**Linear Regression/ Logistic Regression: Additional Examples with Answers**

1. **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

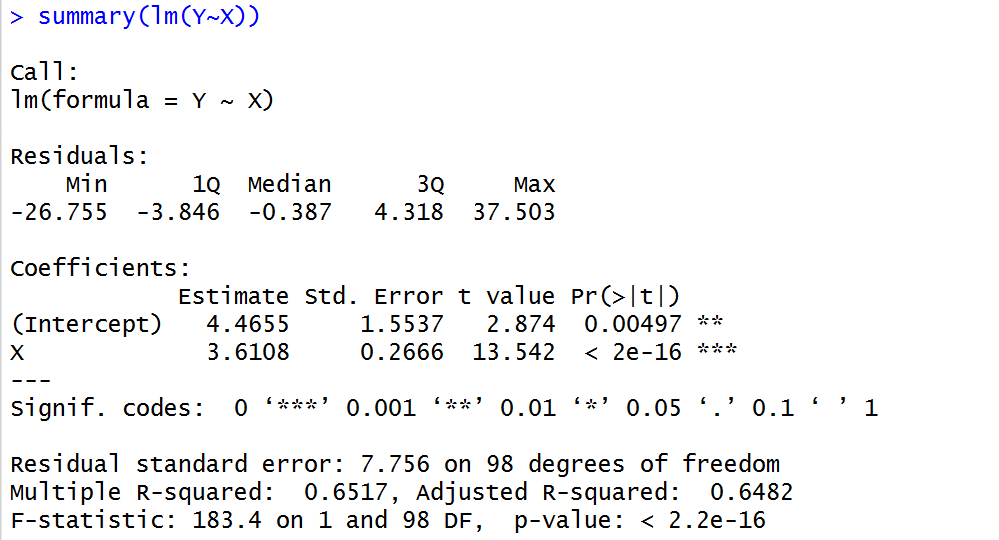
1. **Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X?**

Plot(X,Y)

****

Yes it seems that there is a potential to fit a linear model.

1. **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?**

****

Equation Y=3.61\*X+ 4.46

R2 is equal to 0.65 which means that the model explains 65% of variability.

1. **How the Coefficient of Determination, R2, of the model above is related to the correlation coefficient of X and Y?**

Coefficient of Determination= (Correlation Coefficient)2

R2 is 0.6517 from the previous question. and now

****

Confirming that the two values are equal. i.e.

Coefficient of Determination= (Correlation Coefficient)2

1. **Investigate the appropriateness of using linear regression for this case. Hint: residual analysis. You may also find the following link relevant ☺**

[**http://blog.minitab.com/blog/statistics-and-quality-data-analysis/violations-of-the-assumptions-for-linear-regression-the-trial-of-lionel-loosefit-day-1**](http://blog.minitab.com/blog/statistics-and-quality-data-analysis/violations-of-the-assumptions-for-linear-regression-the-trial-of-lionel-loosefit-day-1)

plot(X,lm(Y~X)$residuals)

Which produces the following plot. We can see that the residuals are higher for higher values of X suggesting non-constant variance of the error term (see slide 28 in lectures). This means that a linear regression model may not be the best choice here.



You can also examine the normality of the error term.

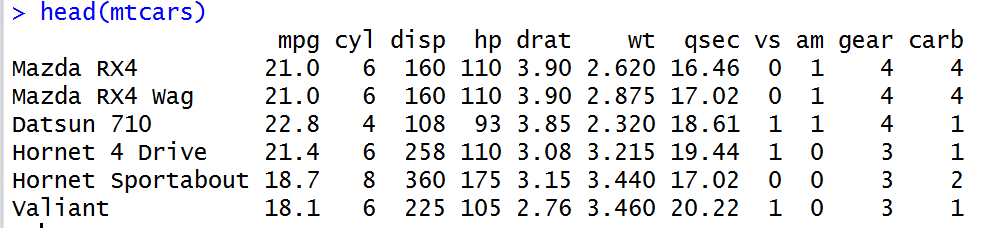
qqnorm(lm(Y~X)$residuals,col="red")

qqline(lm(Y~X)$residuals)

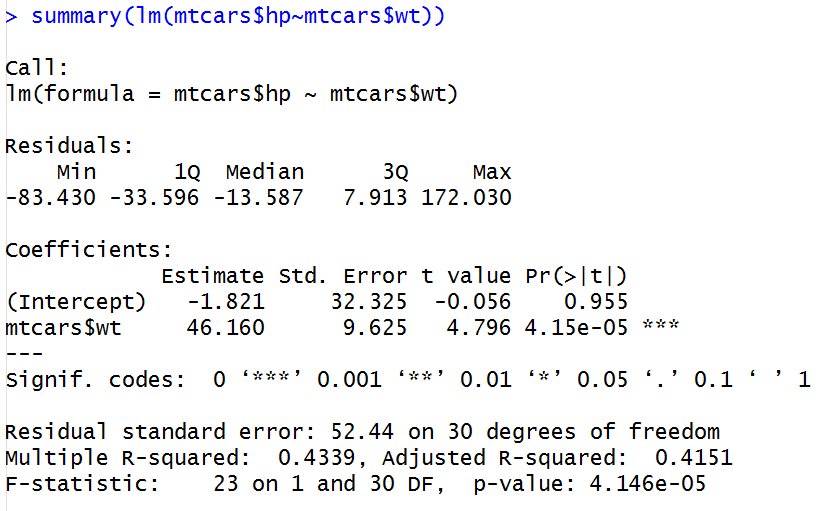
Again we can see that there are few points at the end of the spectrum which cannot be fitted. These are not simply outliers.



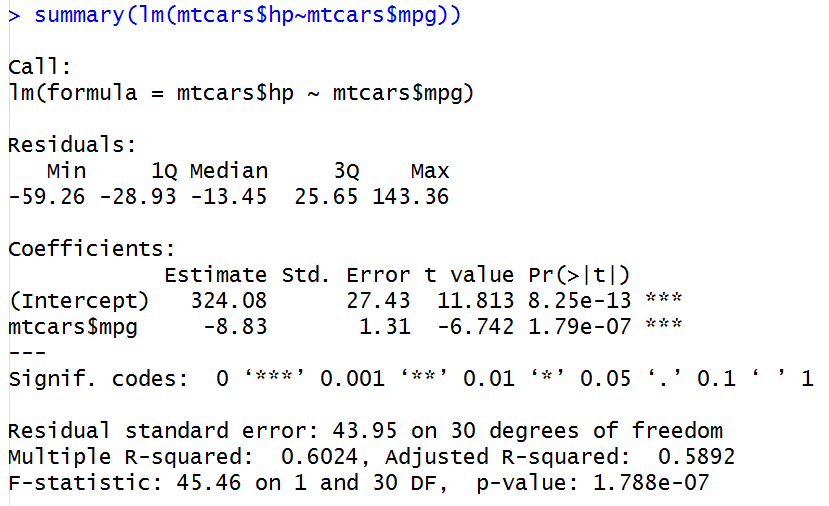
1. **We will use the ‘mtcars’ dataset for this question. The dataset is already included in your R distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows)of the dataset. The description of the dataset can be found** [**here**](https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html)**.**

****

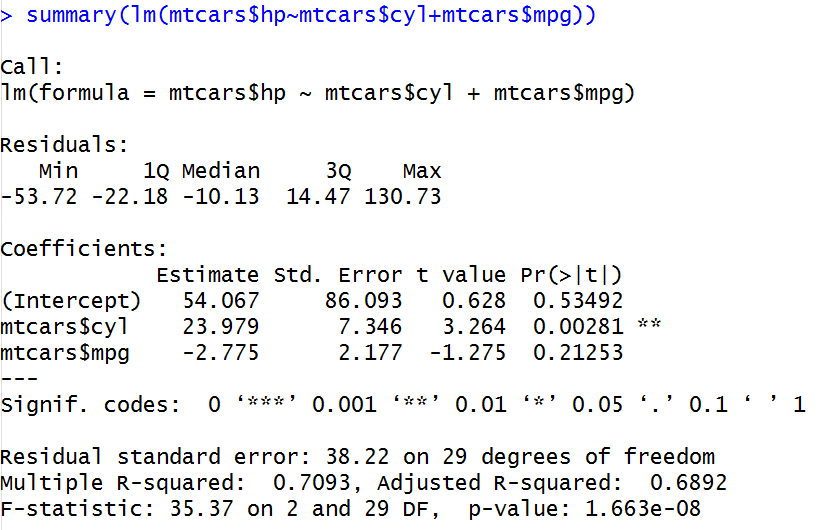
1. **James wants to buy a car! He and his new friend at KSU, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question. (10 marks)**

****

**R2 of the model based on weight is 0.43**

****

**R2 of the model based on mpg is 0.60 hence it is a more accurate model. Chris is right!**

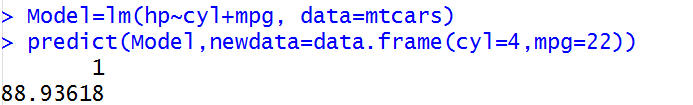
1. **Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp).**
2. **Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22? **

The equation is hp=23.979\*cyl-2.775\*mpg+54.067

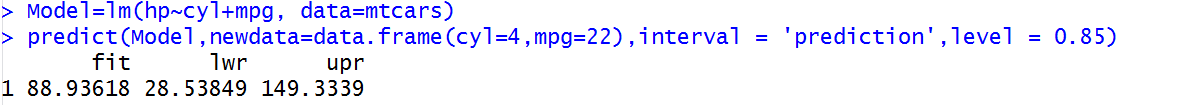
For a car with 4 cyl and mpg=22 we have

hp=23.979\*4-2.775\*22+54.067=88.93

or



1. **Construct an 85% confidence interval of your answer in the above question. Hint: use the predict function**

****

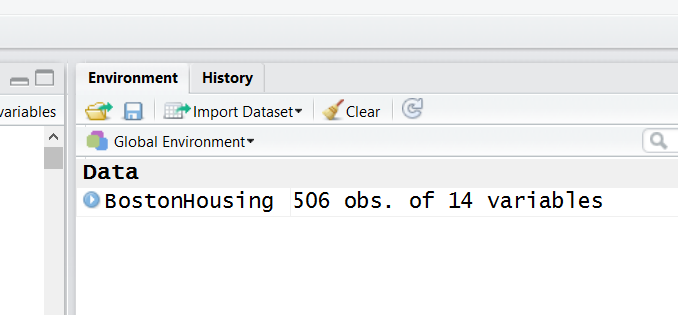
1. **For this question, we are going to use BostonHousing dataset. The dataset is in ‘mlbench’ package, so we first need to instal the package, call the library and the load the dataset using the following commands**

install.packages('mlbench')

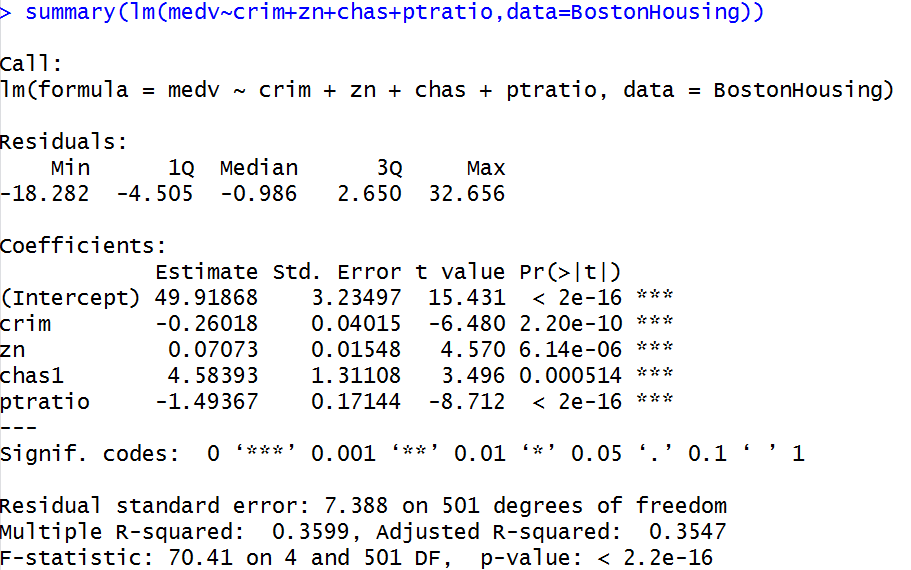
library(mlbench)

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**](https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Boston.html)**. Note the dataset is old, hence low house prices.**

1. **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 Use R2 to judge) **

R2 is 0.359 which is not very impressive.

1. **Use the estimated coefficient to answer these questions?**
2. **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?**

The coefficient chas1 is 4.58 which means that if variable is 1 instead of 0, we will add 4.583 to the estimate of price. That is $4,583 since the price is expressed in $1000. So the house bounding the River is $4,583 more expensive.

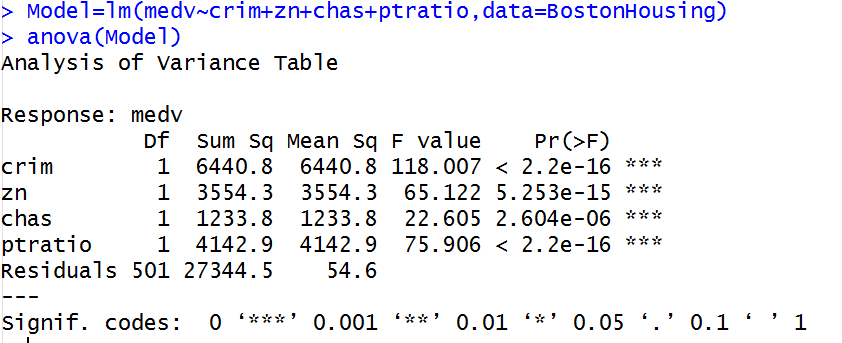
1. **Imagine two houses that are identical in all aspects but in the neighborhood of one of them the pupil-teacher ratio is 15 and in the other one is 18. Which one is more expensive and by how much?**

The coefficient of ptratio is -1.49 which means that for every unit increase in pupil-teacher ratio (i.e. less teachers and more crowded classes in schools), there will be -1.49 unit decrease in the price of the houses of the neighborhood. So the house with pupil-teacher ratio of 18 is 3\*1.49=4.47 unit (i.e. $4,470) cheaper than the house in the neighborhood with pupil-teacher ratio of 15.

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

The p-values of all coefficients are very small so all four variables are statistically significant and have relationship with the house price.

1. **Use the anova analysis and determine the order of importance of these four variables.**

****

Variables with higher Sum Square are more important. In this model the order of importance is 1. crim 2.ptratio 3. Zn and 4. chas

The p-values of all coefficients are very small so all four variables are statistically significant and have relationship with the house price.

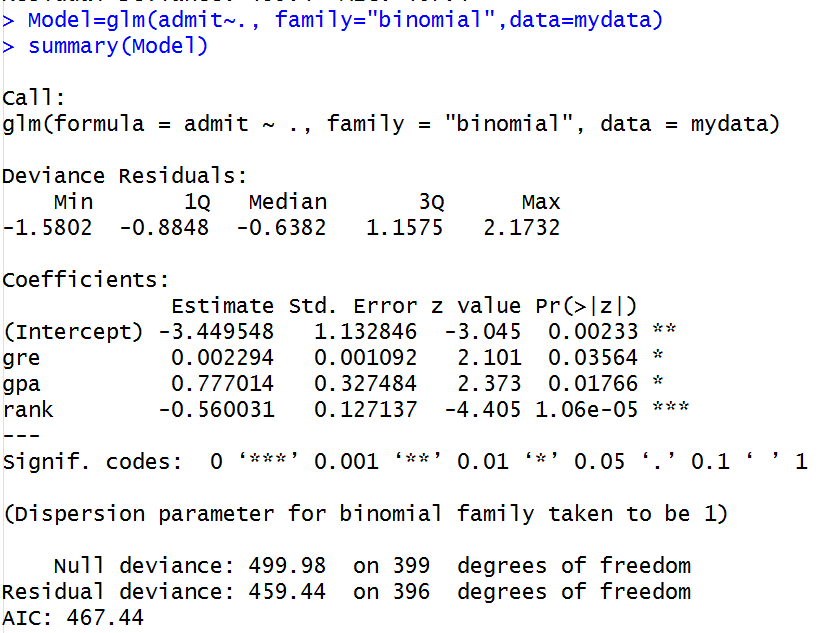
**The following is a Logistic regression Question**

1. **A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don’t admit, is a binary variable. A dataset containing 400 records can be downloaded from the UCLA website using (make sure you are connected to the internet when executing the command below)**

mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")

mydata$admit=as.factor(mydata$admit)

**Now you should have the mydata dataframe in your global environment.**

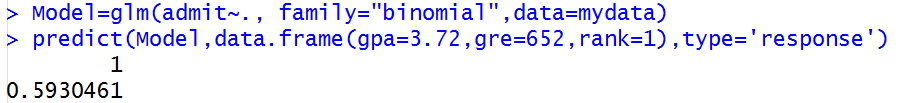
1. **Construct a logistic regression to predict the graduate school admission using GPA, GRE and School Rank. Which variables are statistically important (i.e. related to admission outcome)? **

The p-values of all coefficients are very small so all four variables are statistically significant and have relationship with the house price.

1. **Amongst GRE, GPA and the Rank of the Undergraduate University, which variable is the most important one?**

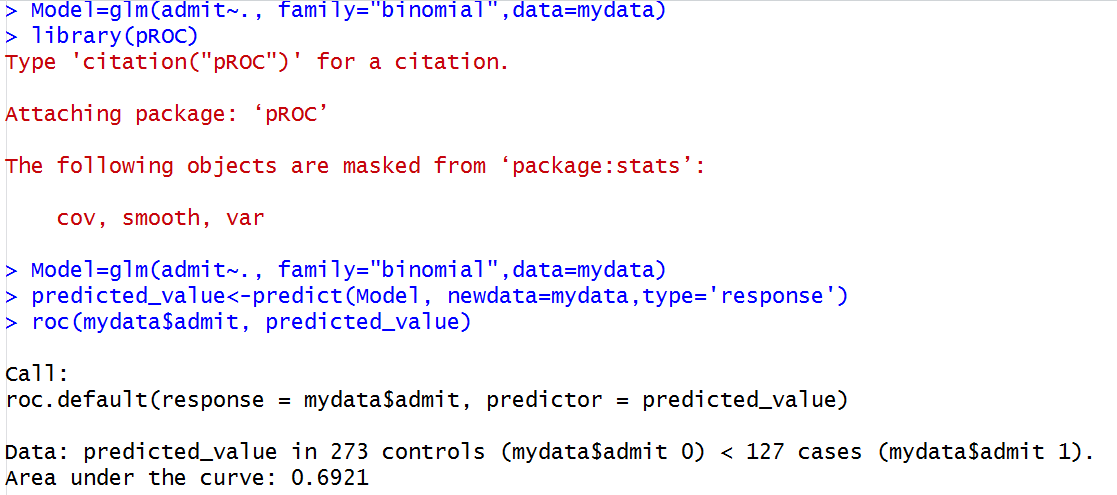
The absolute value of the z-score (or t-score) is the indicator of the variable importance, hence University rank (absolute z-value =4.40) is the most important variable here.

1. **Remember Elizabeth and Hanna from the first assignment? Hannah is now a grown up and has finished her undergrad at the KSU (rank 1 university!) with GPA Score of 3.72 and GRE Score of 652 and is planning to apply for graduate school at KSU. Her mother, Elizabeth, is now worried about her application! What is the probability that Hannah is accepted into the graduate program? Hint: Use the predict() function with type as ‘response’.**

****

She has 59% chance of getting accepted. Perhaps, Elizabeth is right to be worried!

1. **What is the AUC of the model?**



AUC is 0.69.

1. **Chris and James also plan to apply for the graduate school. They have exactly similar situation (same undergrad school and same GPA) but James has a GRE score that is exactly 100 points higher than that of Chris. If Chris has a 75% chance of admission, what is the chance for James?**

The coefficient of the GRE is 0.002294 which means that the logarithm of odds of admission is increased by 0.002294 for every unit increase of the GRE score. The GRE Score of James is 100 units higher that Chris which means that the logarithm of the odds for James is 0.2294 (0.002294\*100) higher than that of Chris.

That is to say

log(Pj/(1-Pj))-log(Pc/(1-Pc))=0.2294

where Pj is the probability of James acceptance, that is unknown and

Pc is is the probability of Chris acceptance, that is 0.75

So log(pj/(1-pj))=log(0.75/0.25)+0.2294=1.328012 =>

pj/(1-pj)=e1.328012 =3.773534 => pj=3.773534-3.773534\*pj => pj= 0.7905 !

James has 79.05% chance of admission.