Grades

Satyam Prakash

10 January 2018

library(readr)  
grades<- read\_csv("G:/Data/Grades/grades.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_integer(),  
## lastname = col\_character(),  
## firstname = col\_character(),  
## gpa = col\_double(),  
## grade = col\_character(),  
## passfail = col\_character()  
## )

## See spec(...) for full column specifications.

dim(grades)

## [1] 105 22

colnames(grades) # this will give us the names of the 22 variables

## [1] "Sr\_No" "id" "lastname" "firstname" "gender"   
## [6] "ethnicity" "year" "lowup" "section" "gpa"   
## [11] "extrc" "review" "quiz1" "quiz2" "quiz3"   
## [16] "quiz4" "quiz5" "final" "total" "percent"   
## [21] "grade" "passfail"

We can see that the data which we have imported has 105 observartions and 22 variables. In these 22 Variables we have continous variables as well as categorial variable. We need to be be concerned with continous or scaled data variables as predictors in order to build our model.Firstly lets study the number of students in diffrent categorial variables

library(psych)

## Warning: package 'psych' was built under R version 3.4.3

table(grades$gender)

##   
## 1 2   
## 64 41

table(grades$ethnicity)

##   
## 1 2 3 4 5   
## 5 20 24 45 11

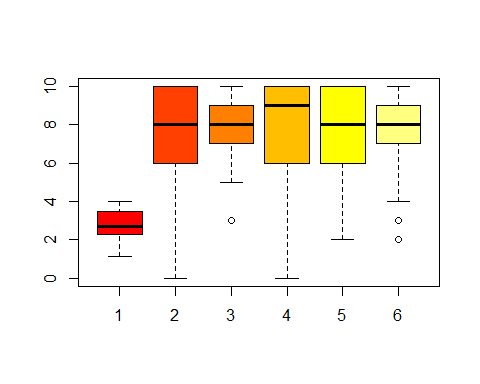
table(grades$passfail)

##   
## F O P   
## 6 1 98

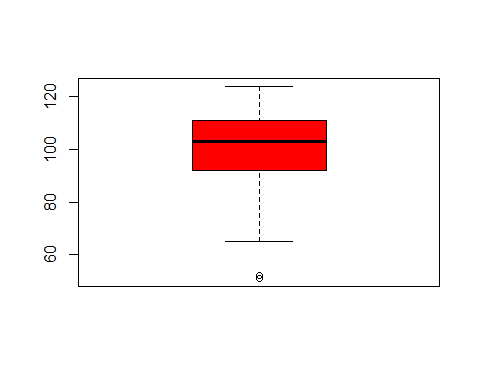
####In the boxplots below we have on X-Axis 1=gpa 2 =quiz1 3=quiz2,4=quiz3,5=quiz4,6=quiz5 & second boxplot is for GPA

These are drawn to check the normality of data in that variable

boxplot(grades$gpa,grades$quiz1,grades$quiz2,grades$quiz3,grades$quiz4,grades$quiz5,col=heat.colors(6))



boxplot(grades$total,col="red")

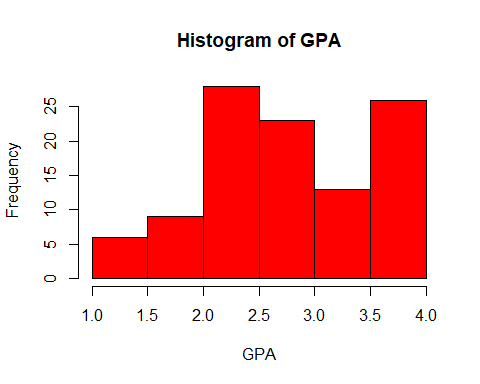


1. Boxplots & Histograms are drawn to check the normality of data for that variable

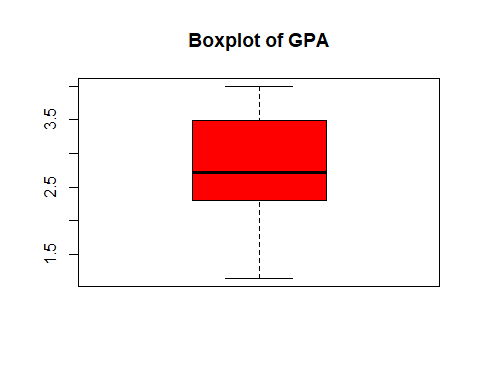
library(psych)  
describe(grades$gpa)

## vars n mean sd median trimmed mad min max range skew kurtosis  
## X1 1 105 2.78 0.76 2.72 2.8 0.76 1.14 4 2.86 -0.05 -0.87  
## se  
## X1 0.07

hist(grades$gpa,xlab="GPA",ylab=" Frequency",main="Histogram of GPA",col="red")



boxplot(grades$gpa,col="red",main= "Boxplot of GPA")

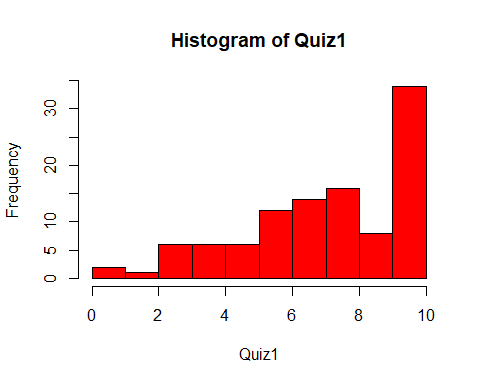


1. Quiz1 Data states that mean is 7.47 sd is 2.48.By the histogram and boxplot we can say that data is skewed towards towards the left that is it is negatively skewed as skewness value is -0.83 also Kurtosis value is 0.04 which is close to 0 hence our data is having heavier tails and sharper peaks an is a leptokurtic distribution.Alsorange of quiz1 is 0 to 10

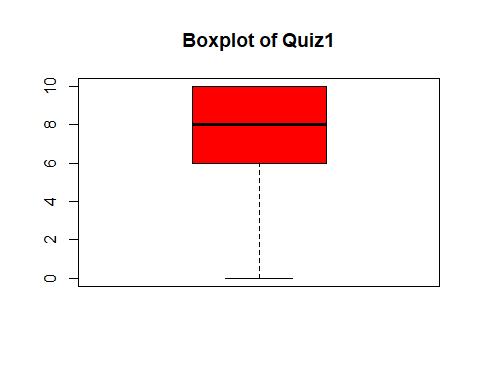
describe(grades$quiz1)

## 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

hist(grades$quiz1,xlab="Quiz1",ylab=" Frequency",main="Histogram of Quiz1",col="red")



boxplot(grades$quiz1,col="red",main= "Boxplot of Quiz1")

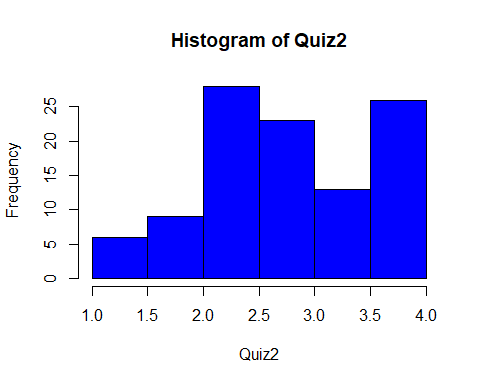


1. Quiz2 Data states that mean value is 7.98 and sd of 1.62. By histogram states that data is skewed little towards the left and boxplot states that data is a almost normally distributed with an outlier in it which is 3 but it is also little skewed towards left.Range of quiz2 is 3 to 10 .Skewness value of -0.64 means a little left skewedkurtosis value of -0.35 means a platykurtic with lighter tails and a flat central peak

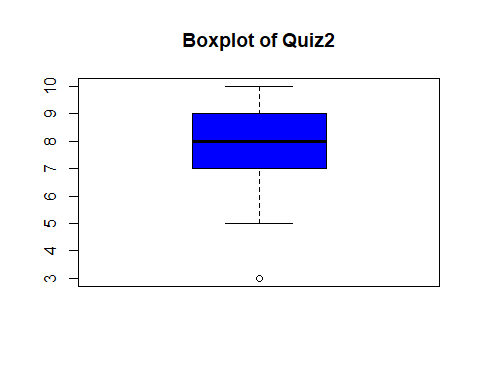
describe(grades$quiz2)

## 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

hist(grades$gpa,xlab="Quiz2",ylab=" Frequency",main="Histogram of Quiz2",col="blue")



boxplot(grades$quiz2,col="blue",main= "Boxplot of Quiz2")

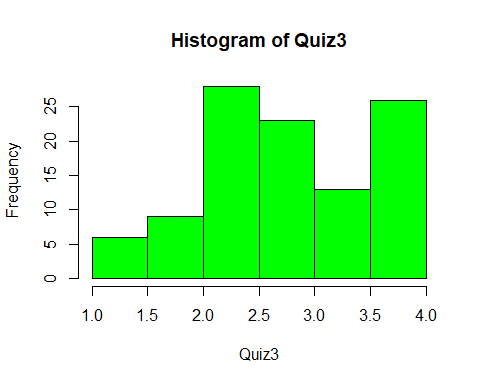


1. Quiz3 Data states that mean value is 7.98 and sd of 2.31(means the observations can have standard dev.of 2.31).By histogram we concluded that data is more towards the left and boxplot states that data is a not normally distributed but completely skewed towards left.Range of quiz3 is 0 to 10 ,Skewness 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

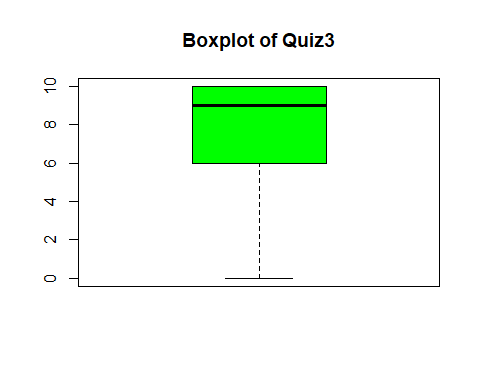
describe(grades$quiz3)

## 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

hist(grades$gpa,xlab="Quiz3",ylab=" Frequency",main="Histogram of Quiz3",col="green")



boxplot(grades$quiz3,col="green",main= "Boxplot of Quiz3")

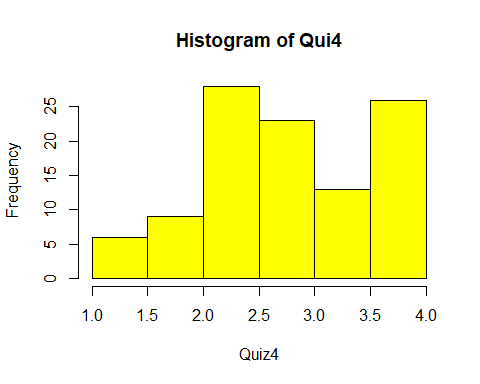


1. Quiz4 Data in quiz4 states that mean value is 7.98 and sd of 2.28(means the observations can have standard dev.of 2.28).By histogram we find that data is more towards the left and boxplot states that data is a not normally distributed but completely skewed towards left.Range of quiz4 is 0 to 10 .Skewness 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

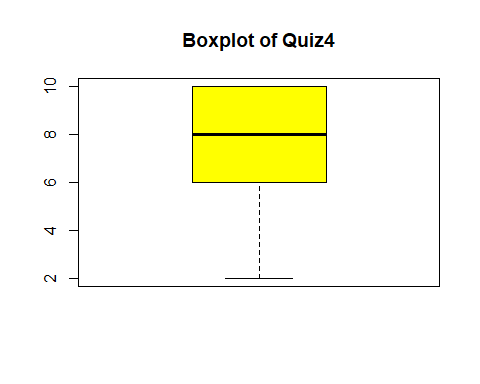
describe(grades$quiz4)

## 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

hist(grades$gpa,xlab="Quiz4",ylab=" Frequency",main="Histogram of Qui4",col="yellow")



boxplot(grades$quiz4,col="yellow",main= "Boxplot of Quiz4")

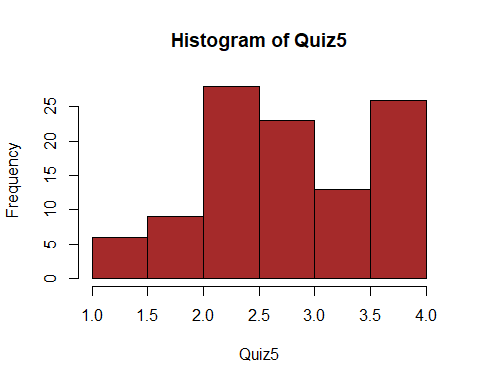


1. Quiz5 Data states that mean value is 7.87 and sd of 1.77 (means the observations can have standard deviation of 1.77) .By the histogram we conclude that data is spread towards the left side more and boxplot states that data is a little more towards the left side but can be said as normally distributed.Range of quiz5 is 0 to 10 ,.Sskewness 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

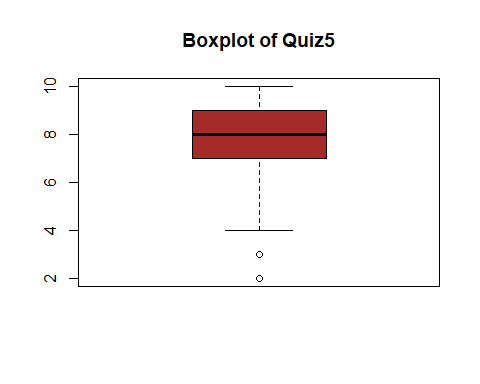
describe(grades$quiz5)

## 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

hist(grades$gpa,xlab="Quiz5",ylab=" Frequency",main="Histogram of Quiz5",col="brown")



boxplot(grades$quiz5,col="brown",main= "Boxplot of Quiz5")

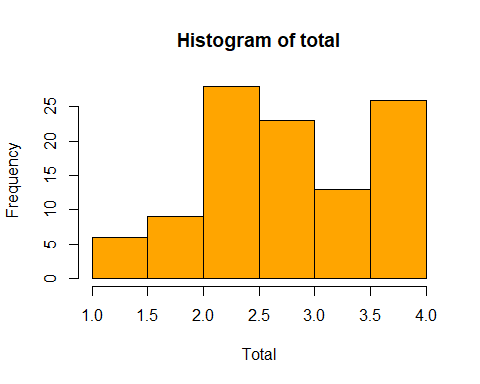


1. Total Data states that mean is 100.57 & sd is 15.3.Also Mean and trimmed mean, sd and mad are very near hence no outliers and no variability, most data on right side hence it is left skewed. Skeweness value of -.081 means data is left skewed or can be negatively skewed.Data is leptokurtic meaning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

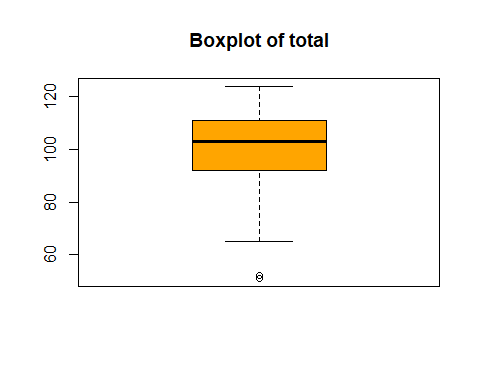
describe(grades$total)

## 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

hist(grades$gpa,xlab="Total",ylab=" Frequency",main="Histogram of total",col="orange")



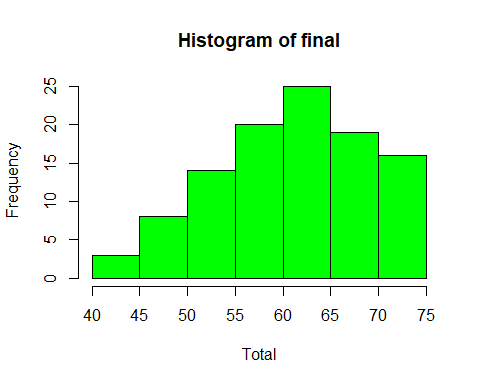
boxplot(grades$total,col="orange",main= "Boxplot of total")

 8. Response Variable Final Data in the final variable states that it has mean of 61.48.final is numeric with range 40-75 which can be seen from min & max.Standard deviation is 7.94 which means that the 68% of the deviations are within the zone of [ 61.47 + 17.94] or [61.48- 17.94] i.e b/w53.54 & 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 after arranging 62 is the median final score of students. 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 whih is 40. trimmed mean of final is 61.74 .which is obtained by removing the observations which are quite far from the other observations or in one way quite far from the mean. 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.Hence the data is platykurtic as per the Kurtosis value

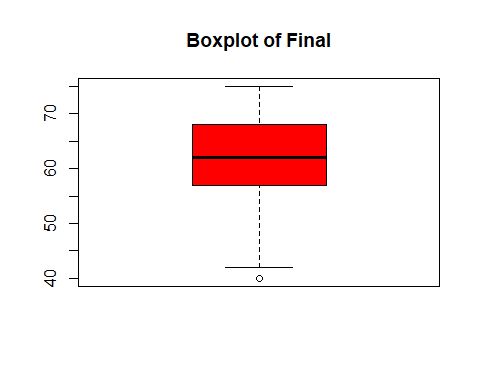
describe(grades$final)

## 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

hist(grades$final,xlab="Total",ylab=" Frequency",main="Histogram of final",col="green")



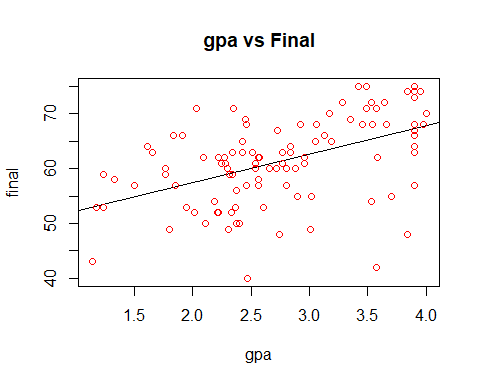
boxplot(grades$final,col="red",main= "Boxplot of Final")



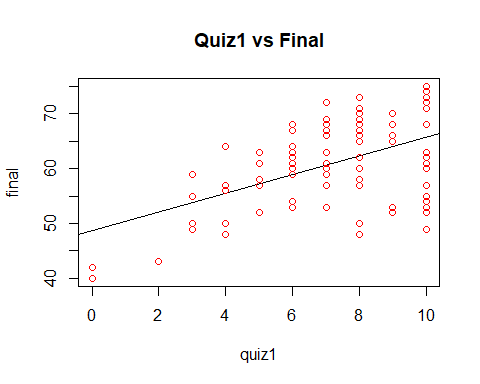
Question 2. Let us describe each of these predictors one by one by plotting their scatter plot and try to understand their relationship with final(y)

Scatter plots Predictor Vs Response Variable

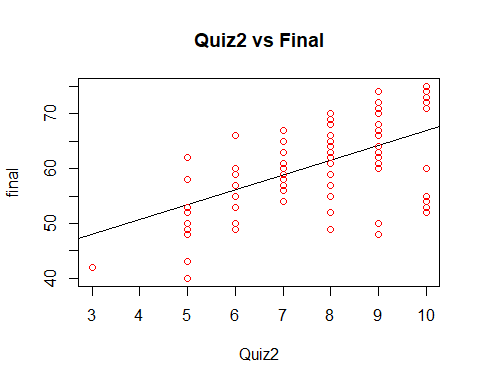
plot(grades$gpa,grades$final,main= "gpa vs Final" , xlab = "gpa", ylab = "final",col="red",abline(lm(final~gpa,data=grades)))



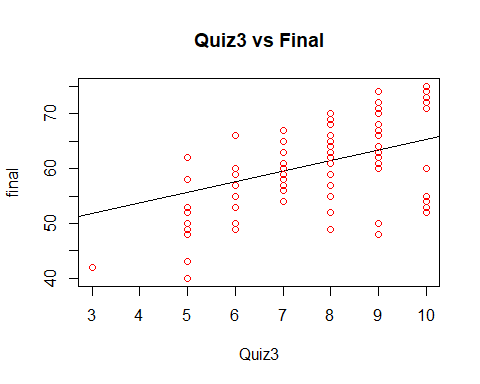
plot(grades$quiz1,grades$final,main= "Quiz1 vs Final" , xlab = "quiz1", ylab = "final",col="red",abline(lm(final~quiz1,data=grades)))



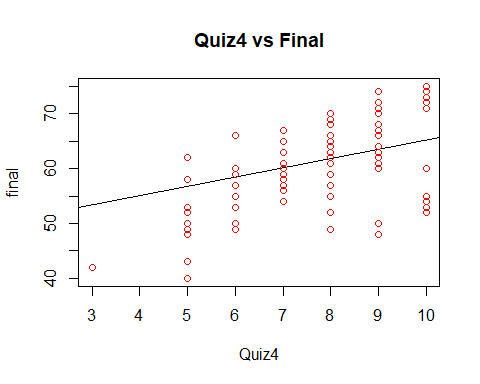
plot(grades$quiz2,grades$final,main= "Quiz2 vs Final" , xlab = "Quiz2", ylab = "final",col="red",abline(lm(final~quiz2,data=grades)))



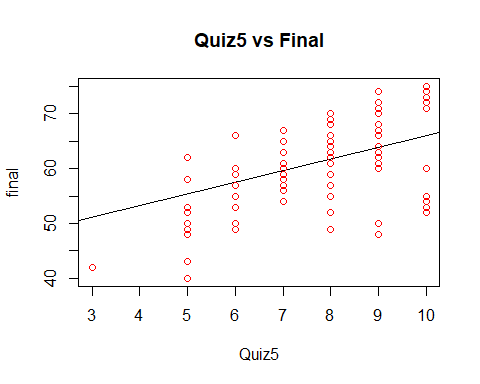
plot(grades$quiz2,grades$final,main= "Quiz3 vs Final" , xlab = "Quiz3", ylab = "final",col="red",abline(lm(final~quiz3,data=grades)))



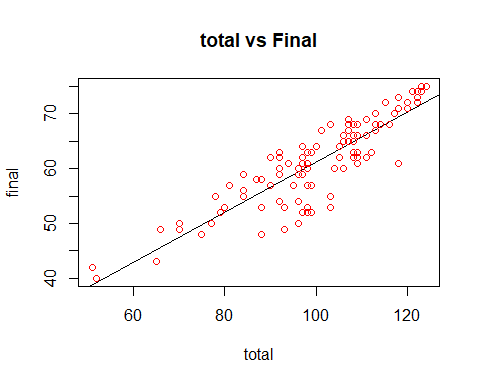
plot(grades$quiz2,grades$final,main= "Quiz4 vs Final" , xlab = "Quiz4", ylab = "final",col="red",abline(lm(final~quiz4,data=grades)))



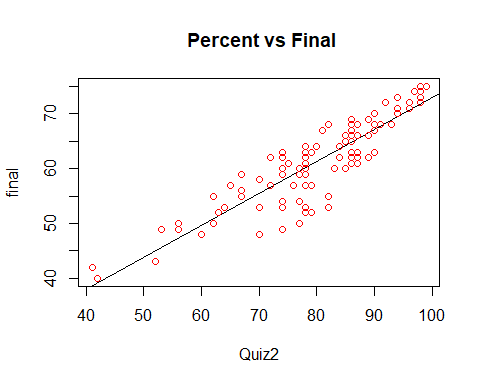
plot(grades$quiz2,grades$final,main= "Quiz5 vs Final" , xlab = "Quiz5", ylab = "final",col="red",abline(lm(final~quiz5,data=grades)))



plot(grades$total,grades$final,main= "total vs Final" , xlab = "total", ylab = "final",col="red",abline(lm(final~total,data=grades)))



plot(grades$percent,grades$final,main= " Percent vs Final" , xlab = "Quiz2", ylab = "final",col="red",abline(lm(final~percent,data=grades)))



Now checking Correlation of each of these predictors with the response variable

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

Hence all the predictors are positively corelated with the final

Question 3

Let us build the linear regression Model

library(car)

## Warning: package 'car' was built under R version 3.4.2

##   
## Attaching package: 'car'

## The following object is masked from 'package:psych':  
##   
## logit

final ~ gpa+quiz1+quiz2+quiz3+quiz4+quiz5

## final ~ gpa + quiz1 + quiz2 + quiz3 + quiz4 + quiz5

fg12345<-lm(final~gpa+quiz1+quiz2+quiz3+quiz4+quiz5,data=grades)  
fg12345

##   
## Call:  
## lm(formula = final ~ gpa + quiz1 + quiz2 + quiz3 + quiz4 + quiz5,   
## data = grades)  
##   
## Coefficients:  
## (Intercept) gpa quiz1 quiz2 quiz3   
## 32.8658 3.6271 0.2986 0.8004 0.9063   
## quiz4 quiz5   
## -0.1428 0.4823

summary(fg12345)

##   
## Call:  
## lm(formula = final ~ gpa + quiz1 + quiz2 + quiz3 + quiz4 + quiz5,   
## data = grades)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.6580 -2.7187 0.7985 3.9664 11.4979   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 32.8658 3.3141 9.917 < 2e-16 \*\*\*  
## gpa 3.6271 0.7879 4.604 1.25e-05 \*\*\*  
## quiz1 0.2986 0.5162 0.578 0.5643   
## quiz2 0.8004 0.6032 1.327 0.1876   
## quiz3 0.9063 0.5217 1.737 0.0855 .   
## quiz4 -0.1428 0.4739 -0.301 0.7638   
## quiz5 0.4823 0.4580 1.053 0.2948   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.831 on 98 degrees of freedom  
## Multiple R-squared: 0.4922, Adjusted R-squared: 0.4611   
## F-statistic: 15.83 on 6 and 98 DF, p-value: 1.224e-12

vif(fg12345)

## gpa quiz1 quiz2 quiz3 quiz4 quiz5   
## 1.107509 5.015198 2.931613 4.433567 3.572136 1.999252

final~gpa+quiz2+quiz3+quiz4+quiz5

## final ~ gpa + quiz2 + quiz3 + quiz4 + quiz5

fg2345<-lm(final~gpa+quiz2+quiz3+quiz4+quiz5,data=grades)  
summary(fg2345)

##   
## Call:  
## lm(formula = final ~ gpa + quiz2 + quiz3 + quiz4 + quiz5, data = grades)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.5494 -2.7210 0.5787 4.0927 11.5517   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 32.49410 3.24023 10.028 < 2e-16 \*\*\*  
## gpa 3.62466 0.78519 4.616 1.17e-05 \*\*\*  
## quiz2 0.82825 0.59928 1.382 0.1701   
## quiz3 1.06375 0.44360 2.398 0.0184 \*   
## quiz4 -0.02019 0.42244 -0.048 0.9620   
## quiz5 0.50431 0.45486 1.109 0.2702   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.812 on 99 degrees of freedom  
## Multiple R-squared: 0.4904, Adjusted R-squared: 0.4647   
## F-statistic: 19.06 on 5 and 99 DF, p-value: 3.089e-13

vif(fg2345)

## gpa quiz2 quiz3 quiz4 quiz5   
## 1.107478 2.912966 3.227419 2.857258 1.985498

final~gpa+quiz3

## final ~ gpa + quiz3

fg3<-lm(final~gpa+quiz3,data=grades)  
fg3

##   
## Call:  
## lm(formula = final ~ gpa + quiz3, data = grades)  
##   
## Coefficients:  
## (Intercept) gpa quiz3   
## 37.541 3.993 1.609

summary(fg3)

##   
## Call:  
## lm(formula = final ~ gpa + quiz3, data = grades)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.5260 -3.2000 -0.0803 4.7388 12.3955   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.5415 2.6802 14.007 < 2e-16 \*\*\*  
## gpa 3.9926 0.7850 5.086 1.67e-06 \*\*\*  
## quiz3 1.6088 0.2598 6.193 1.26e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.929 on 102 degrees of freedom  
## Multiple R-squared: 0.4535, Adjusted R-squared: 0.4428   
## F-statistic: 42.32 on 2 and 102 DF, p-value: 4.133e-14

vif(fg3)

## gpa quiz3   
## 1.06341 1.06341

final~total+quiz3

## 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

vif(ft3)

## total quiz3   
## 3.210517 3.210517

final~quiz2+quiz3

## 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

vif(f23)

## quiz2 quiz3   
## 1.897984 1.897984

final3~quiz3

## final3 ~ quiz3

f3<-lm(final~quiz3,data=grades)  
summary(f3)

##   
## Call:  
## lm(formula = final ~ quiz3, data = grades)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.3759 -4.3759 0.5556 5.6241 11.5556   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 46.0613 2.3312 19.759 < 2e-16 \*\*\*  
## quiz3 1.9315 0.2807 6.881 4.76e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.607 on 103 degrees of freedom  
## Multiple R-squared: 0.3149, Adjusted R-squared: 0.3083   
## F-statistic: 47.35 on 1 and 103 DF, p-value: 4.758e-10

In this model around 31.49% of variance in final is explained by quiz3 alone. Hence We can consider it as an average to good predictor model.

### For prediting 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 prediting 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 5. Use equation writer & Write Regression equation of the best model. Show R Output.

Answer. The Equation for our models are as follows

For Model ft3 the equation of the regression line is given by colnames(grades)

For Model f23 the equation of the regression line is given by final = 39.7129 + (1.5407*quiz2) - (1.1862*quiz3) ####Question6. 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 .Show R Output also.

Answer : Let’s find Durbin Watson Test for our two models For ft3 durbin watson Statistics comes out to be 2.115215

dwt(ft3)

## lag Autocorrelation D-W Statistic p-value  
## 1 -0.07657619 2.115215 0.582  
## Alternative hypothesis: rho != 0

For f23 durbin watson Statistics comes out to be 2.233423

dwt(f23)

## lag Autocorrelation D-W Statistic p-value  
## 1 -0.1209658 2.233423 0.268  
## Alternative hypothesis: rho != 0

#### 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 preditors is as below

For ft3 model

vif(ft3)

## total quiz3   
## 3.210517 3.210517

For f23 model

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: Variance inflation factor

#### Finding the predicted Value of final through both the models made

1. For ft3 model

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

## [1] 49.03415 60.15453 59.65296 65.01969 64.71158 72.55029 65.60005  
## [8] 71.16024 73.24531 75.83195 62.74117 69.77019 64.90503 57.48909  
## [15] 48.33912 60.65610 63.04928 56.48595 65.32781 61.42991 57.10219  
## [22] 66.99010 57.76134 67.29822 44.28365 61.65924 62.81996 71.16024  
## [29] 53.51241 69.77019 57.06632 59.34484 54.47967 63.51498 57.37443  
## [36] 59.34484 57.76134 60.15453 58.45636 73.24531 53.70586 58.26291  
## [43] 70.07831 73.94034 61.93148 42.81481 56.48595 59.65296 71.85526  
## [50] 63.43619 73.24531 65.90817 47.45065 69.18983 38.33655 62.81996  
## [57] 54.59434 59.26605 72.55029 55.79093 55.59748 54.09277 68.18669  
## [64] 62.81996 62.62650 57.37443 57.06632 62.62650 65.21315 58.45636  
## [71] 64.90503 67.10476 63.82310 49.65038 46.87029 72.55029 69.07517  
## [78] 63.51498 59.26605 55.86972 64.90503 66.29508 68.38015 53.78465  
## [85] 66.60319 69.77019 54.28622 63.51498 64.82624 59.76762 69.77019  
## [92] 56.56475 71.16024 64.01655 55.17470 69.57674 63.62964 61.42991  
## [99] 52.39460 62.12493 62.74117 49.65038 48.84069 63.51498 64.01655

1. For f23 model

grades$predf23<-predict(f23)  
grades$predf23

## [1] 55.71974 63.42344 59.98739 60.34196 62.71430 66.98195 65.44121  
## [8] 65.44121 66.98195 65.79578 59.15579 66.98195 63.90047 56.78345  
## [15] 55.71974 59.15579 55.24271 60.69653 59.51036 65.44121 51.32963  
## [22] 63.90047 65.79578 63.06887 52.16123 52.16123 63.90047 66.98195  
## [29] 54.88814 65.44121 61.17356 66.98195 65.44121 63.90047 58.80122  
## [36] 66.98195 59.63282 57.26048 62.71430 66.98195 57.61505 58.44665  
## [43] 61.52813 66.98195 64.25504 47.41656 54.53357 61.52813 65.44121  
## [50] 57.61505 66.98195 61.52813 54.53357 60.34196 46.70741 63.90047  
## [57] 57.26048 57.61505 66.98195 57.61505 57.96962 59.98739 64.25504  
## [64] 65.44121 64.25504 58.80122 62.71430 64.25504 58.44665 61.17356  
## [71] 65.44121 60.34196 64.60961 52.51580 50.97506 66.98195 65.44121  
## [78] 65.44121 54.53357 66.98195 65.44121 65.44121 63.90047 60.81899  
## [85] 63.06887 66.98195 62.71430 65.44121 59.15579 57.96962 65.44121  
## [92] 66.98195 65.44121 64.25504 63.90047 65.79578 58.80122 65.44121  
## [99] 66.98195 62.35973 59.15579 52.51580 54.53357 65.44121 62.71430

#### 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. [testing of slope] Show R Output. ?

1. For Model ft3

grades$errft3<-residuals(ft3)  
grades$errft3

## [1] 3.96585195 -6.15452606 -2.65295606 2.98030856 1.28842493  
## [6] 1.44971142 -2.60005232 -0.16024133 0.75468780 -0.83195308  
## [11] -3.74116694 1.22980593 4.09497131 -2.48909431 3.66087558  
## [16] 0.34390394 -0.04928332 -8.48595431 -2.32780782 -1.42991056  
## [21] -2.10218706 1.00990043 -5.76133881 0.70178405 -1.28364542  
## [26] 0.34076394 0.18004218 -0.16024133 3.48759382 -8.77019407  
## [31] 3.93368482 -4.34483969 -4.47967431 2.48501856 -3.37443156  
## [36] -6.34483969 2.23866119 -0.15452606 4.54363756 0.75468780  
## [41] 2.29414020 0.73709119 -0.07831045 1.05966417 -1.93148056  
## [46] -2.81480729 1.51404569 3.34704394 2.14473505 -2.43619057  
## [51] 1.75468780 3.09183130 0.54935283 -1.18983320 3.66345084  
## [56] 2.18004218 -1.59433706 -0.26604881 -0.55028858 2.20906932  
## [61] 3.40252294 2.90723295 -1.18669320 -0.81995782 1.37349581  
## [66] 2.62556844 4.93368482 -0.62650419 0.78685493 -1.45636244  
## [71] -2.90502869 -2.10476232 -3.82309782 0.34961920 2.12971370  
## [76] 0.44971142 0.92482955 -1.51498144 2.73395119 -2.86972156  
## [81] 1.09497131 1.70492405 -0.38014682 3.21534932 1.39680768  
## [86] 3.22980593 -5.28622068 -2.51498144 -0.82623782 -2.76761881  
## [91] 1.22980593 -4.56474518 0.83975867 2.98344856 -3.17469793  
## [96] 2.42325955 3.37035581 -1.42991056 0.60539657 2.87506581  
## [101] 1.25883306 -0.65038080 1.15930558 -0.51498144 3.98344856

1. For Model f23

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

## [1] -2.7197410 -9.4234430 -2.9873912 7.6580378 3.2856990  
## [6] 7.0180488 -2.4412108 5.5587892 7.0180488 9.2042182  
## [11] -0.1557929 4.0180488 5.0995296 -1.7834541 -3.7197410  
## [16] 1.8442071 7.7572863 -12.6965333 3.4896361 -5.4412108  
## [21] 3.6703655 4.0995296 -13.7957818 4.9311280 -9.1612329  
## [26] 9.8387671 -0.9004704 4.0180488 2.1118573 -4.4412108  
## [31] -0.1735606 -11.9819512 -15.4412108 2.0995296 -4.8012218  
## [36] -13.9819512 0.3671798 2.7395186 0.2856990 7.0180488  
## [41] -1.6150525 0.5533492 8.4718684 8.0180488 -4.2550414  
## [46] -7.4165554 3.4664283 1.4718684 8.5587892 3.3849475  
## [51] 8.0180488 7.4718684 -6.5335717 7.6580378 -4.7074133  
## [56] 1.0995296 -4.2604814 1.3849475 5.0180488 0.3849475  
## [61] 1.0303765 -2.9873912 2.7449586 -3.4412108 -0.2550414  
## [66] 1.1987782 -0.7143010 -2.2550414 7.5533492 -4.1735606  
## [71] -3.4412108 4.6580378 -4.6096124 -2.5158039 -1.9750635  
## [76] 6.0180488 4.5587892 -3.4412108 7.4664283 -13.9819512  
## [81] 0.5587892 2.5587892 4.0995296 -3.8189896 4.9311280  
## [86] 6.0180488 -13.7143010 -4.4412108 4.8442071 -0.9696235  
## [91] 5.5587892 -14.9819512 6.5587892 2.7449586 -11.9004704  
## [96] 6.2042182 8.1987782 -5.4412108 -13.9819512 2.6402700  
## [101] 4.8442071 -3.5158039 -4.5335717 -2.4412108 5.2856990

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

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

## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17  
## [18] 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34  
## [35] 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51  
## [52] 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68  
## [69] 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85  
## [86] 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102  
## [103] 103 104 105

The equations of the model which are built as the best or final Models are 1. For f23 model the Equation of Regression Line is final = 39.7129 + (1.5407*quiz2) + (1.1862*quiz3) 2. For Ft3 model the equation of the Regression line is final = 6.67358 + (.69502*total) - (1.89612*quiz3) predicted values of final models using the above equations are below. These values are giving us the predited value of our model For model ft3,inserting the predicted values in the grades dataset by column creation

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

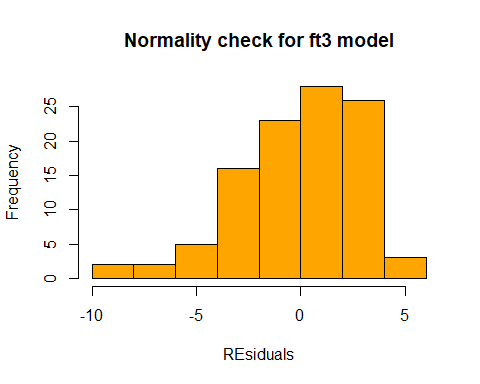
## [1] 49.03415 60.15453 59.65296 65.01969 64.71158 72.55029 65.60005  
## [8] 71.16024 73.24531 75.83195 62.74117 69.77019 64.90503 57.48909  
## [15] 48.33912 60.65610 63.04928 56.48595 65.32781 61.42991 57.10219  
## [22] 66.99010 57.76134 67.29822 44.28365 61.65924 62.81996 71.16024  
## [29] 53.51241 69.77019 57.06632 59.34484 54.47967 63.51498 57.37443  
## [36] 59.34484 57.76134 60.15453 58.45636 73.24531 53.70586 58.26291  
## [43] 70.07831 73.94034 61.93148 42.81481 56.48595 59.65296 71.85526  
## [50] 63.43619 73.24531 65.90817 47.45065 69.18983 38.33655 62.81996  
## [57] 54.59434 59.26605 72.55029 55.79093 55.59748 54.09277 68.18669  
## [64] 62.81996 62.62650 57.37443 57.06632 62.62650 65.21315 58.45636  
## [71] 64.90503 67.10476 63.82310 49.65038 46.87029 72.55029 69.07517  
## [78] 63.51498 59.26605 55.86972 64.90503 66.29508 68.38015 53.78465  
## [85] 66.60319 69.77019 54.28622 63.51498 64.82624 59.76762 69.77019  
## [92] 56.56475 71.16024 64.01655 55.17470 69.57674 63.62964 61.42991  
## [99] 52.39460 62.12493 62.74117 49.65038 48.84069 63.51498 64.01655

##For Model f23,inserting the predicted values in the grades dataset by column creation  
predft3<-predict(ft3)  
grades$predft3<-predft3  
grades$predf23

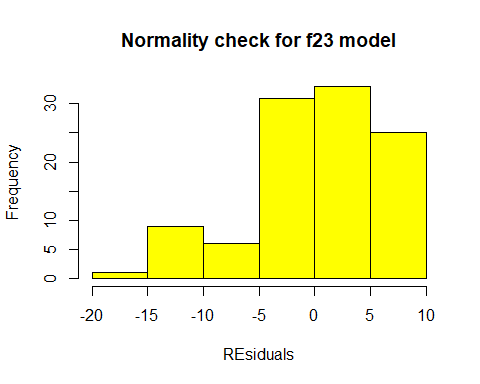
## [1] 55.71974 63.42344 59.98739 60.34196 62.71430 66.98195 65.44121  
## [8] 65.44121 66.98195 65.79578 59.15579 66.98195 63.90047 56.78345  
## [15] 55.71974 59.15579 55.24271 60.69653 59.51036 65.44121 51.32963  
## [22] 63.90047 65.79578 63.06887 52.16123 52.16123 63.90047 66.98195  
## [29] 54.88814 65.44121 61.17356 66.98195 65.44121 63.90047 58.80122  
## [36] 66.98195 59.63282 57.26048 62.71430 66.98195 57.61505 58.44665  
## [43] 61.52813 66.98195 64.25504 47.41656 54.53357 61.52813 65.44121  
## [50] 57.61505 66.98195 61.52813 54.53357 60.34196 46.70741 63.90047  
## [57] 57.26048 57.61505 66.98195 57.61505 57.96962 59.98739 64.25504  
## [64] 65.44121 64.25504 58.80122 62.71430 64.25504 58.44665 61.17356  
## [71] 65.44121 60.34196 64.60961 52.51580 50.97506 66.98195 65.44121  
## [78] 65.44121 54.53357 66.98195 65.44121 65.44121 63.90047 60.81899  
## [85] 63.06887 66.98195 62.71430 65.44121 59.15579 57.96962 65.44121  
## [92] 66.98195 65.44121 64.25504 63.90047 65.79578 58.80122 65.44121  
## [99] 66.98195 62.35973 59.15579 52.51580 54.53357 65.44121 62.71430

#### Question 9. Test the assumption of Normality and interpret your findings. .Show histogram and interpret in maximum 3 lines.

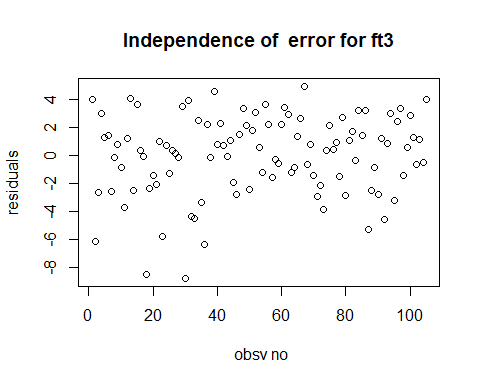
hist(grades$errft3,main = "Normality check for ft3 model", xlab="REsiduals",col="orange")



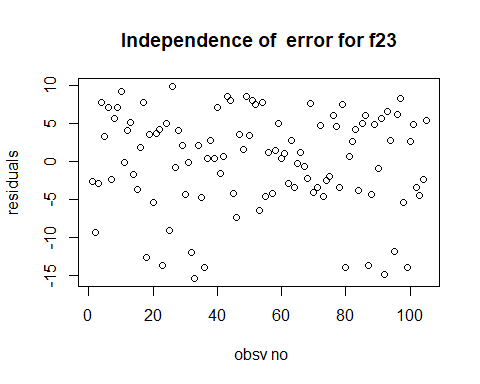
hist(grades$errf23,main = "Normality check for f23 model", xlab="REsiduals",col="yellow")

 ####Question 10. Test the assumption of Independent of observations and interpret in maximum 3 lines

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

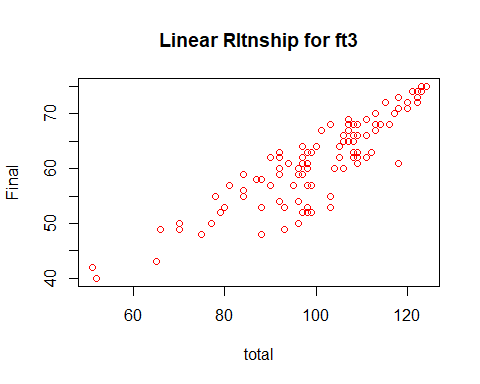


plot(grades$obsno,grades$errf23,main="Independence of error for f23",xlab= " obsv no", ylab="residuals")

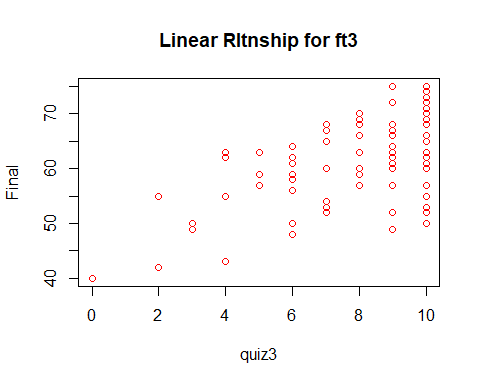
 ####Question 11. Test the assumption of linear relationship and interpret in maximum 3 lines for each predictor . If more than one predictor is used in model then more scatter plots would be required]

Check of linear relationship *For ft3 let us draw scatter pot for final & total and final and quiz3*  
For f23 let us draw scatter pot for final & quiz2 and final and quiz3

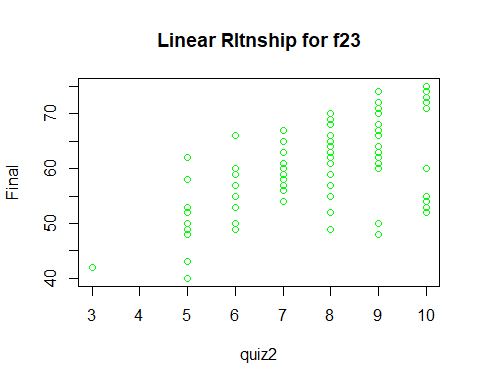
plot(grades$total,grades$final,main="Linear Rltnship for ft3",xlab="total",ylab="Final",col="red")



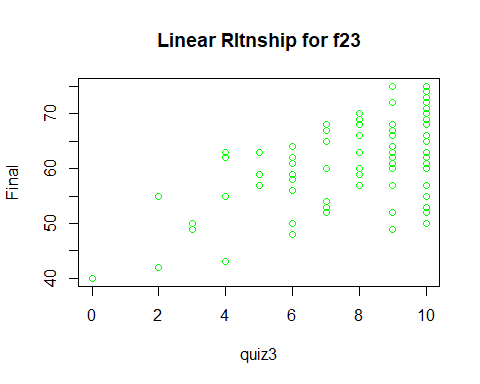
plot(grades$quiz3,grades$final,main="Linear Rltnship for ft3",xlab="quiz3",ylab="Final",col="red")



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



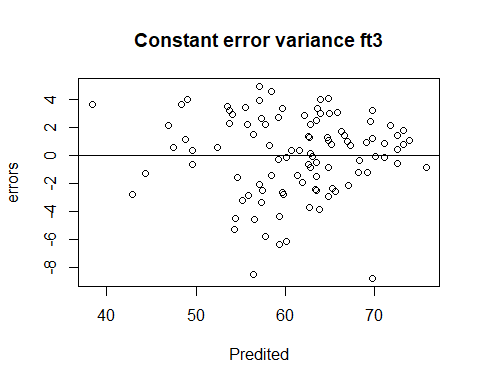
plot(grades$quiz3,grades$final,main="Linear Rltnship for f23",xlab="quiz3",ylab="Final",col="green")



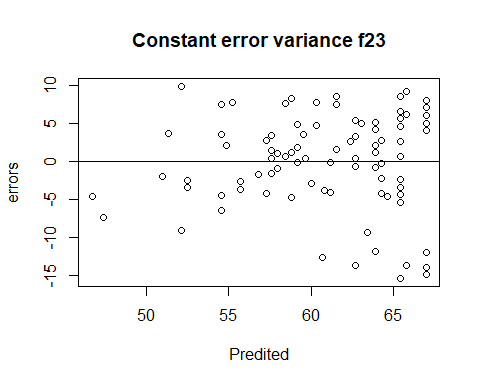
#### Question 12. Test the assumption of Constant Error Variance and interpret in maximum 3 lines

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

plot(grades$predft3,grades$errft3,main="Constant error variance ft3",xlab="Predited",ylab="errors",abline(h=0))



plot(grades$predf23,grades$errf23,main="Constant error variance f23",xlab="Predited",ylab="errors",abline(h=0))



#### Question13. What is Standard Error of Estimate of your model and how do you interpret the same. Show with some hypothetical values of predictors.

final = 6.67358 + (.69502*total) - (1.89612*quiz3) final = 6.67358 + (.69502*80) - (1.89612*7) final= 49.034

Now,Lets find predicted value using f23 model whose equation is given by

final = 39.7129 + (1.5407*quiz2) + (1.1862*quiz3) Lets 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.

#### 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 in 250 words 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 : The given dataset 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 to build a predictive model for predicting final for consumption we have built two linear regression models to predict final .

##### 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:

final=39.7129 + (1.5407×quiz2)+ (1.1862×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

##### Model for predicting final 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 6.192-14. The equation for the model ft3 is given by :

final =6.67358 + (.69502×total) - (1.89612×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 in 500 words.

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 :

final =6.67358 + (.69502×total) - (1.89612×quiz3) final=39.7129 + (1.5407×quiz2) + (1.1862×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

final =6.67358 + (.69502×100) - (1.89612×9) 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

final=39.7129 + (1.5407×8) + (1.1862×9) 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 a 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) = 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 .

If make is Cadillac then the Fitted Regression Equation for Cadillac becomes (in this case both MakeChevrolet and MakePontiac are equal to zero & MakeCadillac = 1 ) ####Price = 1.017e+04 + ((-2.089e-01) \* Mileage) + ((4.361e+03) \* Cylinder)