10)
n  = 100
x  = rnorm(n)
y  = 2*x + rnorm(n, 0, 0.1)

yx = coef(lm(y ~ x + 0))
xy = coef(lm(x ~ y + 0))

print(yx)
```

```
##        x
## 1.995595
```

```
print(xy)
```

```
##         y
## 0.4997899
```

(c).

```r
set.seed(10)
n  = 100
x2  = rnorm(n)
y2  = x2

yx2 = coef(lm(y2 ~ x2 + 0))
xy2 = coef(lm(x2 ~ y2 + 0))

print(yx2)
```

```
## x2
##  1
```

```r
print(xy2)
```

```
## y2
##  1
```

15.

(a).

```r
str(Boston)
```

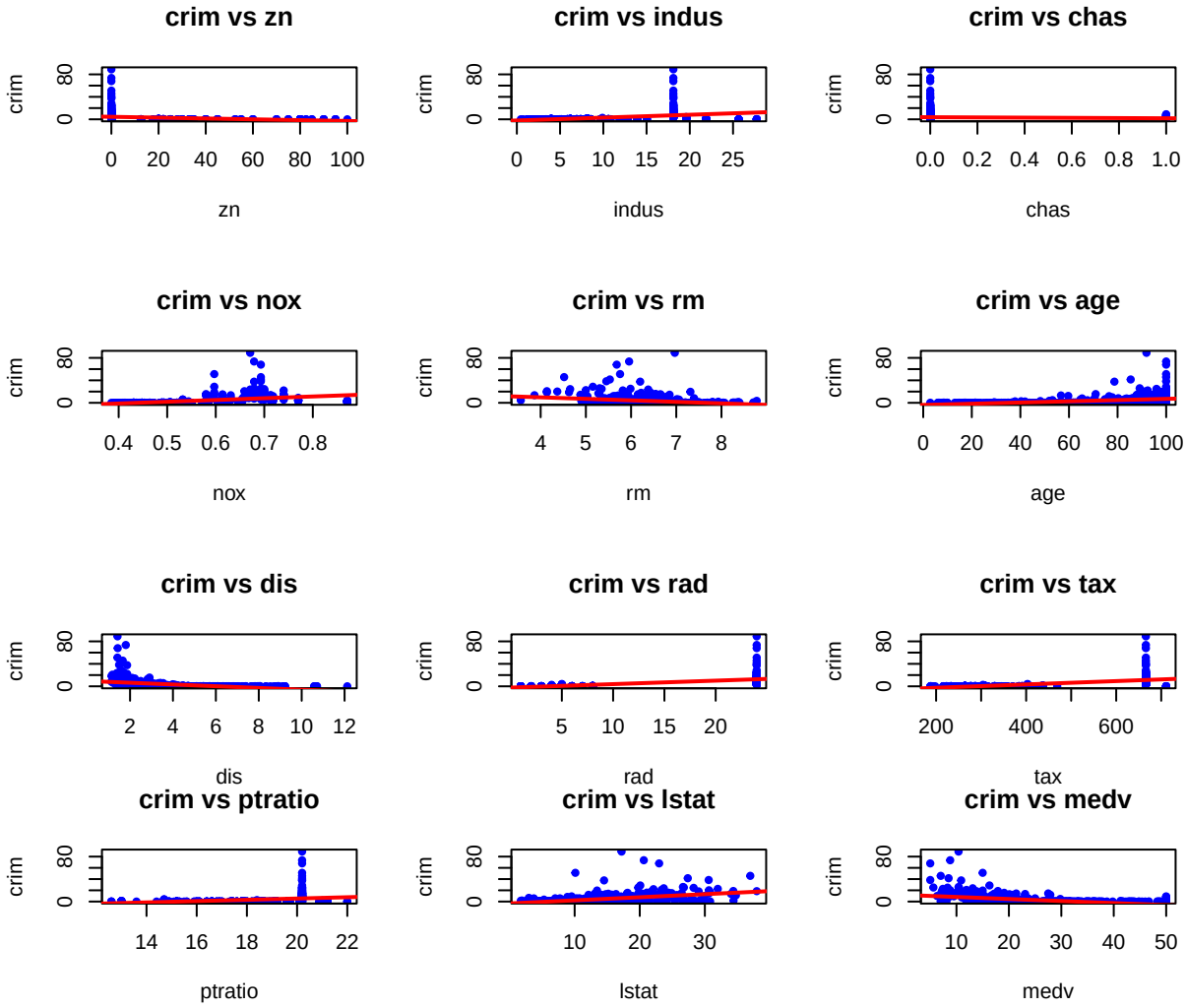
```
## 'data.frame':    506 obs. of  13 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```r
pred = subset(ISLR2::Boston, select = -crim)
boston_model = lapply(pred, function(x) lm(ISLR2::Boston$crim ~ x))
table = do.call(rbind, lapply(boston_model, function(p) coef(summary(p))[2, ]))
colnames(table) = c("Estimate", "Std. Error", "t-value", "p-value")
printCoefmat(table, P.value = TRUE, has.Pvalue = TRUE)
```

```
##           Estimate Std. Error t-value   p-value
## zn      -0.0739350  0.0160946 -4.5938 5.506e-06 ***
## indus    0.5097763  0.0510243  9.9908 < 2.2e-16 ***
## chas    -1.8927766  1.5061155 -1.2567    0.2094
## nox     31.2485312  2.9991904 10.4190 < 2.2e-16 ***
## rm      -2.6840512  0.5320411 -5.0448 6.347e-07 ***
## age      0.1077862  0.0127364  8.4628 2.855e-16 ***
## dis     -1.5509017  0.1683300 -9.2135 < 2.2e-16 ***
## rad      0.6179109  0.0343318 17.9982 < 2.2e-16 ***
## tax      0.0297423  0.0018474 16.0994 < 2.2e-16 ***
## ptratio  1.1519828  0.1693736  6.8014 2.943e-11 ***
## lstat    0.5488048  0.0477610 11.4907 < 2.2e-16 ***
```

```
## medv    -0.3631599  0.0383902 -9.4597 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow = c(3, 3))
for (var in names(ISLR2::Boston)[-which(names(ISLR2::Boston) == "crim")]) {
  fit = lm(crim ~ ISLR2::Boston[[var]], data = ISLR2::Boston)
  plot(ISLR2::Boston[[var]], ISLR2::Boston$crim,
       xlab = var, ylab = "crim",
       main = paste("crim vs", var),
       pch = 20, col = "blue")
  abline(fit, col = "red", lwd = 2)
}
```



Except chas, all other models show a statistically significant association between the predictor and the response as the p-values are less than 0.05. Also we could barely see a trend in the plot of crim vs. chas.

(b).

```r
boston_model2 = lm(data = ISLR2::Boston, crim ~ .)
summary(boston_model2)
```

```
##
## Call:
```

```
## lm(formula = crim ~ ., data = ISLR2::Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.534 -2.248 -0.348  1.087 73.923
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.7783938  7.0818258   1.946 0.052271 .
## zn           0.0457100  0.0187903   2.433 0.015344 *
## indus       -0.0583501  0.0836351  -0.698 0.485709
## chas        -0.8253776  1.1833963  -0.697 0.485841
## nox         -9.9575865  5.2898242  -1.882 0.060370 .
## rm           0.6289107  0.6070924   1.036 0.300738
## age         -0.0008483  0.0179482  -0.047 0.962323
## dis         -1.0122467  0.2824676  -3.584 0.000373 ***
## rad          0.6124653  0.0875358   6.997 8.59e-12 ***
## tax         -0.0037756  0.0051723  -0.730 0.465757
## ptratio     -0.3040728  0.1863598  -1.632 0.103393
## lstat        0.1388006  0.0757213   1.833 0.067398 .
## medv        -0.2200564  0.0598240  -3.678 0.000261 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.46 on 493 degrees of freedom
## Multiple R-squared:  0.4493, Adjusted R-squared:  0.4359
## F-statistic: 33.52 on 12 and 493 DF,  p-value: < 2.2e-16
```

Based on the summary, zn, dis, rad and medv are statistically significant as their p-values are less than 0.05. Therefore, we can reject null hypothesis for these predictors. The R squared of this model is 0.4493, indicating that 44.93% of the variation in crime rate by town can be explained by the predictors. The F-value is 33.52 with a p-value less than 0.05, suggesting the overall model is significant.

(c)

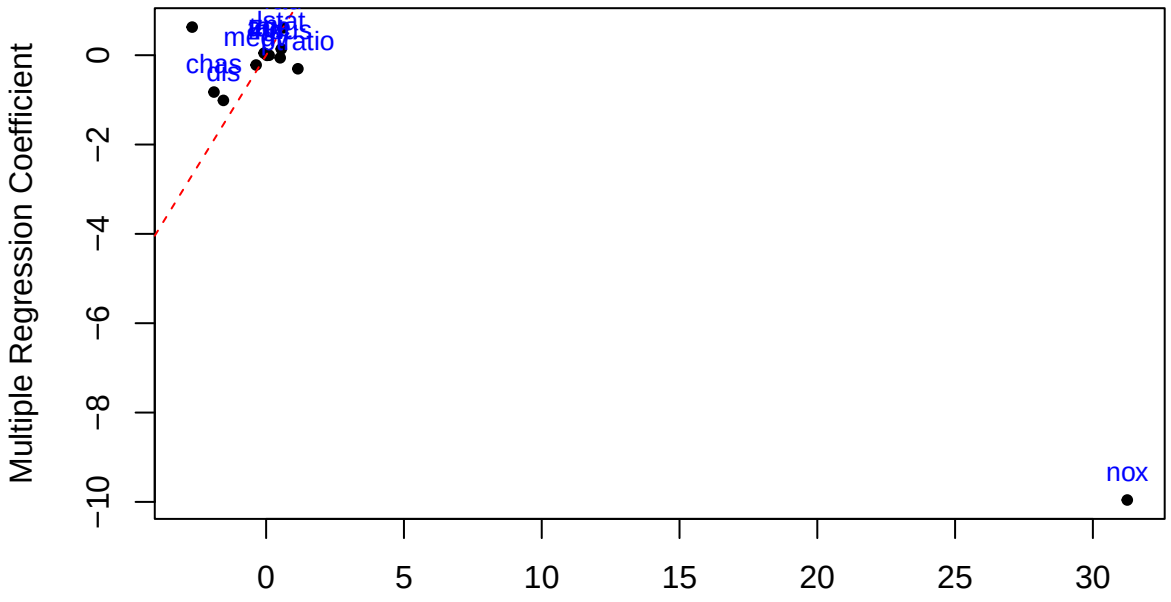
```
pred_order = names(coef(boston_model2))[-1]

uni_slopes = sapply(pred_order, function(v) {
  coef(lm(crim ~ ISLR2::Boston[[v]], data = ISLR2::Boston))[2]
})

multi_slopes = coef(boston_model2)[pred_order]


length(uni_slopes); length(multi_slopes)
```

```
## [1] 12
```

```
## [1] 12
```

```
par(mfrow = c(1, 1))
plot(uni_slopes, multi_slopes,
     xlab = "Univariate Regression Coefficient",
     ylab = "Multiple Regression Coefficient",
     pch = 20)
abline(0, 1, lty = 2, col = "red")
text(uni_slopes, multi_slopes,
```

```
    labels = names(multi_slopes),
    pos = 3,
    cex = 0.8,
    col = "blue")
```



Most predictors that appear significant in simple regressions lose their effect in the multiple regression. Specially nox has a positive relationship with crime rate when doing a simple linear regression but appears to be negatively correlated with crime rate after adding all other predictors in the multiple linear regression.

(d).

```
#To apply quadratic forms, degree must be less than number of unique points, so we will remove chas
preds = subset(pred, select = -chas)
boston_model3 = lapply(names(preds), function(p){
  f = paste0("crim ~ poly(", p,", 3)")
  lm(as.formula(f), data = ISLR2::Boston)
})

for(model in boston_model3)
  printCoefmat(coef(summary(model))
  )
```

```
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    3.61352    0.37219  9.7088 < 2.2e-16 ***
## poly(zn, 3)1 -38.74984    8.37221 -4.6284 4.698e-06 ***
## poly(zn, 3)2  23.93983    8.37221  2.8594  0.004421 **
## poly(zn, 3)3 -10.07187    8.37221 -1.2030  0.229539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       3.6135     0.3300 10.9501 < 2.2e-16 ***
## poly(indus, 3)1  78.5908     7.4231 10.5873 < 2.2e-16 ***
## poly(indus, 3)2 -24.3948     7.4231 -3.2863  0.001086 **
## poly(indus, 3)3 -54.1298     7.4231 -7.2920 1.196e-12 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     3.61352    0.32157 11.2370 < 2.2e-16 ***
## poly(nox, 3)1  81.37202    7.23361 11.2492 < 2.2e-16 ***
## poly(nox, 3)2 -28.82859    7.23361 -3.9854 7.737e-05 ***
## poly(nox, 3)3 -60.36189    7.23361 -8.3446 6.961e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    3.6135     0.3703  9.7584 < 2.2e-16 ***
## poly(rm, 3)1 -42.3794     8.3297 -5.0878 5.128e-07 ***
## poly(rm, 3)2  26.5768     8.3297  3.1906  0.001509 **
## poly(rm, 3)3  -5.5103     8.3297 -0.6615  0.508575
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    3.61352    0.34852 10.3683 < 2.2e-16 ***
## poly(age, 3)1 68.18201    7.83970  8.6970 < 2.2e-16 ***
## poly(age, 3)2 37.48447    7.83970  4.7814 2.291e-06 ***
## poly(age, 3)3 21.35321    7.83970  2.7237   0.00668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)     3.61352    0.32592  11.0870 < 2.2e-16 ***
## poly(dis, 3)1 -73.38859    7.33148 -10.0101 < 2.2e-16 ***
## poly(dis, 3)2  56.37304    7.33148   7.6892 7.870e-14 ***
## poly(dis, 3)3 -42.62188    7.33148  -5.8135 1.089e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     3.61352    0.29707 12.1639 < 2.2e-16 ***
## poly(rad, 3)1 120.90745    6.68240 18.0934 < 2.2e-16 ***
## poly(rad, 3)2  17.49230    6.68240  2.6177  0.009121 **
## poly(rad, 3)3   4.69846    6.68240  0.7031  0.482314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     3.61352    0.30468 11.8599 < 2.2e-16 ***
## poly(tax, 3)1 112.64583    6.85371 16.4358 < 2.2e-16 ***
## poly(tax, 3)2  32.08725    6.85371  4.6817 3.665e-06 ***
## poly(tax, 3)3  -7.99681    6.85371 -1.1668    0.2439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)         3.61352    0.36105 10.0084 < 2.2e-16 ***
## poly(ptratio, 3)1  56.04523    8.12158  6.9008 1.565e-11 ***
## poly(ptratio, 3)2  24.77482    8.12158  3.0505  0.002405 **
## poly(ptratio, 3)3 -22.27974    8.12158 -2.7433  0.006301 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.61352    0.33917 10.6540   <2e-16 ***
## poly(lstat, 3)1  88.06967    7.62944 11.5434   <2e-16 ***
```

```
## poly(lstat, 3)2  15.88816    7.62944  2.0825   0.0378 *
## poly(lstat, 3)3 -11.57402    7.62944 -1.5170   0.1299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      3.61352    0.29203  12.374 < 2.2e-16 ***
## poly(medv, 3)1 -75.05761    6.56915 -11.426 < 2.2e-16 ***
## poly(medv, 3)2  88.08621    6.56915  13.409 < 2.2e-16 ***
## poly(medv, 3)3 -48.03343    6.56915  -7.312 1.047e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the output, there is strong evidence for many variables having non-linear relationships. The cubic term is significant for predictors such as indus, nox, age, dis, ptratio, medv. For zn, rm, rad, lstat, tax, even though the cubic term is not significant, the squared terms have a p-value less than 0.05.