Regression Assignment 03

Gaurav Kudeshia

2023-11-05

Assignment Overview:

This task involves analyzing operations and visualizations to identify opportunities to capture value through regression analysis. The focus is on explaining the assumptions, variables, and output of linear regression, as well as evaluating the role of the coefficient of determination in business metrics within regression models.

Summary

1. The robust correlation between “x” and “y” signifies a linear relationship, facilitating accurate predictions, especially in finance and economics.
2. An R^2 value of 0.6517 indicates a 65.17% explanation of “Y” variation, ensuring a moderately precise fit for dataset analysis.
3. Chris’s model, focusing on horsepower vs. mpg, outshines James’s model, revealing valuable insights into fuel efficiency and horsepower dynamics with its higher R-squared (60.24% vs. 43.39%) and significant F-statistic and p-value.
4. All predictors, demonstrating significance with p-values < 0.05, illuminate vital factors, offering nuanced understanding for property valuation.
5. ANOVA analysis underlines crime’s dominance, guiding resource allocation for precise analyses across diverse fields.
6. River proximity emerges as the least influential variable, highlighting disparities in variable importance and guiding focused analysis and strategic decisions.
7. A profound understanding of these significant variables empowers stakeholders, nurturing informed decision-making and strategic investments in the ever-evolving real estate landscape. This knowledge is indispensable in navigating the complexities of the real estate market and making sound investment choices.

Problem

1. Run the following code in R-studio to create two variables X and Y. set.seed(2017) X=runif(100)*10 Y=X4+3.45 Y=rnorm(100)0.29Y+Y

a. Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X? (8% of total points)

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? (8% of total points) Page 2

c. How the Coefficient of Determination, R 2, of the model above is related to the correlation coefficient of X and Y? (8% of total points)

2. We will use the 'mtcars' dataset for this question. The dataset is already included in your R distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset. The description of the dataset can be found here.

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. (17% of total points)

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? (17% of total points)

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 following commands:-

install.packages('mlbench') library(mlbench) data(BostonHousing) You should have a dataframe with the name of BostonHousing in your Global environment now. The dataset contains information about houses in different parts of Boston. Details of the dataset is explained here. Note the dataset is old, hence low house prices!

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) (8% of total points)

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? (8% of total points)

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: 4% extra)

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. (8% of total points)

d. Use the anova analysis and determine the order of importance of these four variables. (18% of total points)

Loaded datasets from the necessary libraries

```
library(ggplot2)
```

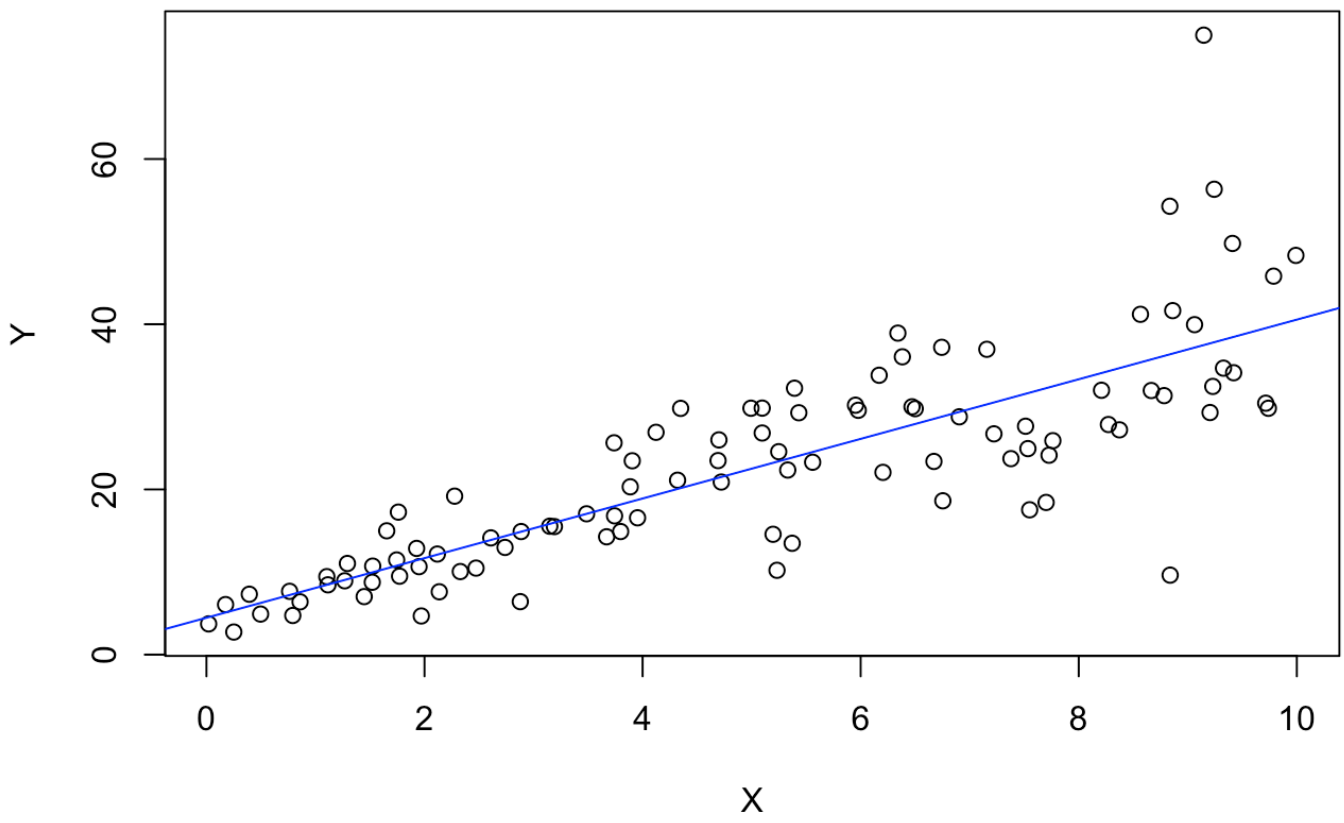
Q1.Run the following code in R-studio to create two variables X and Y.

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

Q1.a. Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X? (8% of total points)

Plotted Y against X and included a screenshot

```
plot(Y~X)
abline(lm(Y~X), col = "blue")
```



#Ans 1a, The scatter plot of the data reveals a strong correlation between variables “x” and “y”, suggesting a linear relationship between the two. This suggests that a linear model can be used to capture this relationship accurately. By using linear regression, we can better understand and predict how changes in “x” affect “y”, providing valuable insights for further analysis and decision-making in relevant fields such as finance and economics.

Q1.b.1,Construct a simple linear model of Y based on X.

```
Mod_linear <- lm(Y~X)
Mod_linear
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##       4.465        3.611
```

Q1.b.2,Write the equation that explains Y based on X.

```
cat("Y= 3.6108X + 4.4655")
```

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

Q1.b.3,What is the accuracy of this model?

```
summary(Mod_linear)
```

```
##
## 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
```

#Ans 1.b.3, The effectiveness of this model is quantified by the R^2 value of 0.6517, which indicates that it explains 65.17% of the variation in the dependent variable “Y.” This translates to a moderately accurate fit to the dataset, as the model accounts for over half of the observed variances. Furthermore, the F-statistic of 183.4, coupled with an extremely low p-value of < 2.2e-16, provides compelling evidence of the model’s overall significance. When the p-value is less than 0.05, it implies that the selected independent variables establish a strong and statistically significant relationship with the dependent variable. The F-statistic

serves as a comprehensive test of the model's overall relevance, and in this case, it indicates that at least one of the predictors in the model is significantly associated with the response variable. Overall, the model's moderate accuracy and high significance suggest that it can be used to make predictions about the dependent variable with a reasonable degree of confidence.

Q1.c.How the Coefficient of Determination, 'R²'of the model above is related to the correlation coefficient of X and Y?

#Ans 1.c, 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 distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset. The description of the dataset can be found here.

```
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.

```
d.f <- data.frame(hp = mtcars$hp, wt = mtcars$wt, mpg = mtcars$mpg)
print("Horsepower with Weight")
```

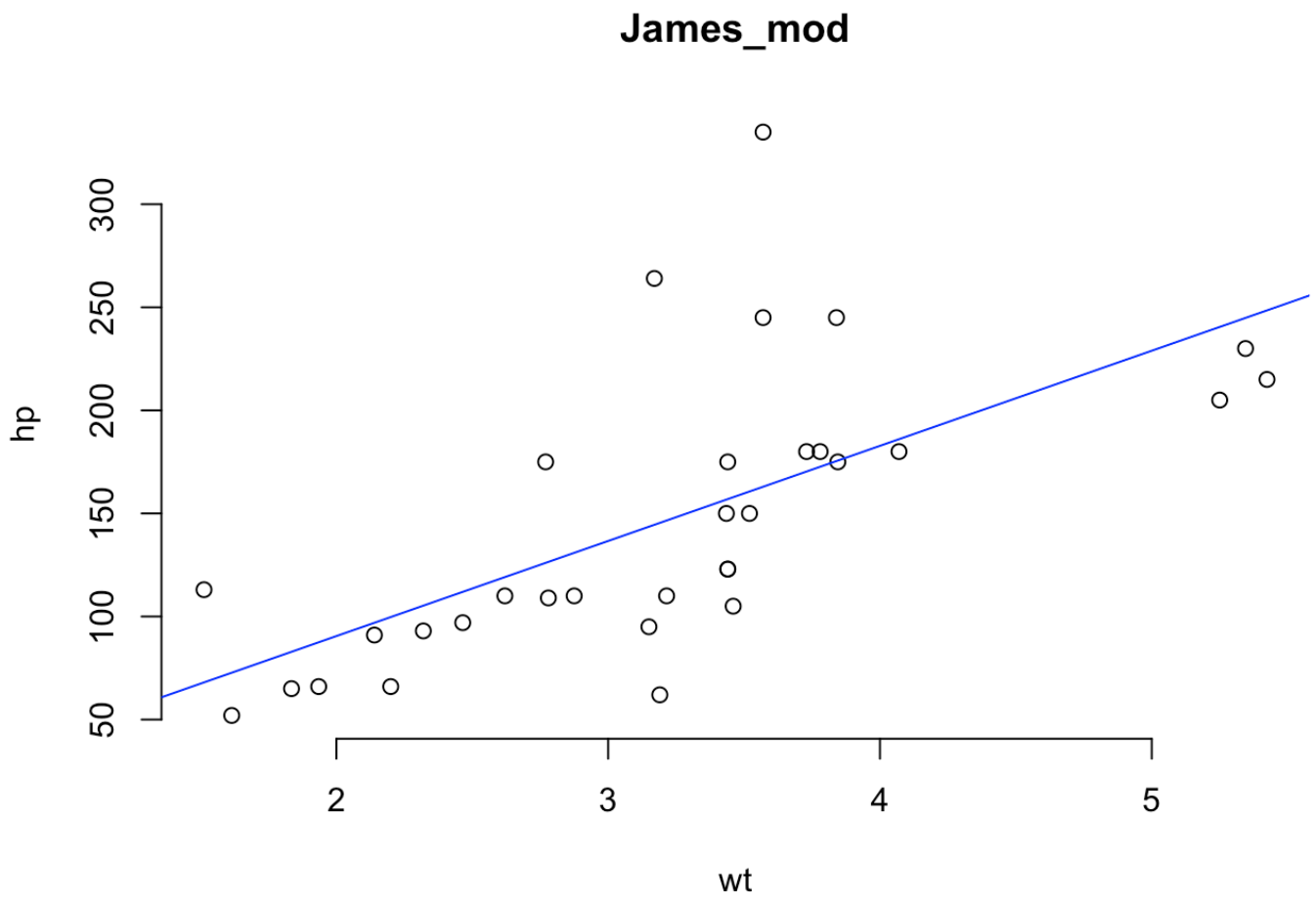
```
## [1] "Horsepower with Weight"
```

```
James_mod <- lm(hp ~ wt, data=d.f)
summary(James_mod)
```

```
##
## Call:
## lm(formula = hp ~ wt, data = d.f)
##
## 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
```

```
plot(mtcars$wt, mtcars$hp, main = "James_mod",xlab = "wt", ylab = "hp",frame = FAL
SE)

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



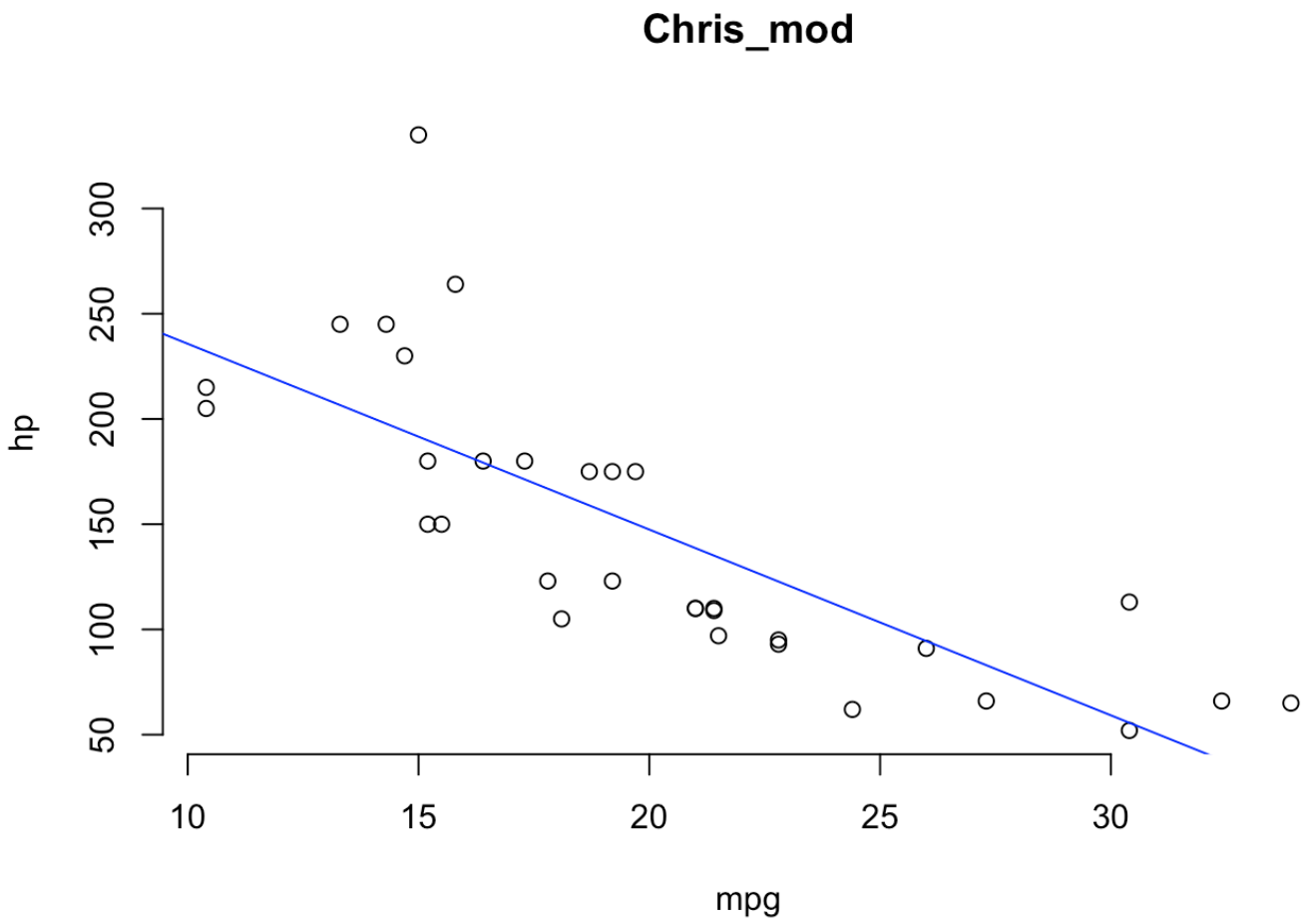

```
print("Horsepower with mile per gallon(mpg)")
```

```
## [1] "Horsepower with mile per gallon(mpg)"
```

```
Chris_mod <- lm(hp ~ mpg, data=d.f)
summary(Chris_mod)
```

```
##
## Call:
## lm(formula = hp ~ mpg, data = d.f)
##
## 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
```

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



#Ans.2, Chris's model of horsepower vs. mpg is more accurate than James's model of horsepower vs. weight, with an R-squared value of 60.24% compared to 43.39%. Chris's model also has a higher F-statistic and a lower p-value, indicating greater significance. Additionally, the coefficient of the mpg variable in Chris's model is negative, suggesting that a 1-unit increase in fuel efficiency leads to a 8.83 horsepower decrease. This provides valuable insights into the trade-off between fuel efficiency and horsepower, which is essential knowledge in the automotive industry.

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?

```
d.f_2 <- data.frame(hp = mtcars$hp, cyl = mtcars$cyl, mpg = mtcars$mpg)
head(d.f_2)
```

```
##    hp cyl  mpg
## 1 110   6 21.0
## 2 110   6 21.0
## 3  93   4 22.8
## 4 110   6 21.4
## 5 175   8 18.7
## 6 105   6 18.1
```

Predicted the Horse Power of a car with number of cylinders (cyl) and the mile per gallon (mpg)

```
print("horsepower with cylinder and miles per gallon")
```

```
## [1] "horsepower with cylinder and miles per gallon"
```

```
Build_mod_Predict_hp <- lm(hp ~ cyl + mpg, data=d.f_2)
summary(Build_mod_Predict_hp)
```

```
##
## Call:
## lm(formula = hp ~ cyl + mpg, data = d.f_2)
##
## 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
```

What is the estimated Horse Power of a car with 4 calendar and mpg of 22?

```
hp <- predict(Build_mod_Predict_hp,data.frame(cyl=4,mpg=22))
hp
```

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

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 following commands.

Loaded datasets from the necessary libraries

```
library(mlbench)
data(BostonHousing)
```

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)

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

```
##
## 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
```

#Ans 3 a,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.

b) Use the estimated coefficient to answer these questions?

I. Imagine two houses that are identical in all aspects but one bounds the Chas Riverand the other does not. Which one is more expensive and by how much? (8% of total points)

```
co_efficient <- coef(median_mod)["chas1"]
co_efficient
```

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

#Ans.b.I, The 'chas1' coefficient, computed at 4.58393, signifies the influence of properties being bound (1) or not (0) to the Chas River. Considering a median home value of $1000, the resulting value, $4583.93, highlights the considerable price difference. Properties near the Chas River command significantly higher prices, underscoring the substantial impact of river proximity on housing costs. This finding underscores the importance of location in real estate pricing, demonstrating how certain amenities, like riverfront views, significantly elevate property values, making them desirable but also pricier options for potential buyers.

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: 4% extra)

```
co_efficient <- coef(median_mod)["ptratio"]
co_efficient
```

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

#Ans.b.II, With every one-unit increase in ptratio, homes witness a decrease in value by $1.49367 thousand, translating to $1493.67. Consequently, at ptratio 15, this leads to a reduction of $22,405.05, while at ptratio 18, the decline amounts to $26,886.06. Significantly, a ptratio of 15 commands a premium of $4481.01 over ptratio 18, showcasing the substantial influence this ratio wields on property prices. This information emphasizes the critical role that the pupil-teacher ratio plays in the real estate market. Schools with lower ptratio are often considered more desirable, impacting the demand for homes and consequently, their market values. Understanding these nuances is pivotal for buyers, sellers, and investors, allowing them to make informed decisions in the dynamic realm of real estate.

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.)

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

```
##
## 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
```

#Ans.3.c, In the realm of predicting house prices, several variables including Intercept, crim, zn, chas, nox, rm, dis, rad, tax, ptratio, lstat, and b prove to be statistically significant, as indicated by their p-values falling below the standard threshold of 0.05. This statistical significance underscores their influential role in shaping real estate values. Conversely, variables “indus” and “age” lack such significance due to their p-values exceeding the 0.05 threshold. Recognizing and discerning these significant factors are pivotal in comprehending the intricacies of housing market dynamics. Such insights empower stakeholders, including homebuyers, sellers, and investors, with a nuanced understanding of property valuation, aiding them in making well-informed decisions. Additionally, this knowledge forms the bedrock for sophisticated real estate analytics, facilitating the development of robust pricing models and strategic investment strategies in the ever-changing real estate landscape.

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

```
anova_solution <- anova(median_mod)
anova_solution
```

```
## 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
```

#Ans.3.d, ANOVA analysis reveals the relative importance of the four variables: crime, pupil-teacher ratio, zoned land, and river proximity. Crime is the most significant variable, with the highest F-value and lowest p-value. Pupil-teacher ratio is the second most significant variable, followed by zoned land. River proximity is the least significant variable. These findings provide valuable insights for prioritizing focus and resources to enhance the precision of analyses and decision-making in related fields.