MINI PROJECT 5

By- Raghav Mathur (rxm162130), Deepak Shanmugam (dxs161930)

Contribution to the Project: Each member contributed equally to the project. We worked together to analyse the predictor variables (qualitative and quantitative), created the model by considering each predictor variables separately and as a set of combinations.

EXERCISE 1:

Section 1: Analysing the data, we came to know that it had 2 categorical variables- '**gleanson'** and '**vesinv'**. So, we considered factors of these variables. The rest were quantitative variables.

We made a linear model of response variable '**psa'**, taking each predictor variable independently, we rejected some of the insignificant variables based on the -value. Next, we compared the ANOVA tables of linear models having different sets of predictor variables and formulated a preliminary model.

We take log () of the response variable (Y) as it seems to have a better coverage of the data distribution of response variable and has a better fit of QQplot with some few outliers compared to the QQplot and Boxplot of Y or sqrt(Y).

Finally, we take lm(**y ~ cancervol + benpros + gleason + vesinv**, data = cancer) as our preliminary model for the data. However, we performed the diagnostics before accepting this model using automatic stepwise model selection procedures based on AIC.

We predicted the y value (psa) for a new data. The new data was formed by taking the mean for quantitative variables and mode for qualitative variables. The result comes out to be **10.176275.**

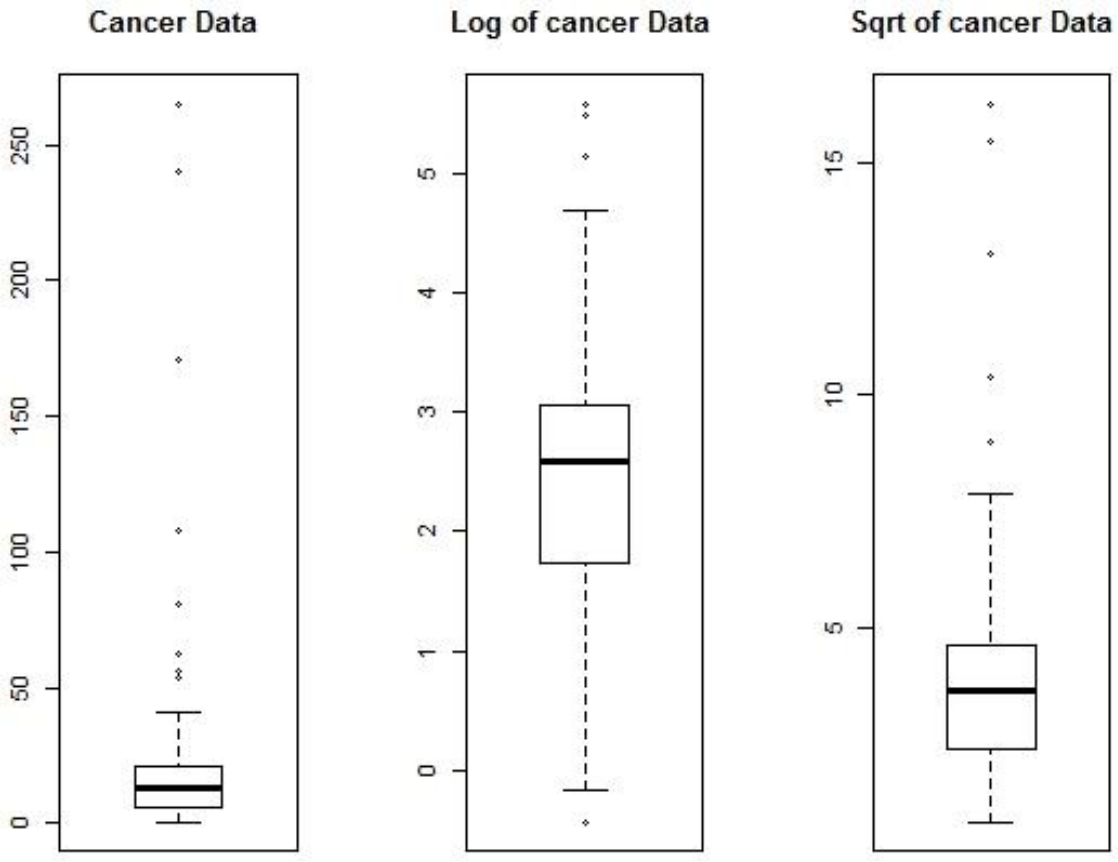

Figure 1: Boxplots of 'psa' Response Variable

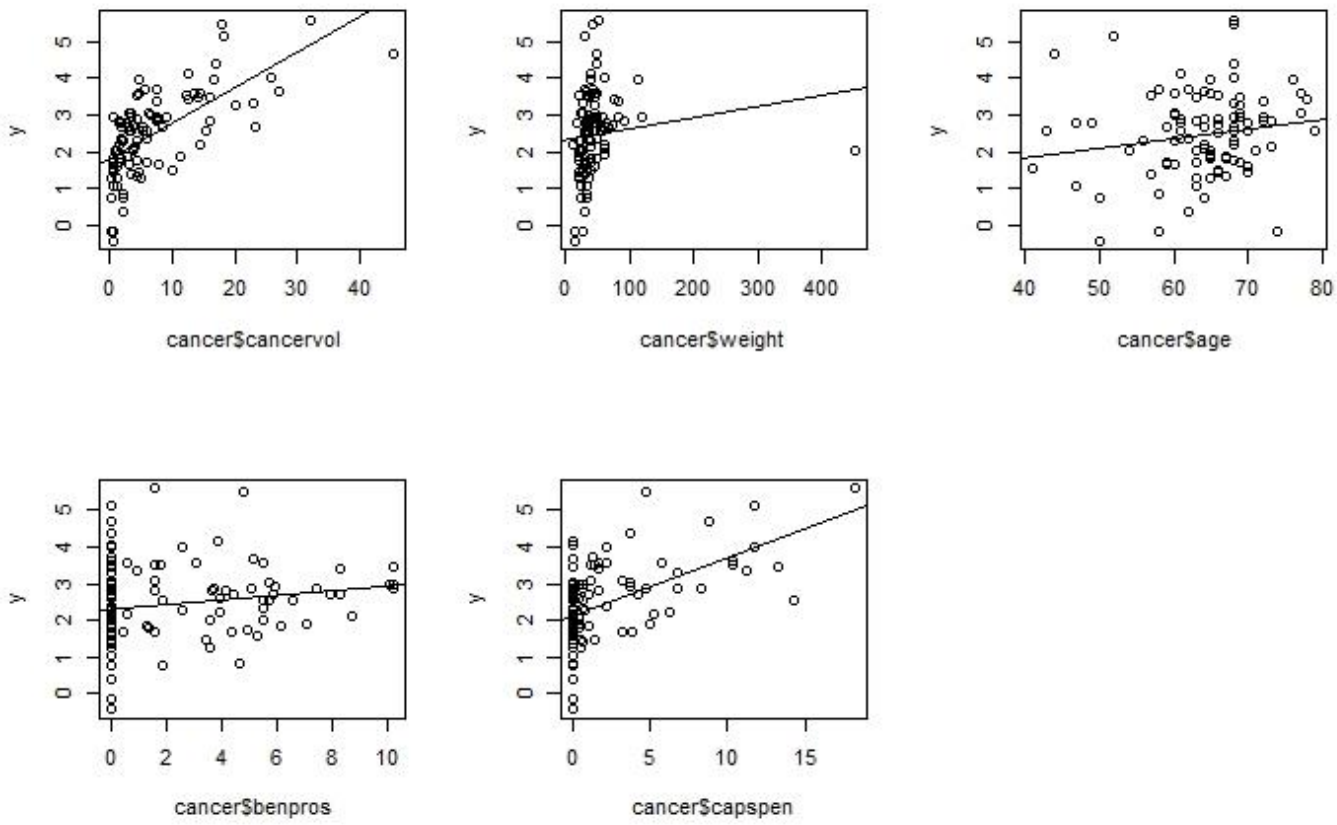

Figure 2: Linear Model taking single Predictor Variable at a time

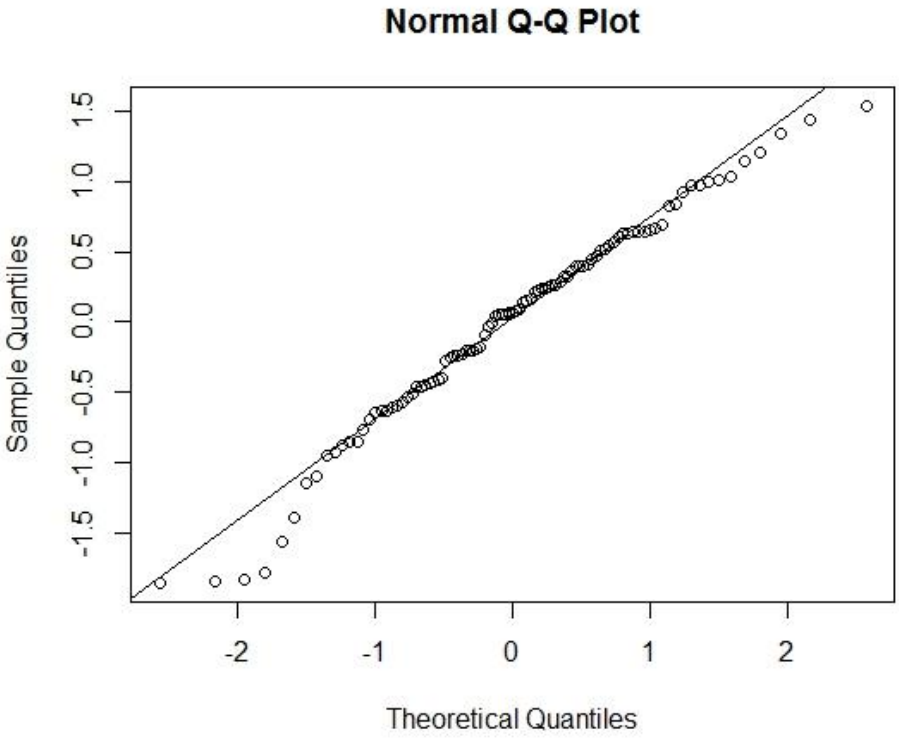

Figure 3: QQplot of final model

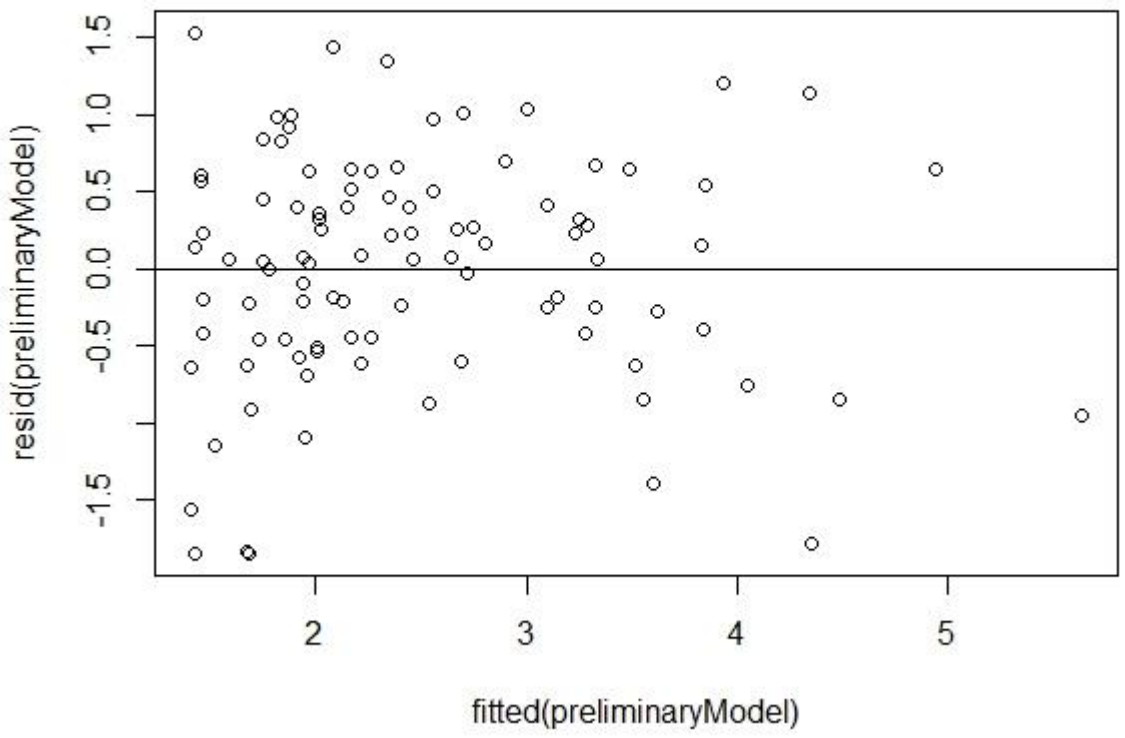

Figure 4: Residual Plot of Preliminary Model

Section 2: (R Code)

```
(a) # Read the cancer data
library(readr)
cancer <- read_csv("C:/Users/mastr/OneDrive/Documents/UTD NOTES/UTD-STATISTICS/mini project 5/prostate_cancer.csv")
str(cancer)
'" Classes 'tbl_df', 'tbl' and \'data.frame\':     97 obs. of  9 variables:
 $ subject  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ psa      : num  0.651 0.852 0.852 0.852 1.448 ...
 $ cancervol: num  0.56 0.372 0.601 0.301 2.117 ...
 $ weight   : num  16 27.7 14.7 26.6 30.9 ...
 $ age      : int  50 58 74 58 62 50 64 58 47 63 ...
 $ benpros  : num  0 0 0 0 0 ...
 $ vesinv   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ capspen  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ gleason  : int  6 7 7 6 6 6 6 6 7 6 ...'"

# Looking at distribution of some predictors
par(mfrow=c(1,3))
boxplot (cancer$psa,main = "Cancer Data ")
boxplot (log(cancer$psa),main = "Log of cancer Data ")
boxplot (sqrt(cancer$psa),main = "Sqrt of cancer Data  ")
```

```
#Log seems to be better as there are minimum outliers and it covers the maximum distribution
#also, later QQplots are better fitted when taking log of (Y)
table(cancer$vesinv)
#0  1
#76 21
table(cancer$gleason)
#6  7  8
#33 43 21

#taking factors of categorical variables
cancer$vesinv=factor(cancer$vesinv)
cancer$gleason=factor(cancer$gleason)
str(cancer)
'''
Classes ‘tbl_df’, ‘tbl’ and \'data.frame\':        97 obs. of  9 variables:
 $ subject  : int  1 2 3 4 5 6 7 8 9 10 ...
$ psa      : num  0.651 0.852 0.852 0.852 1.448 ...
$ cancervol: num  0.56 0.372 0.601 0.301 2.117 ...
$ weight   : num  16 27.7 14.7 26.6 30.9 ...
$ age      : int  50 58 74 58 62 50 64 58 47 63 ...
$ benpros  : num  0 0 0 0 0 ...
$ vesinv   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
$ capspen  : num  0 0 0 0 0 0 0 0 0 0 ...
$ gleason  : Factor w/ 3 levels "6","7","8": 1 2 2 1 1 1 1 1 2 1 ...
'''



# Take log(psa) as response
y <- log(cancer$psa)

# First, let's look at the relationship between response and each predictor variableone by one
par(mfrow=c(2,3))
plot(cancer$cancervol, y)
fit1 <- lm(y ~ cancervol, data = cancer)
abline(fit1)
#we can include this attribute as it has significant positive trend
plot(cancer$weight, y)
fit2 <- lm(y ~ weight, data = cancer)
abline(fit2)
#we can remove weight as data is not distributed properly and it has no significant positive trend
plot(cancer$age, y)
fit3 <- lm(y ~ age, data = cancer)
abline(fit3)
#we can remove age as it has no significant positive trend
plot(cancer$benpros, y)
fit4<- lm(y ~ benpros, data = cancer)
abline(fit4)
summary(fit4)
#we can remove benpros as it has no significant positive trend
plot(cancer$capspen, y)
```

```
fit5<- lm(y ~ capspen, data = cancer)
abline(fit5)
# We see a positive trend

# Next, we definitely expect 'cancervol' to be important predictors.

#We try to make model considering each predictor variable at a time

#considering categoricals variables
fit6.1 <- lm(y ~  gleason, data = cancer)
anova(fit6.1)
'"
Analysis of Variance Table
Response: y
Df Sum Sq Mean Sq F value    Pr(>F)
gleason    2 40.177 20.0887  21.558 1.967e-08 ***
Residuals 94 87.591  0.9318
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 '"

#p-value sigifies that $gleason is significant and can be considered in our model
fit6.2 <- lm(y ~  vesinv, data = cancer)
anova(fit6.2)
'"
Analysis of Variance Table
Response: y
Df Sum Sq Mean Sq F value    Pr(>F)
vesinv     1 40.984  40.984  44.864 1.481e-09 ***
Residuals 95 86.785   0.914
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1'"
#p-value sigifies that $vesinv is significant and can be considered in our model

#Cosidering 2-3 predictor variables at a time,
fit7.1 <- lm(y ~ cancervol , data = cancer)
fit7.2 <- lm(y ~ cancervol + benpros, data = cancer)
anova(fit7.1,fit7.2)
'"
Analysis of Variance Table
Model 1: y ~ cancervol
Model 2: y ~ cancervol + benpros
Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     95 72.605
2     94 64.802  1    7.8034 11.319 0.001111 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1'"
#p-value sigifies that $benpros is significant and can be considered in our model

fit7.3 <- lm(y ~ cancervol+benpros , data = cancer)
fit7.4 <- lm(y ~ cancervol + benpros+capspen, data = cancer)
anova(fit7.3,fit7.4)
```

```
'''
Analysis of Variance Table
Model 1: y ~ cancervol + benpros
Model 2: y ~ cancervol + benpros + capspen
Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     94 64.802
2     93 63.904  1   0.89737 1.3059 0.2561
'''

#p-value sigifies that $capspen is not significant and can be removed from our model
fit7.5 <- lm(y ~ cancervol + benpros , data = cancer)
fit7.6 <- lm(y ~ cancervol + benpros + gleason, data = cancer)
anova(fit7.5, fit7.6)
'''
Analysis of Variance Table

Model 1: y ~ cancervol + benpros
Model 2: y ~ cancervol + benpros + gleason
Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     94 64.802
2     92 58.032  2    6.7695 5.3659 0.006249 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
'''
#p-value sigifies that $gleason is significant and can be considered in our model
fit7.7 <- lm(y ~ cancervol + benpros + gleason, data = cancer)
fit7.8 <- lm(y ~ cancervol + benpros + gleason+vesinv, data = cancer)
anova(fit7.7, fit7.8)
'''
Analysis of Variance Table

Model 1: y ~ cancervol + benpros + gleason
Model 2: y ~ cancervol + benpros + gleason + vesinv
Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     92 58.032
2     91 53.055  1    4.9772 8.5369 0.004389 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
'''
#p-value sigifies that $vesinv is significant and can be considered in our model

#Just verifying weight and age
fit7.9 <- lm(y ~ cancervol + benpros + gleason+vesinv, data = cancer)
fit7.10 <- lm(y ~ cancervol + benpros + gleason+vesinv+age+weight, data = cancer)
anova(fit7.9, fit7.10)

#p-value sigifies that the removal of $weight and $age seems to comply
preliminaryModel <- lm(y ~ cancervol + benpros + gleason+vesinv, data = cancer)

#Therefore, we take lm(y ~ cancervol + benpros + gleason+vesinv, data = cancer) as our
# preliminary model for the data. However, we need to perform the diagnostics
```

```
# before accepting this model. So we use utomatic stepwise model selection procedures
# based on AIC. In the output below '+' means 'add variable' and '-' means 'drop variable.'


# Forward selection based on AIC
fit8.forward <- step(lm(y ~ 1, data = cancer),
                     scope = list(upper = ~cancervol+gleason+vesinv+capspen+weight+age+benpros),
                     direction = "forward")
'''
Start:  AIC=28.72
y ~ 1

Df Sum of Sq     RSS      AIC
+ cancervol  1    55.164  72.605 -24.0986
+ vesinv     1    40.984  86.785  -6.7944
+ gleason    2    40.177  87.591  -3.8967
+ 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
y ~ cancervol

Df Sum of Sq    RSS     AIC
+ benpros  1    7.8034 64.802 -33.128
+ gleason  2    8.2468 64.358 -31.794
+ vesinv   1    6.5468 66.058 -31.265
+ age      1    2.6615 69.944 -25.721
+ weight   1    1.7901 70.815 -24.520
<none>                 72.605 -24.099
+ capspen  1    0.9673 71.638 -23.400

Step:  AIC=-33.13
y ~ cancervol + benpros

Df Sum of Sq    RSS     AIC
+ vesinv   1    7.3339 57.468 -42.778
+ gleason  2    6.7695 58.032 -39.830
<none>                 64.802 -33.128
+ capspen  1    0.8974 63.904 -32.480
+ age      1    0.3961 64.406 -31.723
+ weight   1    0.2057 64.596 -31.436

Step:  AIC=-42.78
y ~ cancervol + benpros + vesinv

Df Sum of Sq    RSS     AIC
+ gleason  2    4.4128 53.055 -46.528
<none>                 57.468 -42.778
```

```
+ weight   1    0.1797 57.288 -41.082
+ capspen  1    0.1460 57.322 -41.025
+ age      1    0.0543 57.413 -40.870

Step:  AIC=-46.53
y ~ cancervol + benpros + vesinv + gleason

Df Sum of Sq    RSS     AIC
<none>              53.055 -46.528
+ capspen  1   0.39210 52.663 -45.248
+ weight   1   0.27344 52.781 -45.029
+ age      1   0.02611 53.029 -44.576
'''
# compared to this AIC(Forward Selection) result, our models complys with it

# Backward elimination based on AIC
fit8.backward <- step(lm(y ~ cancervol+gleason+vesinv+capspen+weight+age+benpros, data = cancer),
                      scope = list(lower = ~1), direction = "backward")
"
Start:  AIC=-41.81
y ~ cancervol + gleason + vesinv + capspen + weight + age + benpros

Df Sum of Sq    RSS     AIC
- age        1    0.0336 52.393 -43.746
- weight     1    0.2804 52.640 -43.290
- capspen    1    0.3843 52.744 -43.099
<none>               52.360 -41.808
- gleason    2    4.7351 57.095 -37.410
- vesinv     1    5.1118 57.471 -34.772
- benpros    1    5.2603 57.620 -34.522
- cancervol  1   11.4641 63.824 -24.603

Step:  AIC=-43.75
y ~ cancervol + gleason + vesinv + capspen + weight + benpros

Df Sum of Sq    RSS     AIC
- weight     1    0.2696 52.663 -45.248
- capspen    1    0.3883 52.781 -45.029
<none>               52.393 -43.746
- gleason    2    4.7535 57.147 -39.322
- vesinv     1    5.0783 57.472 -36.772
- benpros    1    5.6032 57.996 -35.890
- cancervol  1   11.6045 63.998 -26.339

Step:  AIC=-45.25
y ~ cancervol + gleason + vesinv + capspen + benpros

Df Sum of Sq    RSS     AIC
- capspen    1    0.3921 53.055 -46.528
<none>               52.663 -45.248
- gleason    2    4.6590 57.322 -41.025
```

```
- vesinv     1    5.1749 57.838 -38.156
- benpros    1    7.3459 60.009 -34.581
- cancervol  1   11.6437 64.306 -27.872

Step:  AIC=-46.53
y ~ cancervol + gleason + vesinv + benpros

Df Sum of Sq    RSS     AIC
<none>                53.055 -46.528
- gleason    2    4.4128 57.468 -42.778
- vesinv     1    4.9772 58.032 -39.830
- benpros    1    7.2546 60.310 -36.096
- cancervol  1   12.1569 65.212 -28.516
'''

# Both forward/backward
fit8.both <- step(lm(y ~ 1, data = cancer),
                  scope = list(lower = ~1, upper = ~cancervol+gleason+vesinv+capspen+weight+age+benpros),
                  direction = "both")
'''
Start:  AIC=28.72
y ~ 1

Df Sum of Sq     RSS      AIC
+ cancervol  1    55.164  72.605 -24.0986
+ vesinv     1    40.984  86.785  -6.7944
+ gleason    2    40.177  87.591  -3.8967
+ 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
y ~ cancervol

Df Sum of Sq     RSS     AIC
+ benpros    1     7.803  64.802 -33.128
+ gleason    2     8.247  64.358 -31.794
+ 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
- cancervol  1    55.164 127.769  28.725

Step:  AIC=-33.13
y ~ cancervol + benpros

Df Sum of Sq     RSS     AIC
+ vesinv     1     7.334  57.468 -42.778
```

```
+ gleason    2     6.770  58.032 -39.830
<none>                   64.802 -33.128
+ capspen    1     0.897  63.904 -32.480
+ age        1     0.396  64.406 -31.723
+ weight     1     0.206  64.596 -31.436
- benpros    1     7.803  72.605 -24.099
- cancervol  1    59.802 124.603  28.291

Step:  AIC=-42.78
y ~ cancervol + benpros + vesinv

Df Sum of Sq    RSS     AIC
+ gleason    2    4.4128 53.055 -46.528
<none>                   57.468 -42.778
+ weight     1    0.1797 57.288 -41.082
+ capspen    1    0.1460 57.322 -41.025
+ age        1    0.0543 57.413 -40.870
- vesinv     1    7.3339 64.802 -33.128
- benpros    1    8.5905 66.058 -31.265
- cancervol  1   22.7482 80.216 -12.429

Step:  AIC=-46.53
y ~ cancervol + benpros + vesinv + gleason

Df Sum of Sq    RSS     AIC
<none>                   53.055 -46.528
+ capspen    1    0.3921 52.663 -45.248
+ weight     1    0.2734 52.781 -45.029
+ age        1    0.0261 53.029 -44.576
- gleason    2    4.4128 57.468 -42.778
- vesinv     1    4.9772 58.032 -39.830
- benpros    1    7.2546 60.310 -36.096
- cancervol  1   12.1569 65.212 -28.516
'''
# We see that all the direction = backward, forward and both, pick the following model:
# cancervol + benpros + vesinv + gleason
#
# Also, our preliminary model was the same:
# cancervol + benpros + vesinv + gleason


# So we go ahead with this model (y~cancervol + benpros + vesinv + gleason)and perform model diagnostics

# residual plot
plot(fitted(preliminaryModel), resid(preliminaryModel))
abline(h = 0)

# plot of absolute residuals

plot(fitted(preliminaryModel), abs(resid(preliminaryModel)))
```

```
# normal QQ plot
#final qqplot
par(mfrow=c(1,1))
qqnorm(resid(preliminaryModel))
qqline(resid(preliminaryModel))

#In order to predict, we find the mode of categorical variables
vesinv.table = table(cancer$vesinv)
mode.vesinv = names(vesinv.table)[vesinv.table==max(vesinv.table)]

gleason.table = table(cancer$gleason)
mode.gleason = names(gleason.table)[gleason.table==max(gleason.table)]

#Forming a new data frame for prediction and also we are finding the mean of quantitative variables
new.data=data.frame(cancervol=as.numeric(mean(cancer$cancervol)),benpros=as.numeric(mean(cancer$b
enpros)),vesinv=factor(mode.vesinv),gleason=factor(mode.gleason))

#Prediction using predict() function
log.psa=predict(preliminaryModel,newdata=new.data)

pred.psa=exp(log.psa)
```

The prediction result turns out to be
10.176275

This value corresponds to psa variable.