**Mini Project 5**

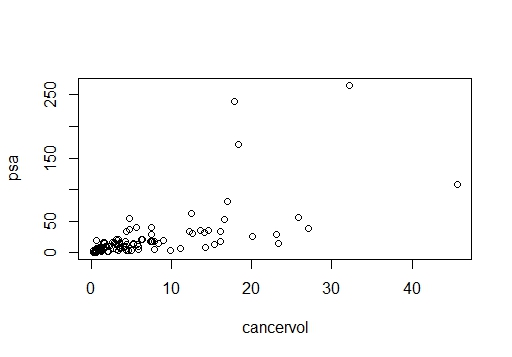
**Group members:** Arnav Sharma (axs144130), Divyanshu Paliwal (dxp151630)

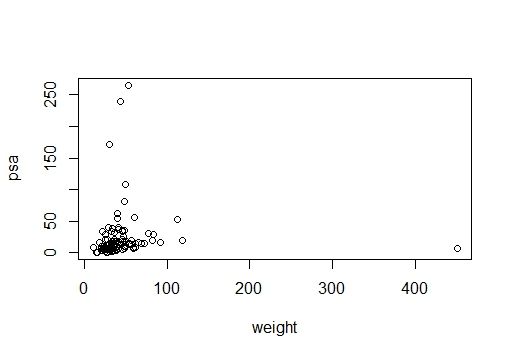
**Contribution of each member:** Both members contributed equally to the following tasks:

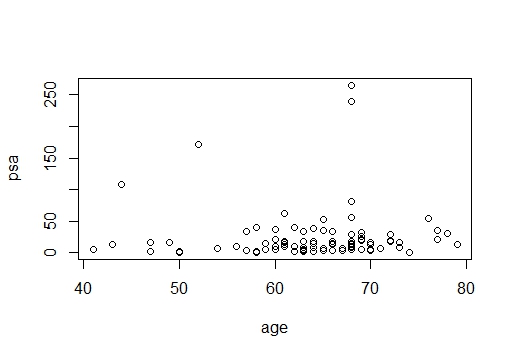
* Writing the R code
* Generating graphs
* Deriving conclusions

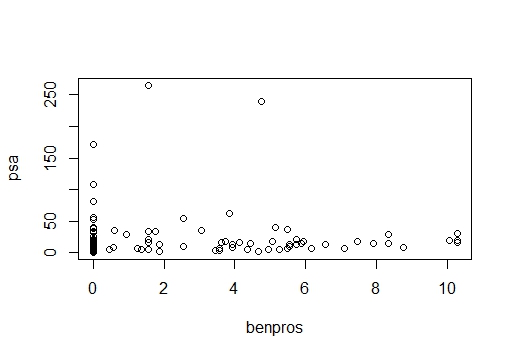
1. Scatter Plots of psa vs all other variables.

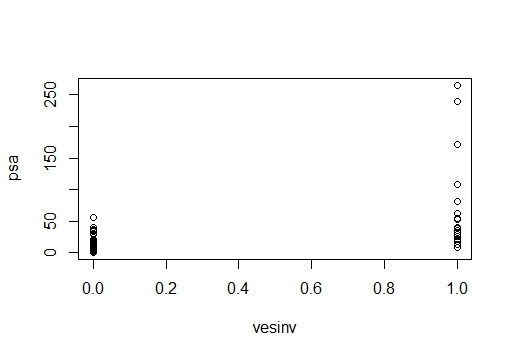
**Cancervol** is the quantitative variable that can be used most effectively to predict PSA level.

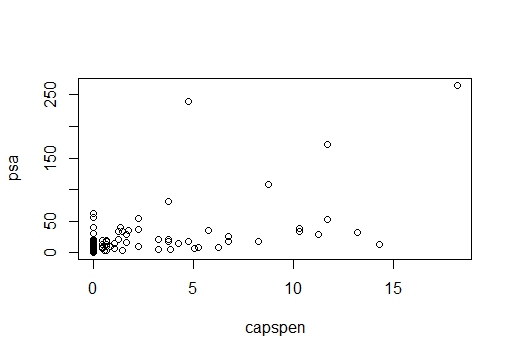


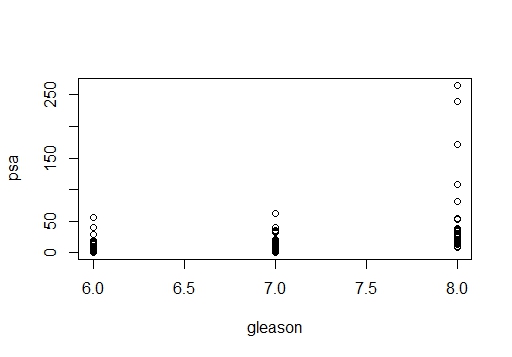












From the scatter plot between PSA level and Cancer Volumn, we can see that when cancer volume increases, the PSA level increases. Therefore, we pick “Cancer Volumn” as a suitable qualitative variable that we think it can be used to predict PSA level using linear regression model.

By fitting a simple linear reqression model, we obtain the linear regression line: Y = 1.125+3.230X. To analyze how well the model is fitting, we use summary() in R, we found the following results:

Residual standard error: 32.03 on 95 degrees of freedom

Multiple R-squared: 0.3896, Adjusted R-squared: 0.3831

F-statistic: 60.63 on 1 and 95 DF, p-value: 8.468e-12

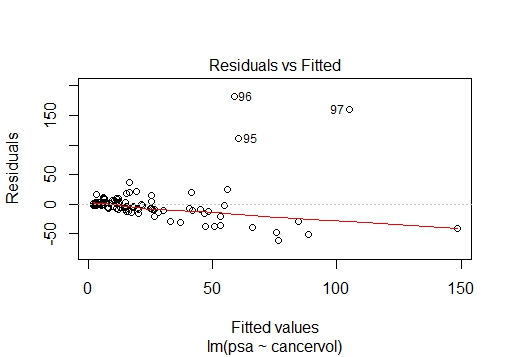
Since p-value of F-test is less than 0.05, so the model we found is significant.

According to F-statistics, we can also say that 60.63 % of the data are well pre-dicted with 95% of significance by our model.

B) Linear model regression diagnostics

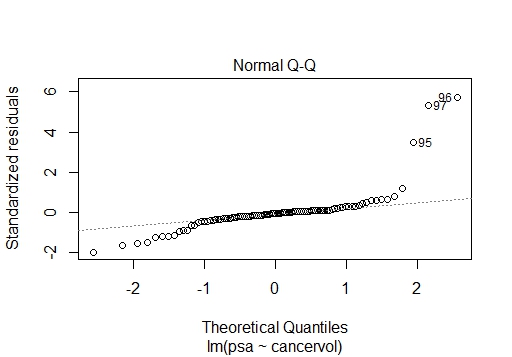
1. Linearity assumption: Y values can be expressed as a linear function of the x variable

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



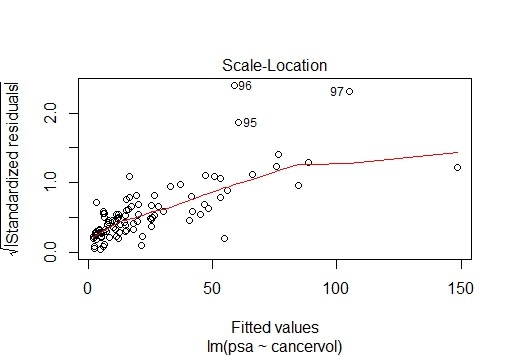
1. Normality assumption: y values(or the error) are normally distributed

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.



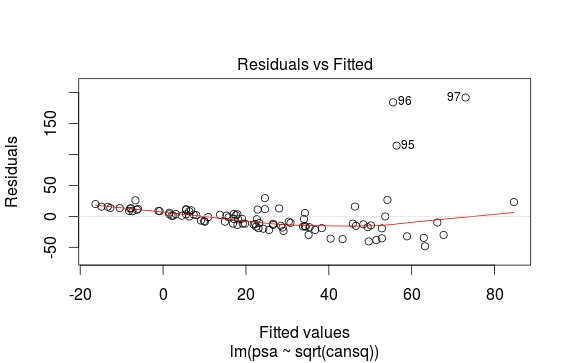
**REMEDY**

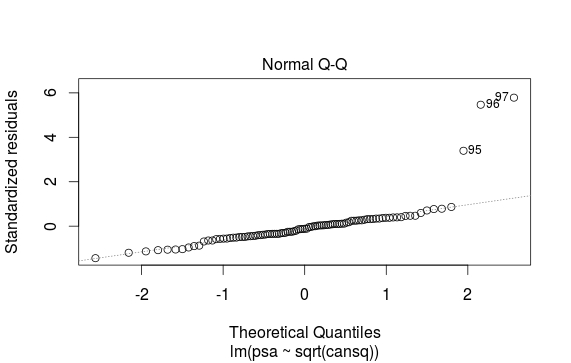
As a remedy, we transform the variable to make the plot linear and satisfy the assumptions.

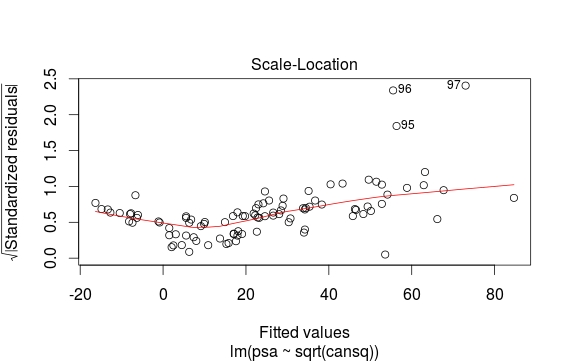
For cancervol data, we take its quad root i.e. sqrt(sqrt(cancervol)).

From the Residual vs Fitted, Normal Q-Q and Scale-Location plot it is clear that variance is almost constant after applying the remedy for increasing variance condition. The residuals are also normal and residual vs fitted line is also more linear. Variance is nearly constant with increase in x.

Thus this linear model meets the above mentioned standard regression 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 = 0.5

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

C) Predicted value of psa for median cancervol = 22.43.

**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")

#developing linear regeression model with cancervol as predictor variable

cansq <- sqrt(cancervol)

fit <- lm(psa~sqrt(cansq), data = d)

summary(fit)

abline(fit)

#building regression diagnostics graphs for our model to check assumptions

plot(fit)

#durbin watson test for checking autocorrelation assumption

dwtest(fit)

#Predict value for the median

X1 <- median(cancervol)

X1 = sqrt(X1)

Y1 = 1.125 + 3.230\*X1

Y1