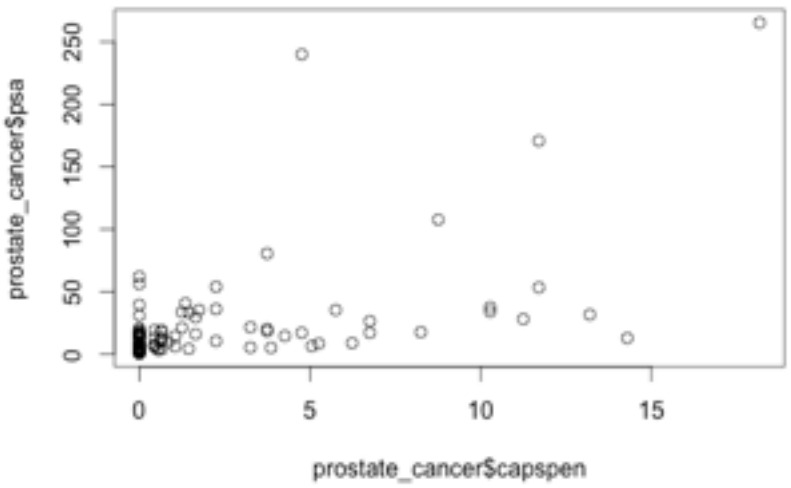
Mini Project: #6
Name: Ajay Vembu

Part - 1:

The next variable which is chosen for building the model is the **capspen**, which is the next best quantitative variable which has a linear trend, positive direction and strong linear relationship.

Below is the scatter plot between the **capspen** and the **psa** variable. Here the predictor variables are the psa and the capspen.



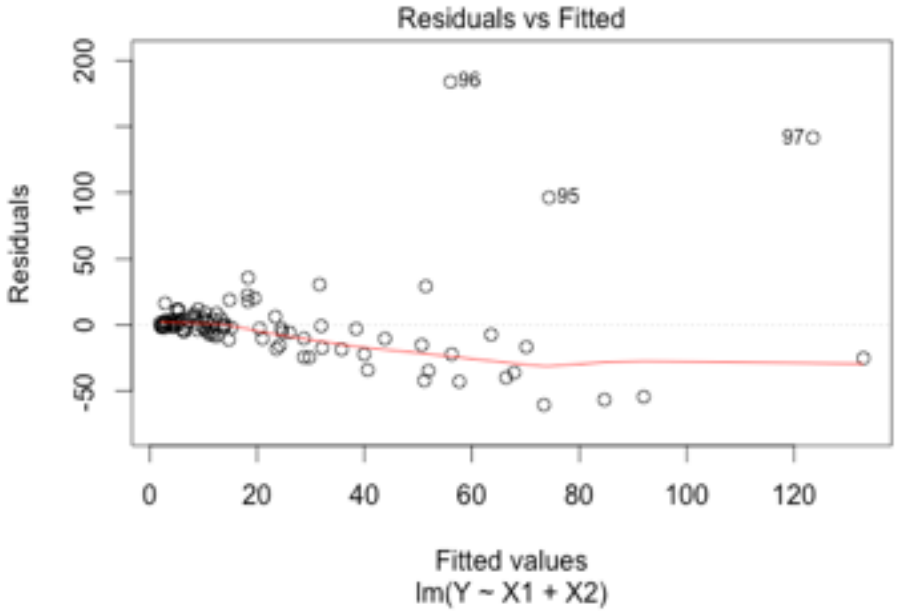
The potential outliers are highlighted.

The linear regression model is constructed without any transformation of the predictor or the response variables.

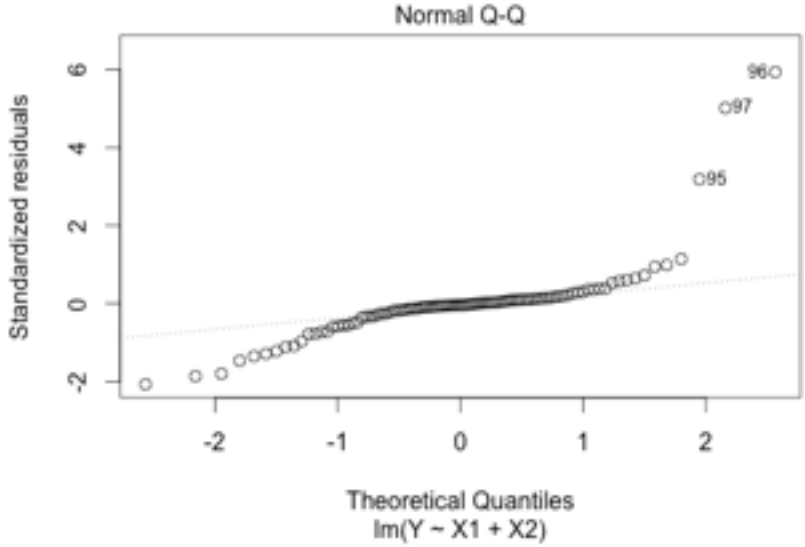
Below are the assumptions made while constructing the regression model,

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 (in the center) indicate that predictor and response are not perfectly linear which is a violation of assumption-2.

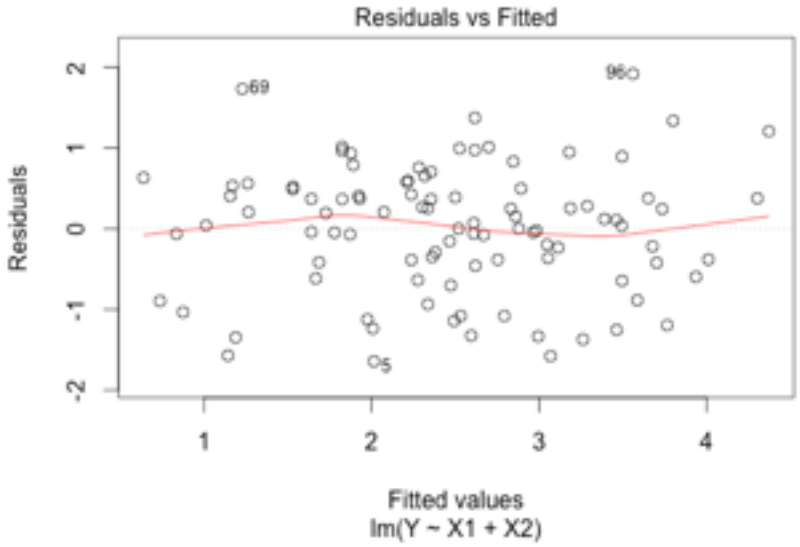


Strong skewness at the tails indicate that residuals are not normally distributed which is a violation of assumption-3.

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

Below are the regression diagnostics after the transformation,

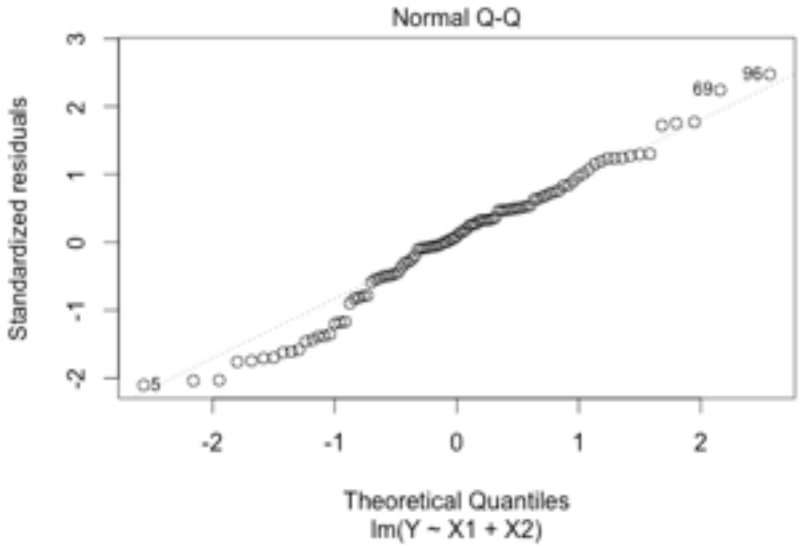
Plot between the residuals and the fitted line for the transformed model,



This plot shows that the violated regression assumptions are countered.

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

QQ plot for the residuals



Fit of the final model

By hypothesis testing,

H0: beta1 = beta2 = 0 (NULL)
H1: beta1 != 0 or beta2 != 0 (ALTERNATIVE - at least one is not 0)

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 at least one of the regression coefficients are not equal to 0.

The adjusted R value is found out to be 0.5355 which explains 53.55% of the variability.

Below is the anova table before adding the variable capspen and after adding the variable capspen to check the model significance,

Output:

Model 1: Y ~ X1
Model 2: Y ~ X1 + X2

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	95	58.968				
2	94	58.114	1	0.8537	1.3809	0.2429

After comparing the two models it is found that the P value has increased which shows us that the addition of variable capspen has not improved the model significantly.

Part - 2:

The predicted value at the median of the cancervol and capspen is found out to have a spa value of 11.86018.

R - code:

```
# to find the scatter plot between psa and capspen

plot (y = prostate_cancer$psa, x = prostate_cancer$capspen)

# the linear model construction

# Before Transformation

Y <- prostate_cancer$psa
X1 <- prostate_cancer$cancervol
X2 <- prostate_cancer$capspen

prostateCancerValMultiBefore <- lm ( formula = Y ~ X1 + X2 )
summary(prostateCancerValMultiBefore)
```

Output:

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-60.346  -8.324  -1.205   4.159 183.843

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3276     4.2861   0.310    0.757
X1            2.4139     0.5655   4.269 4.69e-05 ***
X2            2.4533     1.1779   2.083    0.040 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.48 on 94 degrees of freedom
Multiple R-squared:  0.4165,  Adjusted R-squared:  0.4041
F-statistic: 33.55 on 2 and 94 DF,  p-value: 1.01e-11
```

Here **beta0 = 1.3276, beta1 = 2.4139 and beta2 = 2.4533** and the regression line is **Y = 1.3276 + 2.4139X1 + 2.4533X3** and R-squared value in 0.4041 which explains 40.41 % of variability.

```
# after transformation,
```

Output:

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.64429 -0.42310  0.06919  0.49755  1.91878

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.52299    0.12229  12.454  < 2e-16 ***
X1           0.65531    0.08664   7.564  2.6e-11 ***
X2           0.03172    0.02699   1.175    0.243
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7863 on 94 degrees of freedom
Multiple R-squared:  0.5452,  Adjusted R-squared:  0.5355
F-statistic: 56.33 on 2 and 94 DF,  p-value: < 2.2e-16
```

Here **beta0 = 1.52299, beta1 = 0.65531 and beta2 = 0.03172** and the regression line is **Y = 1.52299 + 0.65531X1 + 0.03172X3** and R-squared value in 0.5355 which explains 53.55 % of variability.

```
# to get the anova table for model significance,

anova (prostateCancerVal,prostateCancerValMultiAfter)

# to get the residual and QQ plots

plot (prostateCancerValMultiBefore)
plot (prostateCancerValMultiAfter)
```

Part - 2:

```
# predicting the final Y value

X1.log.median <- log ( median ( prostate_cancer$cancervol ) )
X2.median <- median ( prostate_cancer$capspen )

x.new <- data.frame( X1 = X1.log.median,  X2 = X2.mdeian )
predictedVal <- predict (prostateCancerValMultiAfter , newdata = x.new)
print ( exp(predictedVal) )
```