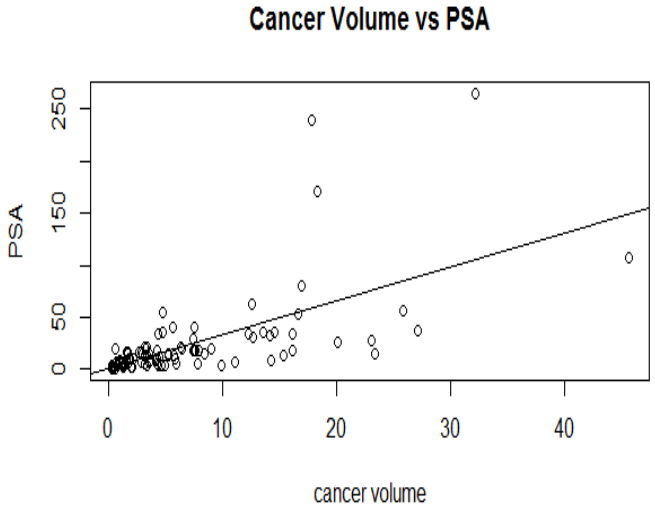
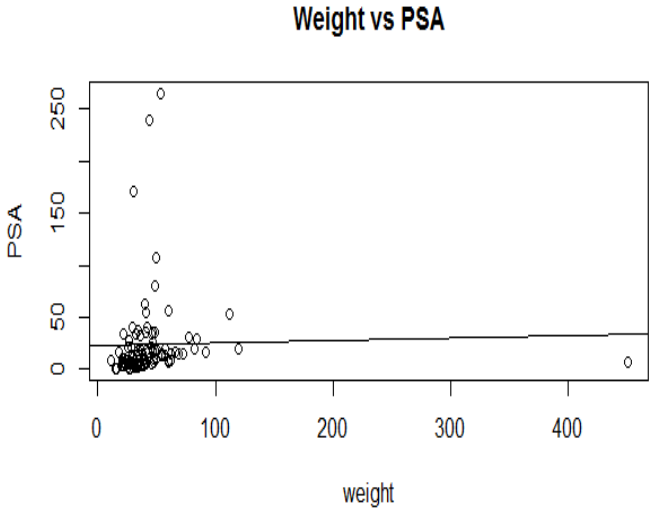
Project 4 **Section: CS 6313.501**

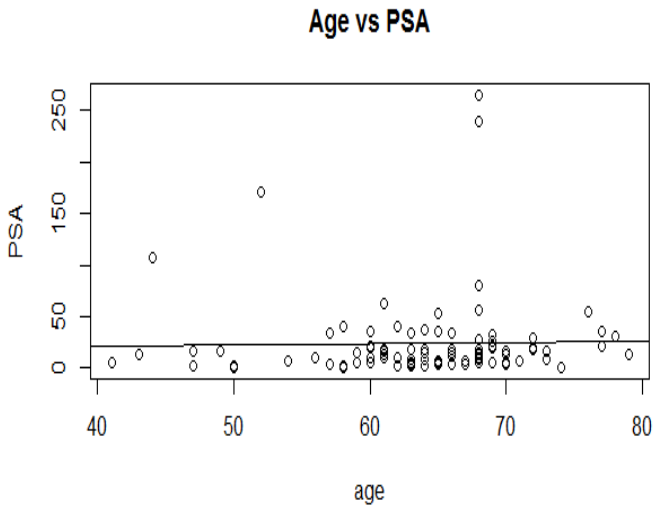
Name : Panchami G. Rudrakshi

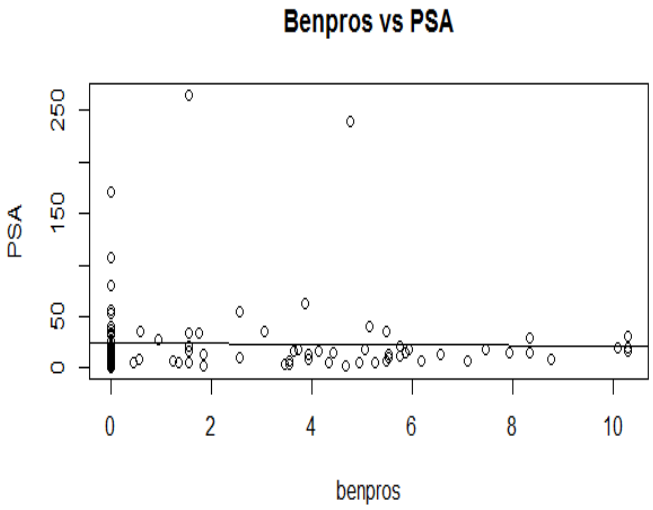
1. Take PSA level is as the response variable. Make scatter plots of PSA level with other variables. Based on these, choose one quantitative variable that you think may be used effectively to predict PSA level. Highlight any potential outliers on the scatter plot of this variable with PSA level.

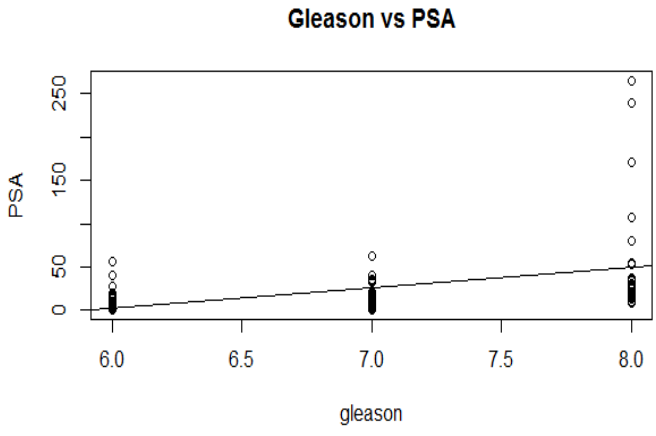
Answer: Scatter plots of PSA level with other variables:

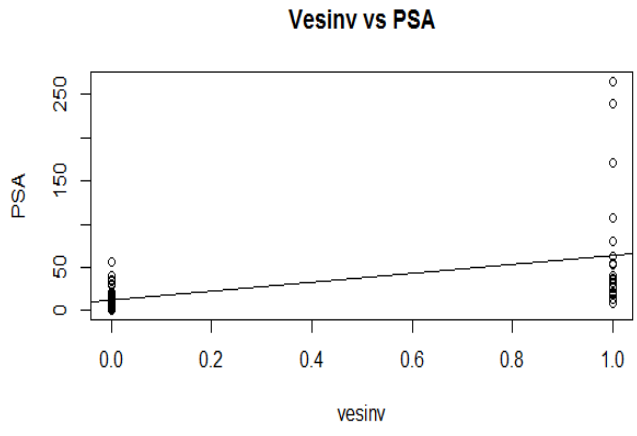


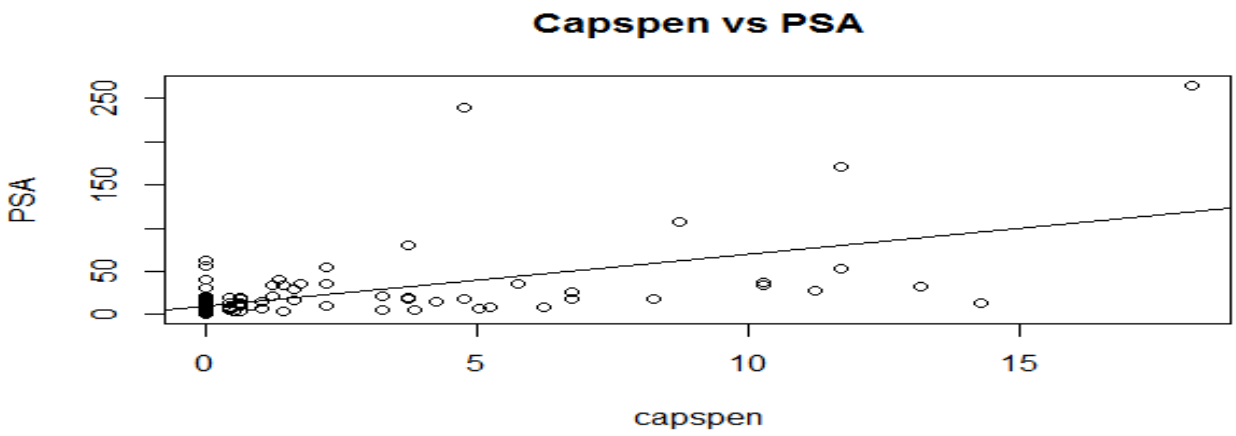












Response: PSA

Predictor: The quantitative variable "**CancerVol**" is the most effective quantitative variable that can be used as predictor of the PSA level. By observing all the correlation values, **CancerVol** is found to have highest correlation value **0.6241506.** It has regression coefficient of **0.3896** and low p- value of **8.468e-12** which is approximately equal to **0**. Hence we choose CancerVol as the quantitative variable.

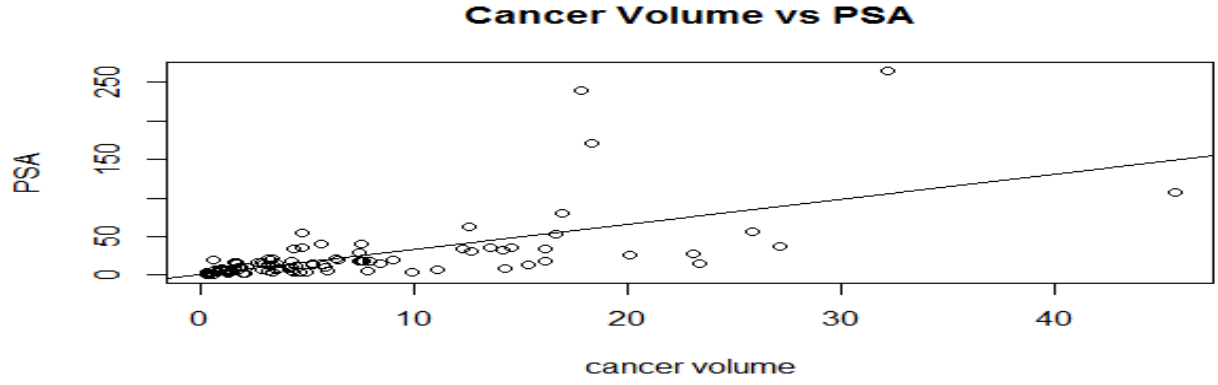
```
Console ~/
> plot(data[,3],data[,2],xlab = "cancer volume", ylab = "PSA", main = "Cancer Volume vs
 PSA")
> abline(lm(data[,2]~data[,3]))
> cor(data[,3],data[,2])
[1] 0.6241506
> plot(data[,4],data[,2],xlab = "weight", ylab = "PSA", main = "Weight vs PSA")
> abline(lm(data[,2]~data[,4]))
> cor(data[,4],data[,2])
[1] 0.02621343
> plot(data[,5],data[,2],xlab = "age", ylab = "PSA", main = "Age vs PSA")
> cor(data[,5],data[,2])
[1] 0.01719938
> abline(lm(data[,2]~data[,5]))
> plot(data[,6],data[,2],xlab = "benpros", ylab = "PSA", main = "Benpros vs PSA")
> cor(data[,6],data[,2])
[1] -0.01648649
> abline(lm(data[,2]~data[,6]))
> plot(data[,7],data[,2],xlab = "vesinv", ylab = "PSA", main = "Vesinv vs PSA")
> cor(data[,7],data[,2])
[1] 0.5286188
> abline(lm(data[,2]~data[,7]))
> plot(data[,8],data[,2],xlab = "capspen", ylab = "PSA", main = "Capspen vs PSA")
> cor(data[,8],data[,2])
[1] 0.5507925
> abline(lm(data[,2]~data[,8]))
> plot(data[,9],data[,2],xlab = "gleason", ylab = "PSA", main = "Gleason vs PSA")
> cor(data[,9],data[,2])
[1] 0.4295798
```

Potential outliers on the scatter plot for Cancer Volume with PSA level:

Looking at the above scatter plot we can say that:

a) There are 2 outliers when cancer volume is between 10 and 20 of Cancer volume at PSA level 150-250.

b) There is an outlier when cancer volume is between30 and 40.



2. Fit a simple linear regression model and carry out regression diagnostics. The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If an assumption is not met, attempt to remedy the situation. Comment on the fit of the final model using appropriate tests and statistics.

```
Call: lm(formula = y ~ x)
Coefficients:
(Intercept)            x
      1.125        3.230
Residuals:
    Min      1Q  Median      3Q     Max
-61.619  -9.023  -1.586   3.151 181.183
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1249     4.3596   0.258    0.797
x             3.2299     0.4148   7.786 8.47e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

Regression model evaluation: It is based on the evaluation of the residuals and key assumption. The following are the assumptions that must be satisfied.

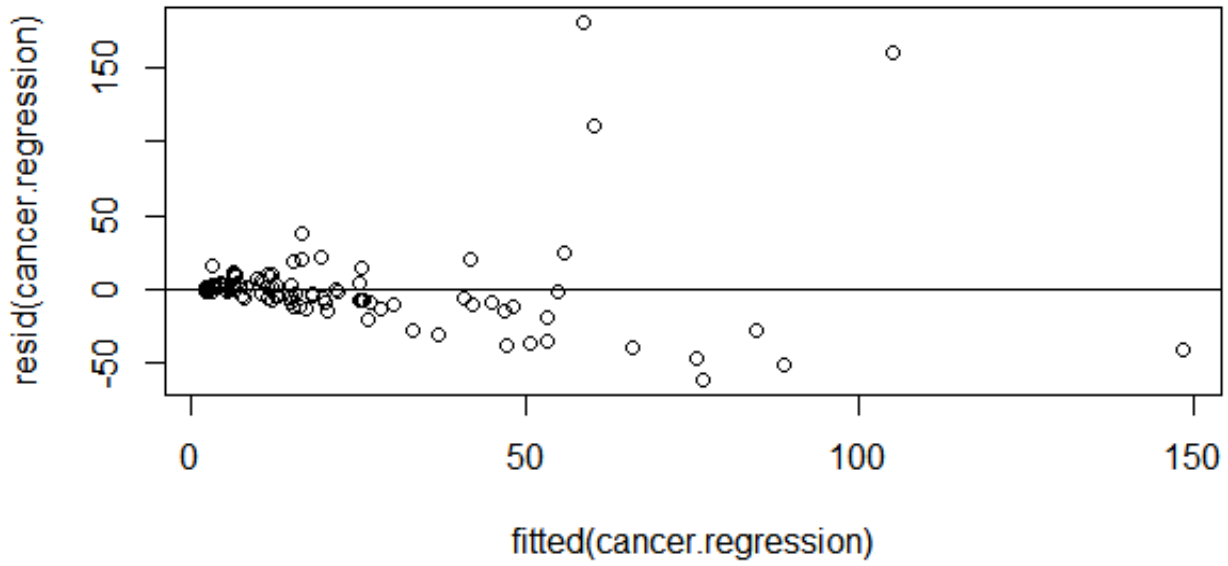
a) Errors have mean zero and constant variance.
b) Errors are normally distributed.
c) Errors are independent.

a) Errors have mean zero and constant variance.

Residual plot is used to test if errors have mean zero and constant variance.

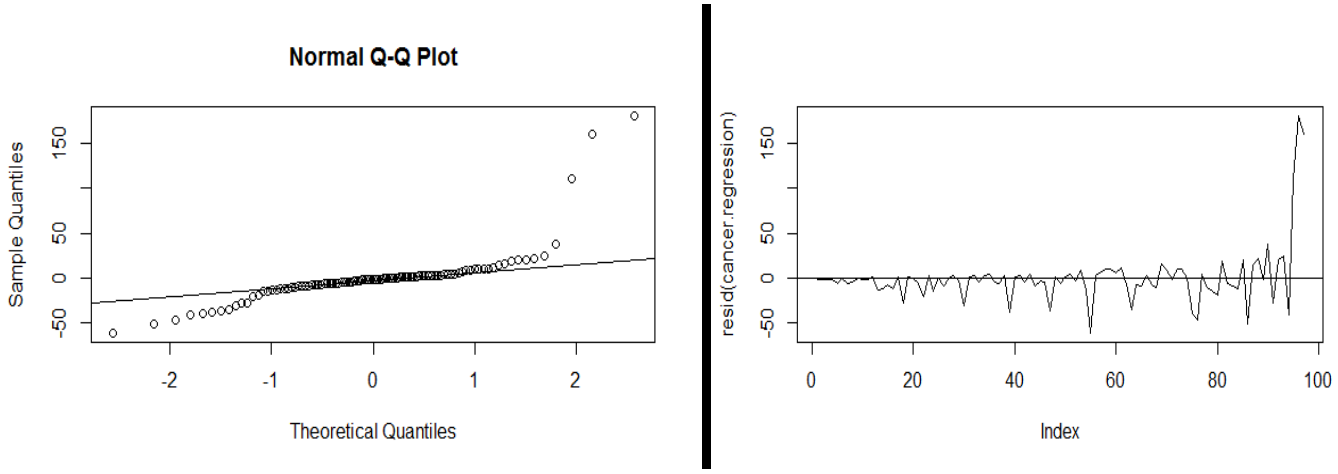
Mean of residual error is approximately zero **(-4.586823e-16**)

mean(resid(cancer.regression))



b) Errors are normally distributed.

We construct Normality Q-Q plot to verify if the errors are normal. We see that there are points diverging from the qq line i.e. the values are not completely on the line, but since most points lie on the line. Hence errors are approximately normally distributed.



c) Errors are independent

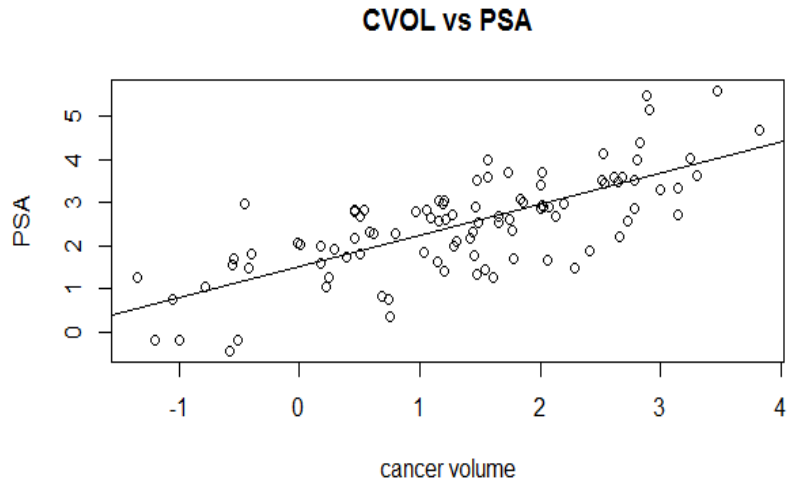
We construct Time series plot to verify if the errors are independent. A trend of up-down can be seen from this plot which shows that the errors are dependent.

The linear regression model cannot be considered as a perfect fit as the R values are not evident and high to be a perfect fit. As all the assumptions do not meet, we apply log transformations on response variable of the regression model to improve the model.

```
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)            x
     1.5092       0.7183
> summary(canc.regnew)
Call:
lm(formula = y ~ x)
Residuals:
    Min      1Q  Median      3Q     Max
-1.6778 -0.4187  0.1012  0.5035  1.9022
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.50923    0.12198   12.37   <2e-16 ***
x            0.71827    0.06822   10.53   <2e-16 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7879 on 95 degrees of freedom
Multiple R-squared:  0.5385,       Adjusted R-squared:  0.5336
F-statistic: 110.8 on 1 and 95 DF,  p-value: < 2.2e-16
```

After the log transformation the new linear model is:

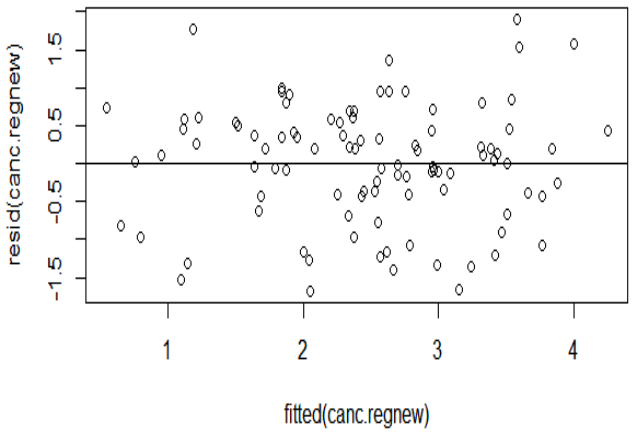


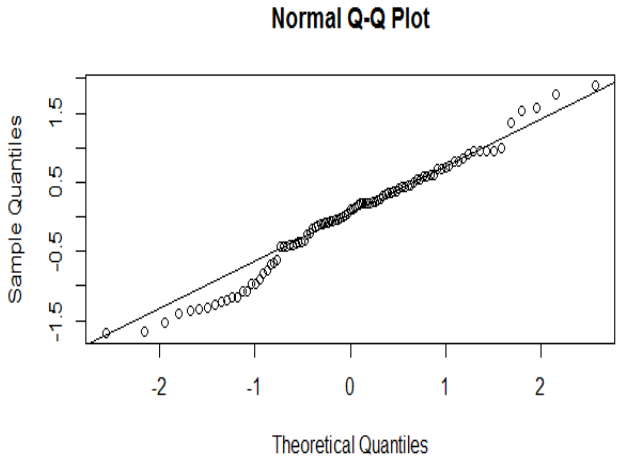
Regression model evaluation after log transformation:

a) **Errors have mean zero and constant variance.**

With log transformation, mean of residual error is zero **(-1.507376e-17**) which is approximately equal to 0 and the value is even closer to 0 than obtained in the previous model.

```
mean(resid(canc.regnew))
#[1]  -1.507376e-17
```



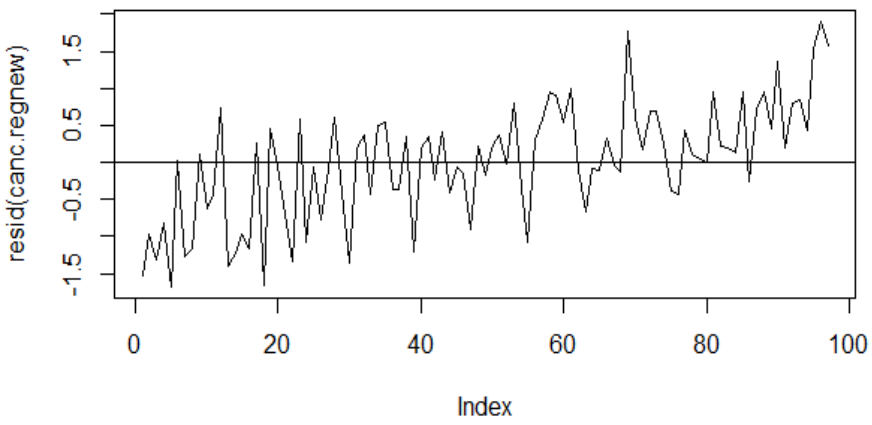


a) **Errors are normally distributed.**

We construct Normality Q-Q plot to verify if the errors are normal. We see that the plot fits approximately normal.

c) Errors are independent

It is difficult to spot a trend in the time series plot and also from the model information we can conclude that the value of R is higher than the value of R obtained in the previous model. Hence, we can conclude that this model is a better fit.



3)

3) Use the final model to predict the PSA level for a patient whose predictor variable value is at the median of the variable.

By using the final model the predict value of PSA level for a patient whose predictor variable at the median of the variable is Exp^(PSA) = **12.81632**

```
x = log(data[,3])
y = log(data[,2])
x.value<-data.frame(x=median(x))
exp(predict(canc.regnew,x.value))
```

R Code:

```
#reading csv
data = read.csv(file.choose(),header = T,sep = ",")
data

#plotting PSA against all the given variables - delete variables that are not
necessary

plot(data[,3],data[,2],xlab = "cancer volume", ylab = "PSA", main = "Cancer
Volume vs PSA")
abline(lm(data[,2]~data[,3]))
cor(data[,3],data[,2])
plot(data[,4],data[,2],xlab = "weight", ylab = "PSA", main = "Weight vs PSA")
abline(lm(data[,2]~data[,4]))
cor(data[,4],data[,2])
plot(data[,5],data[,2],xlab = "age", ylab = "PSA", main = "Age vs PSA")
cor(data[,5],data[,2])
abline(lm(data[,2]~data[,5]))
plot(data[,6],data[,2],xlab = "benpros", ylab = "PSA", main = "Benpros vs PSA")
cor(data[,6],data[,2])
abline(lm(data[,2]~data[,6]))
plot(data[,7],data[,2],xlab = "vesinv", ylab = "PSA", main = "Vesinv vs PSA")
cor(data[,7],data[,2])
abline(lm(data[,2]~data[,7]))
plot(data[,8],data[,2],xlab = "capspen", ylab = "PSA", main = "Capspen vs PSA")
cor(data[,8],data[,2])
abline(lm(data[,2]~data[,8]))
plot(data[,9],data[,2],xlab = "gleason", ylab = "PSA", main = "Gleason vs PSA")
cor(data[,9],data[,2])
abline(lm(data[,2]~data[,9]))

#considering cvol to be the quantitative variable
boxplot(data[,2],data[,3],main = "Boxplot")
x = data[,3] #cancervol
y = data[,2] #PSA

#linear regression model
cancer.regression = lm(y~x)
cancer.regression

#plotting
plot(data[,3],data[,2],xlab = "cancer volume", ylab = "PSA", main = "CVOL vs
PSA",abline(cancer.regression))
z = median(y)
regline = 1.125 + 3.230*z

anova(cancer.regression)
summary(cancer.regression)
confint(cancer.regression)

#residual plot
plot(fitted(cancer.regression), resid(cancer.regression))
abline(h=0)
mean(resid(cancer.regression))

# QQ plot

qqnorm(resid(cancer.regression))
qqline(resid(cancer.regression))
```

```
# Time series plot of residuals

plot(resid(cancer.regression), type="l")
abline(h=0)

# new regression model after log transformations

x = log(data[,3])
y = log(data[,2])

#new linear model
canc.regnew = lm(y~x)
canc.regnew

summary(canc.regnew)

plot(log(data[,3]),log(data[,2]),xlab = "cancer volume", ylab = "PSA", main =
"CVOL vs PSA",abline(canc.regnew))

z = median(y)
regline = 1.5092+ 0.7183*z

#residual
plot(fitted(canc.regnew), resid(canc.regnew))
abline(h=0)

#mean value of residuals

mean(resid(canc.regnew))

# QQ plot

qqnorm(resid(canc.regnew))
qqline(resid(canc.regnew))

# Time series plot of residuals
plot(resid(canc.regnew), type="l")
abline(h=0)

#Predict PSA level
x.value<-data.frame(x=median(x))
exp(predict(canc.regnew,x.value))
```