# MINI PROJECT #6
# Statistical Methods for Data Science
# Group - 46

Anish Joshi(2021591978) and Aneena Manoj(2021623362)

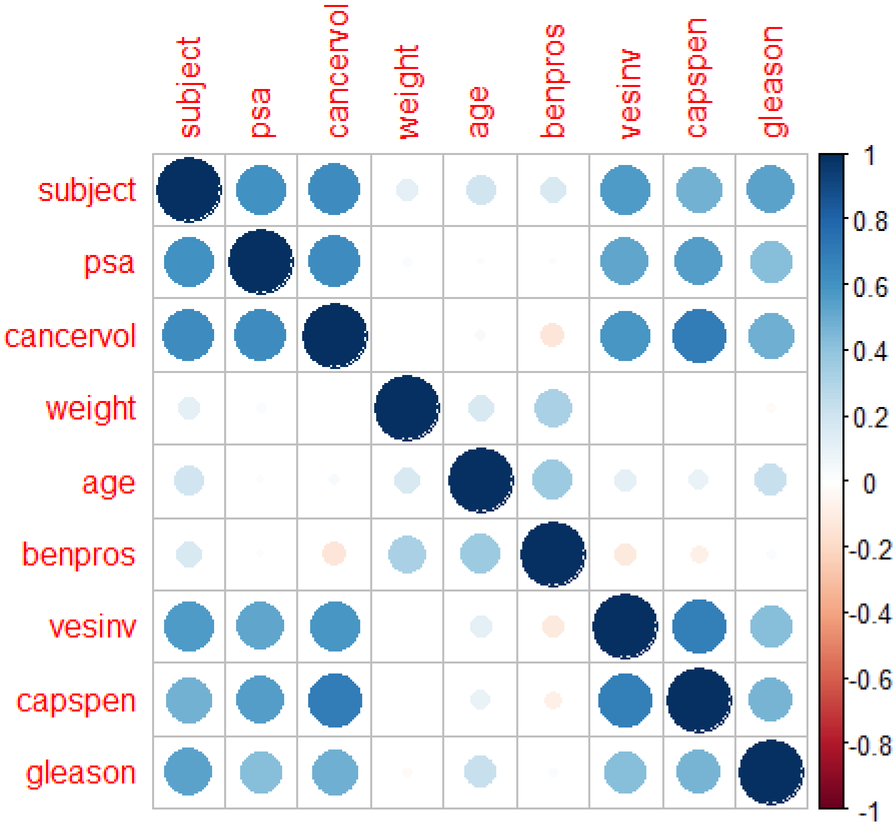
**Contribution –**

Both the team members sat together and analyzed the question and then contributed their self inputs on the question and then the best solution was chosen. The R script was first assessed by the team members who then collaboratively wrote the R code and reported the findings. The findings were then integrated into a document and submitted. Both the team members gave equal contributions and worked together to complete the project.

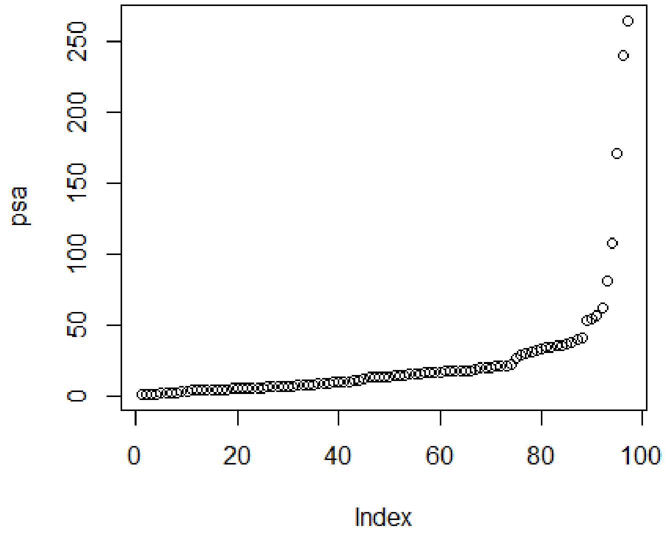
## SECTION - 1

### QUESTION – 1

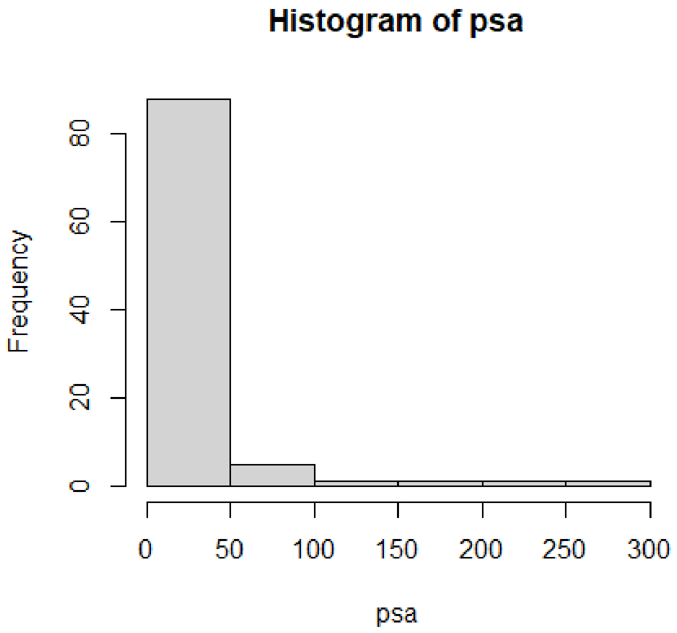
We will start off by first reading the data present in the file using the "read.csv" function. The next step is to form and plot a correlation matrix for the given dataset. The following is the correlation matrix that we get by using the "complot" function which can be used after installing the "complot" library.



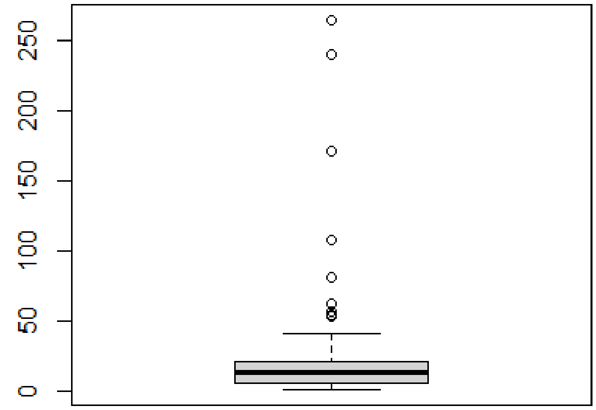
Now, we will attach the data so that we can use the variables involved in it.
The next step is to use some plots to see what the response variable "PSA level" looks like. First, we will plot a scatterplot for PSA.



Now, plot the histogram for PSA.

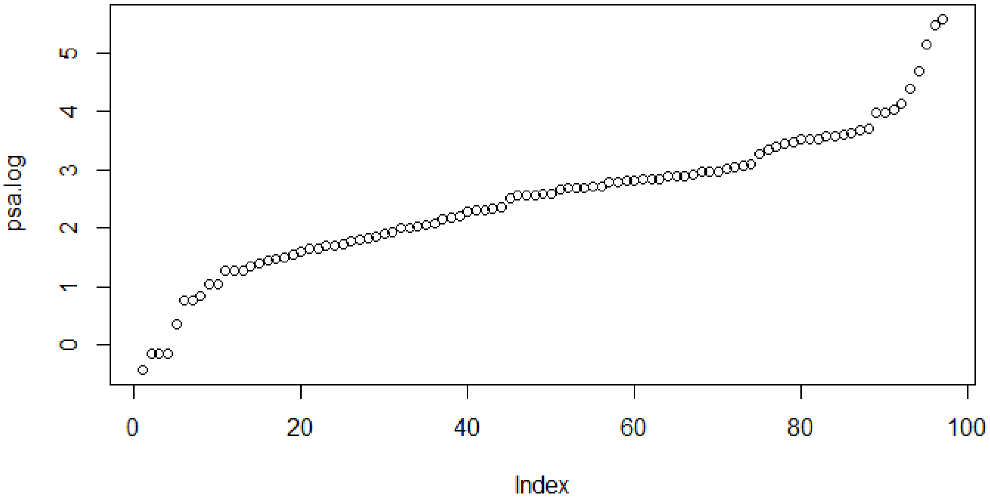


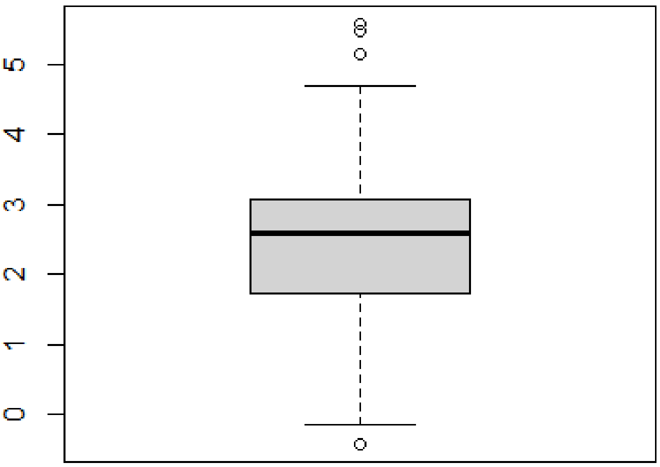
Now we will plot a boxplot for PSA to have a look at the outliers.



When we look at the boxplot, it can be concluded that this dataset contains many outliers. A transformation is required so as to fit it into a linear model. As per the mini-project requirement, we will be applying logarithmic transformation on PSA and then again see how the data can be applied to a linear model.

Following are the new plots for logarithmically transformed PSA values.





As mentioned in the question and in the dataset, “vesting” is a qualitative variable so we use the “as_factor” function to convert the qualitative variable into a factor. This helps in preserving the values and the variable label attributes. R-CODE IN SECTION - 2.

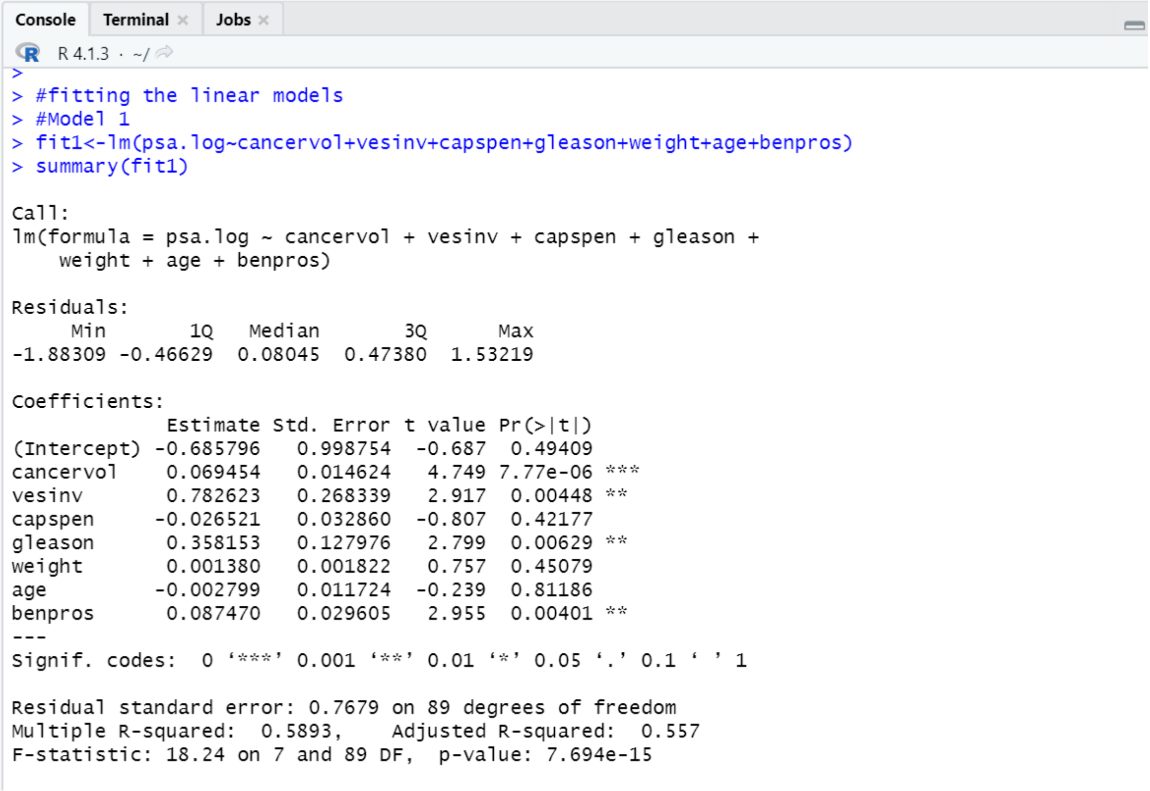
Now we will be fitting the linear models:

**MODEL - 1**

Null Hypothesis (H0): None of the predictors are useful in the prediction of the response variable “PSA”. This means that: Beta0=Beta1=Beta2=....=Beta(n)=0.

Alternative Hypothesis (H1): At least one of the predictors is useful in the prediction of the response variable “PSA”.

Summary for fit1:

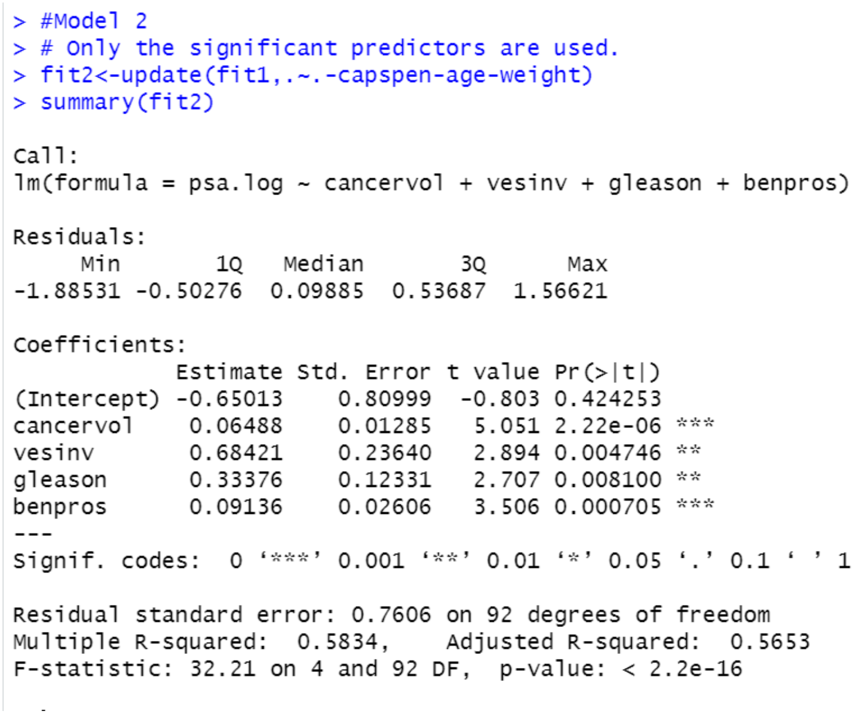


```
>
> #fitting the linear models
> #Model 1
> fit1<-lm(psa.log~cancervol+vesinv+capspen+gleason+weight+age+benpros)
> summary(fit1)

Call:
lm(formula = psa.log ~ cancervol + vesinv + capspen + gleason +
    weight + age + benpros)

Residuals:
     Min       1Q   Median       3Q      Max
-1.88309 -0.46629  0.08045  0.47380  1.53219

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.685796   0.998754  -0.687  0.49409
cancervol    0.069454   0.014624   4.749 7.77e-06 ***
vesinv       0.782623   0.268339   2.917  0.00448 **
capspen     -0.026521   0.032860  -0.807  0.42177
gleason      0.358153   0.127976   2.799  0.00629 **
weight       0.001380   0.001822   0.757  0.45079
age         -0.002799   0.011724  -0.239  0.81186
benpros      0.087470   0.029605   2.955  0.00401 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7679 on 89 degrees of freedom
Multiple R-squared:  0.5893,	Adjusted R-squared:  0.557
F-statistic: 18.24 on 7 and 89 DF,  p-value: 7.694e-15
```

From the above summary, we observe the following predictors: cancervol (p-value:7.77e-06 which is less than 0.05) and is “***”, vesinv (p-value:0.00448 which is less than 0.05), and is “**”, gleason (p-value:0.00629 which is less than 0.05) and is “**”, benpros (p-value:0.00401 which is less than 0.05) and is “**”. All of these predictors are significant. Hence, we reject the Null Hypothesis.

**MODEL - 2**

For this model, we only use the predictors which we found to be significant from fit1.
Summary for fit2:



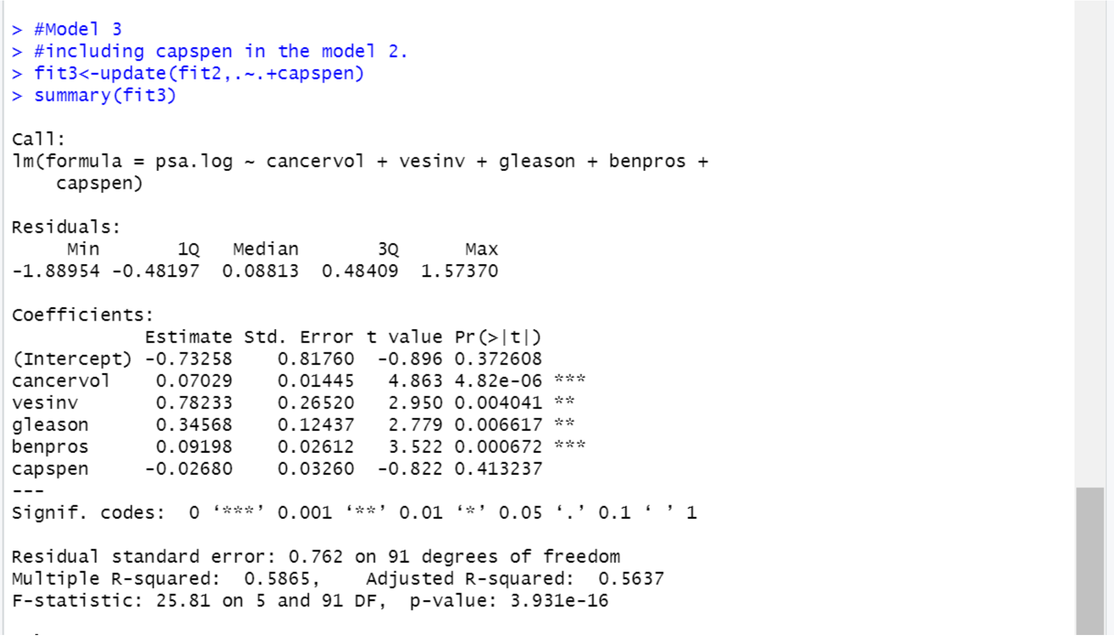
```
> #Model 2
> # Only the significant predictors are used.
> fit2<-update(fit1,.~.-capspen-age-weight)
> summary(fit2)

Call:
lm(formula = psa.log ~ cancervol + vesinv + gleason + benpros)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.88531 -0.50276  0.09885  0.53687  1.56621 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.65013    0.80999  -0.803 0.424253    
cancervol    0.06488    0.01285   5.051 2.22e-06 ***
vesinv       0.68421    0.23640   2.894 0.004746 ** 
gleason      0.33376    0.12331   2.707 0.008100 ** 
benpros      0.09136    0.02606   3.506 0.000705 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7606 on 92 degrees of freedom
Multiple R-squared:  0.5834,	Adjusted R-squared:  0.5653 
F-statistic: 32.21 on 4 and 92 DF,  p-value: < 2.2e-16
```

## MODEL - 3

Now we will include capspen and observe the change in the model and conclude whether capspen is significant in the prediction of the response variable or not.

Including capspen in model 2 for model 3.

Summary for fit3



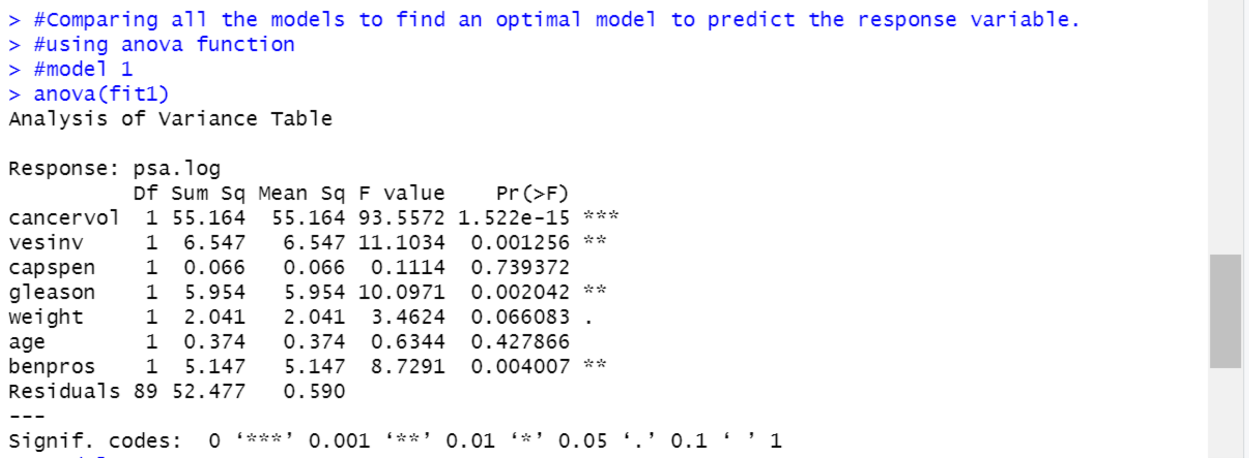
```
> #Model 3
> #including capspen in the model 2.
> fit3<-update(fit2,.~.+capspen)
> summary(fit3)

Call:
lm(formula = psa.log ~ cancervol + vesinv + gleason + benpros + 
    capspen)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.88954 -0.48197  0.08813  0.48409  1.57370 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.73258    0.81760  -0.896 0.372608    
cancervol    0.07029    0.01445   4.863 4.82e-06 ***
vesinv       0.78233    0.26520   2.950 0.004041 ** 
gleason      0.34568    0.12437   2.779 0.006617 ** 
benpros      0.09198    0.02612   3.522 0.000672 ***
capspen     -0.02680    0.03260  -0.822 0.413237    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.762 on 91 degrees of freedom
Multiple R-squared:  0.5865,	Adjusted R-squared:  0.5637 
F-statistic: 25.81 on 5 and 91 DF,  p-value: 3.931e-16
```

Let’s compare the adjusted R square values of model 2 and model 3. For model 2 it is 0.5653 and for model 3 it is 0.5637. If the adjusted R square decreases after the addition of a predictor then that predictor is not significant for the prediction of the response variable.

**COMPARISON OF MODELS –**

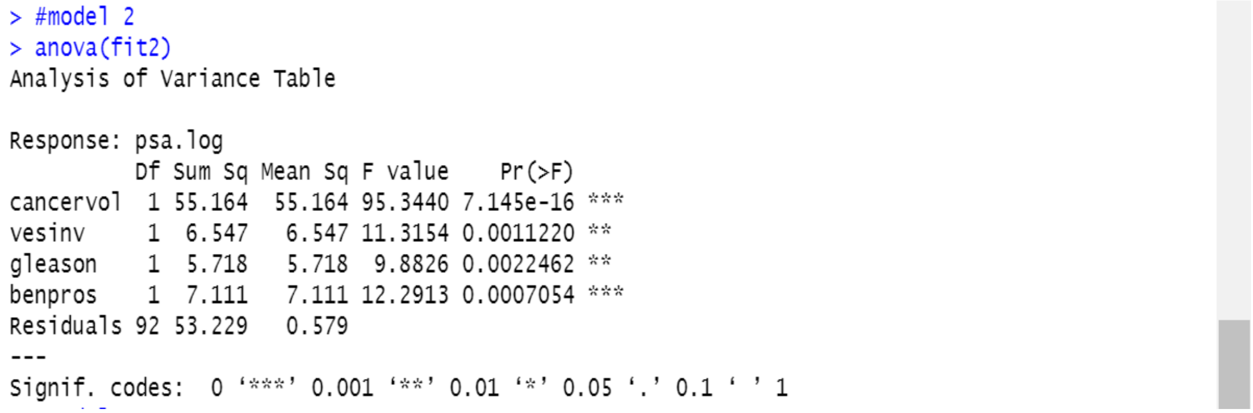
Now we will compare all three models and choose the most optimal model for the prediction of the response variable.

1. Model 1



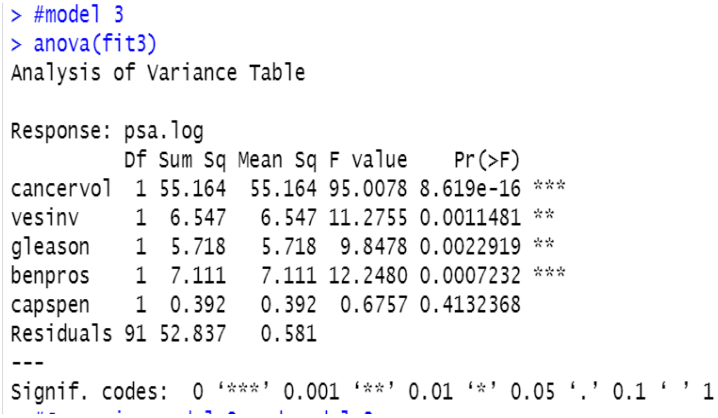
```
> #Comparing all the models to find an optimal model to predict the response variable.
> #using anova function
> #model 1
> anova(fit1)
Analysis of Variance Table

Response: psa.log
          Df Sum Sq Mean Sq F value    Pr(>F)    
cancervol  1 55.164  55.164 93.5572 1.522e-15 ***
vesinv     1  6.547   6.547 11.1034  0.001256 ** 
capspen    1  0.066   0.066  0.1114  0.739372    
gleason    1  5.954   5.954 10.0971  0.002042 ** 
weight     1  2.041   2.041  3.4624  0.066083 .  
age        1  0.374   0.374  0.6344  0.427866    
benpros    1  5.147   5.147  8.7291  0.004007 ** 
Residuals 89 52.477   0.590                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2. Model 2



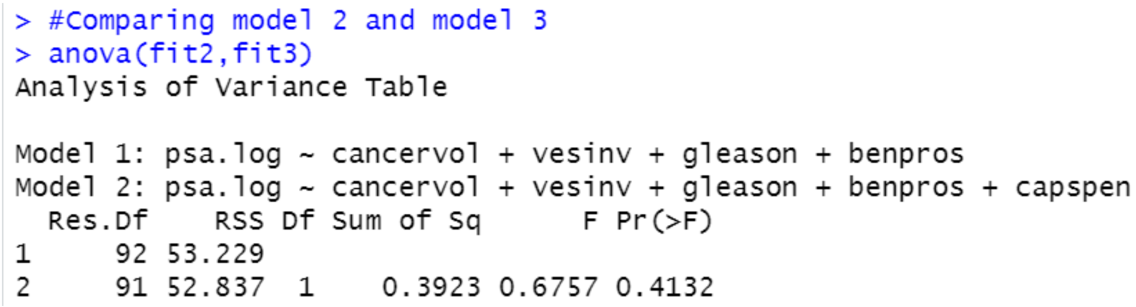
```
> #model 2
> anova(fit2)
Analysis of Variance Table

Response: psa.log
          Df Sum Sq Mean Sq F value    Pr(>F)    
cancervol  1 55.164  55.164 95.3440 7.145e-16 ***
vesinv     1  6.547   6.547 11.3154 0.0011220 ** 
gleason    1  5.718   5.718  9.8826 0.0022462 ** 
benpros    1  7.111   7.111 12.2913 0.0007054 ***
Residuals 92 53.229   0.579                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. Model 3



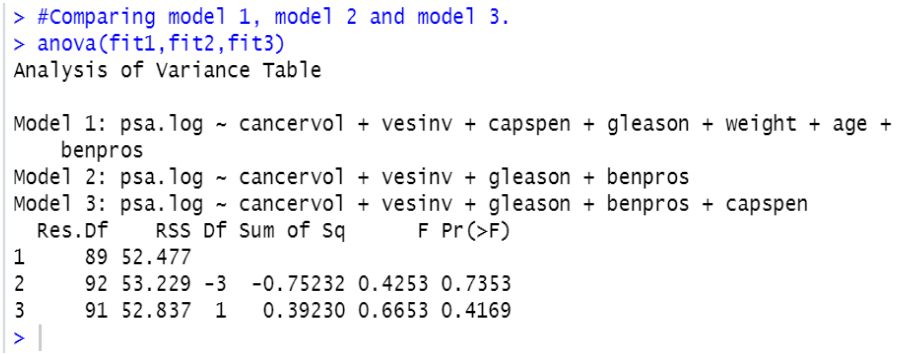
```
> #model 3
> anova(fit3)
Analysis of Variance Table

Response: psa.log
          Df Sum Sq Mean Sq F value    Pr(>F)    
cancervol  1 55.164  55.164 95.0078 8.619e-16 ***
vesinv     1  6.547   6.547 11.2755 0.0011481 ** 
gleason    1  5.718   5.718  9.8478 0.0022919 ** 
benpros    1  7.111   7.111 12.2480 0.0007232 ***
capspen    1  0.392   0.392  0.6757 0.4132368    
Residuals 91 52.837   0.581                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. Model 2 and Model 3



```
> #Comparing model 2 and model 3
> anova(fit2,fit3)
Analysis of Variance Table

Model 1: psa.log ~ cancervol + vesinv + gleason + benpros
Model 2: psa.log ~ cancervol + vesinv + gleason + benpros + capspen
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     92 53.229
2     91 52.837  1    0.3923 0.6757 0.4132
```

5. Model 1, Model 2, and Model 3



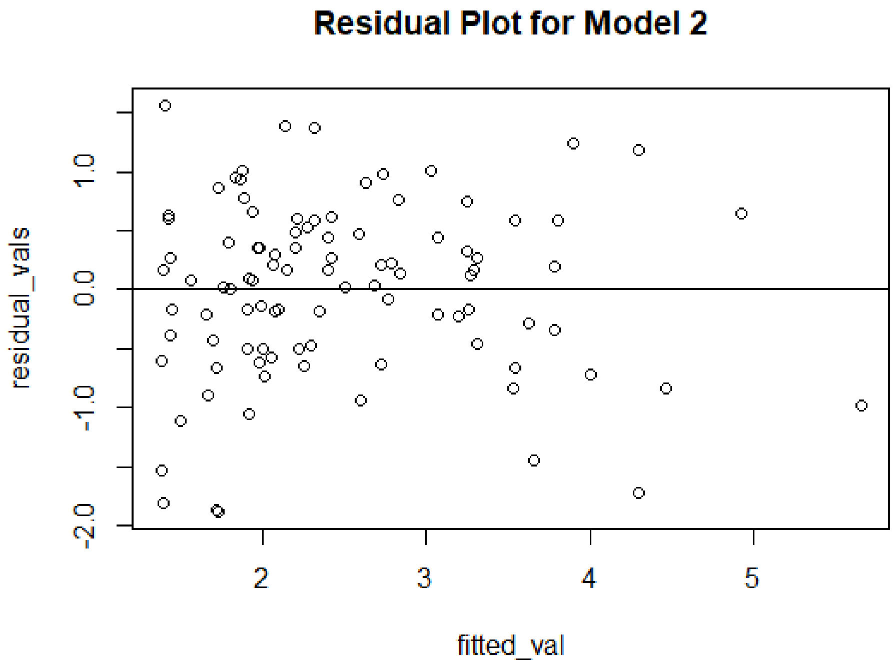
```
> #Comparing model 1, model 2 and model 3.
> anova(fit1,fit2,fit3)
Analysis of Variance Table

Model 1: psa.log ~ cancervol + vesinv + capspen + gleason + weight + age +
    benpros
Model 2: psa.log ~ cancervol + vesinv + gleason + benpros
Model 3: psa.log ~ cancervol + vesinv + gleason + benpros + capspen
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     89 52.477
2     92 53.229 -3  -0.75232 0.4253 0.7353
3     91 52.837  1   0.39230 0.6653 0.4169
> |
```

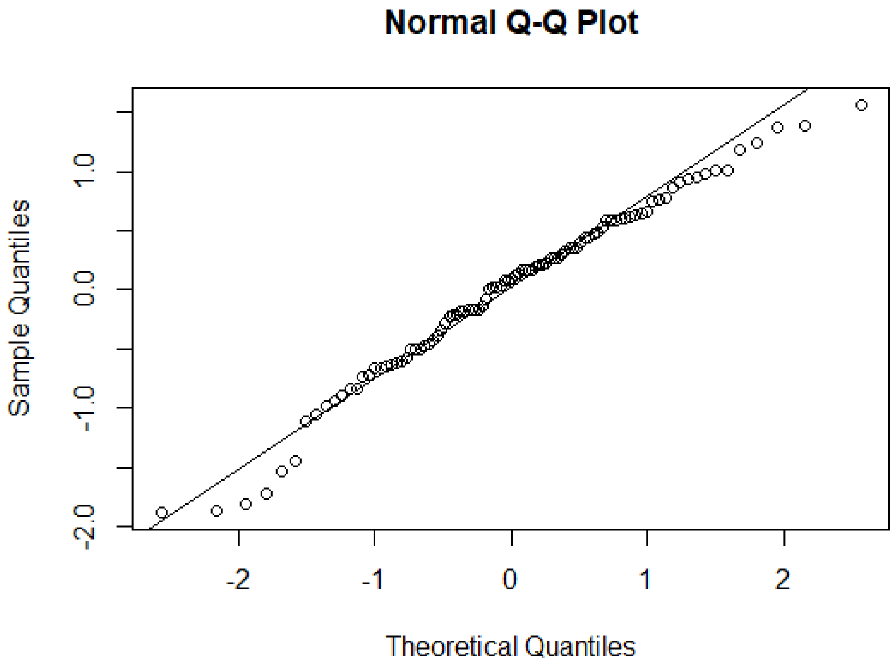
After looking at the three models and comparing them, we can conclude that model 2 is the most optimal linear model which we can use to predict the response variable "PSA".

## EVALUATION OF THE CHOSEN MODEL

### Residual Plot

We will plot the residual values and observe them.



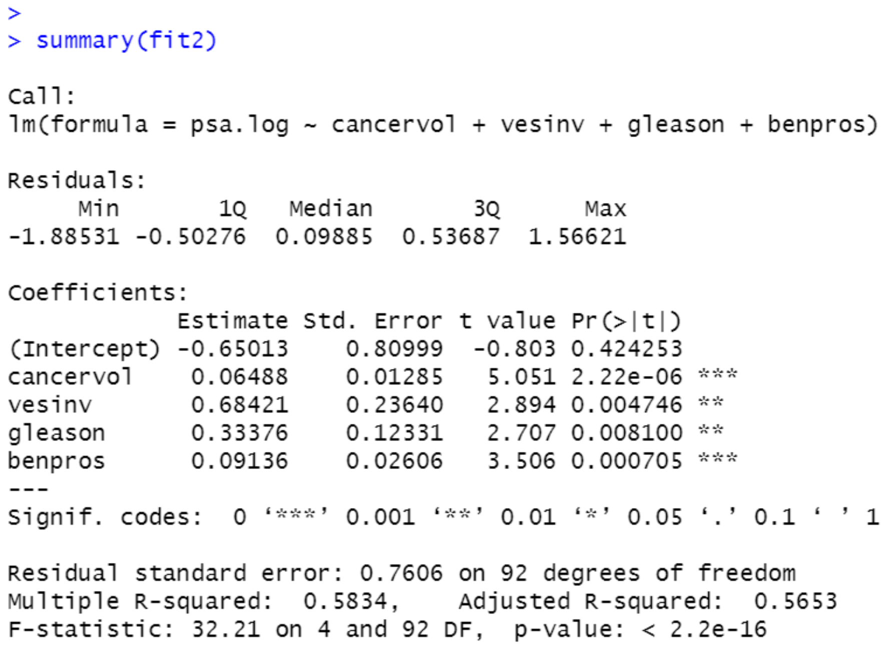


From the above two plots, we can observe that the points are scattered around zero and there's no pattern to them, Hence, the errors have a zero mean and the variance is constant. The errors are normally distributed.

## PREDICTION OF PSA LEVELS USING THE FINAL MODEL

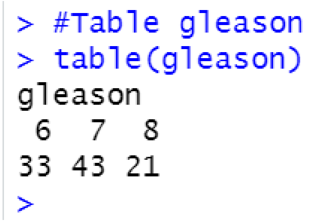
We will again see the summary of the chosen model i.e. Model 2.

The predictor models which we are going to use are cancervol, vesinv, gleason and benpros.

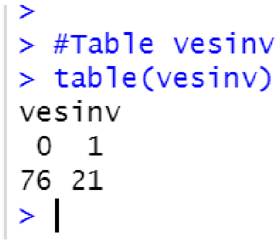


```
>
> summary(fit2)

Call:
lm(formula = psa.log ~ cancervol + vesinv + gleason + benpros)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.88531 -0.50276  0.09885  0.53687  1.56621 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.65013    0.80999  -0.803 0.424253    
cancervol    0.06488    0.01285   5.051 2.22e-06 ***
vesinv       0.68421    0.23640   2.894 0.004746 ** 
gleason      0.33376    0.12331   2.707 0.008100 ** 
benpros      0.09136    0.02606   3.506 0.000705 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7606 on 92 degrees of freedom
Multiple R-squared:  0.5834,	Adjusted R-squared:  0.5653 
F-statistic: 32.21 on 4 and 92 DF,  p-value: < 2.2e-16
```

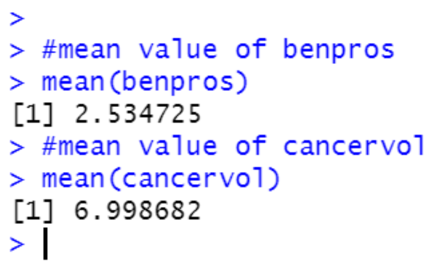
Now, we will have a look at the tables of gleason and vesinv to check the dominant values in both of them.



```
> #Table gleason
> table(gleason)
gleason
 6  7  8 
33 43 21 
>
```



```
>
> #Table vesinv
> table(vesinv)
vesinv
 0  1 
76 21 
> |
```



```
>
> #mean value of benpros
> mean(benpros)
[1] 2.534725
> #mean value of cancervol
> mean(cancervol)
[1] 6.998682
> |
```

The value 7 (43) is dominated in the gleason table and the value 0 (76) is dominated in the vesinv table. The mean of benpros is 2.534725 and the mean of cancervol is 6.998682.
Hence, the predicted value of the response variable "PSA" on the basis of the model  2 is:

$$-0.65013+0.06488\times(6.998682)+0.68421\times(0)+0.33376\times(7)+0.09136\times(2.534725)$$

$$=2.371833$$

This value is equal to the logarithmic value of PSA i.e. psa.log
We will now convert this into the actual value of the PSA level.

$$\log(psa)=2.371833$$

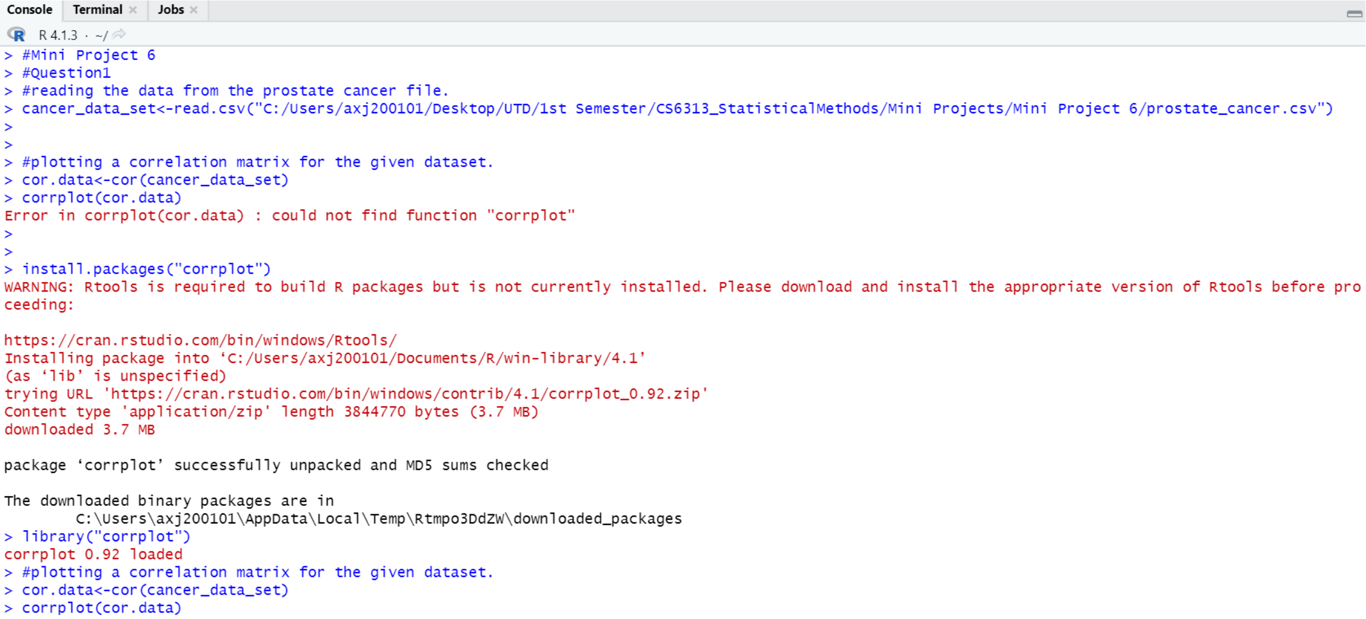
$$psa=exp(2.371833)$$

$$=10.717012$$

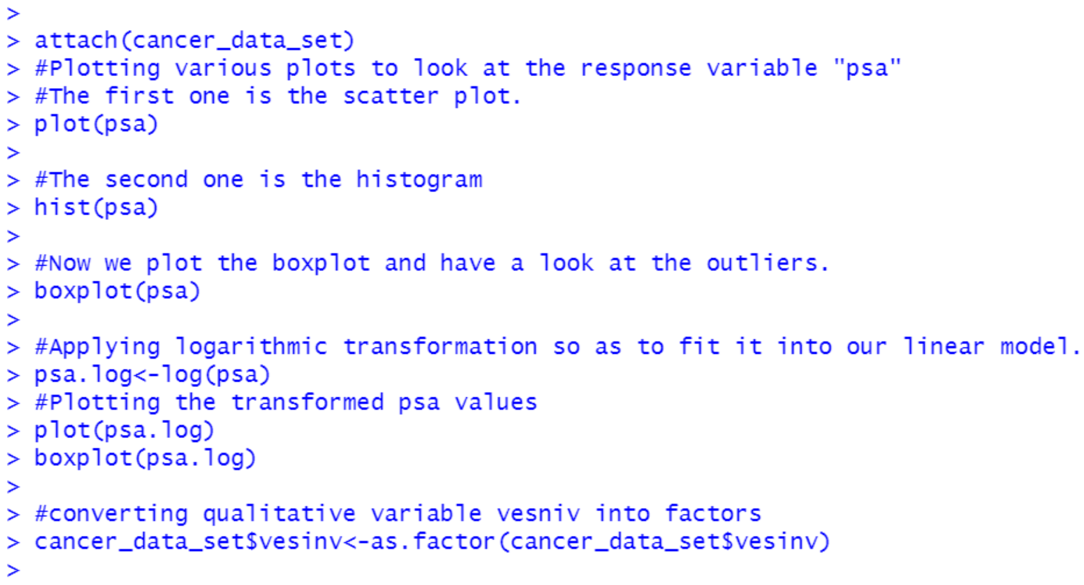
## SECTION - 2

#Mini Project 6
#Question1
#reading the data from the prostate cancer file.
cancer_data_set<-read.csv("C:/Users/axj200101/Desktop/UTD/1st Semester/CS6313_StatisticalMethods/Mini Projects/Mini Project 6/prostate_cancer.csv")

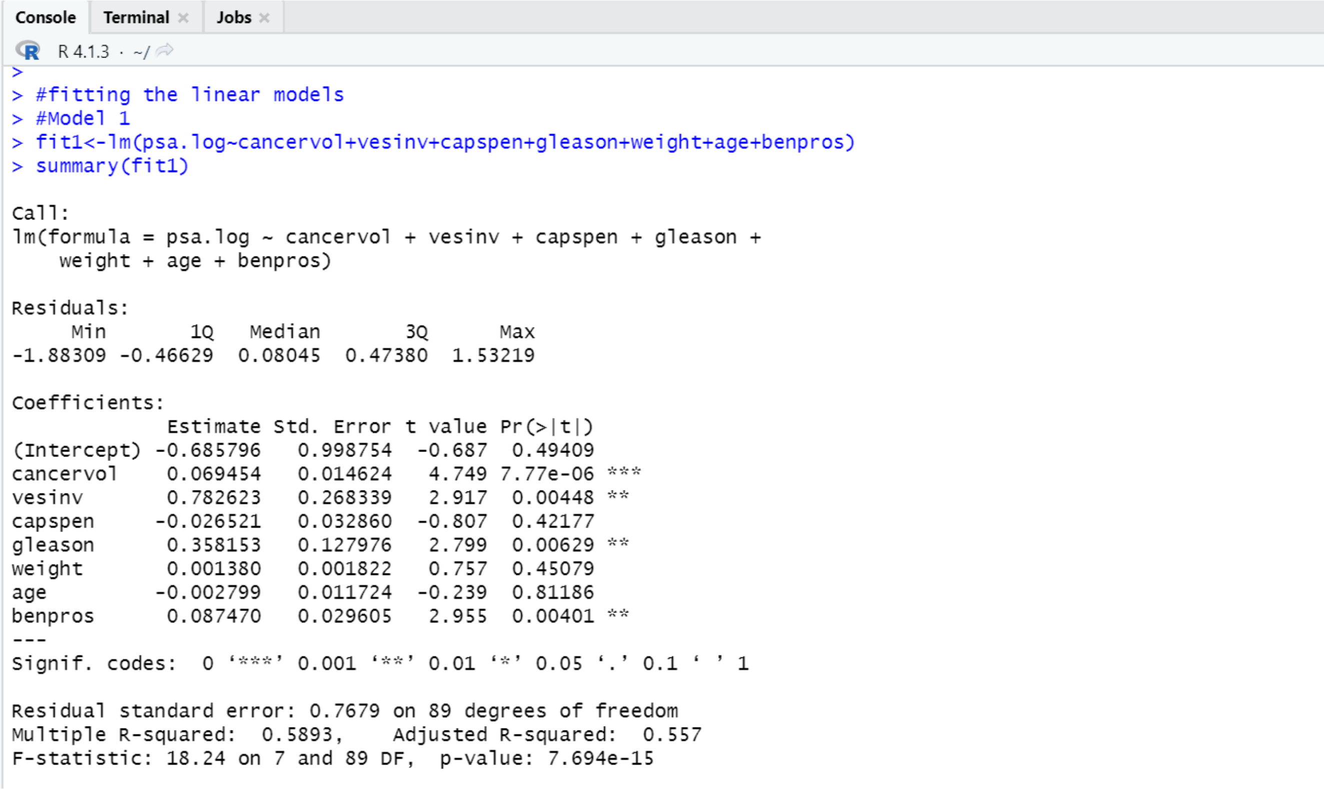
install.packages("corrplot")
library("corrplot")
#plotting a correlation matrix for the given dataset.
cor.data<-cor(cancer_data_set)
corrplot(cor.data)



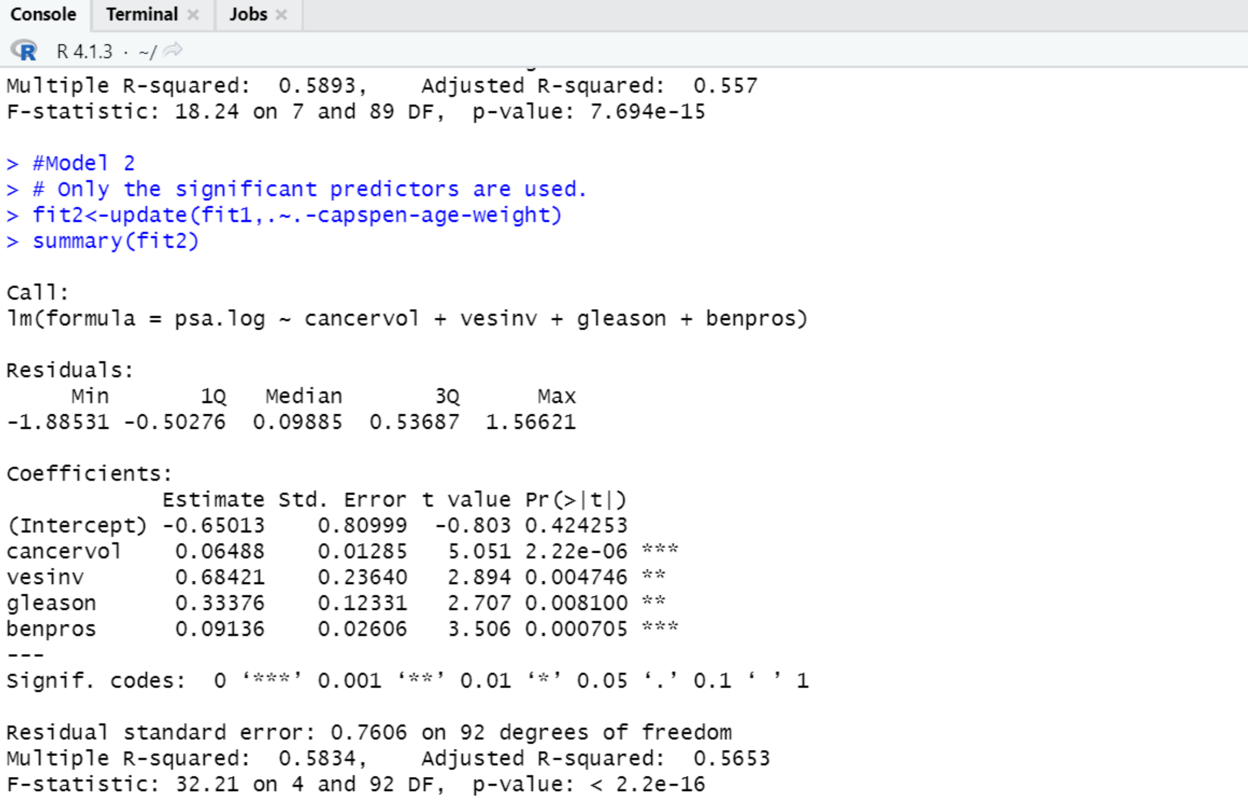
```
> #Mini Project 6
> #Question1
> #reading the data from the prostate cancer file.
> cancer_data_set<-read.csv("C:/Users/axj200101/Desktop/UTD/1st Semester/CS6313_StatisticalMethods/Mini Projects/Mini Project 6/prostate_cancer.csv")
>
>
> #plotting a correlation matrix for the given dataset.
> cor.data<-cor(cancer_data_set)
> corrplot(cor.data)
Error in corrplot(cor.data) : could not find function "corrplot"
>
>
> install.packages("corrplot")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before pro
ceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/axj200101/Documents/R/win-library/4.1’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/corrplot_0.92.zip'
Content type 'application/zip' length 3844770 bytes (3.7 MB)
downloaded 3.7 MB

package ‘corrplot’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\axj200101\AppData\Local\Temp\Rtmpo3DdZw\downloaded_packages
> library("corrplot")
corrplot 0.92 loaded
> #plotting a correlation matrix for the given dataset.
> cor.data<-cor(cancer_data_set)
> corrplot(cor.data)
```



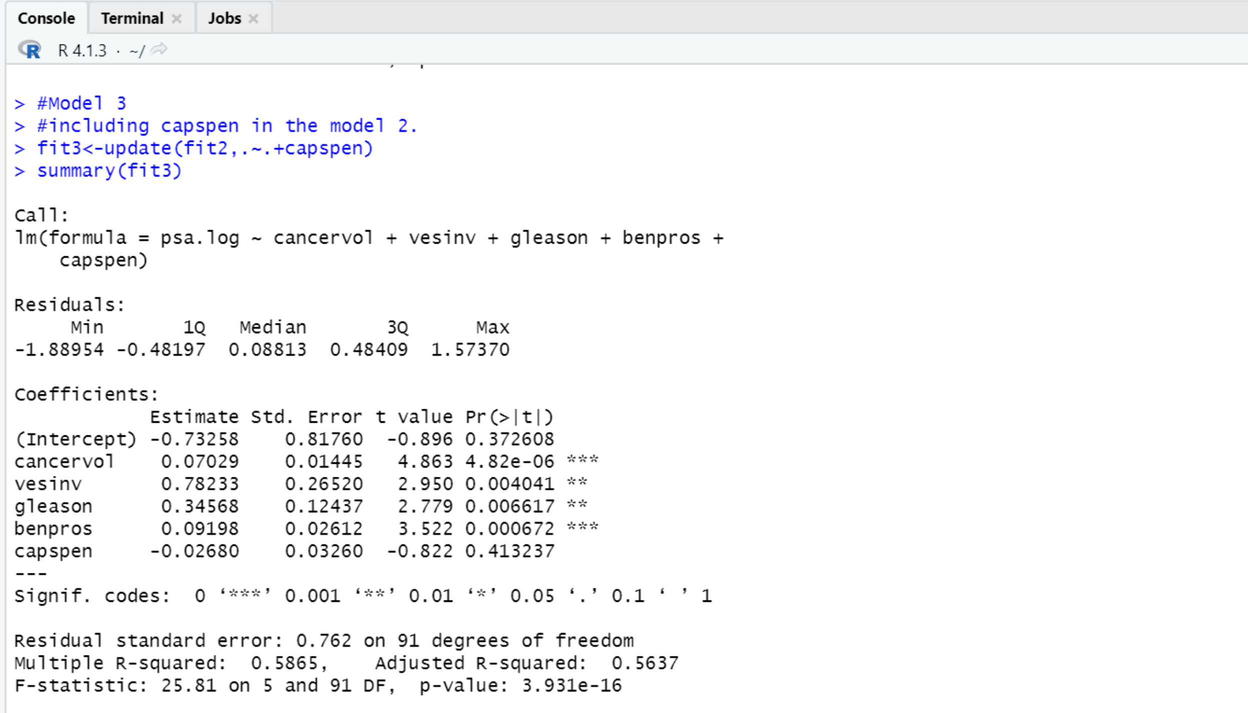
```
>
> attach(cancer_data_set)
> #Plotting various plots to look at the response variable "psa"
> #The first one is the scatter plot.
> plot(psa)
>
> #The second one is the histogram
> hist(psa)
>
> #Now we plot the boxplot and have a look at the outliers.
> boxplot(psa)
>
> #Applying logarithmic transformation so as to fit it into our linear model.
> psa.log<-log(psa)
> #Plotting the transformed psa values
> plot(psa.log)
> boxplot(psa.log)
>
> #converting qualitative variable vesniv into factors
> cancer_data_set$vesinv<-as.factor(cancer_data_set$vesinv)
>
```



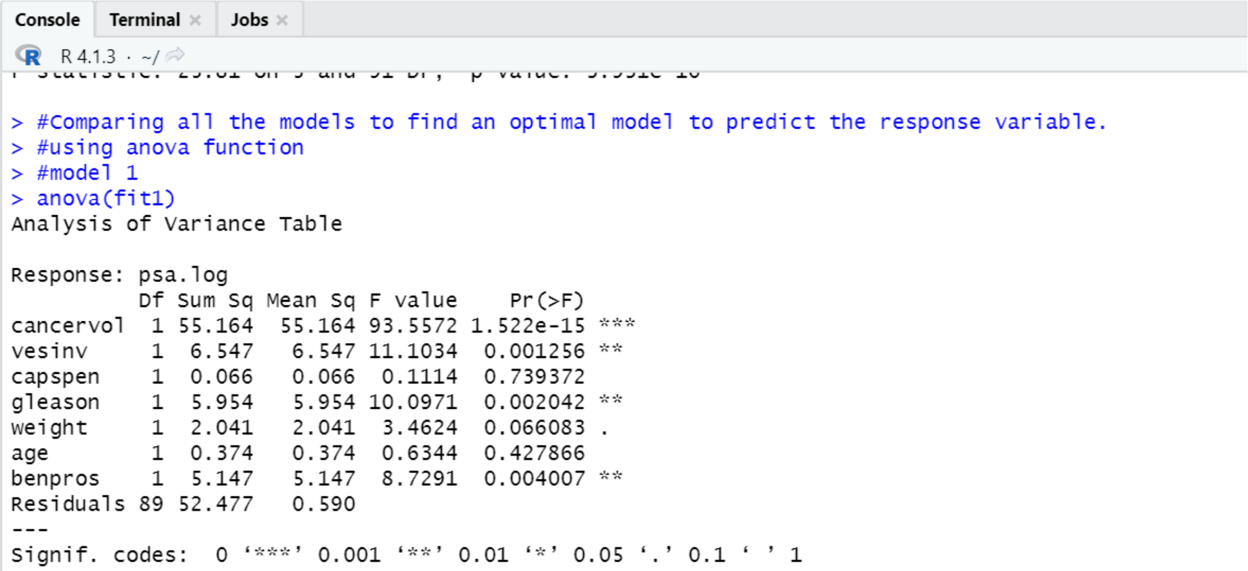
```
>
> #fitting the linear models
> #Model 1
> fit1<-lm(psa.log~cancervol+vesinv+capspen+gleason+weight+age+benpros)
> summary(fit1)

Call:
lm(formula = psa.log ~ cancervol + vesinv + capspen + gleason + 
    weight + age + benpros)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.88309 -0.46629  0.08045  0.47380  1.53219 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.685796   0.998754  -0.687  0.49409    
cancervol    0.069454   0.014624   4.749 7.77e-06 ***
vesinv       0.782623   0.268339   2.917  0.00448 ** 
capspen     -0.026521   0.032860  -0.807  0.42177    
gleason      0.358153   0.127976   2.799  0.00629 ** 
weight       0.001380   0.001822   0.757  0.45079    
age         -0.002799   0.011724  -0.239  0.81186    
benpros      0.087470   0.029605   2.955  0.00401 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7679 on 89 degrees of freedom
Multiple R-squared:  0.5893,	Adjusted R-squared:  0.557 
F-statistic: 18.24 on 7 and 89 DF,  p-value: 7.694e-15
```



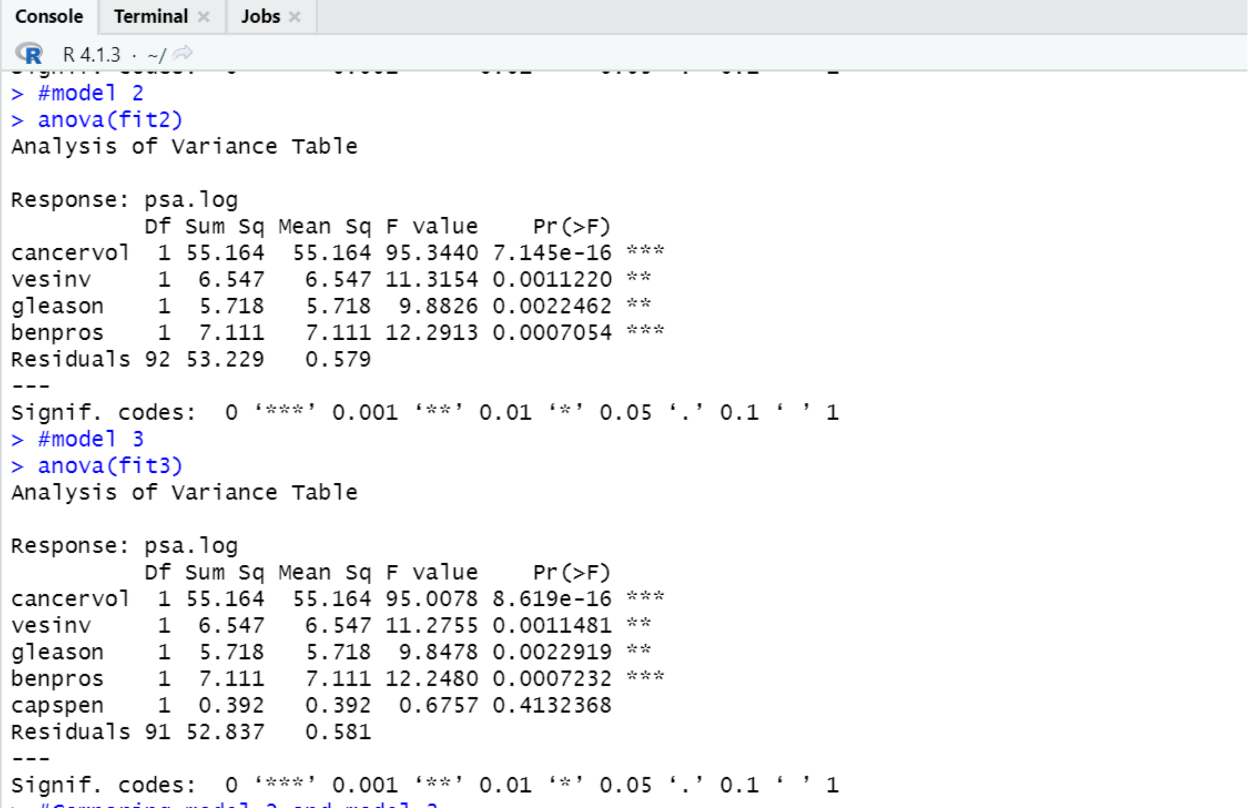
```
Multiple R-squared:  0.5893,	Adjusted R-squared:  0.557 
F-statistic: 18.24 on 7 and 89 DF,  p-value: 7.694e-15

> #Model 2
> # Only the significant predictors are used.
> fit2<-update(fit1,.~.-capspen-age-weight)
> summary(fit2)

Call:
lm(formula = psa.log ~ cancervol + vesinv + gleason + benpros)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.88531 -0.50276  0.09885  0.53687  1.56621 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.65013    0.80999  -0.803 0.424253    
cancervol    0.06488    0.01285   5.051 2.22e-06 ***
vesinv       0.68421    0.23640   2.894 0.004746 ** 
gleason      0.33376    0.12331   2.707 0.008100 ** 
benpros      0.09136    0.02606   3.506 0.000705 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7606 on 92 degrees of freedom
Multiple R-squared:  0.5834,	Adjusted R-squared:  0.5653 
F-statistic: 32.21 on 4 and 92 DF,  p-value: < 2.2e-16
```



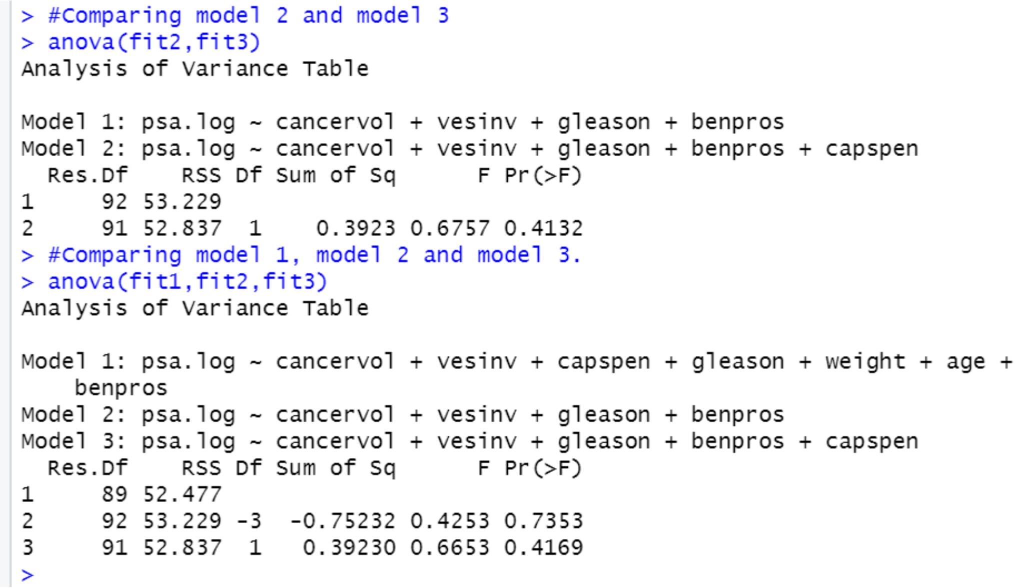
```
> #Model 3
> #including capspen in the model 2.
> fit3<-update(fit2,.~.+capspen)
> summary(fit3)

Call:
lm(formula = psa.log ~ cancervol + vesinv + gleason + benpros + 
    capspen)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.88954 -0.48197  0.08813  0.48409  1.57370 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.73258    0.81760  -0.896 0.372608    
cancervol    0.07029    0.01445   4.863 4.82e-06 ***
vesinv       0.78233    0.26520   2.950 0.004041 ** 
gleason      0.34568    0.12437   2.779 0.006617 ** 
benpros      0.09198    0.02612   3.522 0.000672 ***
capspen     -0.02680    0.03260  -0.822 0.413237    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.762 on 91 degrees of freedom
Multiple R-squared:  0.5865,	Adjusted R-squared:  0.5637 
F-statistic: 25.81 on 5 and 91 DF,  p-value: 3.931e-16
```



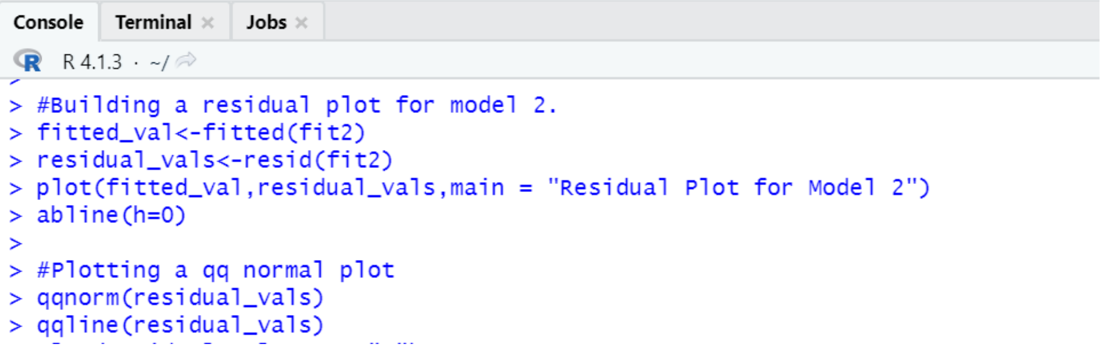
```
F-statistic: 25.81 on 5 and 91 DF,  p-value: 3.931e-16

> #Comparing all the models to find an optimal model to predict the response variable.
> #using anova function
> #model 1
> anova(fit1)
Analysis of Variance Table

Response: psa.log
          Df Sum Sq Mean Sq F value    Pr(>F)    
cancervol  1 55.164  55.164 93.5572 1.522e-15 ***
vesinv     1  6.547   6.547 11.1034  0.001256 ** 
capspen    1  0.066   0.066  0.1114  0.739372    
gleason    1  5.954   5.954 10.0971  0.002042 ** 
weight     1  2.041   2.041  3.4624  0.066083 .  
age        1  0.374   0.374  0.6344  0.427866    
benpros    1  5.147   5.147  8.7291  0.004007 ** 
Residuals 89 52.477   0.590                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



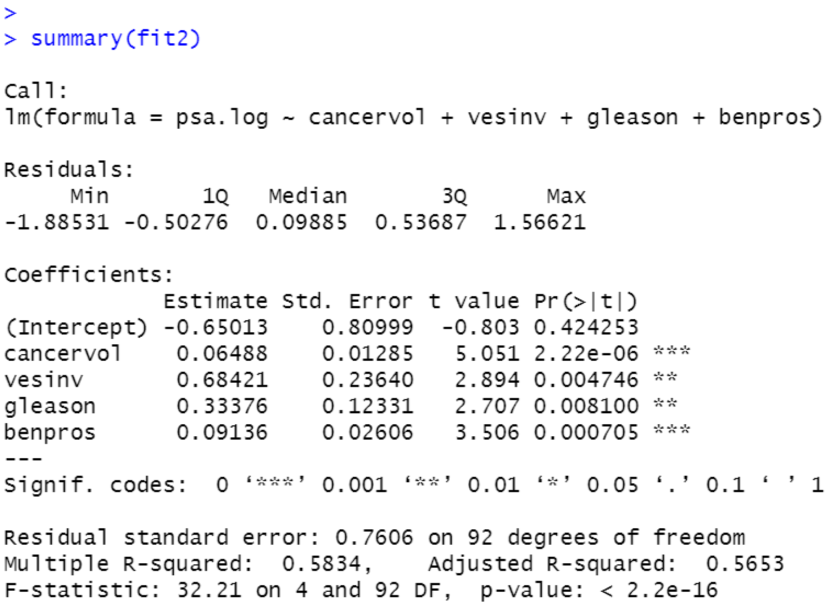
```
> #model 2
> anova(fit2)
Analysis of Variance Table

Response: psa.log
          Df Sum Sq Mean Sq F value    Pr(>F)    
cancervol  1 55.164  55.164 95.3440 7.145e-16 ***
vesinv     1  6.547   6.547 11.3154 0.0011220 ** 
gleason    1  5.718   5.718  9.8826 0.0022462 ** 
benpros    1  7.111   7.111 12.2913 0.0007054 ***
Residuals 92 53.229   0.579                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #model 3
> anova(fit3)
Analysis of Variance Table

Response: psa.log
          Df Sum Sq Mean Sq F value    Pr(>F)    
cancervol  1 55.164  55.164 95.0078 8.619e-16 ***
vesinv     1  6.547   6.547 11.2755 0.0011481 ** 
gleason    1  5.718   5.718  9.8478 0.0022919 ** 
benpros    1  7.111   7.111 12.2480 0.0007232 ***
capspen    1  0.392   0.392  0.6757 0.4132368    
Residuals 91 52.837   0.581                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



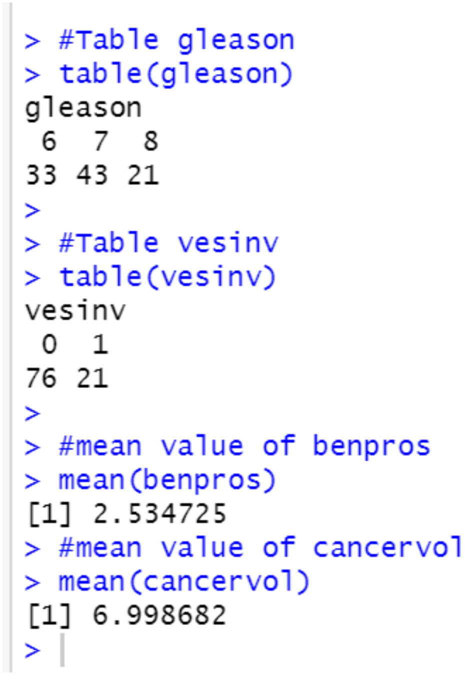
```
> #Comparing model 2 and model 3
> anova(fit2,fit3)
Analysis of Variance Table

Model 1: psa.log ~ cancervol + vesinv + gleason + benpros
Model 2: psa.log ~ cancervol + vesinv + gleason + benpros + capspen
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     92 53.229                           
2     91 52.837  1    0.3923 0.6757 0.4132
> #Comparing model 1, model 2 and model 3.
> anova(fit1,fit2,fit3)
Analysis of Variance Table

Model 1: psa.log ~ cancervol + vesinv + capspen + gleason + weight + age + 
    benpros
Model 2: psa.log ~ cancervol + vesinv + gleason + benpros
Model 3: psa.log ~ cancervol + vesinv + gleason + benpros + capspen
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     89 52.477                           
2     92 53.229 -3  -0.75232 0.4253 0.7353
3     91 52.837  1   0.39230 0.6653 0.4169
> 
```



```
> #Building a residual plot for model 2.
> fitted_val<-fitted(fit2)
> residual_vals<-resid(fit2)
> plot(fitted_val,residual_vals,main = "Residual Plot for Model 2")
> abline(h=0)
>
> #Plotting a qq normal plot
> qqnorm(residual_vals)
> qqline(residual_vals)
```



```
>
> summary(fit2)

Call:
lm(formula = psa.log ~ cancervol + vesinv + gleason + benpros)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.88531 -0.50276  0.09885  0.53687  1.56621 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.65013    0.80999  -0.803 0.424253    
cancervol    0.06488    0.01285   5.051 2.22e-06 ***
vesinv       0.68421    0.23640   2.894 0.004746 ** 
gleason      0.33376    0.12331   2.707 0.008100 ** 
benpros      0.09136    0.02606   3.506 0.000705 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7606 on 92 degrees of freedom
Multiple R-squared:  0.5834,	Adjusted R-squared:  0.5653 
F-statistic: 32.21 on 4 and 92 DF,  p-value: < 2.2e-16
```



```
> #Table gleason
> table(gleason)
gleason
 6  7  8 
33 43 21 
> 
> #Table vesinv
> table(vesinv)
vesinv
 0  1 
76 21 
> 
> #mean value of benpros
> mean(benpros)
[1] 2.534725
> #mean value of cancervol
> mean(cancervol)
[1] 6.998682
> 
```