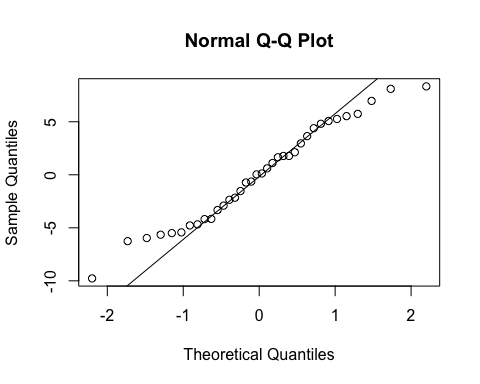
#question 1:  
#importing data  
data1<-read.csv("/Users/Downloads/Assignment/data-table-B8(1).csv",header=T)  
#a.multiple linear regression  
mod<-lm(y~.,data=data1)  
summary(mod)

##   
## Call:  
## lm(formula = y ~ ., data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.7716 -4.1656 0.0802 3.8323 8.3349   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.109e+01 1.669e+00 6.642 1.48e-07 \*\*\*  
## x1 3.501e+02 3.968e+01 8.823 3.38e-10 \*\*\*  
## x2 1.089e-01 9.983e-03 10.912 1.74e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.782 on 33 degrees of freedom  
## Multiple R-squared: 0.8415, Adjusted R-squared: 0.8319   
## F-statistic: 87.6 on 2 and 33 DF, p-value: 6.316e-14

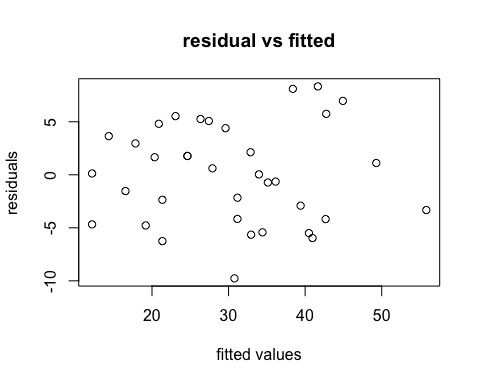
The model is y= 11.09+350.1\*X1 + 0.1089\*X2

#b. Normality plot of residuals  
qqnorm(mod$residuals)  
qqline(mod$residuals)



Yes, there appears to be a problem with the normality assumption. The tails deviate significantly from the 45 degree line.

#c.residual vs fitted  
plot(y=mod$residuals,x=mod$fit,xlab="fitted values",ylab="residuals",main="residual vs fitted")



The plot of residual versus fitted values appears to have more negative values as compared to positive ones. There is some deviation from a totally random distribution of points.

#d. influence analysis  
library(MASS)  
## COOKS DISTANCE   
cd=cooks.distance(mod)   
cd

## 1 2 3 4 5   
## 4.479634e-02 3.085529e-03 2.610378e-03 2.543510e-02 2.257348e-02   
## 6 7 8 9 10   
## 1.816571e-02 5.519484e-03 6.942400e-04 2.018962e-02 5.944222e-02   
## 11 12 13 14 15   
## 3.133287e-05 2.164992e-02 1.018252e-02 2.131706e-02 2.164130e-02   
## 16 17 18 19 20   
## 1.831760e-06 8.114930e-02 3.109642e-02 5.714600e-03 2.148520e-03   
## 21 22 23 24 25   
## 1.803195e-04 2.021487e-03 3.830608e-02 4.026314e-02 2.148520e-03   
## 26 27 28 29 30   
## 7.493094e-03 1.486473e-02 2.319863e-01 6.778439e-02 5.942277e-02   
## 31 32 33 34 35   
## 3.073922e-03 4.701174e-02 1.015499e-03 1.147647e-01 1.221554e-01   
## 36   
## 9.431870e-02

plot(cd,main="plot of cook's distance")



There is no point which has Cook’s distance greater than 2. Therefore, no point unduly impacts the regression coefficients.

## DFFITS   
dfts=dffits(mod)   
dfts

## 1 2 3 4 5   
## -0.367024948 -0.094902505 0.087310653 0.278019198 0.261357136   
## 6 7 8 9 10   
## 0.233066373 0.127125921 -0.044953597 -0.243919024 -0.421710448   
## 11 12 13 14 15   
## 0.009547369 0.253430381 0.173192697 0.253168632 0.255548409   
## 16 17 18 19 20   
## 0.002308412 -0.497560507 -0.305836342 -0.129444324 0.079231343   
## 21 22 23 24 25   
## 0.022909170 -0.076931030 -0.342566184 -0.352117971 0.079231343   
## 26 27 28 29 30   
## -0.149415982 -0.212277797 -0.888397441 -0.455061347 0.426259099   
## 31 32 33 34 35   
## 0.094652580 -0.373209407 -0.054373760 0.607850194 0.629088389   
## 36   
## 0.543747087

plot(dfts,main="plot of dffits")



The 34th point , 35th point and the 36th point have a dffits value greater than the cutoff value of 0.4714. Therefore according to this analysis, they unduly impact the parameters.

#e. Residual analysis  
e=residuals(mod) ## RESIDUAL  
std\_e=stdres(mod) ## STANDARDIZED RESIDUAL  
std\_e

## 1 2 3 4 5   
## -1.037115175 -0.333758043 0.356275584 1.187266673 1.129176925   
## 6 7 8 9 10   
## 0.946369325 0.461534357 -0.141070312 -0.650555446 -0.955185630   
## 11 12 13 14 15   
## 0.027428737 0.799919557 0.641316052 1.035126037 1.089371414   
## 16 17 18 19 20   
## 0.008015465 -1.237884530 -1.041331230 -0.508984388 0.379286883   
## 21 22 23 24 25   
## 0.128876651 -0.458446375 -1.289781318 -1.351031103 0.379286883   
## 26 27 28 29 30   
## -0.882640059 -1.154169443 -2.186886007 -1.255401518 1.266420897   
## 31 32 33 34 35   
## 0.248696768 -0.773278110 -0.161131924 1.783666308 1.835215764   
## 36   
## 1.541125067

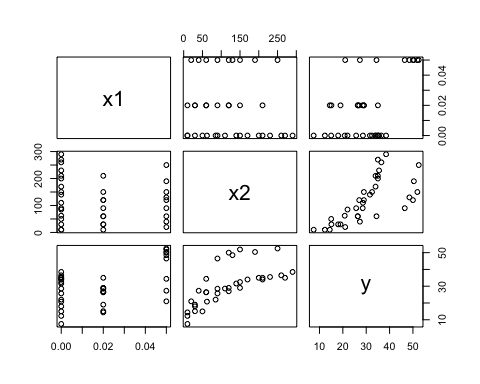
None of the standardized residuals are greater than 3 therefore none of them are outliers.

r=studres(mod) ## STUDENTIZED RESIDUAL   
max(abs(r))

## [1] 2.328851

The studentized residuals also don’t show any outlier.

#f appropriate transformation  
plot(data1)

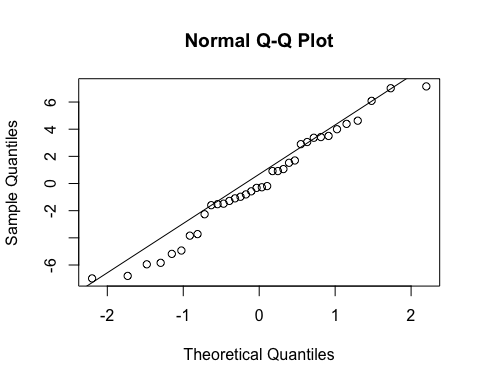


It can be observed from the relationship of x2 and Y that the assumption of linearity is not valid. Instead, the relation of Y with x2 appears to be that of a squareroot.

mod\_new=lm(y~x1+I(sqrt(x2)),data1)  
summary(mod\_new)

##   
## Call:  
## lm(formula = y ~ x1 + I(sqrt(x2)), data = data1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.987 -1.762 -0.299 3.134 7.151   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.0508 1.9584 0.537 0.595   
## x1 333.6476 32.6876 10.207 9.66e-12 \*\*\*  
## I(sqrt(x2)) 2.2928 0.1667 13.752 3.23e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.957 on 33 degrees of freedom  
## Multiple R-squared: 0.8915, Adjusted R-squared: 0.8849   
## F-statistic: 135.6 on 2 and 33 DF, p-value: < 2.2e-16

qqnorm(mod\_new$residuals)  
qqline(mod\_new$residuals)

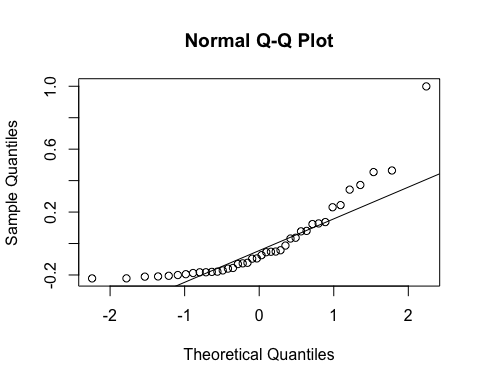


The normal probability plot shows that the residuals are much more closer to a normal distribution and there is not departure from the straight line towards the tails.

#question 2:  
#importing data  
data2<-read.csv("/Users/Downloads/Assignment/data-table-B10(1).csv",header=T)  
#a.multiple linear regression  
mod1<-lm(y~.,data=data2)  
summary(mod1)

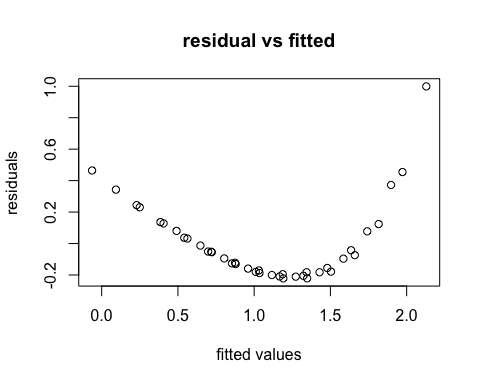
##   
## Call:  
## lm(formula = y ~ ., data = data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.22179 -0.18102 -0.08439 0.09111 0.99908   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.679439 0.143532 4.734 3.20e-05 \*\*\*  
## x1 1.407331 0.196925 7.147 1.81e-08 \*\*\*  
## x2 -0.015629 0.001428 -10.948 3.67e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2593 on 37 degrees of freedom  
## Multiple R-squared: 0.822, Adjusted R-squared: 0.8124   
## F-statistic: 85.46 on 2 and 37 DF, p-value: 1.351e-14

#b. Normality plot of residuals  
qqnorm(mod1$residuals)  
qqline(mod1$residuals)



The normal probability plot shows that the assumption of normality of residuals is not valid and they deviate from the straight line around the tail.

#c.residual vs fitted  
plot(y=mod1$residuals,x=mod1$fit,xlab="fitted values",ylab="residuals",main="residual vs fitted")



The residual vs fitted plot clearly shows that the assumption of linearity is not valid because there exists a parabolic dependence of residuals with the fitted values.

#d Press Statistic  
library(MASS)  
hat=hatvalues(mod1)  
e=residuals(mod1)  
pr=e/(1-hat)  
PRS\_STAT=sum(pr^2)  
PRS\_STAT

## [1] 3.112132

The press statistic is quite high. A lower press statistic implies that the model will be good in predicting values for new data.

#e. influence analysis  
## COOKS DISTANCE   
cd1=cooks.distance(mod1)   
plot(cd1,main="plot of cook's distance")



It can be observed that none of the Cook’s distance values are greater than 2. Therefore, none of the values affect the parameters unduly.

#f. Residual analysis  
e1=residuals(mod1) ## RESIDUAL  
std\_e1=stdres(mod1) ## STANDARDIZED RESIDUAL  
std\_e1

## 1 2 3 4 5 6   
## 4.12506460 1.85047600 0.49856753 -0.29664261 -0.71373741 -0.88426286   
## 7 8 9 10 11 12   
## -0.88839737 -0.75641754 -0.53154771 -0.22802726 1.50648326 0.30940133   
## 13 14 15 16 17 18   
## -0.38086603 -0.71936570 -0.82571664 -0.78539422 -0.62750427 -0.37476341   
## 19 20 21 22 23 24   
## -0.05206262 0.32487322 -0.17301052 -0.62046469 -0.81095613 -0.82198801   
## 25 26 27 28 29 30   
## -0.70633351 -0.49382219 -0.20051681 0.14472775 0.54428159 0.98765431   
## 31 32 33 34 35 36   
## -0.75768234 -0.79721339 -0.69250138 -0.49296725 -0.20967123 0.12691113   
## 37 38 39 40   
## 0.51366805 0.93351970 1.40088652 1.92524618

The first standardized residual is greater than 4 which hints to an outlier. This point should be observed carefully.