REGRESSION-ASSIGNMENT-BA-03

ANURODH-SINGH

2023-11-03

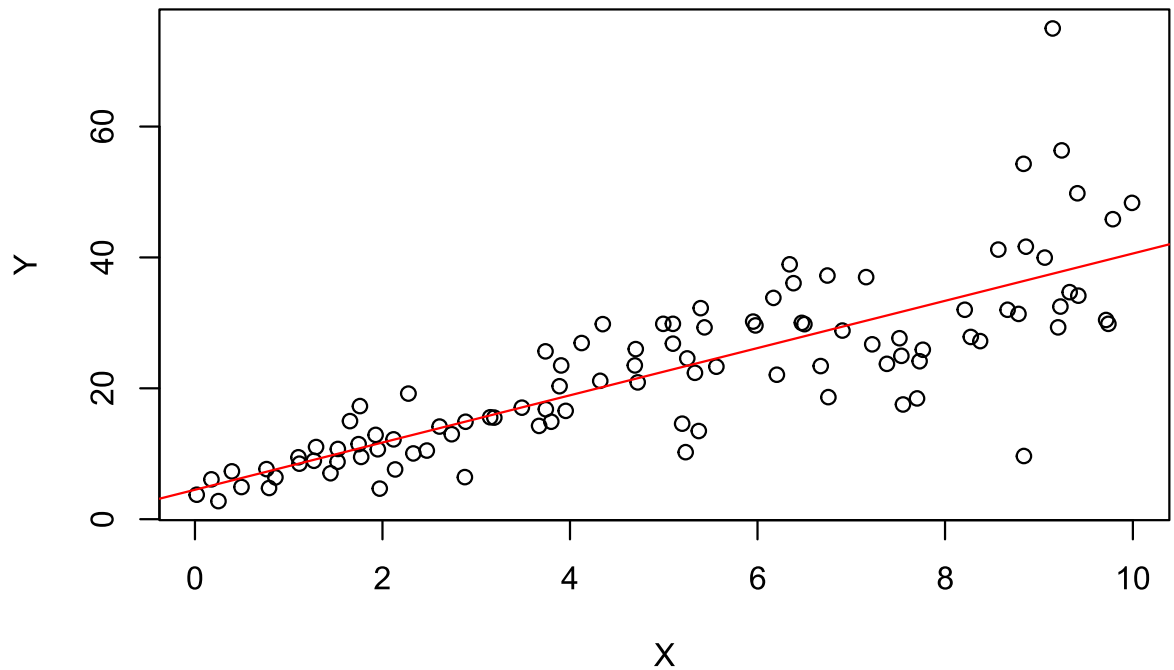
#Summary

- A strong correlation between the variables "x" and "y" indicates a linear relationship using the equation Y=X*4+3.45, providing the potential for precise predictions and valuable insights, particularly in domains such as finance and economics.
- An R^2 value of 0.6517 obtained using using the 'corr' function suggests that the model explains approximately 65.17% of the variation in "Y," indicating a moderately accurate fit for the analysis of the dataset.
- Now using the 'mtcars' dataset as reference Chris's model(model2), which examines the relationship between horsepower and miles per gallon (mpg), outperforms James's model, which relates horsepower to vehicle weight(model1), Chris's model boasts a higher R-squared value (60.24% versus 43.39%), along with a significant F-statistic and p-value, offering valuable insights into the dynamics of fuel efficiency and horsepower.
- Using this model the estimated Horse Power of a car with 4 cylinders and mpg of 22 is 88.93618 horsepower.
- All predictors are statistically significant, as their p-values are less than 0.05. This underscores the importance of these factors in fostering a nuanced understanding of property valuation.
- Through ANOVA analysis, it becomes evident that crime holds a dominant position, exhibiting the highest F-value. This insight guides resource allocation for precise analyses across various fields.
- Notably, the proximity of properties to rivers emerges as the least significant variable, underscoring the disparities in variable importance. This finding paves the way for a more focused analysis and strategic decision-making.
- We now used BostonHousing dataset. The dataset is in 'mlbench' package,so we first installed the package, called the library and the loaded the dataset.
- A thorough understanding of these significant variables empowers stakeholders, enabling informed decision-making and strategic investments in the constantly evolving real estate landscape.

```r
library(ggplot2)
```

```r
#Q1. #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
```

```r
#Q1.a. Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu youca
plot(Y~X)
abline(lm(Y~X), col = "red")
```



We can see there exists correlation between the variables "x" and "y" from the plot.A best fit line can also be seen hence linear model would be a good fit.

```
#Q1.b.Construct a simple linear model of Y based on X.
model <- lm(Y~X)

#Writing the equation that explains Y based on X
cat("Y= 3.6108X + 4.4655")
```

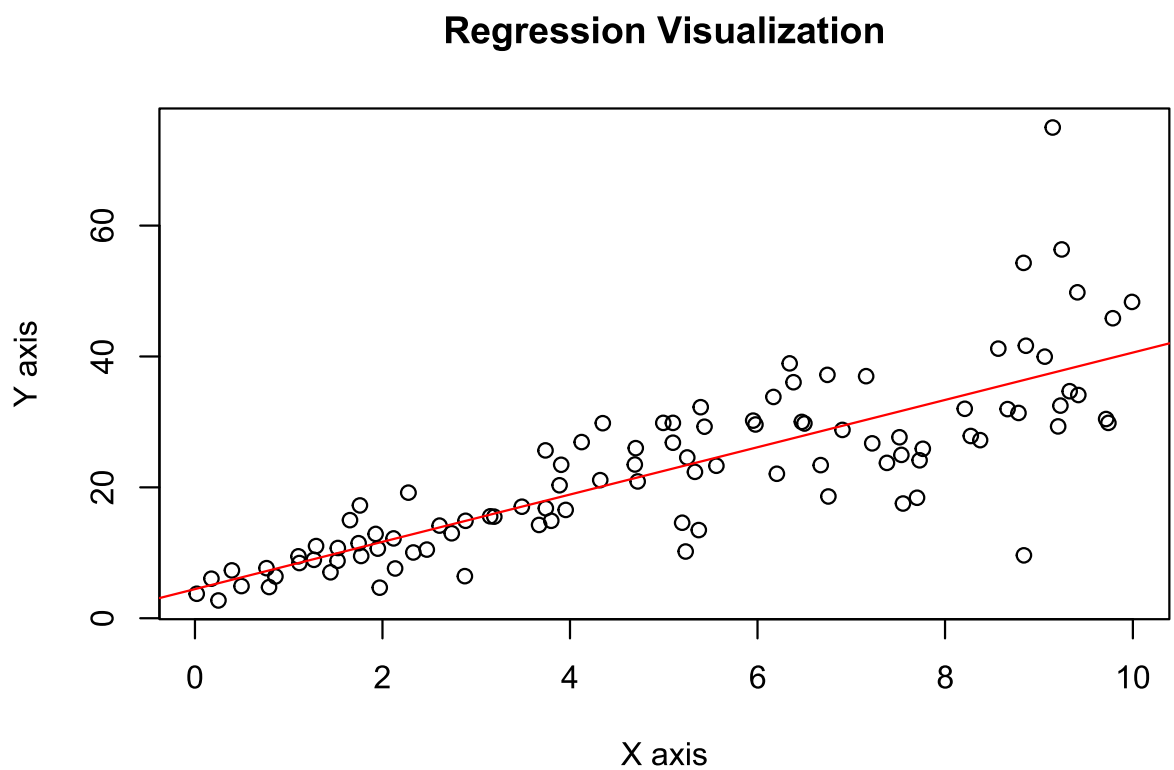
```
## Y= 3.6108X + 4.4655
```

```
#.What is the accuracy of this model?
summary(model)
```

```
##
## 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")
```



The accuracy of this model is represented by the (R^2) value, which is 0.6517 that is 65.17% variability with the dependent variable "Y". In this case, the model explains approximately 65.17% of the variance, suggesting a moderately accurate fit to the data. The F-statistic is 183.4, with a very low p-value (< 2.2e-16). If the p-value remains less than 0.05, it indicates that the independent variables chosen are very prominent and strongly establish the significant relationship with the dependent variable. The F-statistic tests the overall significance of the model. In this case, the model is highly significant, indicating that at least one of the predictors in the model is associated with the response variable.

Q1.c.How the Coefficient of Determination, 'R^2'of the model above is related to the correlation coefficient of X and Y?

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

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

Ans.The (R^2) value of 0.6517 indicates that approximately 65.17% of the variability in the dependent variable "Y" is explained by the independent variable "X." R^2 measures how much of the variance in the dependent variable is accounted for by the independent variable. In this context, the model provides a moderately accurate fit to the data by explaining around 65.17% of the variance in "Y."

```
#Q.2.We will use the 'mtcars' dataset for this question. The dataset is already included in your R dist
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
```

Q2.a.James wants to buy a car. He and his friend, Chris, have different opinions about the HorsePower (hp) of cars. James think the weight of a car (wt) can be used to estimate the HorsePower 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.

```
P1 <- data.frame(hp = mtcars$hp, wt = mtcars$wt, mpg = mtcars$mpg)
head(P1)
```

```
##    hp    wt  mpg
## 1 110 2.620 21.0
## 2 110 2.875 21.0
## 3  93 2.320 22.8
## 4 110 3.215 21.4
## 5 175 3.440 18.7
## 6 105 3.460 18.1
```

```
cat("when horsepower compared with weight ")
```

```
## when horsepower compared with weight
```

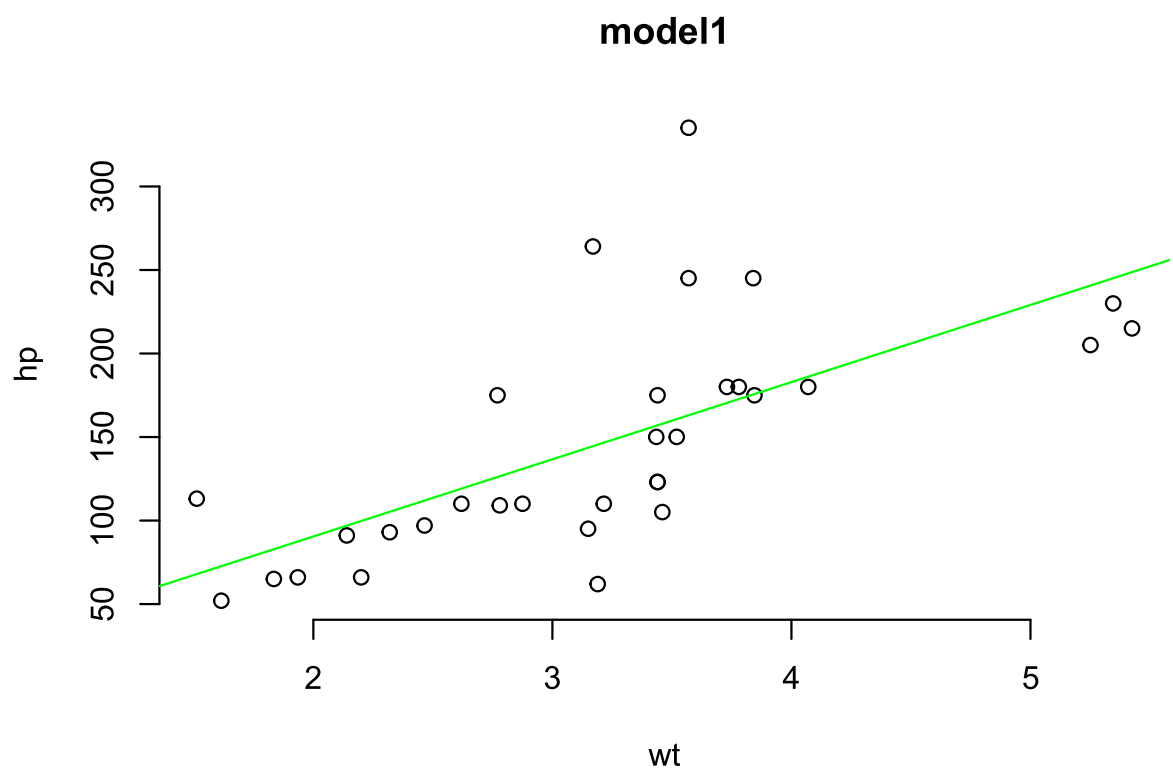
```
model1 <- lm(hp ~ wt, data=P1)
summary(model1)
```

```
##
## Call:
## lm(formula = hp ~ wt, data = P1)
##
## 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
```

```r
plot(mtcars$wt, mtcars$hp, main = "model1",xlab = "wt", ylab = "hp",frame = FALSE)

# Adding regression line
abline(lm(mtcars$hp ~ mtcars$wt), col = "green")
```




```r
cat("when horsepower compared with miles per gallon(mpg) ")
```

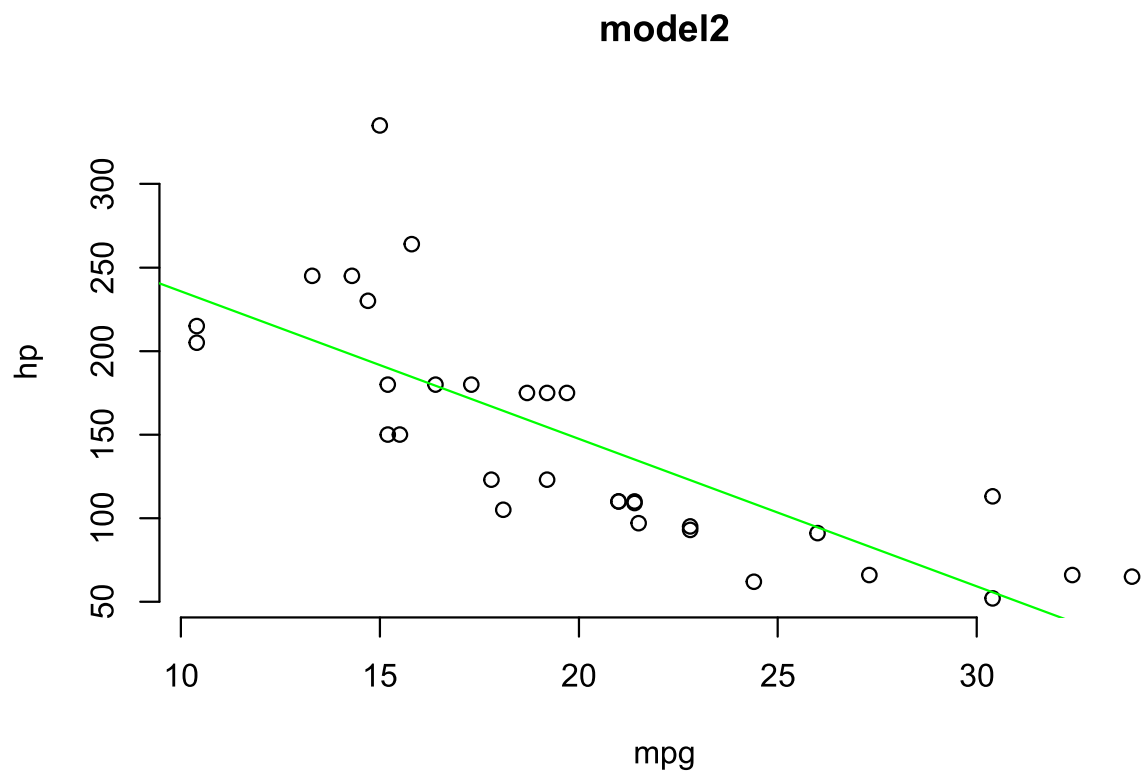
```
## when horsepower compared with miles per gallon(mpg)
```

```r
model2 <- lm(hp ~ mpg, data=P1)
summary(model2)
```

```
##
## Call:
## lm(formula = hp ~ mpg, data = P1)
##
## 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
```

```r
plot(mtcars$mpg, mtcars$hp, main = "model2",
     xlab = "mpg", ylab = "hp",frame = FALSE)
# Adding regression line
abline(lm(mtcars$hp ~ mtcars$mpg), col = "green")
```



As per the results of linear regression models where in model 1(when 'horsepower' compared with 'weight') and model 2(when 'horsepower' compared with 'mpg') It is quite clear that Chris is correct and seems to suggest a better indicator "mpg" for the HorsePower (hp) of cars because Model 2 has a higher R-squared value (0.6024 or 60.24%) compared to Model 1 (0.4339 or 43.39%). This suggests that Model 2 explains a larger proportion of the variance in the dependent variable. Model 2 has a higher F-statistic (45.46) with an extremely low p-value (1.788e-07), indicating that the model as a whole is highly significant.Model 1 has a lower F-statistic (23) with a significant but higher p-value (4.146e-05).However, the coefficients of

independent variable indicate that because mpg has negative value of b1 (-8.83) means that when we increase 1 unit of fuel, it can decrease 8.83 hp of Horse power. In contrast, when the weight increase by 1 unit, the Power of horse will increase 46.16 unit.

Q2.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?

```
P2 <- data.frame(hp = mtcars$hp, cyl = mtcars$cyl, mpg = mtcars$mpg)
#head(P2)
cat("when horsepower compared with cylinder and miles per gallon ")
```

```
## when horsepower compared with cylinder and miles per gallon
```

```
model3 <- lm(hp ~ cyl + mpg, data=P2)
summary(model3)
```

```
##
## Call:
## lm(formula = hp ~ cyl + mpg, data = P2)
##
## 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
```

```
Estimate_hp <- predict(model3,data.frame(cyl=4,mpg=22))
Estimate_hp
```

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

Using this model the estimated Horse Power of a car with 4 cylinders and mpg of 22 is 88.93618 horsepower.

Q.3.For this question, we are going to use BostonHousing dataset. The dataset is in 'mlbench' package,so we first need to instal the package, call the library and the load the dataset using the followingcommands

```
#install.packages('mlbench')
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.3.2
```

```r
data(BostonHousing)
#You should have a dataframe with the name of BostonHousing in your Global environment now.
```

Q.3.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 R 2)

```r
model4 <- lm(medv ~ crim + zn + ptratio + chas, data = BostonHousing )
summary(model4)
```

```
##
## 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
```

The model might be considered reasonably accurate because the model indicates that all predictors are statistically significant, as their p-values are very close to zero (p-value < 0.005). This suggests that these predictors have a significant impact on the dependent variable (medv). The residual standard error is 7.388. A lower residual standard error indicates a better fit of the model to the data. In this case, it's relatively low, which is a positive aspect. The R-squared value is 0.3599. It measures the proportion of variance in the dependent variable (medv) explained by the predictors. In this model, approximately 35.99% of the variance is explained. The adjusted R-squared is 0.3547. It accounts for the number of predictors in the model and is slightly lower than the R-squared. A higher adjusted R-squared is generally preferred. The F-statistic tests the overall significance of the model. A high F-statistic (70.41) with an extremely low p-value ($< 2.2e-16$) indicates that the model as a whole is highly significant. Hence this model is statistically significant, and the predictors have low p-values, suggesting they are relevant for predicting the dependent variable. The model might be considered reasonably accurate.

3.b(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?

```r
coefficient_chas1 <- coef(model4)["chas1"]
coefficient_chas1
```

```
##    chas1
## 4.583926
```

3.b(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?

```
coefficient_chas1 <- coef(model4)["ptratio"]
coefficient_chas1
```

```
##   ptratio
## -1.493673
```

Q.3.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.)

```
model5 <- lm(medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + lstat + b, d
summary(model5)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + indus + chas + nox + rm + age +
##     dis + rad + tax + ptratio + lstat + b, data = BostonHousing)
##
## 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
## chas1        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 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## b            9.312e-03  2.686e-03   3.467 0.000573 ***
## ---
## 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
```

Statistically Important Variables based on the p-values (related to the house price) are Intercept, crim, zn, chas, nox, rm, dis, rad, tax, ptratio, lstat and b. These variables are statistically important because their p-values are less than 0.05. Variables "indus" and "age" have p-values above the typical significance level of 0.05 and are considered not statistically significant in this model.

Q.3.d.Use the anova analysis and determine the order of importance of these four variables.

```r
anova.result <- anova(model4)
print(anova)
```

```
## function (object, ...)
## UseMethod("anova")
## <bytecode: 0x000001a946f35990>
## <environment: namespace:stats>
```

Based on the ANOVA analysis results: crim has the highest F-value (118.00683) and the lowest p-value (7.902220e-25), indicating the highest importance among the four variables. ptratio is the second most important with an F-value of 86.28724 and a very low p-value (4.738745e-19). zn is the third most important with an F-value of 65.12189 and a very low p-value (5.252886e-15). chas is the least important among the four variables, with an F-value of 12.22407 and a relatively higher p-value (5.136898e-04).