**Mini Project 6**

**Group members:** Arnav Sharma (axs144130), Divyanshu Paliwal (dxp151630)

**Contribution of each member:** Both members contributed equally to the following tasks:

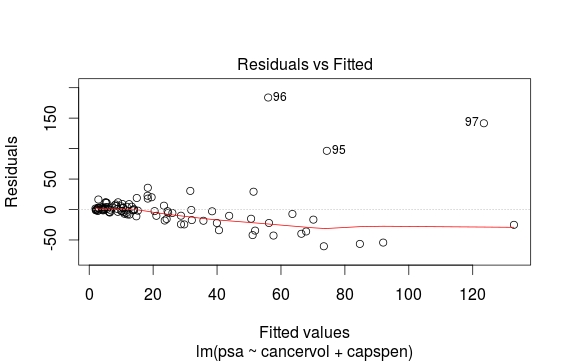
* Writing the R code
* Generating graphs
* Deriving conclusions

A) **The two variables used in the model are cancervol and capspen.**

Linear model regression diagnostics

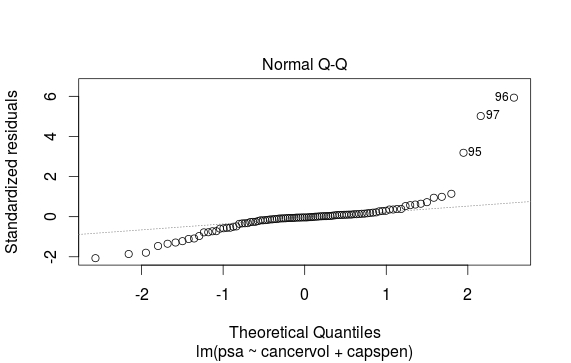
1. Linearity assumption: Y values can be expressed as a linear function of the x variable

Fairly flat regression line obtained in residual vs fitted graph suggests that the linearity assumption is followed in this model.



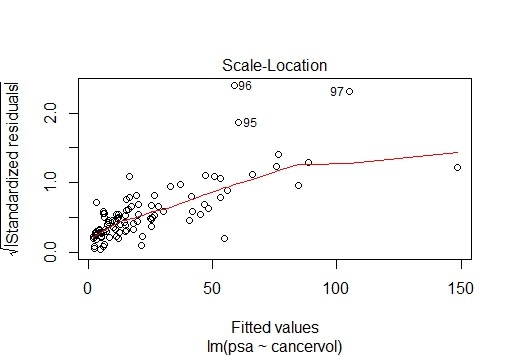
1. Normality assumption: y values(or the error) are normally distributed

From the graph between standardized residuals vs theoretical quantiles it is clear that the residuals are almost normally distributed.



1. Constant Variance assumption: Variation of observations around the regression line (the residual SE) is constant.

Fromm the Scale-Location plot it is clear that variance is not constant. Variance is increasing with increase in x. Thus this linear model does not follow the constant variance assumption.

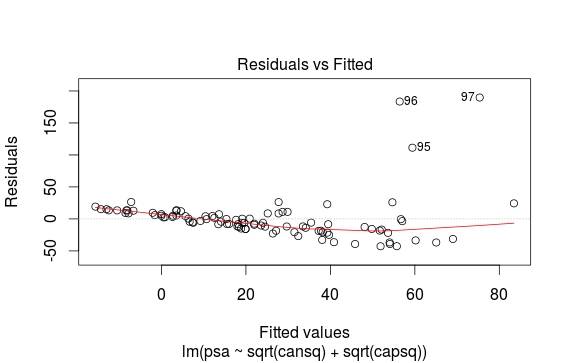


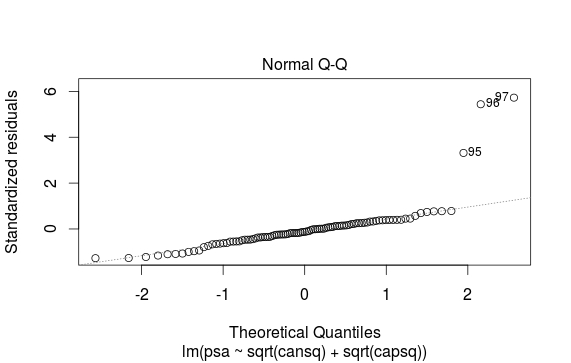
**REMEDY**

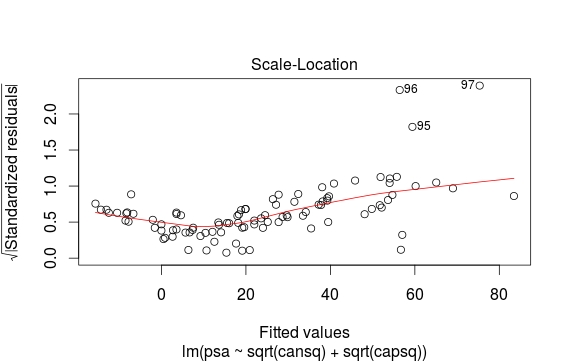
As a remedy, we do a quad root of both predictors cancervol and capspen i.e. sqrt(sqrt(cancervol)) and sqrt(sqrt(capspen)).

From the Scale-Location plot it is clear that variance is almost constant after applying the remedy for increasing variance condition. Variance is nearly constant with increase in x.

Thus this linear model follows the constant variance assumption.







1. Auto-Correlation Assumption: [autocorrelation](http://www.investopedia.com/terms/a/autocorrelation.asp) in the residuals

From the value of Durbin-Watson statistic obtained by using dwtest() function in R, it is clear that a positive autocorrelation exists between the residuals.

Durbin-Watson (DW) statistic = 0.51677

DW statistic near to zero means there is a positive correlation between the residuals.

C) Predicted value of psa for median of cancervol and capspen is 8.248768

**APPENDIX**

#read data file

d<-read.csv("D:/UTD Stats/mini\_project-5/prostate\_cancer.csv",header = TRUE)

#scatterplots between PSA and other variables

plot(cancervol, psa)

plot(weight, psa)

plot(age, psa)

plot(benpros, psa)

plot(vesinv, psa)

plot(capspen, psa)

plot(gleason, psa)

##ScatterPlot(cancervol, psa, col.out30 = "red")

cansq <- sqrt(cancervol)

############

#developing linear regeression model with cancervol and capspen as predictor variables

capsq <- sqrt(capspen)

fitnew <- lm(psa~sqrt(cansq) + sqrt(capsq), data = d)

plot(fitnew)

summary(fitnew)

abline(fitnew)

#durbin watson test for checking autocorrelation assumption

dwtest(fitnew)

#predicting value of the median of cancervol and capspen

X1 <- median(sqrt(cansq))

X2 <- median(sqrt(capsq))

Y <- 3.63788 + 3.23139\*X1 - 0.03951\*X2

Y