

# Mini Project # 6

Qingyu Lan, Lakshmi Priyanka Selvaraj

## Contribution of each group member :

Both members worked on the questions together.

## Section 1. Answers to the specific questions asked

### 1. Question 1

Step 1: To build a linear model we need to analyse the linear relationship between the predictors and the response variable.

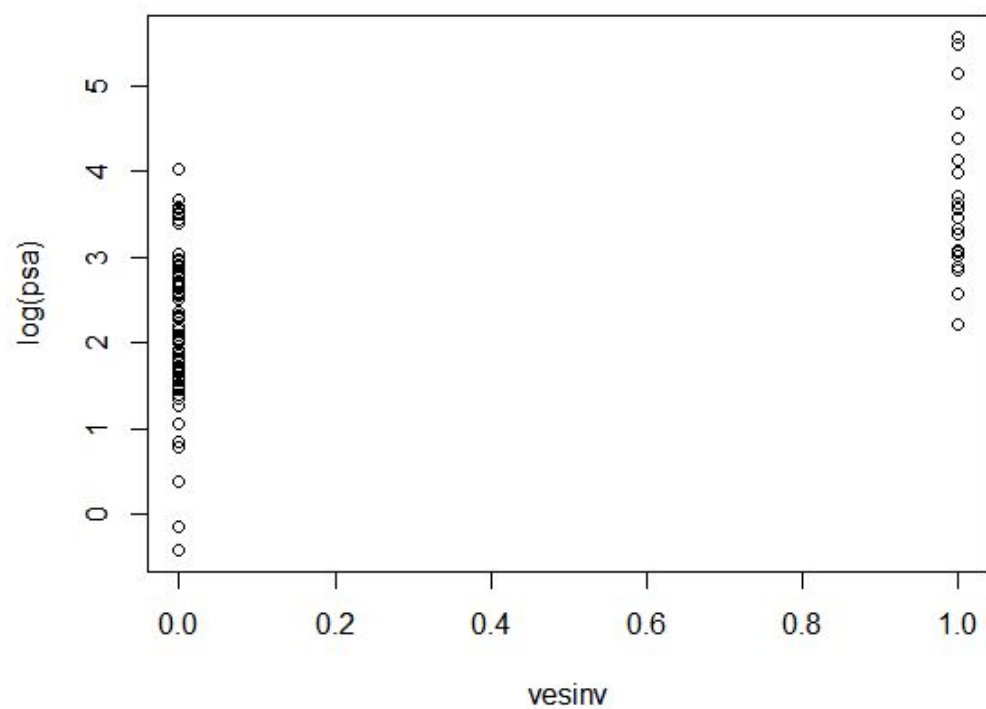
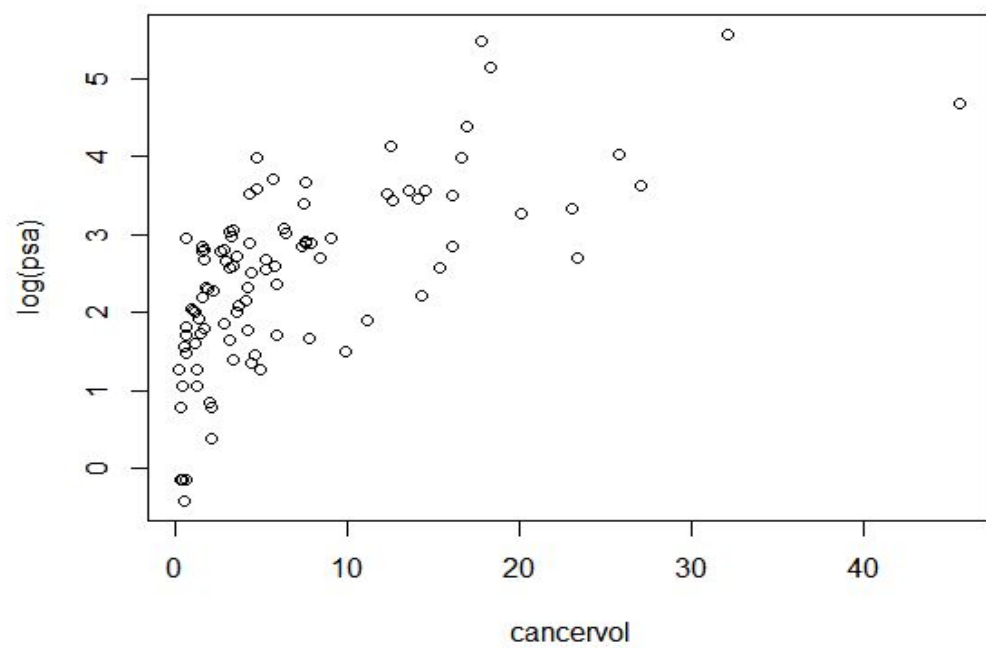
Given: psa is the response variable and the rest of the variables are predictors.

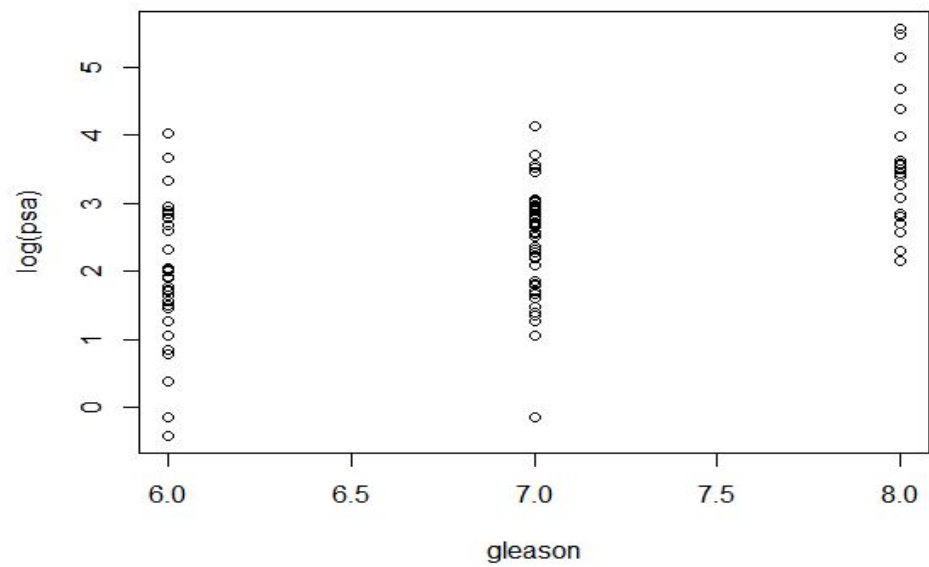
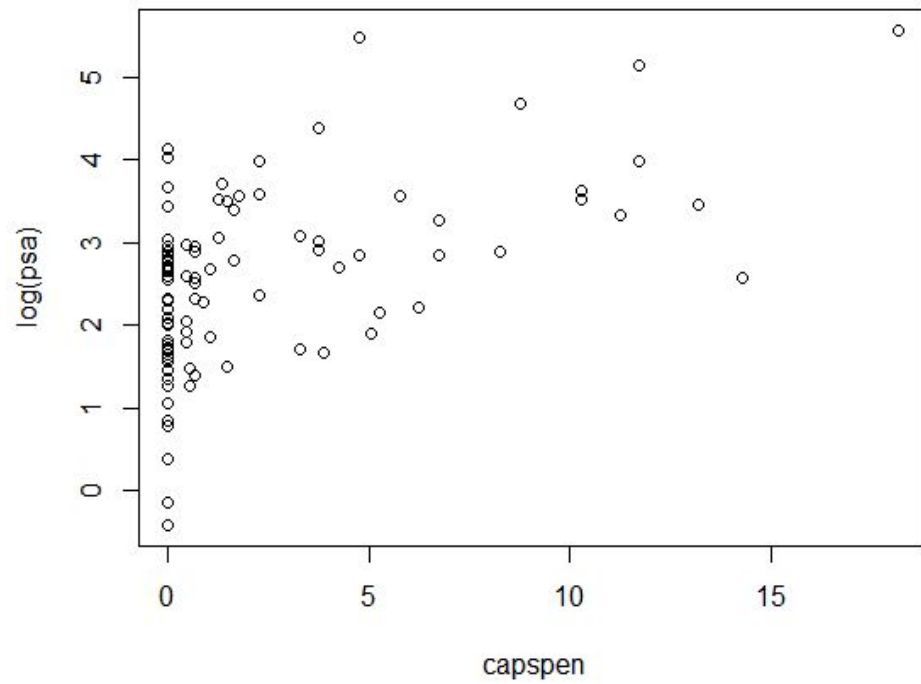
We plot each of the predictors versus the response variable. Also, we apply log transformation for the response variable, as this transformation improves the linear relationship.

For better understanding of the linear relationship between the variables, we have taken the correlation between the variables, which casts light upon the positive/negative linear relationship between variables. Upon, observation, there were four variables whose correlation was above 0.5 and seemed worthwhile considering.

The variables/predictors are cancervol, vesinv, capspen, gleason.

Now we will look at their scatterplots and correlations.





```
> cor(cancervol,log(psa))
[1] 0.6570739
> cor(vesinv,log(psa))
[1] 0.5663641
> cor(capspen,log(psa))
[1] 0.5180231
> cor(gleason,log(psa))
[1] 0.5390167
> |
```

Vesinv is a factor/qualitative variable and gleason is a quantitative variable.

As we can see, all the scatterplots show not much of a linear trend, hence we have chosen the correlation values to guide us in choosing the variables that might have an effect on the response variable.

Also, further we want to be sure about ruling out the other variables and their effects on the response variable, hence, we build a linear model to observe their p-values.

Variables we want to rule out: Weight, Age, Benpros

Null Hypothesis: Slope values for weight, age and benpros are zero

Alternative Hypothesis: At Least one among these variables have a slope which is non-zero

```
> fit1 <- lm(log(psa)~weight+age+benpros)
> anova(fit1)
Analysis of Variance Table

Response: log(psa)
          Df Sum Sq Mean Sq F value Pr(>F)
weight     1   1.893   1.89301   1.4414 0.2330
age         1   2.951   2.95084   2.2468 0.1373
benpros     1   0.786   0.78558   0.5982 0.4412
Residuals  93 122.139   1.31333
```

We can clearly observe that all the three variables have a p-value  $>0.05$  which goes on to show that we can accept the null hypothesis and reject the alternative hypothesis. When we try to build the actual model, we can avoid these three variables.

Next step is to build the linear model with the other variables available. Stepwise selection with BIC is more realistic and tries to build a decent model with minimal number of variables. Hence, we use this method.

```

> nullmd = lm(log(cancer_data$psa)~1, data= cancer_data)
> step(nullmd,scope= list(lower=~1, upper=~cancervol + as.factor(vesinv) +
+                           capspen+gleason), k = log(97))
Start: AIC=31.3
log(cancer_data$psa) ~ 1

              Df Sum of Sq    RSS    AIC
+ cancerrvol  1    55.164   72.605 -18.9492
+ as.factor(vesinv) 1    40.984   86.785 -1.6449
+ gleason      1    37.122   90.647  2.5788
+ capspen     1    34.286   93.482  5.5663
<none>                127.769  31.2993

Step: AIC=-18.95
log(cancer_data$psa) ~ cancerrvol

              Df Sum of Sq    RSS    AIC
+ gleason      1     8.247   64.358 -26.070
+ as.factor(vesinv) 1     6.547   66.058 -23.541
<none>                72.605 -18.949
+ capspen     1     0.967   71.638 -15.675
- cancerrvol  1    55.164  127.769  31.299

Step: AIC=-26.07
log(cancer_data$psa) ~ cancerrvol + gleason

              Df Sum of Sq    RSS    AIC
+ as.factor(vesinv) 1     4.0178  60.340 -27.7480
<none>                64.358 -26.0697
+ capspen          1     0.1685  64.190 -21.7493
- gleason           1     8.2468  72.605 -18.9492
- cancerrvol        1    26.2887  90.647  2.5788

Step: AIC=-27.75
log(cancer_data$psa) ~ cancerrvol + gleason + as.factor(vesinv)

              Df Sum of Sq    RSS    AIC
<none>                60.340 -27.748
- as.factor(vesinv)  1     4.0178  64.358 -26.070
+ capspen            1     0.3013  60.039 -23.659
- gleason             1     5.7179  66.058 -23.541
- cancerrvol          1    12.7041  73.044 -13.789

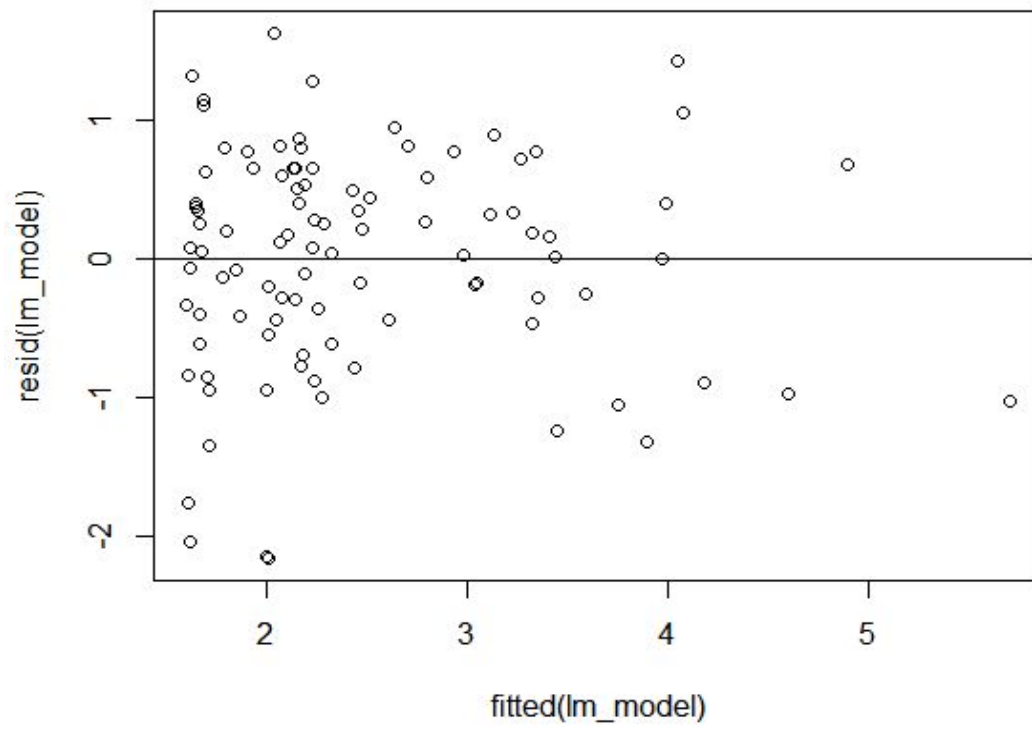
Call:
lm(formula = log(cancer_data$psa) ~ cancerrvol + gleason + as.factor(vesinv),
    data = cancer_data)

Coefficients:
              (Intercept)              cancerrvol              gleason  as.factor(vesinv)1
                -0.72120                0.05981                0.38491                0.62117

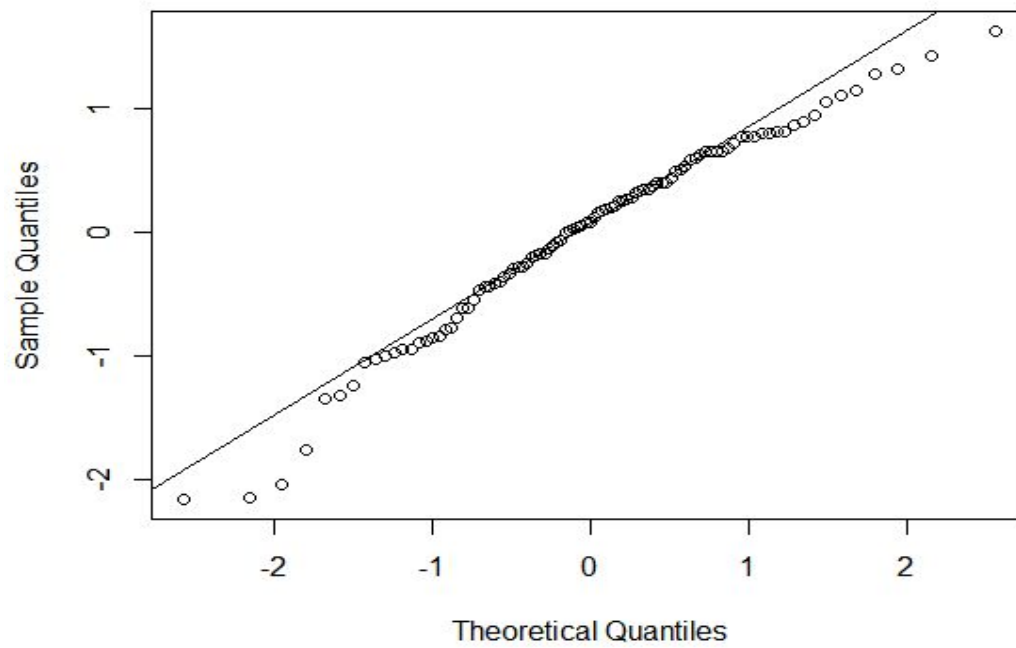
```

This model has suggested us to use three predictors cancerrvol, gleason and vesinv(factor variable).

To verify if it is a good model we try to look at the fitted values and the residuals. We plot the fitted value against the residuals and also look at the QQplot for the residuals



**Normal Q-Q Plot**



We observe that towards both the tails, the residuals seem to be way below the normality line.

We also want to try a stepwise model with AIC.

```
> nullmd = lm(log(cancer_data$psa)~1, data= cancer_data)
> #step 2 WE are going to try stepwise selection with AIC
> step(nullmd, scope= list(lower=~1, upper=~cancervol
+                               + as.factor(vesinv) +
+                               capspen +gleason), k = 2)
Start:  AIC=28.72
log(cancer_data$psa) ~ 1
```

	Df	Sum of Sq	RSS	AIC
+ cancervol	1	55.164	72.605	-24.0986
+ as.factor(vesinv)	1	40.984	86.785	-6.7944
+ gleason	1	37.122	90.647	-2.5707
+ capspen	1	34.286	93.482	0.4169
<none>			127.769	28.7246

```
Step:  AIC=-24.1
log(cancer_data$psa) ~ cancervol
```

	Df	Sum of Sq	RSS	AIC
+ gleason	1	8.247	64.358	-33.794
+ as.factor(vesinv)	1	6.547	66.058	-31.265
<none>			72.605	-24.099
+ capspen	1	0.967	71.638	-23.400
- cancervol	1	55.164	127.769	28.725

```
Step:  AIC=-33.79
log(cancer_data$psa) ~ cancervol + gleason
```

	Df	Sum of Sq	RSS	AIC
+ as.factor(vesinv)	1	4.0178	60.340	-38.047
<none>			64.358	-33.794
+ capspen	1	0.1685	64.190	-32.048
- gleason	1	8.2468	72.605	-24.099
- cancervol	1	26.2887	90.647	-2.571

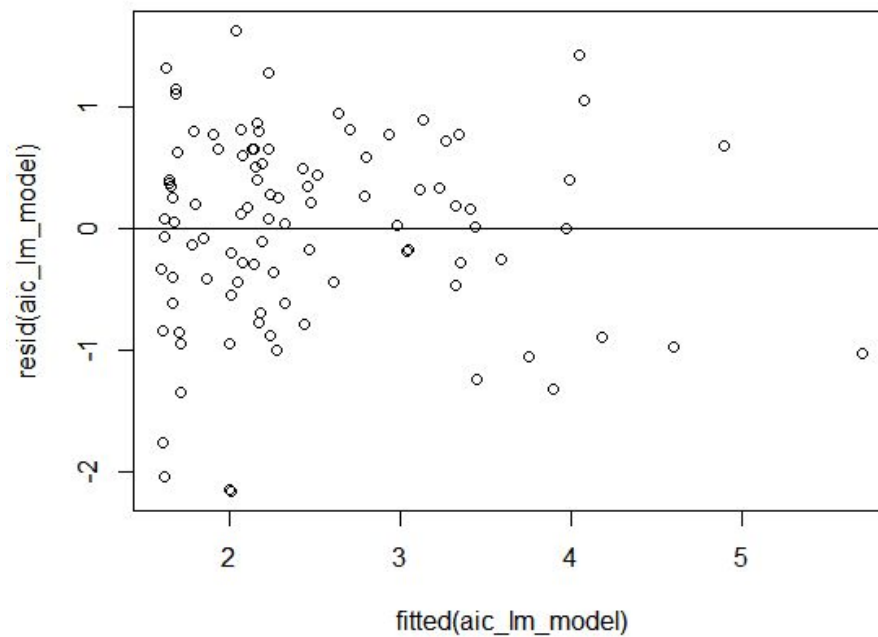
```
Step:  AIC=-38.05
log(cancer_data$psa) ~ cancervol + gleason + as.factor(vesinv)
```

	Df	Sum of Sq	RSS	AIC
<none>			60.340	-38.047
+ capspen	1	0.3013	60.039	-36.532
- as.factor(vesinv)	1	4.0178	64.358	-33.794
- gleason	1	5.7179	66.058	-31.265
- cancervol	1	12.7041	73.044	-21.513

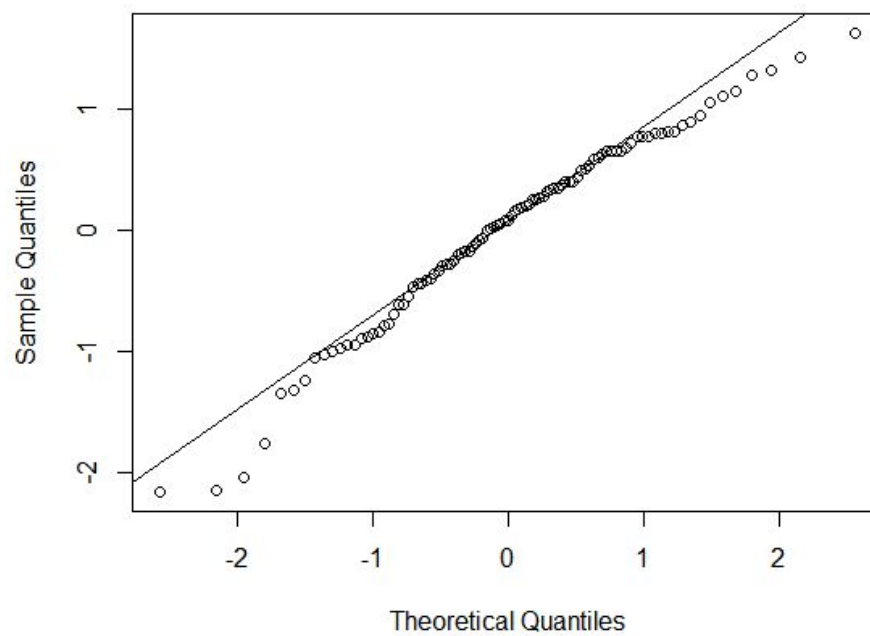
```
Call:
lm(formula = log(cancer_data$psa) ~ cancervol + gleason + as.factor(vesinv),
    data = cancer_data)

Coefficients:
      (Intercept)      cancervol      gleason  as.factor(vesinv)1
      -0.72120         0.05981         0.38491         0.62117
```

Below are the residual vs fitted scatterplot, and qqplot for the residuals.



**Normal Q-Q Plot**



These plots are very similar to the Stepwise selection model calculated with BIC.

Though we initially ruled out three variables, we would like to build a stepwise selection with all variables to see if there is any significant improvement in the model result.



```
> step(nullmd,scope=list(lower=~1,upper=~weight+age+benpros+cancervol+
+                             gleason+as.factor(vesinv)+capspen))
```

```
Start: AIC=28.72
```

```
log(cancer_data$psa) ~ 1
```

	Df	Sum of Sq	RSS	AIC
+ cervixvol	1	55.164	72.605	-24.0986
+ as.factor(vesinv)	1	40.984	86.785	-6.7944
+ gleason	1	37.122	90.647	-2.5707
+ capspen	1	34.286	93.482	0.4169
+ age	1	3.688	124.080	27.8831
+ benpros	1	3.166	124.603	28.2911
<none>			127.769	28.7246
+ weight	1	1.893	125.876	29.2767

```
Step: AIC=-24.1
```

```
log(cancer_data$psa) ~ cervixvol
```

	Df	Sum of Sq	RSS	AIC
+ gleason	1	8.247	64.358	-33.794
+ benpros	1	7.803	64.802	-33.128
+ as.factor(vesinv)	1	6.547	66.058	-31.265
+ age	1	2.662	69.944	-25.721
+ weight	1	1.790	70.815	-24.520
<none>			72.605	-24.099
+ capspen	1	0.967	71.638	-23.400
- cervixvol	1	55.164	127.769	28.725

```
Step: AIC=-33.79
```

```
Step: AIC=-33.79
```

```
log(cancer_data$psa) ~ cervixvol + gleason
```

	Df	Sum of Sq	RSS	AIC
+ benpros	1	6.2827	58.075	-41.758
+ as.factor(vesinv)	1	4.0178	60.340	-38.047
+ weight	1	2.0334	62.325	-34.908
<none>			64.358	-33.794
+ age	1	0.9611	63.397	-33.253
+ capspen	1	0.1685	64.190	-32.048
- gleason	1	8.2468	72.605	-24.099
- cervixvol	1	26.2887	90.647	-2.571

```
Step: AIC=-41.76
```

```
log(cancer_data$psa) ~ cervixvol + gleason + benpros
```

	Df	Sum of Sq	RSS	AIC
+ as.factor(vesinv)	1	4.8466	53.229	-48.211
<none>			58.075	-41.758
+ weight	1	0.4006	57.675	-40.429
+ capspen	1	0.1863	57.889	-40.069
+ age	1	0.0059	58.070	-39.768
- benpros	1	6.2827	64.358	-33.794
- gleason	1	6.7262	64.802	-33.128
- cervixvol	1	29.9589	88.034	-3.407

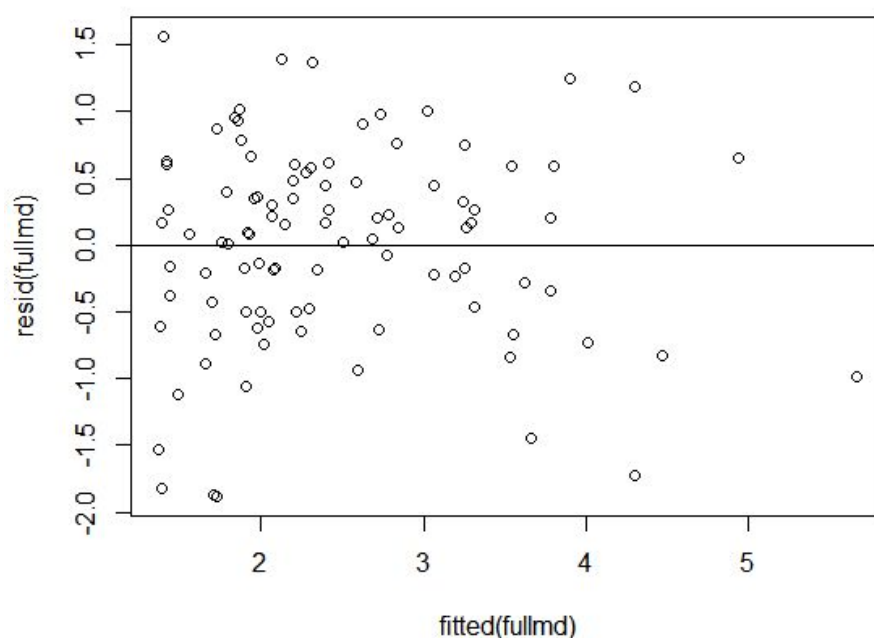
```
Step: AIC=-48.21
log(cancer_data$psa) ~ cancervol + gleason + benpros + as.factor(vesinv)
```

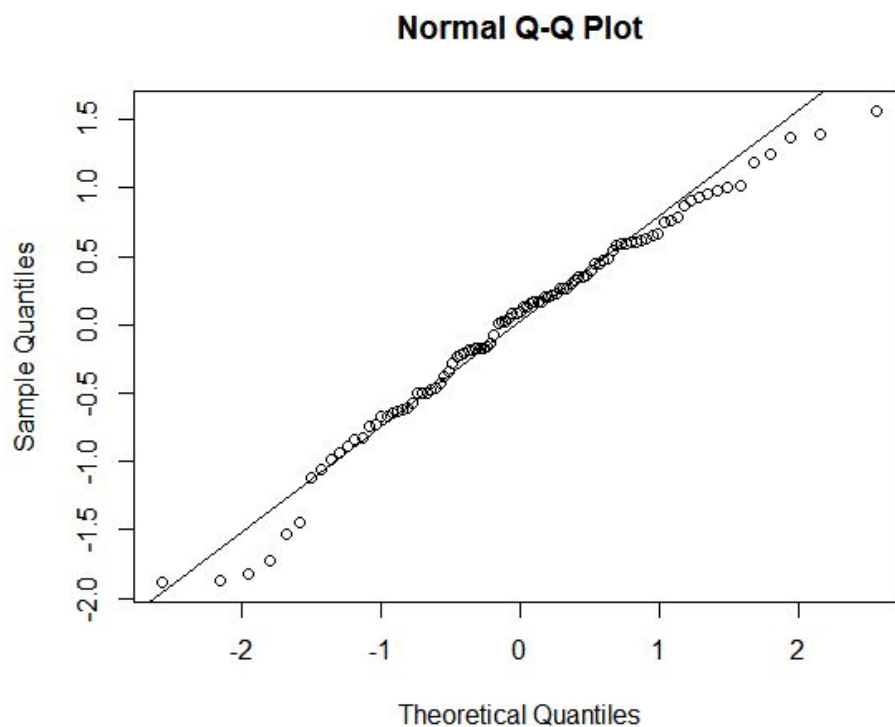
	Df	Sum of Sq	RSS	AIC
<none>			53.229	-48.211
+ capspen	1	0.3923	52.837	-46.928
+ weight	1	0.3306	52.898	-46.815
+ age	1	0.0250	53.204	-46.256
- gleason	1	4.2389	57.468	-42.778
- as.factor(vesinv)	1	4.8466	58.075	-41.758
- benpros	1	7.1115	60.340	-38.047
- cancervol	1	14.7580	67.987	-26.473

```
Call:
lm(formula = log(cancer_data$psa) ~ cancervol + gleason + benpros +
    as.factor(vesinv), data = cancer_data)
```

```
Coefficients:
      (Intercept)          cancervol          gleason          benpros
      -0.65013           0.06488           0.33376           0.09136
as.factor(vesinv)1
      0.68421
```

We can observe that the three variables we initially rejected are not included in the model, but all the other variables are available. To see if this is a better model, we look at the fitted vs residual scatter plots and the qq plot for residuals.





We can clearly see from the qqplot, the residuals show a significantly better normal plot and hence, the final model is inclusive of these four predictors(cancervol,gleason,benpros,vesinv)

The regression result is:

```

Coefficients:
      (Intercept)      cancervol      gleason      benpros
          -0.65013           0.06488           0.33376           0.09136
as.factor(vesinv)1
           0.68421

```

The predicted value for psa according to specifications given is

```

> pred_1 <- mean(cancervol)
> pred_2 <- mean(gleason)
> pred_3 <- mean(benpros)
> table(vesinv)
vesinv
 0  1
76 21
> pred_4 <- 0
> x_new <- data.frame(cancervol=pred_1, gleason=pred_2, benpros=pred_3, vesinv= pred_4)
> predict(fullmd, newdata = x_new)
           1
2.330541

```

Equation =

**-0.65013+(0.06488\*pred\_1)+(0.33376\*pred\_2)+(0.09136\*pred\_3)+(0.68421\*pred\_4)**

-----  
**Predicted\_Psa = 2.330541**

## Section 2: R code.

R code for question 1

```
#mini project 6
#Team members
#Qingyu Lan
#Lakshmi Priyanka Selvaraj

#Q1
rm(list =ls())
setwd("D:/UTD/Fall_2020/1. Statistical Methods of Data Science/Project work")

cancer_data <- read.csv("prostate_cancer.csv")
head(cancer_data)
str(cancer_data)

#First and foremost is we try to plot and observe linear trends
attach(cancer_data)
plot(cancervol,log(psa))
cor(cancervol,log(psa))
plot(weight,log(psa))
cor(weight,log(psa))
plot(age,log(psa))
cor(age,log(psa))
plot(benpros,log(psa))
cor(benpros,log(psa))
plot(vesinv,log(psa))
cor(vesinv,log(psa))
plot(capspen,log(psa))
cor(capspen,log(psa))
plot(gleason,log(psa))
cor(gleason,log(psa))

#None of the above variables show a linear trend with the psa-response variable

cor(cancervol,psa)
```

#The variables Cancervol and psa seem to have a strong positive correlation though

```
table(cancer_data$vesinv)
table(cancer_data$gleason)
```

```
#We are going to try some individual variables
fit1 <- lm(log(psa)~weight+age+benpros)
```

```
summary(fit1)
```

```
anova(fit1)
```

```
step(nullmd,scope=list(lower=~1,upper=~weight+age+benpros))
```

#Step 1 I am going to try Stepwise Selection with BIC

```
nullmd = lm(log(cancer_data$psa)~1, data= cancer_data)
nullmd
```

```
step(nullmd,scope= list(lower=~1, upper=~cancervol + as.factor(vesinv) +
                        capspen +gleason), k = log(97))
```

```
lm_model <- lm(formula = log(cancer_data$psa) ~ cancervol + gleason + as.factor(vesinv),
               data = cancer_data)
```

```
fitted(lm_model)
resid(lm_model)
```

```
plot(fitted(lm_model),resid(lm_model))
abline(h=0)
```

```
qqnorm(resid(lm_model))
qqline(resid(lm_model))
```

#step 2 WE are going to try Stepwise Selection with AIC

```
step(nullmd, scope= list(lower=~1, upper=~cancervol
                        + as.factor(vesinv) +
                        capspen +gleason), k = 2)
```

```
aic_lm_model <- lm(formula = log(cancer_data$psa) ~ cancervol + gleason +
                  as.factor(vesinv),
                  data = cancer_data)
```

```
plot(fitted(aic_lm_model),resid(aic_lm_model))
abline(h=0)
```

```
qqnorm(resid(aic_lm_model))
qqline(resid(aic_lm_model))
```

```
#Stepwise selection with BIC
nullmd = lm(log(cancer_data$psa)~1, data= cancer_data)
step(nullmd,scope=list(lower=~1,upper=~weight+age+benpros+cancervol+
                      gleason+as.factor(vesinv)+capspen))
```

```
fullmd <- lm(formula = log(cancer_data$psa) ~ cancervol + gleason + benpros +
             as.factor(vesinv), data = cancer_data)
```

```
plot(fitted(fullmd),resid(fullmd))
abline(h=0)
```

```
qqnorm(resid(fullmd))
qqline(resid(fullmd))
```

```
#Prediction 3 quantitative variables, 1 qualitative variable
pred_1 <- mean(cancervol)
pred_2 <- mean(gleason)
pred_3 <- mean(benpros)
table(vesinv)
```

```
pred_4 <- 0
```

```
x_new <- data.frame(cancervol=pred_1, gleason=pred_2, benpros=pred_3, vesinv= pred_4)
```

```
predict(fullmd, newdata = x_new)
```