Assignment 3

```
library(ggplot2)
```

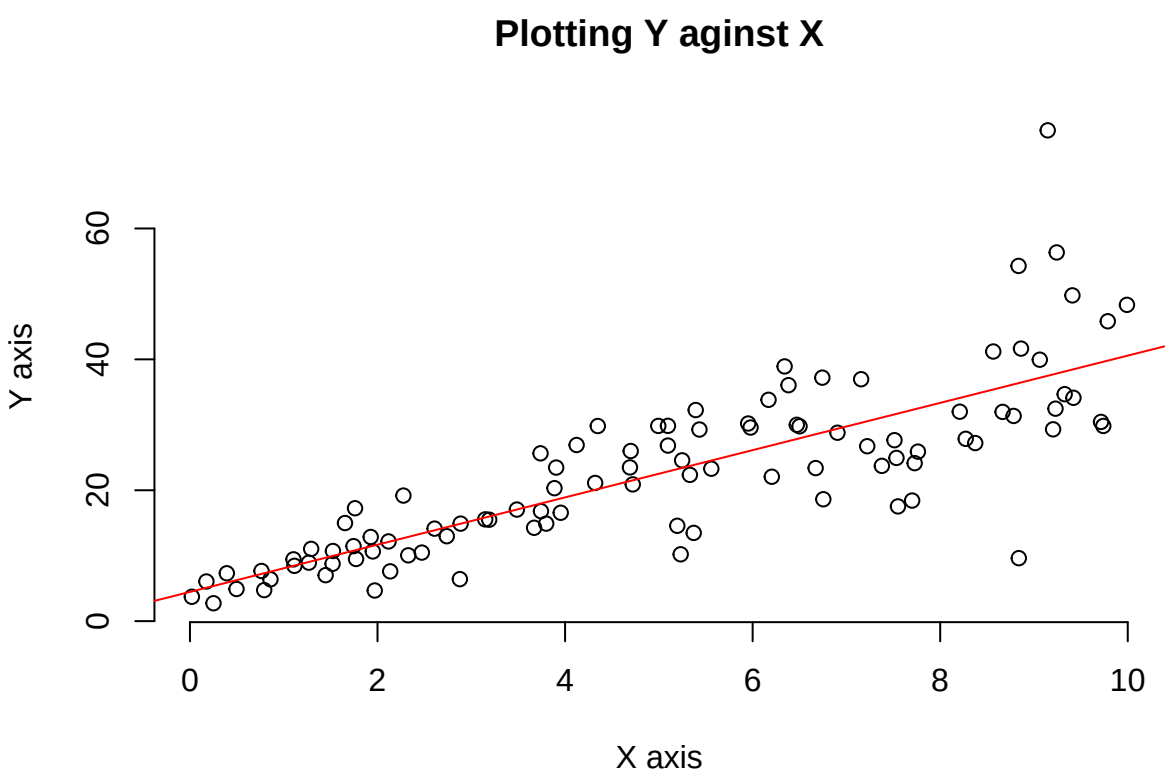
```
# Run the following code in R-studio to create two variables X and Y.

set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y
```

a)Plot Y against X.Based on the plot do you think we can fit a linear model to explain Y based on X?

```
# Plot Y against X.

plot(X,Y, main = "Plotting Y aginst X", xlab = "X axis", ylab = "Y axis",frame = FALSE)
# Add regression line
abline(lm(Y~X), col = "red")
```



Interpretation:

We can see there exists correlation between the variables "x" and "y" from the plot. Hence linear model would be a good fit.

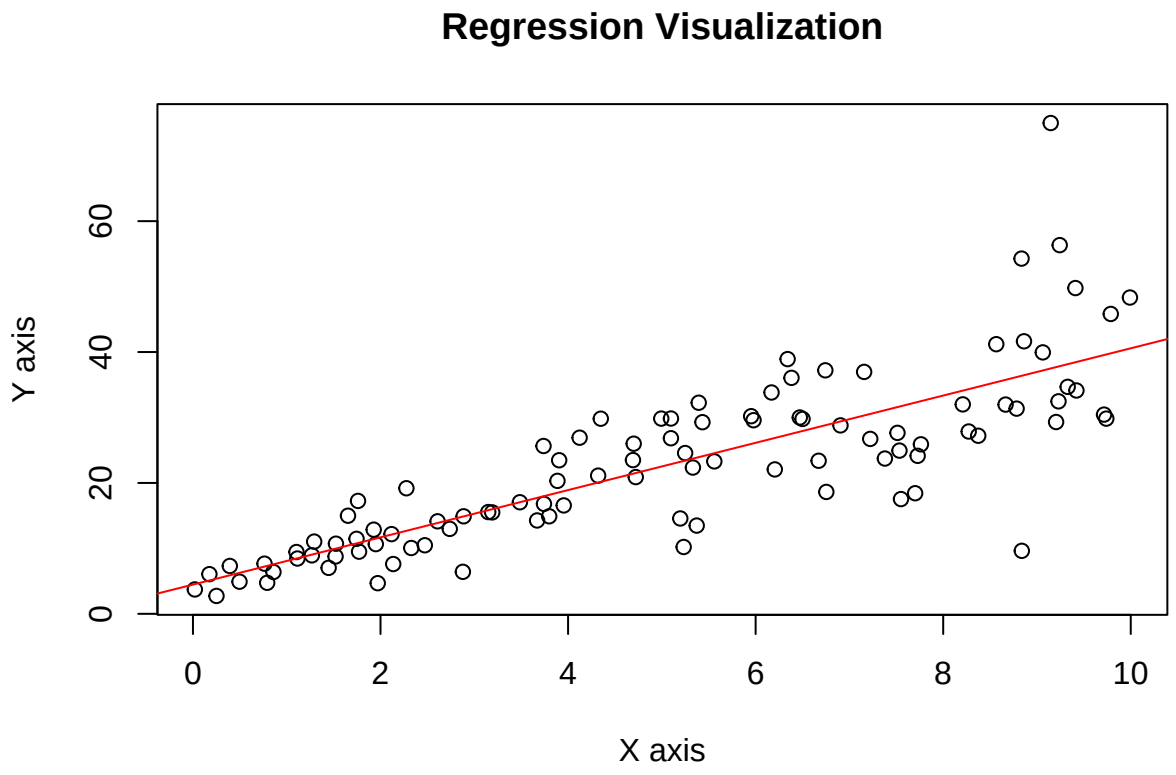
b)Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What is the accuracy of this model?

```
# Y=4.4655+3.6108*X
# Accuracy is 0.6517 or 65%
linear_mod <- lm(Y~X)
summary(linear_mod)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -26.755  -3.846  -0.387   4.318  37.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4655     1.5537   2.874  0.00497 **
## X             3.6108     0.2666  13.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.756 on 98 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6482
## F-statistic: 183.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
# Regression visualization
plot(X, Y, xlab = "X axis",
     ylab = "Y axis",
     main = "Regression Visualization")
abline(4.4655, 3.6108, col = "red")
```



c)How the Coefficient of Determination, R2, of the model above is related to the correlation coefficient of X and Y?

```r
cor(X,Y)^2
```

```
## [1] 0.6517187
```

Interpretation:

The coefficient of determination, R^2, of the above model is similar to the coefficient of X and Y, which is around 65%

Question 2:

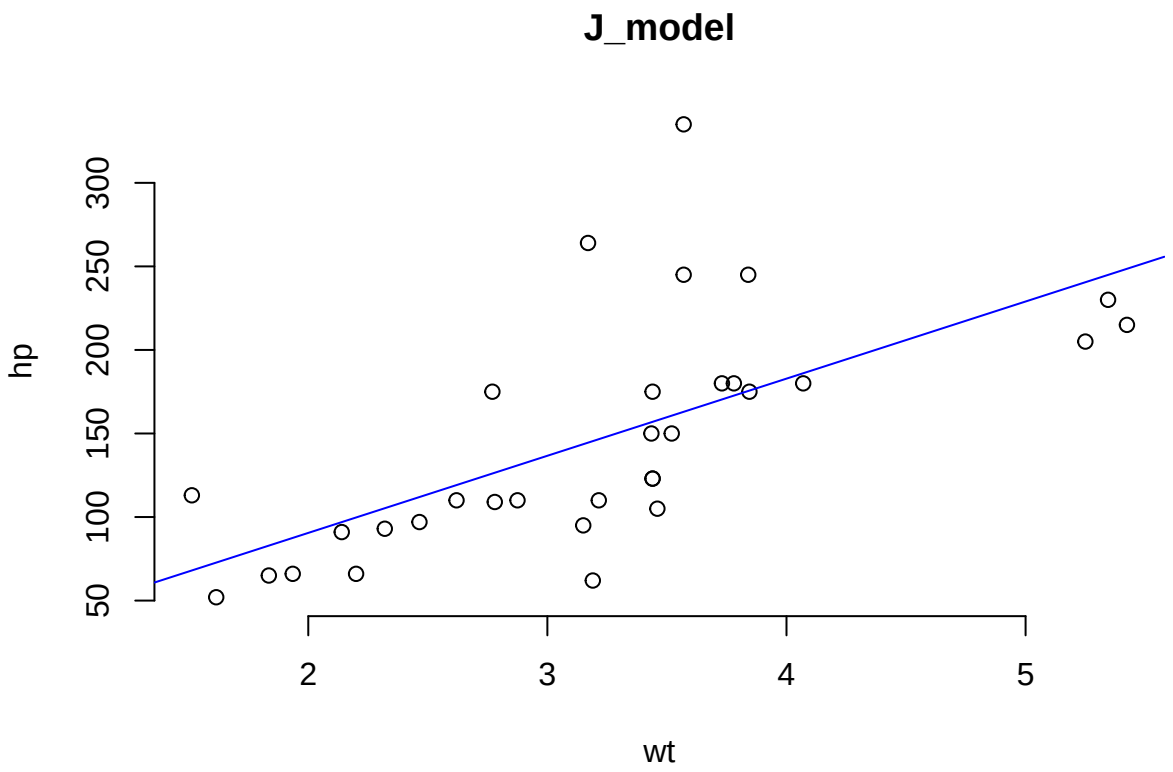
a)James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question.

```r
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Simple Linear model according to James

```r
plot(mtcars$wt, mtcars$hp, main = "J_model",
     xlab = "wt", ylab = "hp",frame = FALSE)
# Add regression line
abline(lm(mtcars$hp ~ mtcars$wt), col = "blue")
```



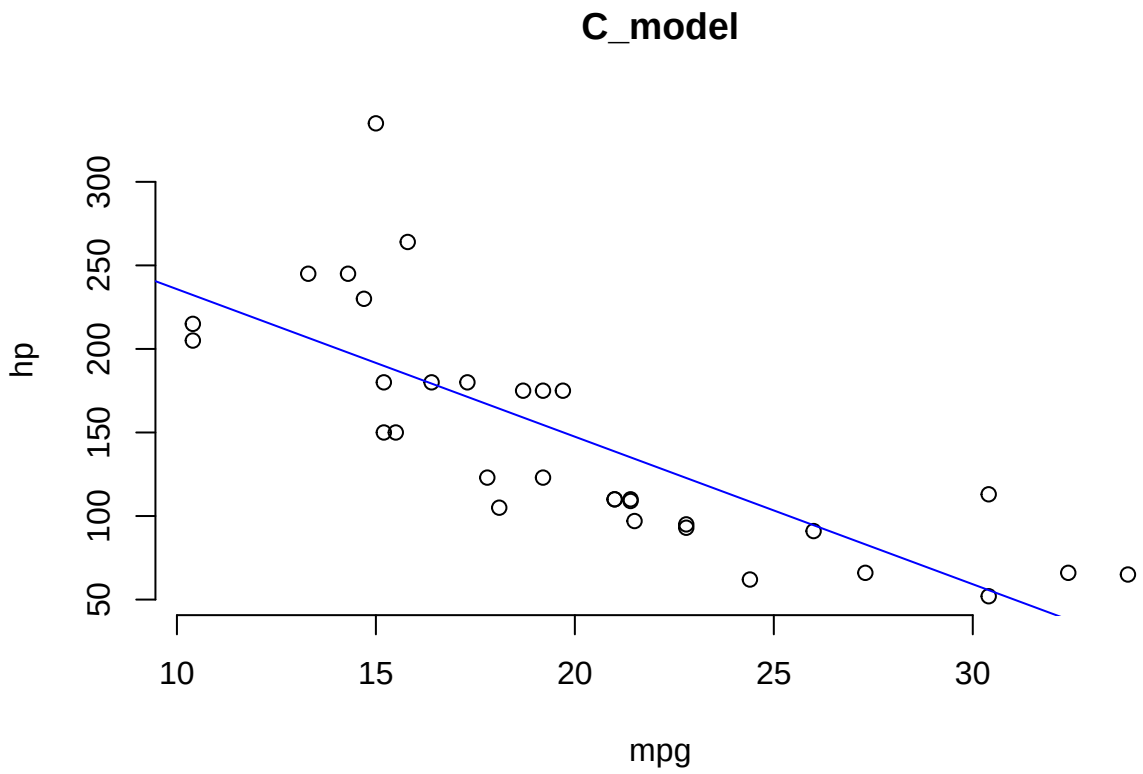

```
J_model <- lm(formula = hp~wt, data = mtcars)
summary(J_model)
```

```
##
## Call:
## lm(formula = hp ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.430 -33.596 -13.587   7.913 172.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.821     32.325  -0.056    0.955
## wt            46.160      9.625   4.796 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.44 on 30 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4151
## F-statistic:    23 on 1 and 30 DF,  p-value: 4.146e-05
```

```
# Accuracy of J_model is 0.4339
```

Simple Linear model according to Chris

```r
plot(mtcars$mpg, mtcars$hp, main = "C_model",
     xlab = "mpg", ylab = "hp",frame = FALSE)
# Add regression line
abline(lm(mtcars$hp ~ mtcars$mpg), col = "blue")
```




```r
C_model <- lm(formula = hp~mpg, data = mtcars)
summary(C_model)
```

```
##
## Call:
## lm(formula = hp ~ mpg, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.26 -28.93 -13.45  25.65 143.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   324.08      27.43  11.813 8.25e-13 ***
## mpg            -8.83       1.31  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.95 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
```

```
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

```
# Accuracy of C_model is 0.6024
```

Interpretation:

We can see the value of C_model is fairly accurate. so,Chris' opinion is right (i.e)the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp).

b)Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp). Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22?

```
H_model <- lm(formula = hp~cyl+mpg, data = mtcars)
summary(H_model)
```

```
##
## Call:
## lm(formula = hp ~ cyl + mpg, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.72 -22.18 -10.13  14.47 130.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.067     86.093   0.628  0.53492
## cyl           23.979      7.346   3.264  0.00281 **
## mpg           -2.775      2.177  -1.275  0.21253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.22 on 29 degrees of freedom
## Multiple R-squared:  0.7093, Adjusted R-squared:  0.6892
## F-statistic: 35.37 on 2 and 29 DF,  p-value: 1.663e-08
```

```
Esti_hp <- predict(H_model,data.frame(cyl=4,mpg=22))
Esti_hp
```

```
##        1
## 88.93618
```

Interpretation:

The estimated Horse Power of a car with cyl = 4 and mpg = 22 is 88.93

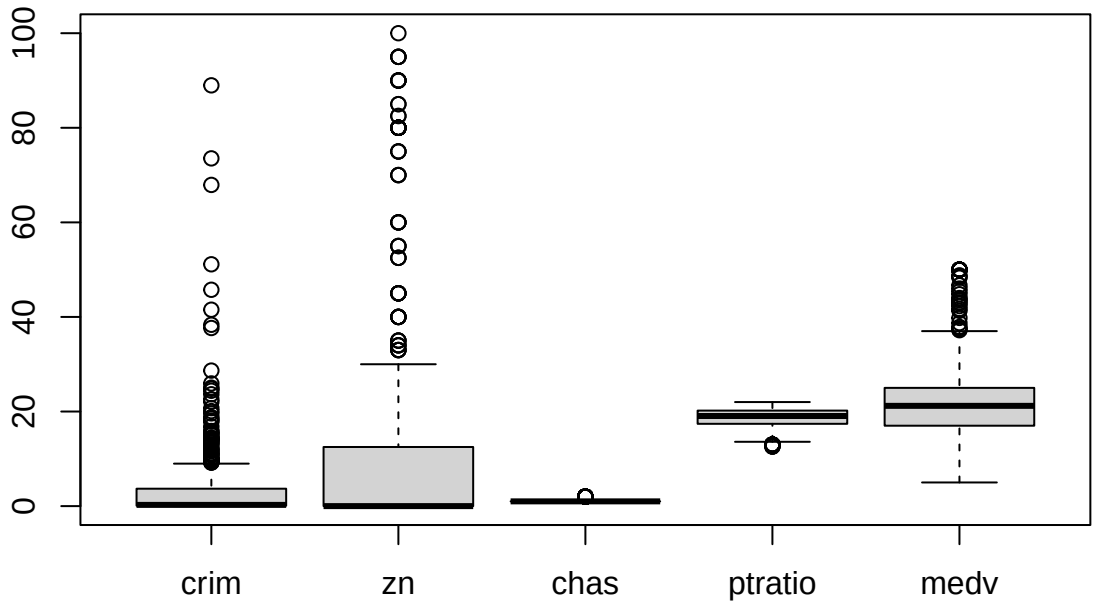
Question 3

```
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.2.2
```

```
data("BostonHousing")
View(BostonHousing)

# Plotting all the variable using box plot to observe how the values of the various variables in the da
boxplot(BostonHousing[,c(1,2,4,11,14)])
```



a) Build a model to estimate the median value of owner-occupied homes(medv) based on the following variables:crime crate (crim),proportion of residential land zoned for lots over 25,000 sq.ft(zn), the local pupil-teacher ratio(ptratio) and weather the whether the tract bounds Chas River(chas). Is this an accurate model?(Hint check R2)

```
set.seed(125)
owner_model <- lm(formula = medv~crim+zn+ptratio+chas,data = BostonHousing)
summary(owner_model)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + ptratio + chas, data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.91868    3.23497  15.431  < 2e-16 ***
## crim        -0.26018    0.04015  -6.480 2.20e-10 ***
## zn           0.07073    0.01548   4.570 6.14e-06 ***
## ptratio     -1.49367    0.17144  -8.712  < 2e-16 ***
## chas1        4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
```

```
owner_model
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + ptratio + chas, data = BostonHousing)
##
## Coefficients:
## (Intercept)         crim           zn      ptratio        chas1
##    49.91868     -0.26018      0.07073     -1.49367      4.58393
```

Interpretation:

The accuracy of owner_model is 0.3599, which means the model is not accurate enough.

b) Use the estimated coefficient to answer these questions?

I. Imagine two houses that are identical in all aspects but one bounds the Chas River and the other does not. Which one is more expensive and by how much?

Answer:

The estimated coefficient of chas1 is 4.58393. chas is the factor of two variable 0 and 1, one bound Chas River is 1 and if it doesn't it is 0. It is given that the median value of owner-occupied homes is 1000 dollars. when multiplied with coefficient(4.58393*1000), the result is 4583.93$ which is expensive.

II. Imagine two houses that are identical in all aspects but in the neighborhood of one of them the pupil-teacher ratio is 15 and in the other one is 18. Which one is more expensive and by how much? (Golden Question)

Answer:

It is clear that for every single unit increase in ptratio, price of houses is decreased by 1.49367 (i.e) 1493.67 (in thousands).If ptratio is 15, then it will be decrease of 15 * 1493.67= 22405.05. Likely, if ptratio is 18 then it will be a decrease of18*1493.67 = 26886.06. Finally, if ptratio of 15 expensive by $4481.01 when comapred to ptratio of 18.

c) Which of the variables are statistically important (i.e. related to the house price)? Hint: use the p-values of the coefficients to answer.

Answer:

The P-values are not equal to 0. so, we can reject the null hypothesis and conclude that there is no relationship between house price and other factors in the model. Hence, each variable has statistical significance.

d) Use the anova analysis and determine the order of importance of these four variables.

```
anova(owner_model)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## crim        1  6440.8  6440.8 118.007 < 2.2e-16 ***
## zn          1  3554.3  3554.3  65.122 5.253e-15 ***
```

```
## ptratio     1  4709.5  4709.5  86.287 < 2.2e-16 ***
## chas        1   667.2   667.2  12.224 0.0005137 ***
## Residuals 501 27344.5    54.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

As we can see, the crim variable explains substantially more variability(sum squared) than the other variables. This could be explained by the model being greatly enhanced by the addition of the crim. However, residuals demonstrate that a significant fraction of the variability is unaccounted for.

The order of importance is crim,ptratio,zn,chas