Linear Regression Project

Guidelines:

a. Write your answer in R Markdown and answer below each question
b. Attach R Script along with your answer
c. Show RELEVANT R output in your answers (R Markdown will do this)
d. You can take more space for your answers if needed.
e. This assignment would require a thorough/solid reading from Ken Black book and net roughly of 20 hrs by each participant (not collective hrs of study!)

Task

Refer file **grades**.csv

The school principal wants to build a predictive model for predicting final for his consumption. As a principal he is very keen to have good scores by his students. He has given this data file to you with a request to suggest an appropriate model.

You are required to build at least 4 models with different sets of predictors (independent variables). Selection of sets of predictor/s is upon you. Different sets of predictors can be a single variable or more than one variable. However, selection of predictor/s should be based on some logic. For example, for predicting final score of students, roll number cannot be a logical predictor.

You will analyze all 4 models based on following points and recommend the best model to the Principal.

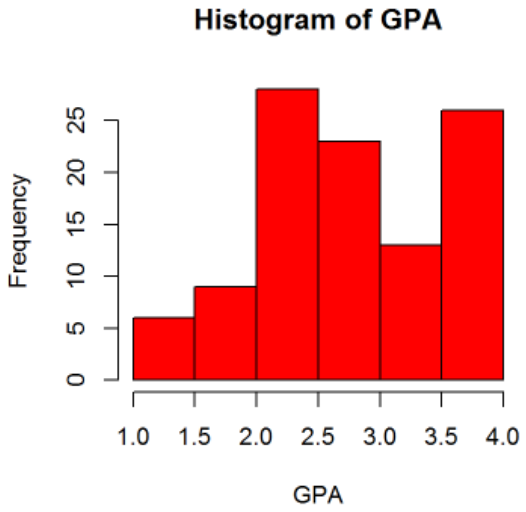
Question:1 Describe data of response variable and predictors in terms of key summary statistics like mean, mode, median, standard deviation, range, skewness and kurtosis. Show histogram and box plots also for each variables..Each variable to be explained in 30 words maximum.

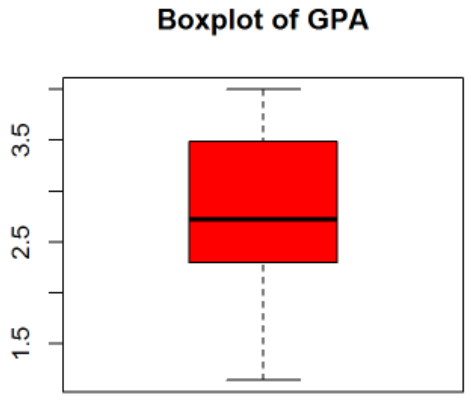
Answer

Box plots & Histograms are drawn to check the normality of data for that variable

1. GPA

Data of GPA says that mean of GPA is2.78 & sd of GPA is 0.76 Also Mean and trimmed mean, sd and mad are very near hence no outliers, more data on right side and data is platykurtic (meaning it has lighter tails & flat central peak) as shown by kurtosis value. Also skewness value is -0.05 it means it is a bit left skewed or negatively skewed.

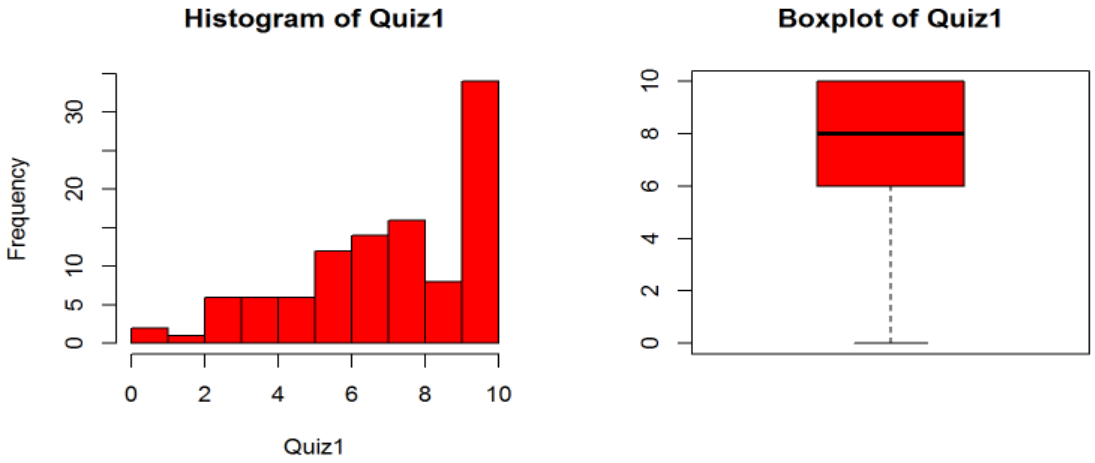




2. Quiz1

Data states that mean is 7.47 sd is 2.48.By the histogram and boxplot we can say that data is skewed towards the left that is it is negatively skewed as skewness value is -0.83 also Kurtosis value is 0.04 which is close to 0 hence our data is having heavier tails and sharper peaks an is a leptokurtic distribution.

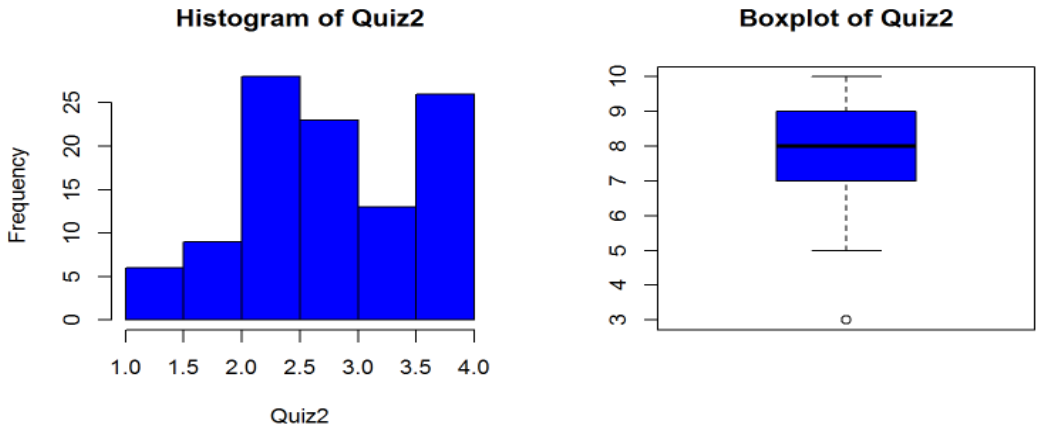
```
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis
## X1    1 105 7.47 2.48      8    7.76 2.97   0  10    10 -0.83     0.04
##      se
## X1 0.24
```



3. **Quiz2**

Data states that mean value is 7.98 and sd of 1.62. By histogram states that data is skewed little towards the left and boxplot states that data is a almost normally distributed with an outlier in it which is 3 but it is also little skewed towards left. Range of quiz2 is 3 to 10 .Skewness value of -0.64 means a little left skewed .kurtosis value of -0.35 means a platykurtic with lighter tails and a flat central peak

```
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis
## X1    1 105 7.98 1.62      8    8.12 1.48   3  10     7 -0.64    -0.35
##      se
## X1 0.16
```

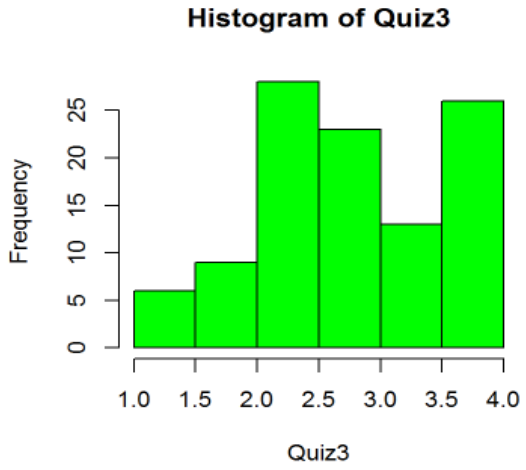


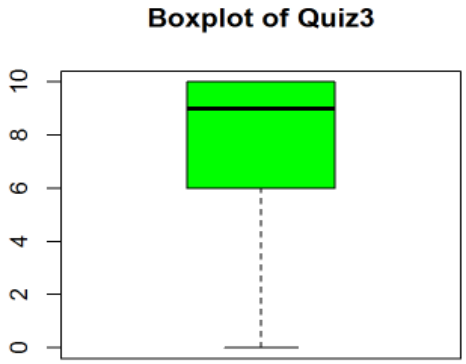
4. **Quiz3**

Data states that mean value is 7.98 and sd of 2.31(means the observations can have standard dev.of 2.31).By histogram we concluded that data is more towards the left and boxplot states that data is a not normally distributed but completely

skewed towards left. Range of quiz3 is 0 to 10 ,Skewness value of -1.1 means data is left skewed .Also kurtosis value of 0.59 means a leptokurtic with heavier tails and a high sharp peak.

```
##    vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 105 7.98 2.31      9    8.34 1.48   0  10    10 -1.1     0.59 0.23
```

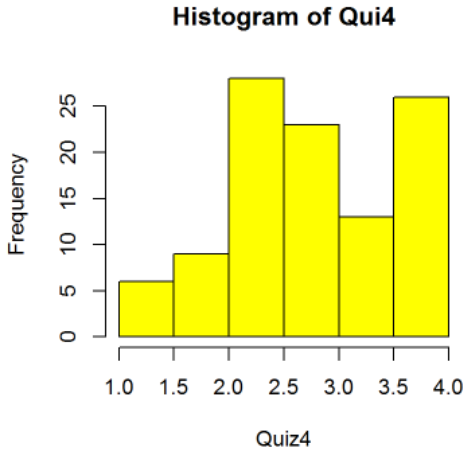


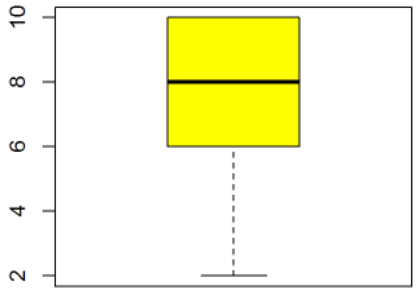


5. Quiz4

Data in quiz4 states that mean value is 7.98 and sd of 2.28(means the observations can have standard dev.of 2.28).By histogram we find that data is more towards the left and boxplot states that data is a not normally distributed but completely skewed towards left. Range of quiz4 is 0 to 10 .Skewness value of -0.89 means data is left skewed .kurtosis value of -0.09 means a platykurtic with lighter tails and a flat central peak

```
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis
## X1    1 105  7.8 2.28      8    8.11 2.97   2  10     8 -0.89    -0.09
##      se
## X1 0.22
```

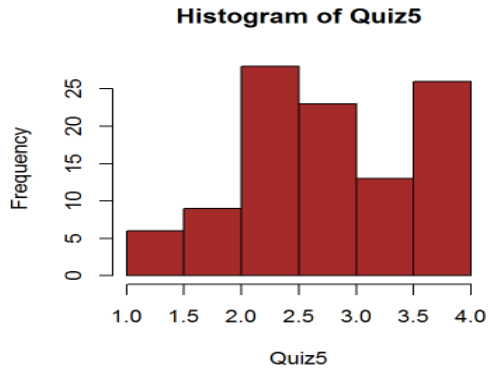


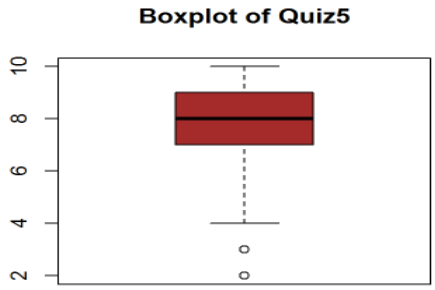


6. **Quiz5**

Data states that mean value is 7.87 and sd of 1.77 (means the observations can have standard deviation of 1.77) .By the histogram we conclude that data is spread towards the left side more and boxplot states that data is a little more towards the left side but can be said as normally distributed. Range of quiz5 is 0 to 10 ,.Skewness value of -0.69 means data is a bit left skewed .kurtosis value of 0.16 means a leptokurtic with heavier tails and a sharper central peak than no distribution .

```
##    vars   n mean   sd median trimmed  mad min max range  skew kurtosis
## X1    1 105 7.87 1.77      8    8.02 1.48   2  10     8 -0.69     0.16
##       se
## X1 0.17
```

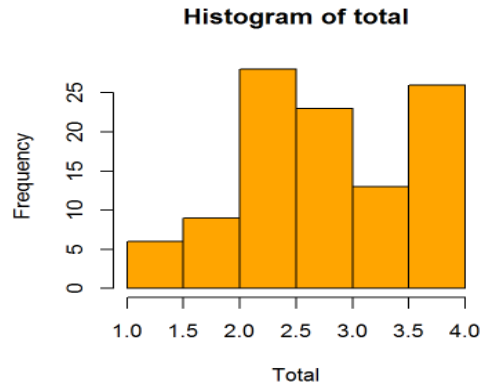


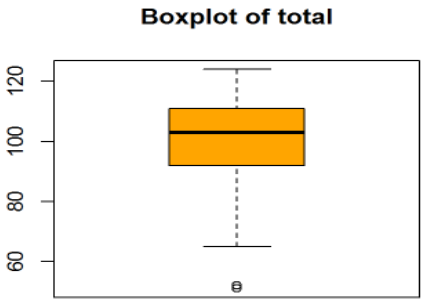


7. **Total**

Data states that mean is 100.57 & sd is 15.3.Also Mean and trimmed mean, sd and mad are very near hence no outliers and no variability, most data on right side hence it is left skewed. Skewness value of -.081 means data is left skewed or can be negatively skewed. Data is leptokurtic meaning heavier tails and sharper central peak as shown in kurtosis value.

```
##    vars   n   mean   sd median trimmed   mad min max range  skew kurtosis
## X1    1 105 100.57 15.3    103   101.8 13.34  51 124    73 -0.81     0.77
##       se
## X1 1.49
```

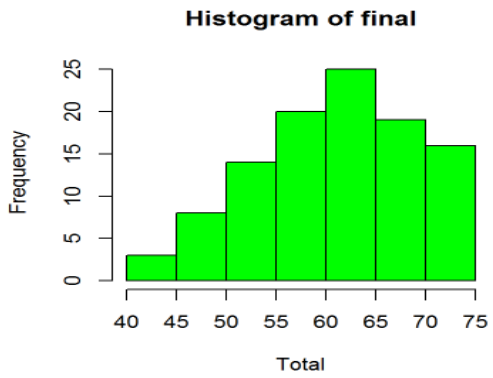


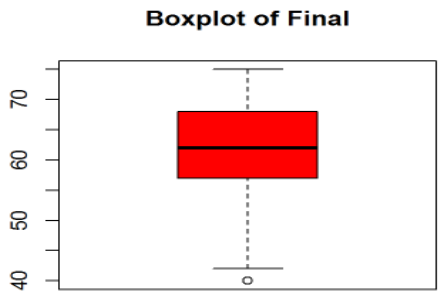


8. Response Variable Final

Data in the final variable states that it has mean of 61.48.Numeric range 40-75 .Standard deviation is 7.94 which means that the 68% of the deviations are in the zone of **[61.47 + 1*7.94] or [61.48- 1*7.94]** i.e. b/w **53.54 & 69.42** whereas **95%** of the final observations are in the range **[61.48 +- (2 *7.94)] i.e. b/w 45.6 & 77.36**. Median of final is **62**. By Histogram we can say that final data is almost normally distributed around its mean but a little skewed towards left and there is an outlier in the data. Trimmed mean of final is 61.74 .which is obtained by removing the observations which are quite far from the other observations. Skewness value of final data is -0.33 which means that the data little towards the left due to outliers present. Kurtosis value of -0.42 means that distribution is with light and thinner tails and its central peak is lower and broader when compared with normal distribution, the data is platykurtic

```
##    vars   n  mean   sd median trimmed mad min max range  skew kurtosis
## X1    1 105 61.48 7.94     62   61.74 8.9  40  75    35 -0.33    -0.42
##      se
## X1 0.78
```





Question 2 :

How predictor/s is related to response variable (final)? [plot scatter diagram followed by correlation test]

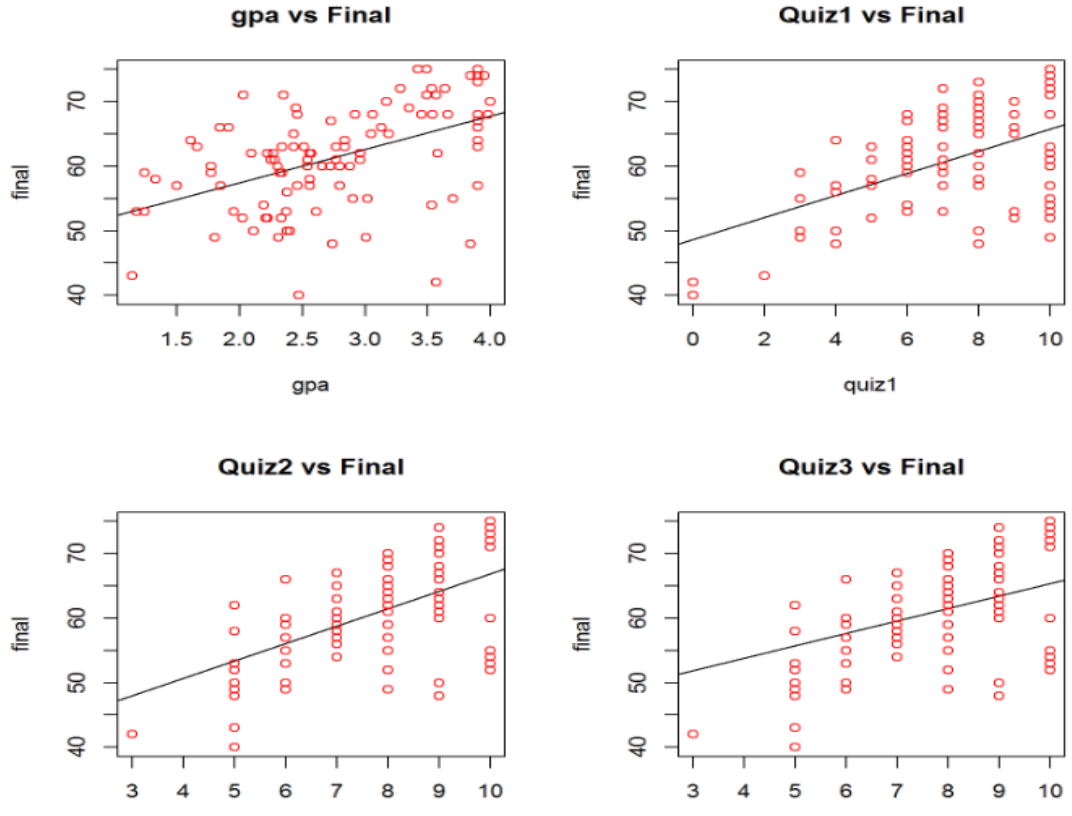
Present diagram/s and correlations in the following space. Before diagrams explain relationship.

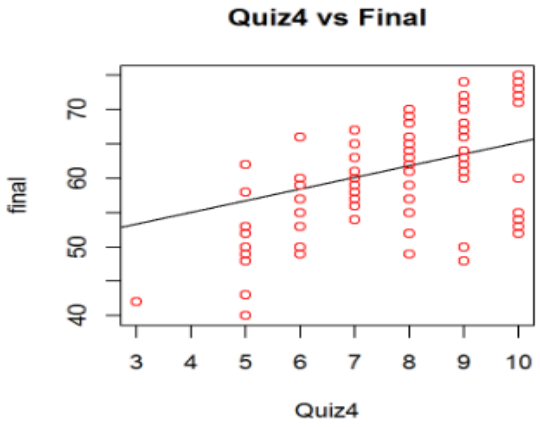
Answer: Response Variable or Dependent Variable is : **final**

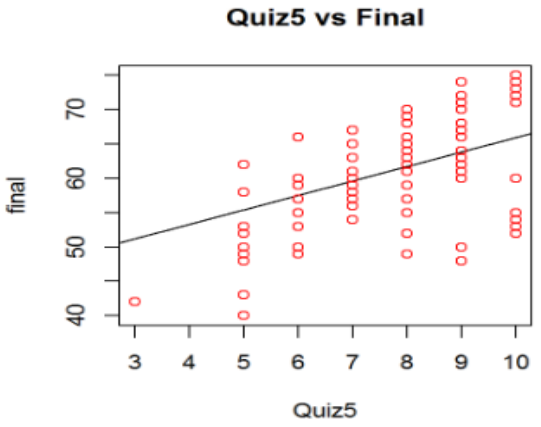
The scaled variables which can be considered as predictors(x-axis) for predicting the response variable final(dependent variable-y axis) can be

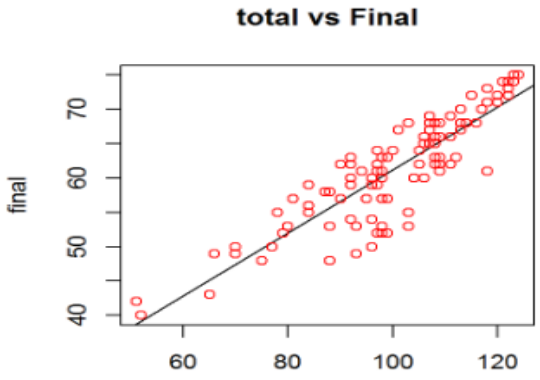
1. gpa
2. quiz1
3. quiz2
4. quiz3
5. quiz4
6. quiz5
7. total
8. percent

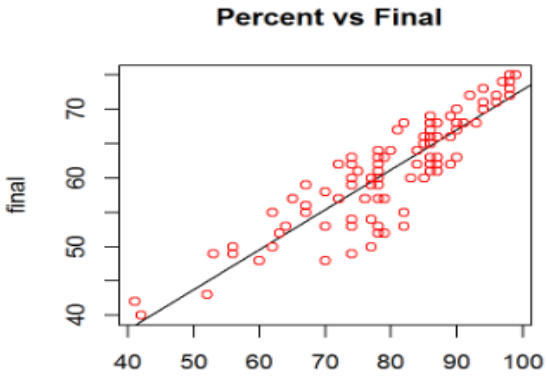
Now let us describe each of these predictors one by one by plotting their scatter plot and try to understand their relationship with final(y)












```
> cor(grades$gpa,grades$final) #correlation between  gpa & final
[1] 0.498055
> cor(grades$quiz1,grades$final) #correlation between  quiz1 & final
[1] 0.5350754
> cor(grades$quiz2,grades$final) #correlation between  quiz2 & final
[1] 0.5518668
> cor(grades$quiz3,grades$final) #correlation between  quiz3 & final
[1] 0.5611773
> cor(grades$quiz4,grades$final) #correlation between  quiz4 & final
[1] 0.4878348
> cor(grades$quiz5,grades$final) #correlation between  quiz5 & final
[1] 0.4715109
> cor(grades$total,grades$final) #correlation between  total & final
[1] 0.8826091
> cor(grades$percent,grades$final)   #correlation between  percent & final
[1] 0.8895457
>
```

It can be understood from data that every predictor is positively correlated with the response variable. More over the scatter plot of all the variables separately show that there is some-way or the other a linear relationship between the response variable and predictor variable.

We can notice that total & percent are the two variables which are highly positively related with final .This is quite obvious as final is a part of the total & percent. Higher the final score higher is total or percent. According to the Scatter plots we can say that the final and predictors are related to each other through a linear line Hence the linearity assumption is true for all.

Question 3: What are R Square and Adjusted R Square of your final model? Show R Output and explain in 3 or 4 lines. Explain the difference between R Square and Adjusted R Square. Which one is superior and why? Explain.

Answer:

Various models can be built based on the predictors available .Let's start building Linear Regression Models . The two best Models built are with

1. final ~ total+quiz3

```
> ft3<-lm(final~total+quiz3,data=grades)
> ft3

Call:
lm(formula = final ~ total + quiz3, data = grades)

Coefficients:
(Intercept)        total        quiz3
      6.674        0.695       -1.892

> summary(ft3)

Call:
lm(formula = final ~ total + quiz3, data = grades)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7702 -1.5943  0.4497  2.1800  4.9337

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.67358    2.11542   3.155  0.00211 **
total        0.69502    0.03282  21.180  < 2e-16 ***
quiz3       -1.89162    0.21754  -8.696 6.19e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.857 on 102 degrees of freedom
Multiple R-squared:  0.8731,    Adjusted R-squared:  0.8706
F-statistic: 350.8 on 2 and 102 DF,  p-value: < 2.2e-16
```

The Best Linear Regression model with condition can have total as a predictor of final scores for the students (who have got scores in all the quiz1-5, final and have gpa available-hence total available with them) or (future students- who dont have any score available with them and an assumption based on estimated total and quiz3 score will predict final score for the students) . Principal can predict using this model that a student with total & quiz3 must have a final score in a particular range. The Linear Regression of total combined with quiz3 will give us explanation of **87.31%** of variance in final (y). For future students who don't have total score available with them cannot use this model to predict final score obviously as they don't have total sores available with them. But Prinipal an take an idea that for future students for this much total & quiz3

score student final score can be predicted and will lie in a particular range as predicted by the model with 87.31% variance in final explained

2. final ~ quiz2+quiz3

```
> f23<-lm(final~quiz2+quiz3,data=grades)
> summary(f23)

Call:
lm(formula = final ~ quiz2 + quiz3, data = grades)

Residuals:
    Min      1Q  Median      3Q     Max
-15.441  -3.720   1.030   4.931   9.839

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.7129     3.1398  12.648  < 2e-16 ***
quiz2         1.5407     0.5311   2.901  0.00456 **
quiz3         1.1862     0.3735   3.176  0.00198 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.381 on 102 degrees of freedom
Multiple R-squared:  0.3671,    Adjusted R-squared:  0.3547
F-statistic: 29.59 on 2 and 102 DF,  p-value: 7.359e-11
```

In this model considering quiz2 & quiz3 exams which happens before the final exams can together help us explain 36.71% variance in the final As p value of both is less than LOS(0.05) so we can say that both quiz2 & quiz3 are significantly related to final(dependent variable) .

For predicting final the two best models selected are ft3 & f23

1. ft3 will be used when predicting with total and quiz3: with explaining 87.31% of variance in final with r square value of .8731 and adjusted R square value of .8706

2. f23 will be predicting with quiz2 & quiz3: explaining 36.71% of variance is R with R-squared value of .8731 and adjusted R square value of .3547

The Formula for R-squared Value is given by SSR/SST which can be reduced to R-Squared = 1 - SSE/SST

Adj-R Squared = [1 - (SSE/degree of freedom of residuals)/(SST/df of model)]

Hence the formula is reduced to

Adj-R Squared = [1 - (SSE/n-k-1)/(SST n-1)]

Out of R-Squared value & Adjusted R-Squared value, Adjusted R squared value can be considered more superior, As it takes into consideration the predictors of the model while calculating the variance for the dependent variable .R-Squared value suppose that every independent variable is responsible for some variation

in the dependent variable whereas Adjusted R - Squared Value gives the percentage for those independent variables which in actual effects the Dependent variable. R-squared measures the proportion of the variation in the dependent variable (Y) explained by the independent variables (X) for a linear regression model whereas, Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. Although there is very less difference in the value of both, so we can consider any one for our consideration.

Question 4 : How do you interpret significance value of *F*-statistics?Show R Output. [Fitness of model]

Answer :

Significance value of F Statistics basically indicates the fit of the Overall Linear regression model built.It tells us whether the model provides a better fit to the data than a model that has no independent variable (in which we take the mean as the only data predicting point) .This test is basically the Hypothesis test for relationship.

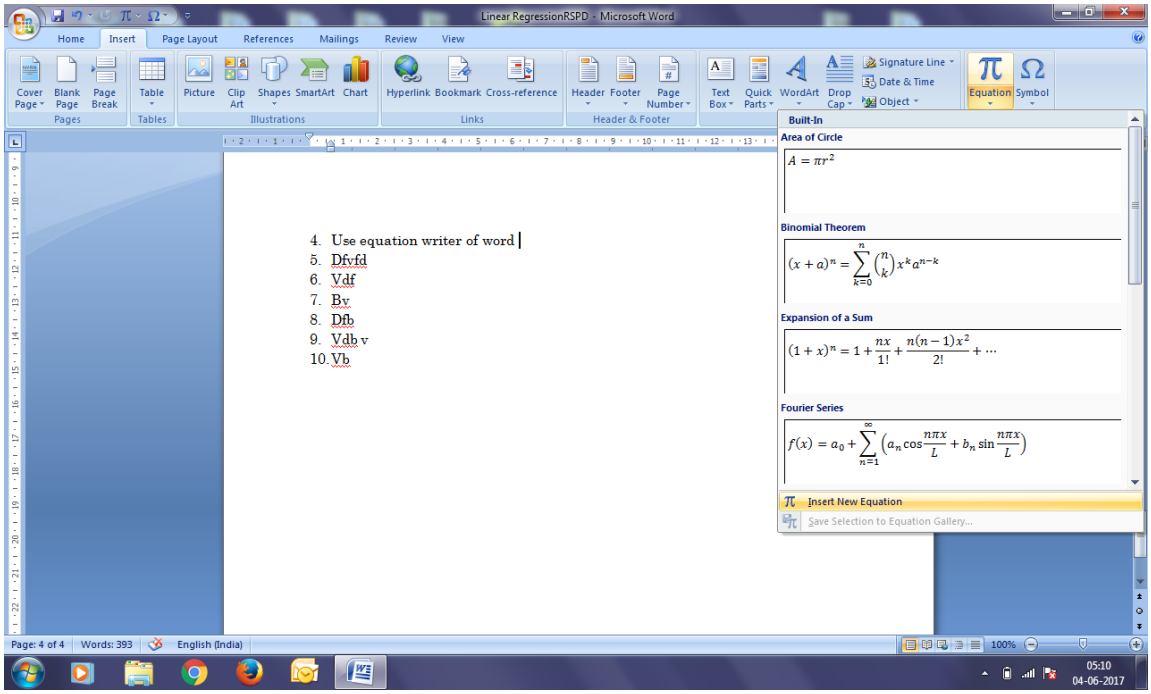
The F-test uses its p-value for overall significance has 2 hypothesis : H0: there is no significant relationship i.e Slope of population is "0".Ha: there is significant relationship .i.e slope of Population is not "0" We compare the p-value of the F-test with LOS(alpha) i.e Level of significance . If p value is less than alpha we can come to the conclusion that the regression model fits the data better as H0 will be rejected and Ha will be accepted. It also mean the model is adding significant predictability for variable y by x. But if p-value of F-test is more than Level of significance we can conclude that our regression model is not significant . Now we have two models for whih the assumptions are to be checked the name of the models are f23 & ft3

```
For f23 our F-test has p-value which is equal to 7.359e-11 which is
almost '0' hence we reject our null Hypothesis for that model is
rejected meaning F-statistic value of 29.59 is significant fit for
our model
Residual standard error: 6.381 on 102 degrees of freedom
Multiple R-squared:  0.3671,    Adjusted R-squared:  0.3547
F-statistic: 29.59 on 2 and 102 DF,  p-value: 7.359e-11

For ft3 our F-test has p-value which is equal to 2.2e-16 which is
almost '0' hence we reject our null Hypothesis for that model is
rejected meaning F-statistic value of 350.8 is significant fit for
our model
```

```
Residual standard error: 2.857 on 102 degrees of freedom
Multiple R-squared:  0.8731,    Adjusted R-squared:  0.8706
F-statistic: 350.8 on 2 and 102 DF,  p-value: < 2.2e-16
```

Question 5 : Use equation writer of word [Insert → Equation → Insert New Equation and write Regression equation of the best model. Show R Output.



Equation for " **ft3**" model → $final = 6.67358 + (.69502 \times total) - (1.89612 \times quiz3)$

Equation for " **f23**" model -> $final = 39.7129 + (1.5407 \times quiz2) + (1.1862 \times quiz3)$

Question 6 : What is Durbin Watson Statistics of your model? How DWS is interpreted? Show how do you find dL and dU and design four boundaries in the sample diagram.. [Explore about Durbin Watson Statistics and table. Table is used for finding dL and dU based on which you will design limits. You need to impose your DWS value in the diagram and decide about presence of autocorrelation] .Show R Output also.

Answer

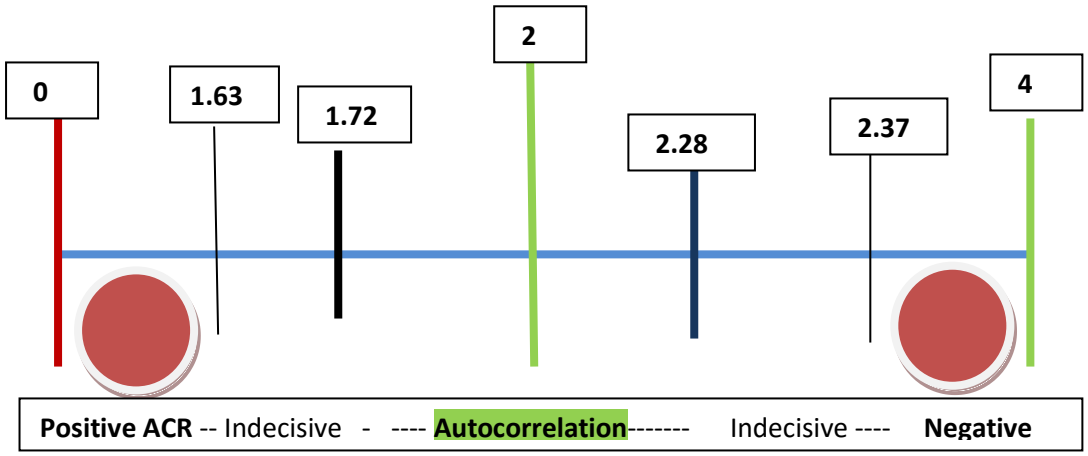
Let's find Durbin Watson Test for our two models

```
> dwt(ft3)  # For ft3 durbin watson Statistics comes out to be 2.115215
 lag Autocorrelation D-W Statistic p-value
   1     -0.07657619      2.115215   0.546
 Alternative hypothesis: rho != 0
> dwt(f23)  #For f23 durbin watson Statistics comes out to be 2.233423
 lag Autocorrelation D-W Statistic p-value
   1      -0.1209658      2.233423   0.246
 Alternative hypothesis: rho != 0
```

The DWS statistics is the number that tests the autocorrelation in the errors/residuals from a regression analysis. It's value will always lie between "0 & 4"" .DWS value of '2' means there is no autocorrelation in the sample,whereas value towards "zero" from '2' means there is +ve autocorrelation & "towards 4"" from '2' means -ve autocorrelation. Autocorrelation is a characteristic of data in which the correlation between the values of the same variables is based on related objects. In this dataset Autocorrelation for variable 'final' it means that the marks of student 2 in final depends on the marks of student 1 in final and that of student 3 in final depends on the marks of student 2 in final and so on. If such autocorrelation occurs we should not build a linear regression model for our model and we just need to move to time analysis Model.

The formula for DWS is [summation {E(i)- E(i-1)^2} / summation {E(i)^2}] , which is quite complex .Here **ei** means the expected value based on the regression equation which we have generated . As number of predictors in our model is 2 hence we need to look at the table for DWS with n= 105 & k=2 to obtain dl & du. We need to find these value using the table with alpha=0.05. We get the value of dl= 1.63 & du = 1.72

The Durbin Watson table scale for our models ft3 & f23 is as shown below. Both models have 2 predictors so the scale of DWS would be same for both models.



How to interpret Durbin Statistics Value

After finding the dl & du value we place them on the scale of Durbin Watson .If the value of DWS we found or our model is lying between the zone of "No auto correlation" than we can be sure of "no autocorrelation" & hence we can proceed with the Linear Regression Model. If the value of DWS lies in Positive or Negative range of the scale than for sure Linear Regression model cannot be applied as it shows there is relationship between the observation values. Now if it comes in Indecisive zone the benefit of the doubt goes with the Researcher. Most of the times, he continues with the Linear Regression model results. As our Durbin statistics is "2.11 for ft3 model" & "2.23 for f23 model" both of which lies in No-Autocorrelation region we can say that there is no-autocorrelation and thus we can proceed further with the results of linear regression models.

Question 7: What is VIF for each predictor/s? How do you interpret VIF or what VIF signifies? Max 5 lines. [VIF (Variance Inflation Factor). Show R Output.

Answer

VIF means Variance Inflation factor This inflation factor is basically used to calculate the multi-collinearity between the predictors .If VIF of 2 or more predictors/variables is more than 5 generally it means they are highly correlated, we can remove any one/more of them which we feel is less significant for the model .It is a simple approach to identify collinearity among explanatory variables. VIF calculations are straightforward and easily comprehensible; the higher the value, the higher the collinearity. A VIF for a single explanatory variable is obtained using the r-squared value of the regression of that variable

against all other explanatory variables. In our model ft3 & f23 VIFs for predictors is as below

➔ **vif for our Models For ft3 & f23 models**

```
> vif(ft3)
   total    quiz3
3.210517 3.210517
> vif(f23)
   quiz2    quiz3
1.897984 1.897984
```

We can also calculate VIF for any predictor by interchanging the it with the response variable and getting the value of R-squared & then plugging it in the formula for Variance Inflation factor given below:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Question 8 : How do you interpret the significance of slope of predictors based on sig. Value or p-value associated with *t*-statistics of each predictor/s. Show R Output.

Answer:

The interpretation of significance of the slope of the predictors based on significance value or p-value associated with the t statistics is comparing it with the level of Significance.

A t-test is always associated with a hypothesis testing .In the Regression analysis a t-test is done for testing the slope of the predictor whether it has any significant relation with the dependent variable or not .and this t-test has p value associated with it which is compared with the LOS (alpha) So this interpretation tells us whether slope of predictor has some linear relationship with response variable or not.

The hypothesis associated with the testing of the slope is as follows for Simple Regression. Ho: Beta1=0, there is no significant relationship of slope of predictor. Ha: Beta1 != zero, there is significant relationship with slope of that predictor

The hypothesis associated with the testing of the slope is as follows for multiple Regression. Ho: Beta1=0, there is no significant relationship. Ha: at least one

slope of one predictor is not equal to “zero”, there is significant relationship with slope of that predictor

Significance value or p-value for our t-test is used for the above hypothesis testing for slope test. If we get to know that the p-value > LOS(alpha=0.05) then null hypothesis will be accepted and it will mean that slope of that variable has no significant effect on the dependent variable & if we p-value< LOS(alpha=0.05) ,we reject the null hypothesis meaning that there is significant relationship between them

For our Model “ft3” the value of Pr(>|t|) or p-value for our predictor total is 2e-16 & for quiz3 is 6.19e-14 .Both almost equal to “zero” and less than LOS meaning both the predictors have significant slope for our model (p-value< LOS (alpha=0.05))

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.67358    2.11542   3.155  0.00211 **
total        0.69502    0.03282  21.180  < 2e-16 ***
quiz3       -1.89162    0.21754  -8.696 6.19e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For our Model “f23” the value of Pr(>|t|) or p-value for our predictor quiz2 is -.00456 & for quiz3 is 0.00198 .Both less than LOS meaning both the predictors have significant slope for our model.(p-value< LOS (alpha=0.05))

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.7129     3.1398  12.648  < 2e-16 ***
quiz2         1.5407     0.5311   2.901  0.00456 **
quiz3         1.1862     0.3735   3.176  0.00198 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Before moving to assumptions questions let us add few columns to our data set

```
## error values of final through the models are below. These values are giving us the residuals of our model



grades$errft3<-residuals(ft3)
grades$errft3 # For Model ft3


grades$errf23<-residuals(f23)
grades$errf23

# Adding observation no.s against each row in the data set grades

grades$obsno<-c(1:105)
grades$obsno


#For model ft3,inserting the predicted values in the grades dataset  by column creation

predft3<-predict(ft3)
grades$predft3<-predft3
grades$predft3

#For Model f23,inserting the predicted values in the grades dataset  by column creation

predft3<-predict(ft3)
grades$predft3<-predft3
grades$predf23
```

Question 9: Test the assumption of Normality and interpret your findings. Show histogram and interpret.

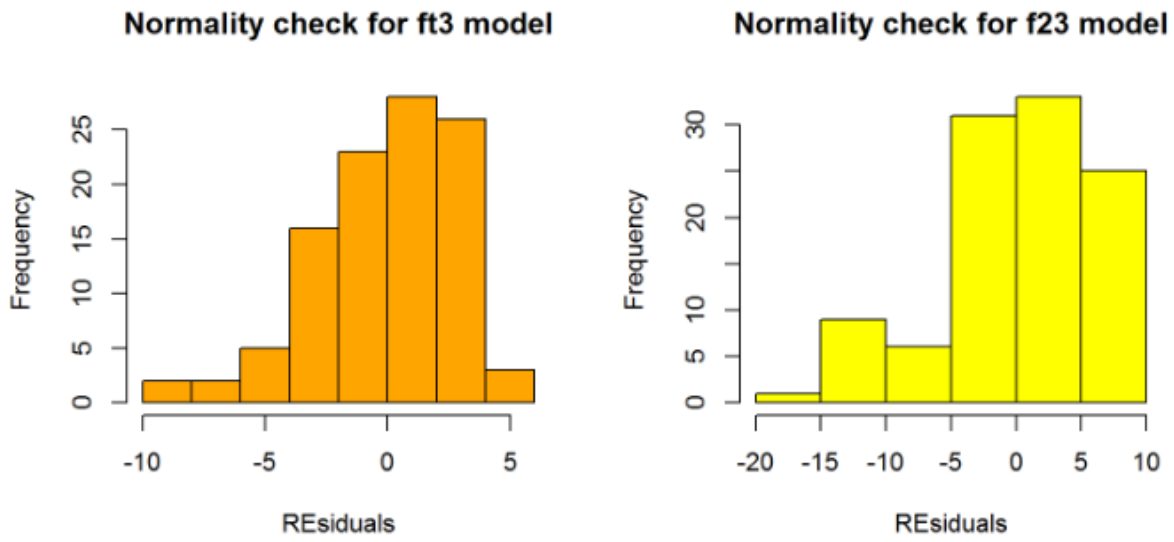
Answer :

Assumptions Test for ft3 & f23 Models

1. Normality test for ft3 & f23

*For Ft3 & F23 let us draw histogram for both the models of error

```
hist(grades$errft3,main = "Normality check for ft3 model", xlab="REsiduals",col="orange")
hist(grades$errf23,main = "Normality check for f23 model", xlab="REsiduals",col="yellow")
```



In both the models the data is almost normally distributed but little skewed towards the left. It means it is a bit left skewed. But we can consider that the assumption of normality does hold.

Question 10: Test the assumption of Independent of observations and interpret the plot.

Answer:

2. Independent of observations

*For Ft3 & F23 let us draw scatter pot for residuals & the Observations the models of errors

```
plot(grades$obsno,grades$errft3,main="Independence of  error for ft3",xlab= " obsv no", ylab="residuals")
plot(grades$obsno,grades$errf23,main="Independence of  error for f23",xlab= " obsv no", ylab="residuals")
```

We can observe from the graphs for both the models the observations are every independent from each other as we can see the spread of the data points which are widely spread indicating the independence of observations for the model ft3 and f23 .

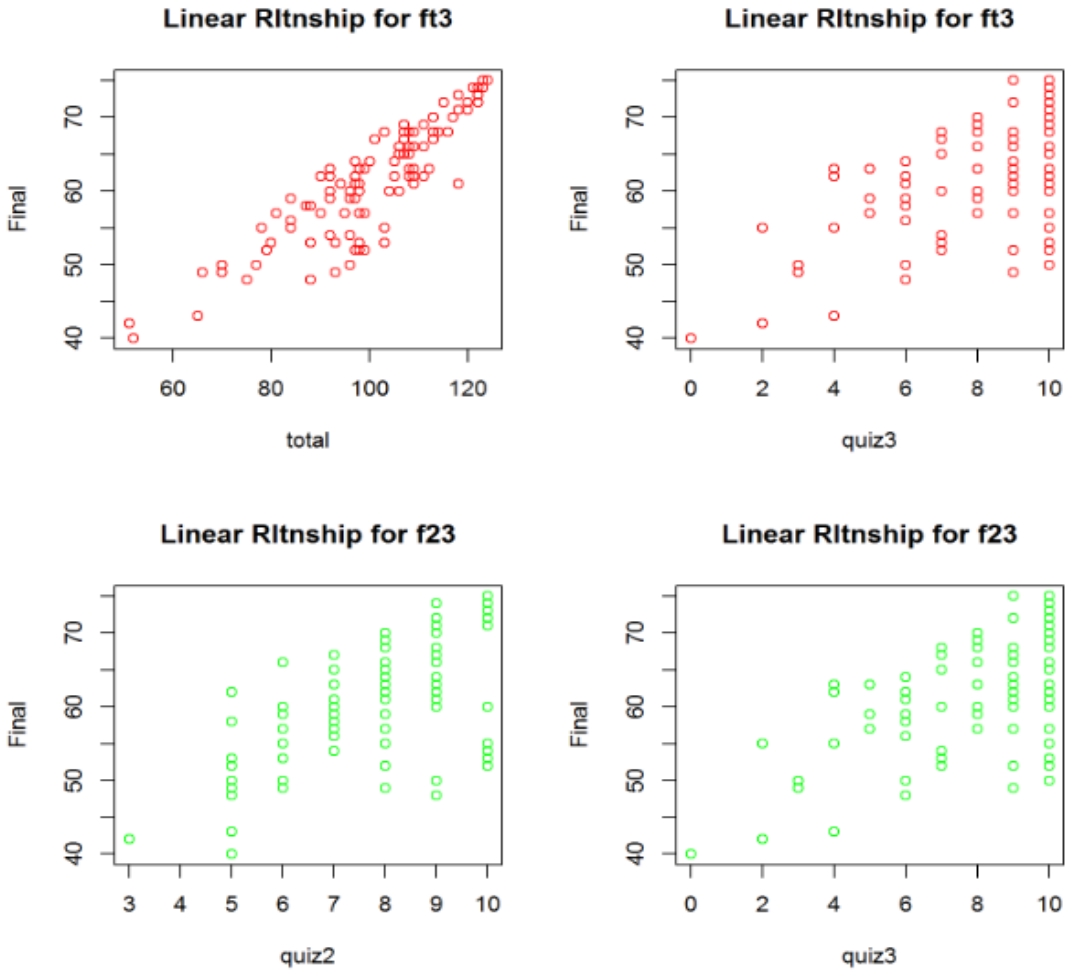
Question 11 :Test the assumption of linear relationship and interpret for each predictor .If more than one predictor is used in model then more scatter plots would be required

Answer :

3. Check of linear relationship

```
*For ft3 let us draw scatter plot for final & total and final and quiz3
*For f23 let us draw scatter plot for final & quiz2 and final  and quiz3
```

```
plot(grades$total,grades$final,main="Linear Rltnship for ft3",xlab="total",ylab="Final",col="red")
plot(grades$quiz3,grades$final,main="Linear Rltnship for ft3",xlab="quiz3",ylab="Final",col="red")
plot(grades$quiz2,grades$final,main="Linear Rltnship for f23",xlab="quiz2",ylab="Final",col="green")
plot(grades$quiz3,grades$final,main="Linear Rltnship for f23",xlab="quiz3",ylab="Final",col="green")
```



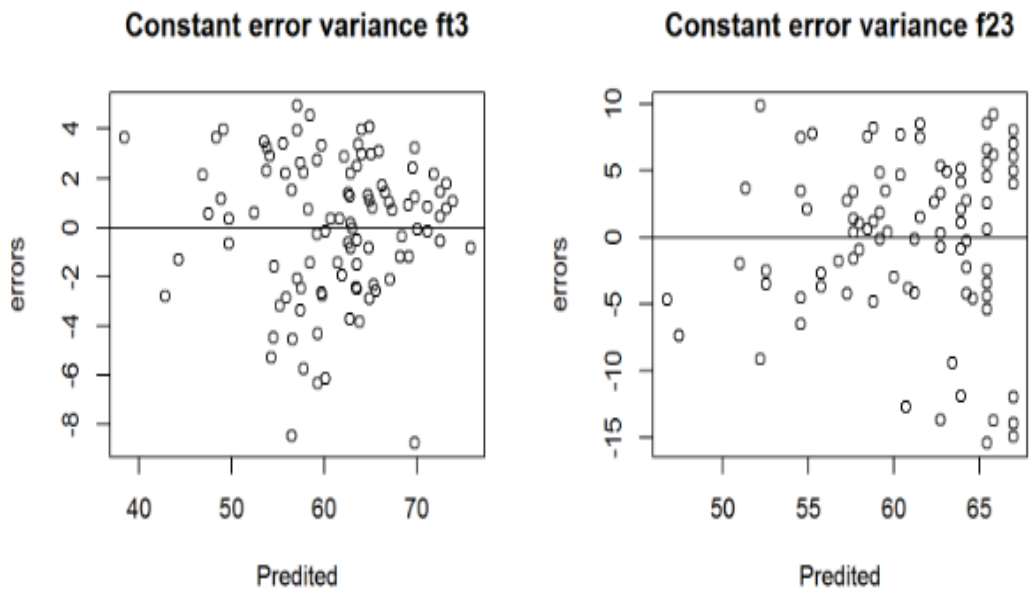
The assumption of linear relationship can be seen in all the predictors of both the models .Hence we can say that the assumption of linear Relationship or linearity is valid for both the models ft3 & f23.

Question 12: Test the assumption of Constant Error Variance and interpret

Answer:

4. Constant Error Variance : Homoscedacity

```
*For ft3 let us draw  scatter pot  for residuals   & Predicted
*For f23 let us draw  scatter pot  for residuals   & Predicted
```



As we can see in the scatter plot that both the plots have a great spread and looks data is randomly spread throughout hence our data Is having constant error variance in both the models ft3 & f23 .

Question No 13 : What is Standard Error of Estimate of your model and how do you interpret the same. Show with some hypothetical values of predictors.

Answer:

The standard error of estimate is a measure of the accuracy of predictions. In a regression line, the smaller the standard error of the estimate is, the more accurate the predictions are.

This standard error of estimate is the used to give us the confidence intervals at a particular level of confidence intervals. It is never good to give predictions in the Point estimate form so we give them in a range using standard error of estimate.

Say, you want to predict final for a given total & quiz3 marks using ft3 model equation

```
final =     6.67358 + (.69502*total) - (1.89612*quiz3)
Let's find for total= 5 & quiz3=7
final =     6.67358 + (.69502*80) - (1.89612*7)
final= 49.034
```

We can also find the upper & lower range or the confidence interval using Standard error of estimate & the predicted values for ft3 model Ul of final & Ll of final

Now, let's find predicted value using f23 model whose equation is given by

```
final =     39.7129 + (1.5407*quiz2) + (1.1862*quiz3)

Let's find for quiz2= 5 & quiz3=7

final =     39.7129 + (1.5407*5) + (1.1862*7)

final= 55.7198
```

We can also find the upper & lower range or the confidence interval using Standard error of estimate & the predicted values of f23 model Ul of final & Ll of final.

For Model ft3 : Standard Error of Estimate is 2.857

```
Residual standard error: 2.857 on 102 degrees of freedom
Multiple R-squared:  0.8731,    Adjusted R-squared:  0.8706
F-statistic: 350.8 on 2 and 102 DF,  p-value: < 2.2e-16
```

For Model f23 : Standard Error of Estimate is 6.381

```
Residual standard error: 6.381 on 102 degrees of freedom
Multiple R-squared:  0.3671,    Adjusted R-squared:  0.3547
F-statistic: 29.59 on 2 and 102 DF,  p-value: 7.359e-11
```

Question no 14: Congratulation! You have done a marvellous job indeed and build your first predictive model. M just reminding that regression model is somewhere 50% of a data analyst routine job and has great importance in practical world.

Now write a summary of your findings which you will show to your reporting manager (before forwarding the model to your client/Principal in this case). This time, no R Output and minimum pictures are needed. Mind it, your reporting manager is a senior statistician/data scientist and do not have time to go into your entire work. He will prefer to read meaningful, to the point and technically correct summary! Here is your chance to impress your boss!

Answer :

Grades.csv consists of 105 observations and 22 variables (mostly categorical or nominal) .The data tells about the details of the students giving gender, ethnicity, class sections ,marks in 5 quizzes ,marks in final, total marks, gpa and percentage and also tell whether the student is pass or fail .

As we the main motive of the project was to build a predictive model for predicting final for consumption I have built two linear regression models to predict final.

i) Model to predict final using quiz2 & quiz3

One of our Linear Regression Model predicts the final score using the performance in quiz2 & quiz3 combined. The name of this model is f23 This multiple regression model helps us explain the variance in the final with 36.71% confidence and the remaining 63.29% is due to the other factors like quiz1 quiz2 quiz3 quiz4 quiz5 gpa or total etc. The model predictors have significant slope with the dependent or the response variable final which is checked by the significance value for t-test of the predictors i.e the p-value. Moreover the Variance Inflation factor for both the variables is 1.89 which is in acceptance zone meaning these two predictors are not highly correlated .Also the DWS for the model is 2.233 which lies in the no autocorrelation zone meaning our linear regression is a right model for predicting . Standard error of estimate of our model is 6.381 means that with this much standard deviation range we would be able to predict our final range with 95% confidence level. The equation of the model of predicting final using quiz2 & quiz3 with 36.71% variance in final explained is given as below:

$$\boldsymbol{final} = \mathbf{39.7129} + (\mathbf{1.5407} \times \boldsymbol{quiz2}) + (\mathbf{1.1862} \times \boldsymbol{quiz3})$$

The f-test significance value i.e p-value of –f-test is less than LOS meaning that the test of overall model is significant with the final dependent variable.

The assumptions of our models like normality, Linear relationship ,constant error variance are almost satisfied with the data provided and calculated .so we can say that our Model I right way to predict final using quiz2 & quiz3 scores

ii) Model for predicting final is using the variables total and quiz 3

The name of the model is ft3. This multiple regression model helps us explain the variance in the final with 87.31% confidence and the remaining 12.69% is due to the other factors like quiz1 quiz2 quiz3 quiz4 quiz5 gpa alone or grouped together. The model predictors have significant slope with the dependent or the response variable final which is checked by the significance value for t-test of the predictors i.e the p-value of t test for model. P-value for t-test is 2e-16 and for quiz3 it is 6.192e-14.

The equation for the model ft3 is given by :

$$\boldsymbol{final} = \mathbf{6.67358} + (\mathbf{.69502} \times \boldsymbol{total}) - (\mathbf{1.89612} \times \boldsymbol{quiz3})$$

Variance Inflation factor for both the variables is 3.21 which is in acceptance zone meaning these two predictors are not highly correlated. Also the DWS for the model is 2.113 which lies in the no autocorrelation zone meaning our linear regression is a right model for predicting .Standard error of estimate of our model is 2.857 means that with this much standard deviation range we would be able to predict our final range with a particular LOS defined. The f-test significance value i.e p-value of –f-test is less than LOS meaning that the test of overall model is significant with the final dependent variable. The assumptions of our models like normality, Linear relationship ,constant error variance are very well satisfied with the data provided and calculated .so we can say that our Model I right way to predict final using quiz2 & quiz3 scores

This model can be used by the principal if she doesn't have final score with her. and she already have total score with her . Or also this model is used to predict a final score using the predictors combination of total and quiz3 with 87.31% variance in the model.

The assumptions of our models like normality, Linear relationship ,constant error variance are almost satisfied with the data provided and calculated .so we can say that our Model I right way to predict final using quiz2 & quiz3 scores

Question No 15: This is final stroke! Besides your boss, your client is equally or rather more important to you!
Your challenge is this that the Principal/client is not statistics savvy! You need to summarize your work/findings in a non-statistical manner or in a lay man manner and this is indeed challenging. However, no way out and you have to do it in a simple but impressive manner (impressive to client!). Write down summary.

Answer:

There are two models build to predict the final score of the total either one can predict using quiz1 marks & quiz2 marks together or using total and quiz3 marks together .For both the models equations have been developed and by plugging in the values one can obtain the predicted final score.

The equations derived by the models of the regression plane are :

$$\boldsymbol{final} = \mathbf{6.67358} + (\mathbf{.69502} \times \boldsymbol{total}) - (\mathbf{1.89612} \times \boldsymbol{quiz3})$$

$$\boldsymbol{final} = \mathbf{39.7129} + (\mathbf{1.5407} \times \boldsymbol{quiz2}) + (\mathbf{1.1862} \times \boldsymbol{quiz3})$$

For Example : if a student gets 100 marks in total & 8 marks in quiz3 .The predicted final score will for model named ft3 will be

$$\boldsymbol{final} = \mathbf{6.67358} + (\mathbf{.69502} \times \mathbf{100}) - (\mathbf{1.89612} \times \mathbf{9})$$

Using the equation we get the final value as **59.11**. So basically for a total of 100 and quiz3 score of 8 final predicted score as per model is 59.11. This predicted value of final is calculated with a spread or variance of 87.31% using these two predictors .

Similarly a second model name f23 which predicts the final score using the quiz2 & quiz3 score will predict final score for quiz2=8 and quiz3=9 ,equals to **62.7143** with a 36.71% variance in final which is explained using these two predictors

$$\boldsymbol{final} = \mathbf{39.7129} + (\mathbf{1.5407} \times \mathbf{8}) + (\mathbf{1.1862} \times \mathbf{9})$$

As we can understand these predicted final values are point estimates ,so we need to find out the range/intervals of final that will help us predict the scores to a better extent keeping some deviations in mind .These are called confidence intervals. They are calculated using the standard error of estimate. For our

model which uses total and quiz3 for predicting final sore has a standard error of estimate as **2.857** & the model which uses quiz2 and quiz3 for predicting final sore has a standard error of estimate as **6.381**

For example: If we want to predict a confidence interval for a predicted score of 59.11 which is obtained using our model and we will use the formula for model named **ft3**

- **Upper limit of Confidence interval** = Predicted score + (z Critical value at confidence level selected * Standard error of estimate for model)

  = 59.11 + (z critical value at 95% confidence level * 2.857)

  = 59.11 + (1.96* 2.857)

  = 64.70

- **Lower limit of Confidence interval** = Predicted score - (Critical value at confidence level selected * Standard error of estimate)

  = 59.11 - (z critical value at 95% confidence level * 2.857)

  = 59.11 - (1.96* 2.857)

  = 53.40

Similarly If we want to predict a confidence interval for a predicted score of 62.7143 which is obtained using quiz2 & quiz3 scores we will use the below formula for model named **f23**

- **Upper limit of Confidence interval** = Predicted score + (z Critical value at confidence level selected * Standard error of residuals for model)

  = 62.7143 + (z critical value at 95% confidence level * 2.857)

  = 62.7143+ (1.96* 6.381)

  = 62.7143 +

  = 75.22

- **Lower limit of Confidence interval** = Predicted score - (Critical value at confidence level selected * Standard error of estimate)

  = 62.7143 - (z critical value at 95% confidence level * 2.857)

  = 62.7143 - (1.96* 6.381)

  = 62.7143 – 12.506

= 50.20

This is as simple explanations to predict the final scores based on the best models which can be built using predictor variables available.

The Model named ft3 which uses total as a predictor can help us predict final only if either total score is available with or total sore is assumed and based on it the final score obtained by model can be predicted .

Question No 16: Now time to show case your work to rest of the world! Prepare a website as per the sample attached which is only a guideline. Apply your creativity and make it really impressive.

Answer:

The link of the website for the project is as mention below-:

www.mhtdsm.wixsite.com/linearregression