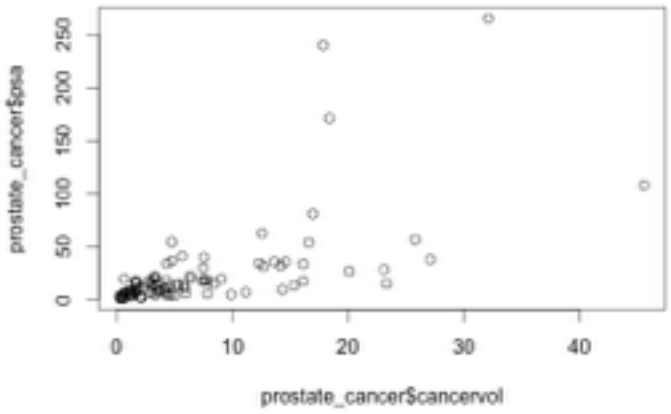
Mini Project: #5
Name: Ajay Vembu

Part - 1:

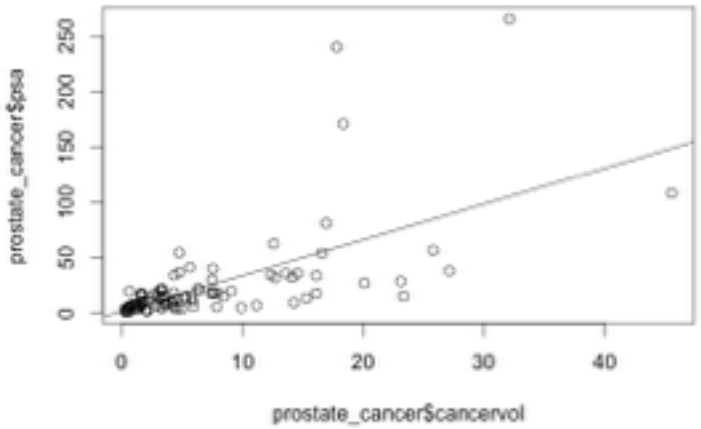
The scatter plots are plotted between all the variables of the data and psa and inferred that the quantitative variable **cancervol** is a significant predictor for the response variable **psa**.



The potential outliers here are the three points highlighted in the graph.

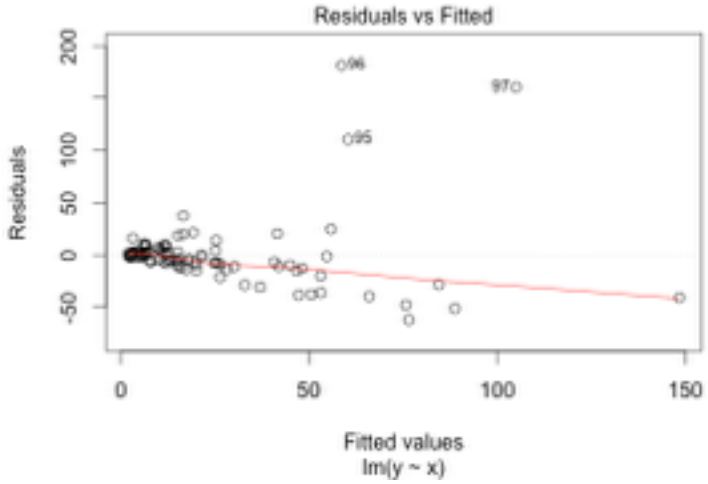
Part - 2:

The linear regression model is constructed and below is the regression line,



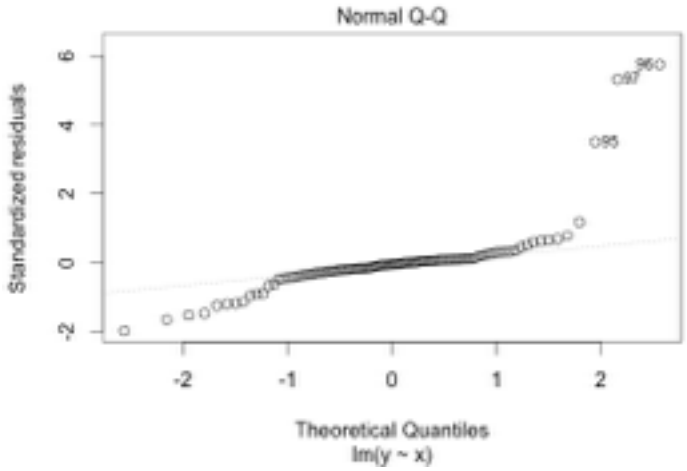
Two plots are made to check three key regression assumptions,
Assumption-1: The residuals and the fitted values should not be related (should be constant)
Assumption-2: The relation between the predictor and the response variable should be linear.
Assumption-3: The residuals should be normally distributed.

Below is the plot between the residuals and the fitted values,



The above plot shows that residuals and the fitted values are related, that is the residuals decreases are the fitted values increases which is a violation of the assumption-1 and the red line indicate that predictor and response are not perfectly linear which is a violation of assumption-2.

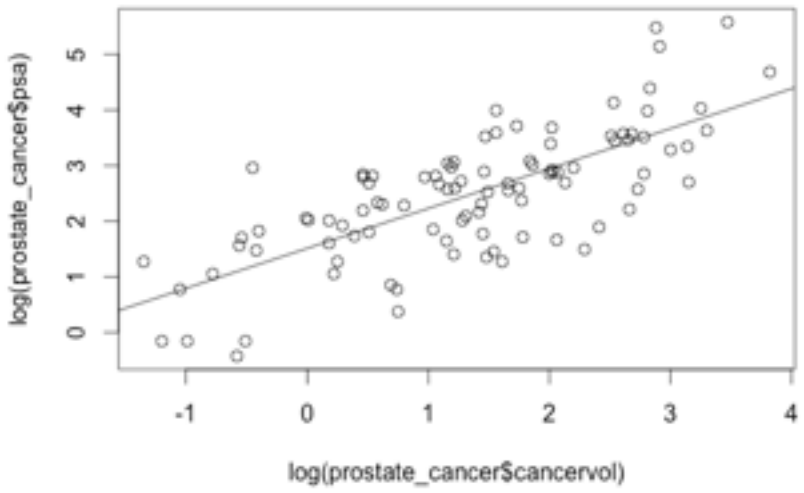
Below is the plot QQ plot for the residuals,



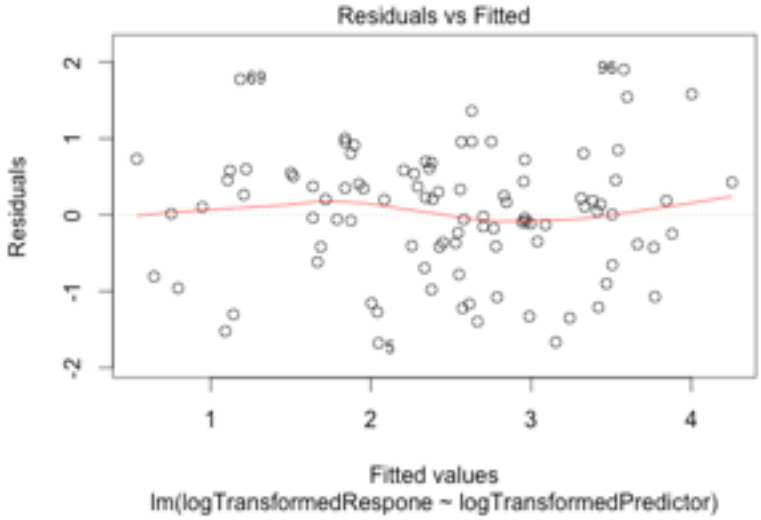
Strong skewness at the tails indicate that residuals are not normally distributed which is a violation of assumption-3.

The remedy attempted for this is the log transformation of the response variable and the predictor variable.

Below is the regression model fitted after the transformation,



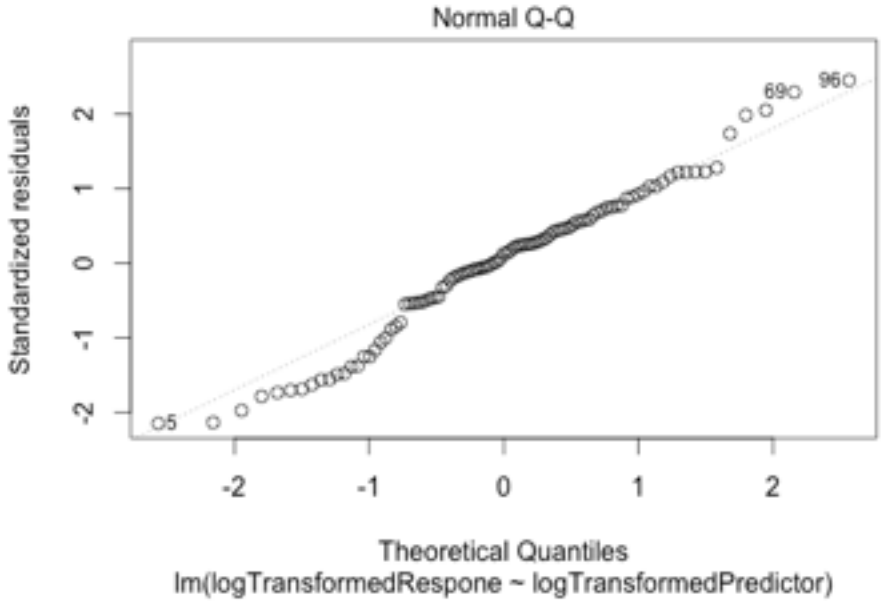
Below is the plot between the residuals and the fitted line for the transformed model,



This plot shows that the violated regression assumptions are countered.

The residuals are not changing for the fitted values and the red line indicate that the relationship between the residuals and fitted values are linear.

QQ plot for the residuals



This shows that the residuals are approximately normal and counters the assumption violation.

Fit of the final model

By hypothesis testing,

H0: beta0 = 0 (NULL)
H1: beta0 != 0 (ALTERNATIVE)

The computed pValue is found to be 2.2e-16 which is at the 2.5% level of significance we can reject the NULL hypothesis and accept the ALTERNATIVE that beta0 not equal to 0

The adjusted R value if found out to be 0.5336 which explains 53.36% of the variability.

Part - 3:

The predicted psa value found out to be 12.81632 at the median of the cancervol.

RCode:

Part - 1:

```
# the scatter plot between all the variables and the psa

plot  (x=prostate_cancer$cancervol,y=prostate_cancer$psa);
plot  (x=prostate_cancer$weight,y=prostate_cancer$psa);
plot  (x=prostate_cancer$age,y=prostate_cancer$psa);
plot  (x=prostate_cancer$benpros,y=prostate_cancer$psa);
plot  (x=prostate_cancer$vesinv,y=prostate_cancer$psa);
plot  (x=prostate_cancer$capspen,y=prostate_cancer$psa);
plot  (x=prostate_cancer$gleason\y=prostate_cancer$psa);
```

Part - 2:

```
# Before transformation

x <- prostate_cancer$cancervol
y <- prostate_cancer$psa

# the model
prostateCancerVal <- lm ( formula = y ~ x )

summary(prostateCancerVal)
```

Output:

```
# before transformation

Call:
lm(formula = y ~ x)

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
```

Here **beta0 = 1.1249 and beta1 = 3.2299** and the regression line is
Y = 1.1249 + 3.2299X and R-squared value in 0.3831 which explains 38.31 % of variability.

```
# after transformation

logTransformedRespone <- log ( y )
logTransformedPredictor <- log ( x )

# the log transformed fitted model

prostateCancerVal <- lm ( formula = logTransformedRespone ~ logTransformedPredictor )

summary(prostateCancerVal)
```

Output:

```
Call:
lm(formula = logTransformedRespone ~ logTransformedPredictor)

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 ***
logTransformedPredictor  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
```

Here **beta0 = 1.50923 and beta1 = 0.71827** and the regression line is
Y = 1.50923 + 0.71827X and R-squared value in 0.5336 which explains 53.36 % of variability.

Part - 3:

```
# predict the data

x.log.median <- log ( median(x) )
x.new <- data.frame( logTransformedPredictor =  x.log.median  )
predictedVal <- predict (prostateCancerVal , newdata = x.new)
print (exp(predictedVal))
```