IMT 573: Problem Set 6 - Regression

LEE CHEN HSIN

Due: Tuesday, November 16, 2021

Collaborators:

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset5.Rmd` file from Canvas. Open `problemset5.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset5.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licenses as Creative Commons (CC-BY-SA). This means you have to attribute any code you refer from SO.
4. Partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. But please **DO NOT** submit pages and pages of hard-to-read code and attempts that is impossible to grade. That is, avoid redundancy. Remember that one of the key goals of a data scientist is to produce coherent reports that others can easily follow. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object dont' exist
# if you run this on its own it with give an error
```

6. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the knitted PDF file to `ps6_ourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.
7. Collaboration is often fun and useful, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.

Setup In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(MASS) # Modern applied statistics functions
```

Housing Values in Suburbs of Boston

In this problem we will use the Boston dataset that is available in the `MASS` package. This dataset contains information about median house value for 506 neighborhoods in Boston, MA. Load this data and use it to answer the following questions.

```
data(Boston)
```

1. Describe the data and variables that are part of the `Boston` dataset. Tidy data as necessary.

```
summary(Boston)
```

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

```
#crim means per capital crime rate by town.
#zn means proportion of residential land zoned for lots over 25,000 sq.ft.
#indus means proportion of non-retail business acres per town.
#chas means Charles River dummy variable.
#nox means nitrogen oxides concentration.
#rm means average number of rooms per dwelling.
#age means proportion of owner-occupied units built prior to 1940.
#dis means weighted mean of distances to five Boston employment centres.
#rad means index of accessibility to radial highways.
#tax means full-value property-tax rate per $10,000.
#ptratio means pupil-teacher ratio by town.
#black means where BkBk is the proportion of blacks by town.
#lstat means lower status of the population.
#medv means median value of owner-occupied homes.

#There is no need for tidying data since there are no NA, missing values or
#weird mean or median values in each variable.
```

2. Consider this data in context, what is the response variable of interest?

```
#MEDV - Median value of owner is the response variable of interest
```

3. For each predictor, fit a simple linear regression model to predict the response. In which of the models is there a statistically significant association between the predictor and the response? Describe your results.

```
f1=lm(formula=medv~crim,data=Boston)
summary(f1)
```

```
##
## Call:
## lm(formula = medv ~ crim, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.957  -5.449  -2.007   2.512  29.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.03311    0.40914   58.74   <2e-16 ***
## crim        -0.41519    0.04389   -9.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.484 on 504 degrees of freedom
## Multiple R-squared:  0.1508, Adjusted R-squared:  0.1491
## F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
cor(Boston$crim,Boston$medv,method = c("pearson"))
```

```
## [1] -0.3883046
```

```
f2=lm(formula=medv~zn,data=Boston)
summary(f2)
```

```
##
## Call:
## lm(formula = medv ~ zn, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.918  -5.518  -1.006   2.757  29.082
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.91758    0.42474  49.248   <2e-16 ***
## zn           0.14214    0.01638   8.675   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.587 on 504 degrees of freedom
## Multiple R-squared:  0.1299, Adjusted R-squared:  0.1282
## F-statistic: 75.26 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
cor(Boston$zn,Boston$medv,method = c("pearson"))
```

```
## [1] 0.3604453
```

```
f3=lm(formula=medv~indus,data=Boston)
summary(f3)
```

```
##
## Call:
## lm(formula = medv ~ indus, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.017  -4.917  -1.457   3.180  32.943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.75490    0.68345   43.54   <2e-16 ***
## indus       -0.64849    0.05226  -12.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.057 on 504 degrees of freedom
## Multiple R-squared:  0.234,  Adjusted R-squared:  0.2325
## F-statistic:   154 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
cor(Boston$indus,Boston$medv,method = c("pearson"))
```

```
## [1] -0.4837252
```

```
f4=lm(formula=medv~chas,data=Boston)
summary(f4)
```

```
##
## Call:
## lm(formula = medv ~ chas, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.094  -5.894  -1.417   2.856  27.906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.0938     0.4176  52.902  < 2e-16 ***
## chas          6.3462     1.5880   3.996 7.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.064 on 504 degrees of freedom
## Multiple R-squared:  0.03072,    Adjusted R-squared:  0.02879
## F-statistic: 15.97 on 1 and 504 DF,  p-value: 7.391e-05
```

```
cor(Boston$chas,Boston$medv,method = c("pearson"))
```

```
## [1] 0.1752602
```

```
f5=lm(formula=medv~nox,data=Boston)
summary(f5)
```

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

```
## lm(formula = medv ~ nox, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.691  -5.121  -2.161   2.959  31.310
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   41.346      1.811   22.83   <2e-16 ***
## nox          -33.916      3.196  -10.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.323 on 504 degrees of freedom
## Multiple R-squared:  0.1826, Adjusted R-squared:  0.181
## F-statistic: 112.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
cor(Boston$nox,Boston$medv,method = c("pearson"))
```

```
## [1] -0.4273208
```

```
f6=lm(formula=medv~rm,data=Boston)
summary(f6)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
cor(Boston$rm,Boston$medv,method = c("pearson"))
```

```
## [1] 0.6953599
```

```
f7=lm(formula=medv~age,data=Boston)
summary(f7)
```

```
##
## Call:
## lm(formula = medv ~ age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.097  -5.138  -1.958   2.397  31.338
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.97868    0.99911  31.006   <2e-16 ***
## age         -0.12316    0.01348  -9.137   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.527 on 504 degrees of freedom
## Multiple R-squared:  0.1421, Adjusted R-squared:  0.1404
## F-statistic: 83.48 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
cor(Boston$age,Boston$medv,method = c("pearson"))
```

```
## [1] -0.3769546
```

```
f8=lm(formula=medv~dis,data=Boston)
summary(f8)
```

```
##
## Call:
## lm(formula = medv ~ dis, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.016  -5.556  -1.865   2.288  30.377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.3901     0.8174  22.499  < 2e-16 ***
## dis           1.0916     0.1884   5.795 1.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 504 degrees of freedom
## Multiple R-squared:  0.06246,    Adjusted R-squared:  0.0606
## F-statistic: 33.58 on 1 and 504 DF,  p-value: 1.207e-08
```

```
cor(Boston$dis,Boston$medv,method = c("pearson"))
```

```
## [1] 0.2499287
```

```
f9=lm(formula=medv~rad,data=Boston)
summary(f9)
```

```
##
## Call:
## lm(formula = medv ~ rad, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.770  -5.199  -1.967   3.321  33.292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.38213    0.56176  46.964   <2e-16 ***
## rad         -0.40310    0.04349  -9.269   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.509 on 504 degrees of freedom
## Multiple R-squared:  0.1456, Adjusted R-squared:  0.1439
## F-statistic: 85.91 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
cor(Boston$rad,Boston$medv,method = c("pearson"))
```

```
## [1] -0.3816262
```

```
f10=lm(formula=medv~tax,data=Boston)
summary(f10)
```

```
##
## Call:
## lm(formula = medv ~ tax, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.091  -5.173  -2.085   3.158  34.058
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.970654   0.948296   34.77   <2e-16 ***
## tax         -0.025568   0.002147  -11.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.133 on 504 degrees of freedom
## Multiple R-squared:  0.2195, Adjusted R-squared:  0.218
## F-statistic: 141.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
cor(Boston$tax,Boston$medv,method = c("pearson"))
```

```
## [1] -0.4685359
```

```
f11=lm(formula=medv~ptratio,data=Boston)
summary(f11)
```

```
##
## Call:
## lm(formula = medv ~ ptratio, data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8342  -4.8262  -0.6426   3.1571  31.2303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.345      3.029   20.58   <2e-16 ***
## ptratio       -2.157      0.163  -13.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.931 on 504 degrees of freedom
## Multiple R-squared:  0.2578, Adjusted R-squared:  0.2564
```

```
## F-statistic: 175.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
cor(Boston$ptratio,Boston$medv,method = c("pearson"))
```

```
## [1] -0.5077867
```

```
f12=lm(formula=medv~black,data=Boston)
summary(f12)
```

```
##
## Call:
## lm(formula = medv ~ black, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.884  -4.862  -1.684   2.932  27.763
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.551034   1.557463   6.775 3.49e-11 ***
## black        0.033593   0.004231   7.941 1.32e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.679 on 504 degrees of freedom
## Multiple R-squared:  0.1112, Adjusted R-squared:  0.1094
## F-statistic: 63.05 on 1 and 504 DF,  p-value: 1.318e-14
```

```
cor(Boston$black,Boston$medv,method = c("pearson"))
```

```
## [1] 0.3334608
```

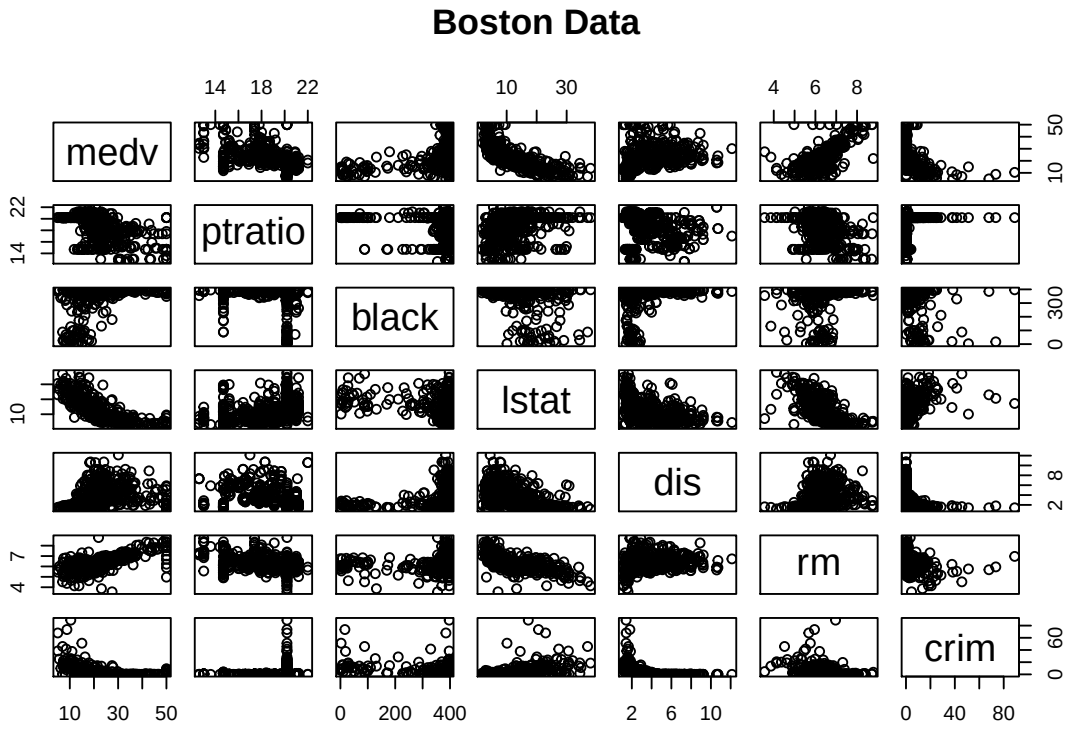
```
f13=lm(formula=medv~lstat,data=Boston)
summary(f13)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
cor(Boston$lstat,Boston$medv,method = c("pearson"))
```

```
## [1] -0.7376627
```

```
pairs(~ medv + ptratio + black + lstat + dis + rm + crim, data = Boston, main = "Boston Data")
```




```
#lstat, rm and ptratio have a a statistically significant association
```

4. Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

```
model <- lm(medv ~ crim+ zn + indus+chas+nox+rm+age+dis+rad+tax+ptratio+black+lstat, data = Boston)
summary(model)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + indus + chas + nox + rm + age +
##     dis + rad + tax + ptratio + black + lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
```

```
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

```
#crim,zn,chas,nox,rm,dis,rad,tax,ptratio,black,lstat of p-value<0.05
#can reject the null hypothesis H0
```

5. How do your results from (3) compare to your results from (4)? *Hint: You need to compare the coefficients across the two models and report on all the changes that you observe and the reason why.*

```
#Take p-value as an example to illustrate the Estimate.

#For my result from(4), a multiple regression model,the p-value of crim is
#0.001087 and for my result from(3), a simple repression model, the p-value of
#crim is 2.2e-16, has been increased

#For my result from(4), a multiple regression model,the p-value of zn is
#0.000778 and for my result from(3), a simple repression model, the p-value of
#zn is 2.2e-16, has been increased


#For my result from(4), a multiple regression model,the p-value of indus is
#0.738288 and for my result from(3), a simple repression model, the p-value of
#zn is 2.2e-16, has been increased


#For my result from(4), a multiple regression model,the p-value of chas is
#0.001925 and for my result from(3), a simple repression model, the p-value of
#zn is 7.391e-05, has been increased


#For my result from(4), a multiple regression model,the p-value of nox is
#4.25e-06 and for my result from(3), a simple repression model, the p-value of
#zn is 2.2e-16, has been increased


#For my result from(4), a multiple regression model,the p-value of rm is
#2e-16 and for my result from(3), a simple repression model, the p-value of
#zn is 2.2e-16, has been decreased


#For my result from(4), a multiple regression model,the p-value of age is
#0.958229 and for my result from(3), a simple repression model, the p-value of
#zn is 2.2e-16, has been increased
```

```
#For my result from(4), a multiple regression model,the p-value of dis is
#6.01e-13 and for my result from(3), a simple repression model, the p-value of
#zn is 1.207e-08, has been increased


#For my result from(4), a multiple regression model,the p-value of rad is
#5.07e-06 and for my result from(3), a simple repression model, the p-value of
#zn is 2.2e-16, has been increased


#For my result from(4), a multiple regression model,the p-value of tax is
#0.001112 and for my result from(3), a simple repression model, the p-value of
#zn is 2.2e-16, has been increased


#For my result from(4), a multiple regression model,the p-value of ptratio is
#1.31e-12 and for my result from(3), a simple repression model, the p-value of
#zn is 2.2e-16, has been decreased


#For my result from(4), a multiple regression model,the p-value of black is
#0.000573 and for my result from(3), a simple repression model, the p-value of
#zn is 1.318e-14, has been increased


#For my result from(4), a multiple regression model,the p-value of lstat is
#2e-16 and for my result from(3), a simple repression model, the p-value of
#zn is 2.2e-16, has been decreased

#This is because in multiple regression model, variables are not independent so
#they may be affected by each other and lead to the change of p-value
```