**Module Assignment**

**Module 4**

**QMB-6304 Foundations of Business Statistics**

**#Debjani Sarma**

**rm(list=ls())**

**library("rio")**

**library("moments")**

**getwd()**

Write a simple R script to execute the following:

**Preprocessing**

1. Load the file “6304 Module 4 Assignment Data.xlsx” into R. This file contains information on 99 used automobiles for sale.

**my.cars = import("6304 Module 4 Assignment Data.xlsx")**

1. Using the numerical portion of your U number as a random number seed, take a random sample of 30 cars using the method presented in class. This will be the your primary data for your assignment.

**> set.seed(24173877)**

**> my.sample=my.cars[sample(1:nrow(my.cars),30),]**

**Analysis**

Using your primary data frame:

1. Show the results of the str() command.

**>attach(my.sample)**

**> str(my.sample)**

**'data.frame': 30 obs. of 4 variables:**

**$ price : num 16699 13420 11177 26995 5983 ...**

**$ mileage: num 66296 5783 34758 403 74398 ...**

**$ age : num 6 1 2 0 9 2 6 18 6 3 ...**

**$ make : chr "Lexus" "Nissan" "Nissan" "Honda" ...**

1. Conduct a simple regression using the price variable as the dependent variable and mileage as the independent variable.

**my.simpleLR = lm(price ~ mileage , data=my.sample)**

1. Building on your results in Part 2 give clear written interpretations of your model’s beta coefficients and associated p values. Make certain your beta coefficient interpretations are in terms of the case provided.

**> summary(my.simpleLR)**

**Call:**

**lm(formula = my.sample$price ~ my.sample$mileage, data = my.sample)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-10006 -2111 -1030 1978 12107**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 1.491e+04 1.168e+03 12.763 3.43e-13 \*\*\***

**my.sample$mileage -6.420e-02 1.332e-02 -4.821 4.52e-05 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 4144 on 28 degrees of freedom**

**Multiple R-squared: 0.4536, Adjusted R-squared: 0.4341**

**F-statistic: 23.25 on 1 and 28 DF, p-value: 4.52e-05**

**> confint(my.simpleLR)**

**2.5 % 97.5 %**

**(Intercept) 1.252016e+04 1.730714e+04**

**my.sample$mileage -9.147933e-02 -3.692505e-02**

Since there is a negative slope (-6.420e-02), it means that for each mile driven , the car price is expected to decrease by $0.06420.

**Beta coefficient and p-values:**

**Intercept:** Since the p-value is extremely low (**3.43e-13**), we reject the null hypothesis (intercept of true regression model is 0) and we have strong evidence to accept the alternate hypothesis (intercept of true regression model is not 0).

**Slope:** Since the p-value is extremely low (**4.52e-05**), we reject the null hypothesis (slope of true regression model is 0) and we have strong evidence to accept the alternate hypothesis (slope of true regression model is not 0). This also indicates that there is a significant relationship between mileage and price.

**Confidence Interval:**

**Intercept:** Estimated beta coefficient of intercept is 1.491e+04. However, we are 95% confident that the true beta coefficient of the intercept of the regression model lies between 1.252016e+04 and 1.730714e+04. This means that when the mileage is zero, the expected price of the car is between $12520.16 and $173071.14.

**Slope:** Estimate beta coefficient of intercept is 6.420e-02. However, we are 95% confident that the true beta coefficient of the slope of the regression model lies between -9.147933e-02 and -3.692505e-02. This indicates for each unit increase in mileage, the price of the car is expected to decrease by between $0.0914 and $0.0369.

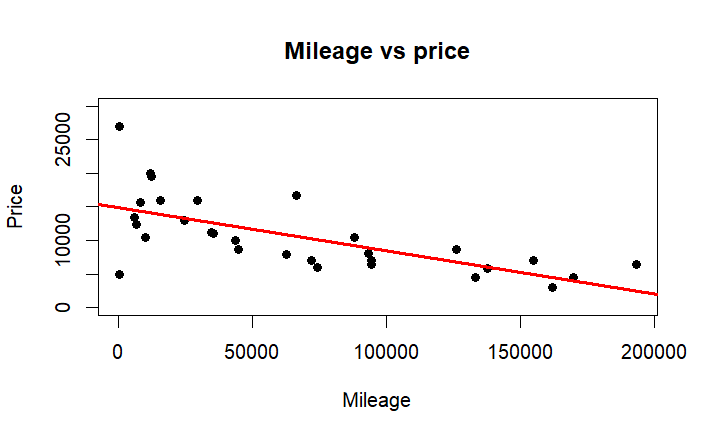
**R-squared:** 0.4536 suggests that 45.36% of the variability in car price is explained by mileage.

1. Building on your results in Part 2 determine and state whether your model is in conformity with the LINE assumptions of regression.

**> plot(my.sample$mileage, my.sample$price, main="Mileage vs price", pch=19,**

**+ xlab="Mileage", ylab="Price", ylim=c(0,30000))**

**> abline(my.simpleLR, col="red",lwd=3)**



The negative slope indicates that as mileage increases, car price decreases. Where there is some scatter around the regression line, the general trend is linear. This supports the linearity assumptions of the model.

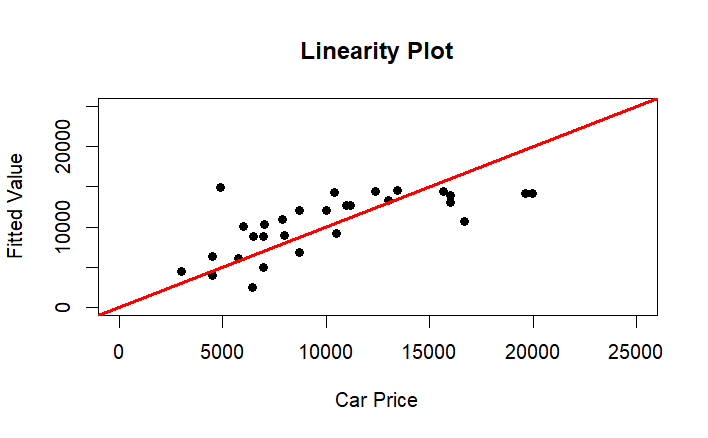
**> plot(my.sample$price, my.simpleLR$fitted.values, pch=19,**

**+ xlab="Car Price", ylab="Fitted Value",**

**+ xlim=c(0,25000),ylim=c(0,25000),**

**+ main="Linearity Plot")**

**> abline(0,1,col="red",lwd=3)**



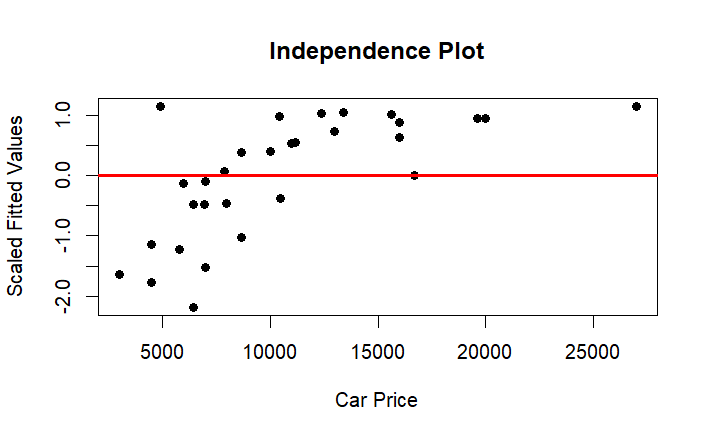
The points generally follow a linear trend, suggesting that we are in conformity with the assumption of linearity.

**plot(my.sample$price, scale(my.simpleLR$fitted.values), pch=19,**

**xlab="Car Price", ylab = "Scaled Fitted Values",**

**main="Independence Plot")**

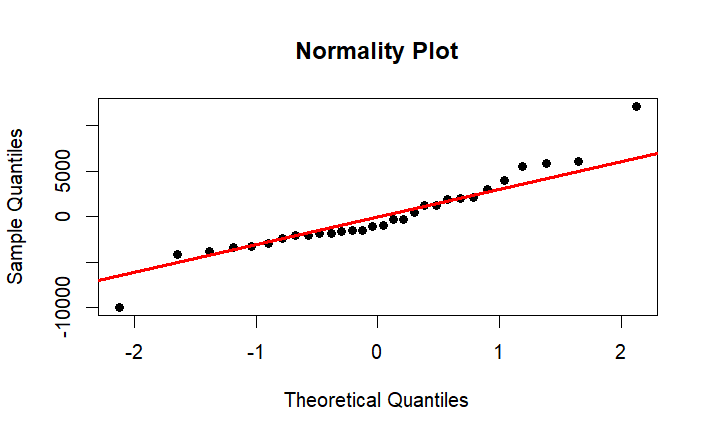
**abline(0,0,col="red",lwd=3)**



Since there is no clear patten or trend in the above plot, we are in conformity with assumption of independence.

**> qqnorm(my.simpleLR$residuals, main="Normality Plot", pch=19)**

**> qqline(my.simpleLR$residuals, col="red", lwd=3)**



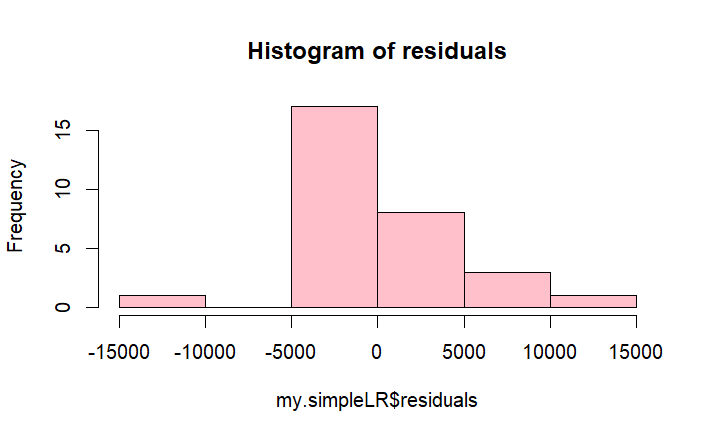
**> hist(my.simpleLR$residuals, col="pink",main="Histogram of residuals")**

**> skewness(my.simpleLR$residuals)**

**[1] 0.6140125**

**> kurtosis(my.simpleLR$residuals)**

**[1] 4.699025**



The skewness is 0.6140125 indicating slightly right skewness.

The kurtosis is 4.699025 which means that the residuals are a bit more peaked than a normal distribution. The residuals are centered around zero (as per histogram) but have some outliers confirming the skewness and kurtosis.

Overall, the residuals are approximately normally distributed, with only slight deviations.

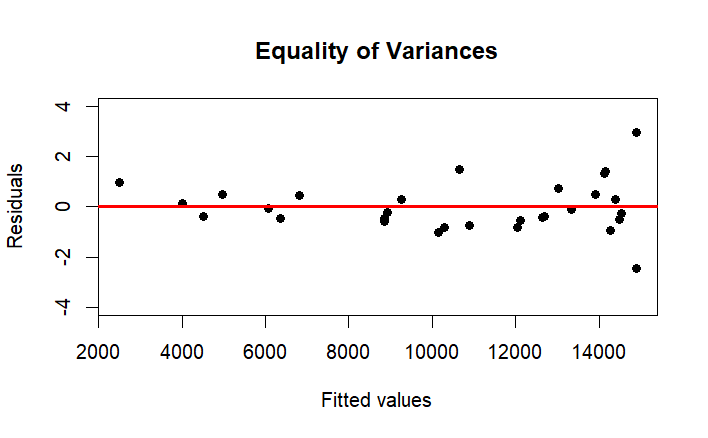
We are in conformity with assumptions of normality.

**> plot(my.simpleLR$fitted.values, scale(my.simpleLR$residuals), pch=19,**

**+ xlab="Fitted values", ylab="Residuals", main="Equality of Variances",**

**+ ylim=c(-4,4))**

**> abline(0,0, col="red",lwd=3)**



In the above plot, some variability can be seen, which might be influenced by outliers. The residuals appear to be fairy constant in their spread, suggesting that the homoscedasticity assumption is satisfied. Therefore, we are in conformity with assumption of equality of variances.

1. A two-year-old Honda CR-V with approximately 25,000 miles is offered for sale. What does your model predict the price of this car should be?

**> attach(my.sample)**

**> newdata=data.frame(mileage=25000)**

**> predicted\_price = predict(my.simpleLR, newdata, interval = "none")**

**> predicted\_price**

**1**

**13308.59**

1. The vehicle shown below is a 1934 Rolls Royce Phantom II Continental Owens Drophead Sedanca Coupe, a custom-built vehicle which is likely one of a kind and highly collectable. Its odometer shows 1275 miles. The last time the vehicle sold was for a price in excess of $600,000. Explain why it would be inappropriate to use your model to predict the price of this car.

A close-up of a black car

Description automatically generated

**> rolls.royce=data.frame(mileage=1275)**

**> predicted.price1 = predict(my.simpleLR, rolls.royce, interval = "none")**

**> predicted.price1**

**1**

**14831.79**

The 1934 Rolls Royce Phantom II Continental Owens Drophead Sedanca Coupe is one of a kind,

highly collectable vehicle with only 1275 miles and a price of $600,000, making it an outlier in terms of age, mileage , price compared to the vehicles used for this current linear model. As a custom-build car, it is not well represented in a model that is based on standard vehicles and primarily focuses on the mileage to predict the price. Therefore, the price predicted as per mileage is $14831.79, which is way less than $600,000. The unique factors such as rarity, collectible – is not captured by the model. Therefore, applying this model to predict the price of the Rolls Royce would be inappropriate and would lead to unreliable and inaccurate predictions.

Your deliverable will be a single MS-Word file showing 1) the R script which executes the above instructions and 2) the results of those instructions. The first line of your script file should be a “#” comment line showing your name as it appears in Canvas. Results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. **This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.**