Mini Project # 6

Names of group members: (Group-18)

Deepika Mamidipelly

Preethi Pasunuri

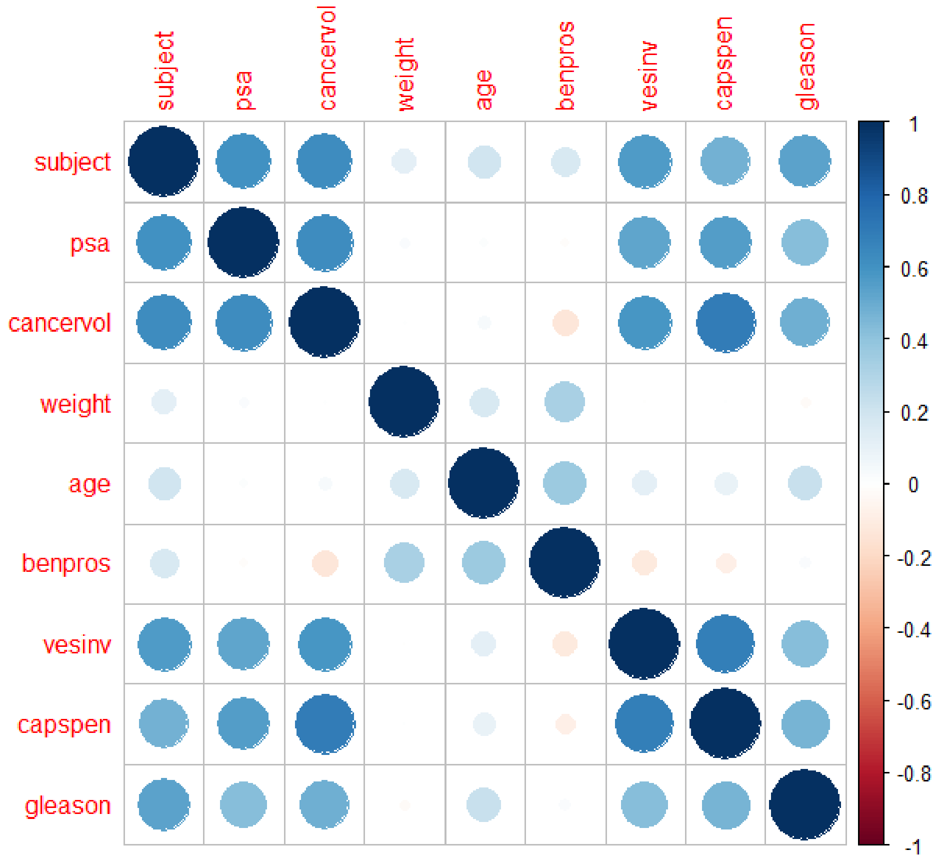
Contribution of each group member

Both group members contributed equally to the inputs for both questions and best is chosen among them to solve these problems. Collaboratively learned R, ran the scripts, and assessed the results. Some of the scripts written by Deepika were analysed and finalized by Preethi and similarly scripts written by Preethi were assessed and finalized by Deepika. Both group members distribute equal amounts of report documentation, which is then integrated into a single final document. Members of the group worked diligently to meet all the project criteria.

Question-1)

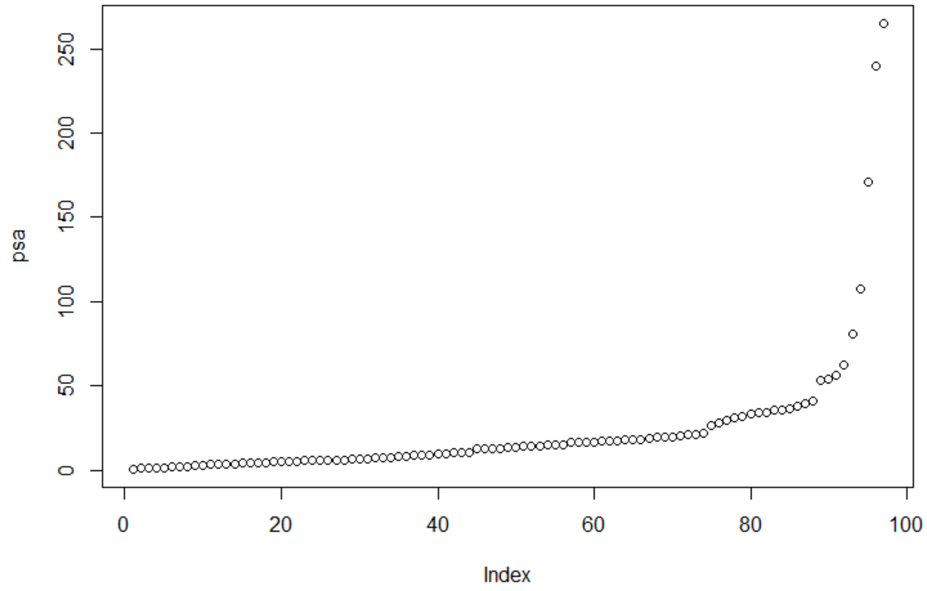
Solution-1:

First load the data and then plot correlation matrix.

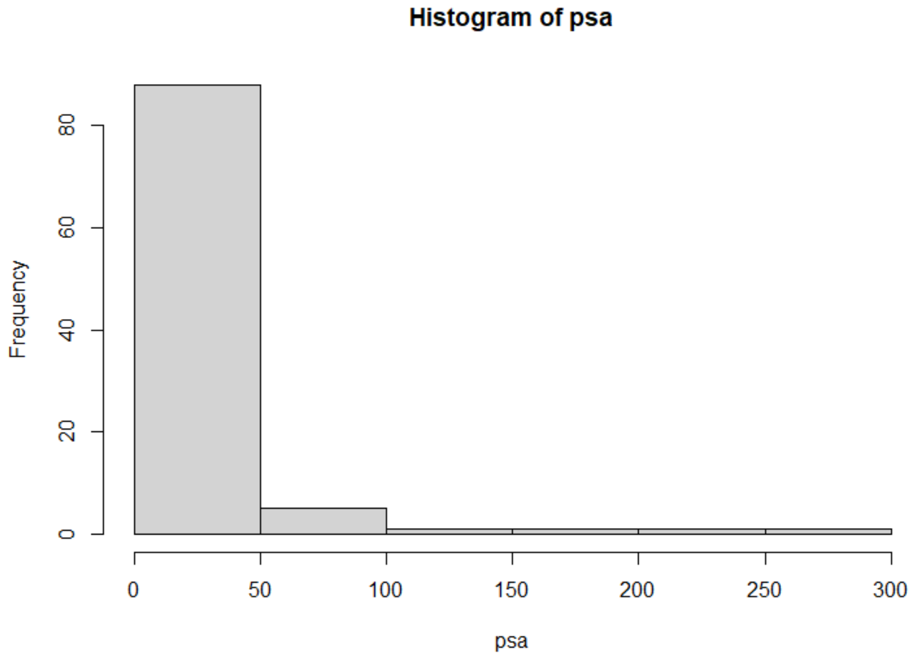


Attaching the data so that we can use the variables

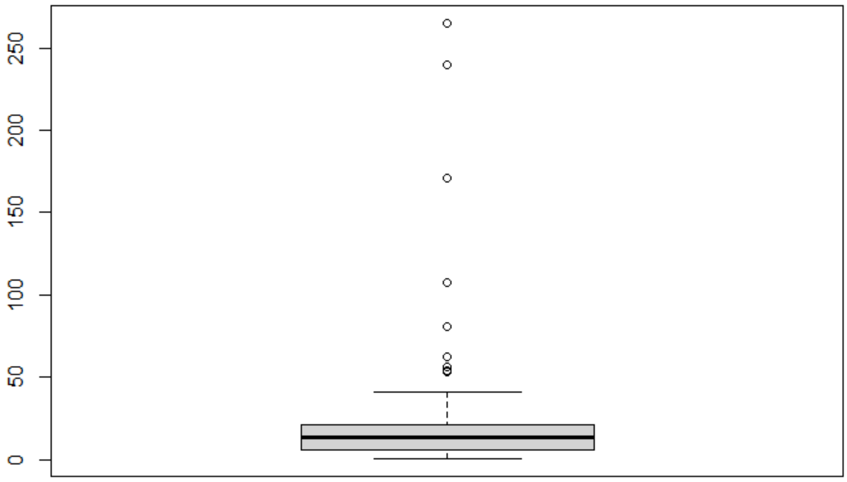
Plotting scattered plot for psa:



Plotting histogram for psa:



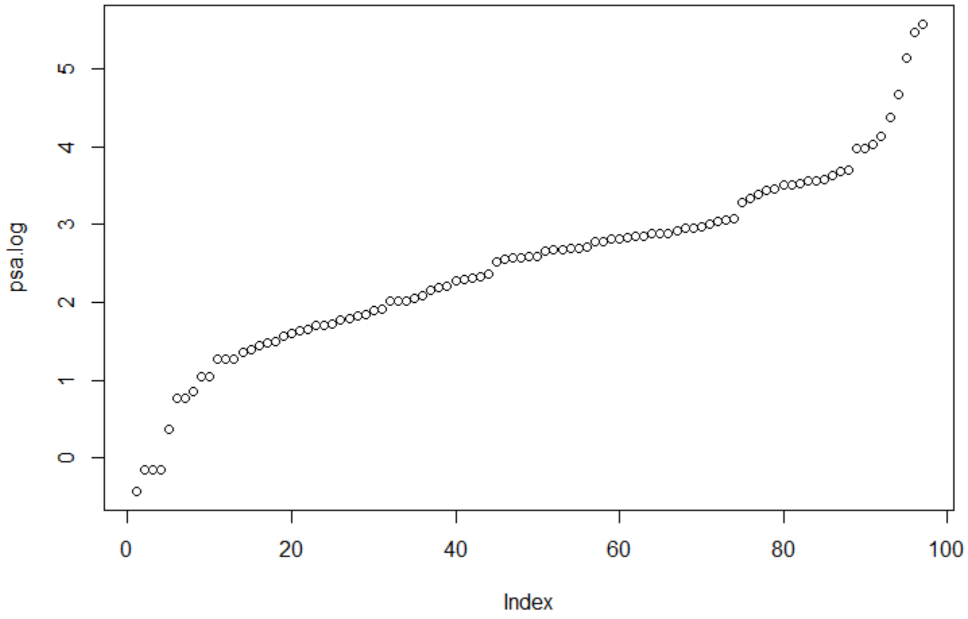
Plotting boxplot for psa:



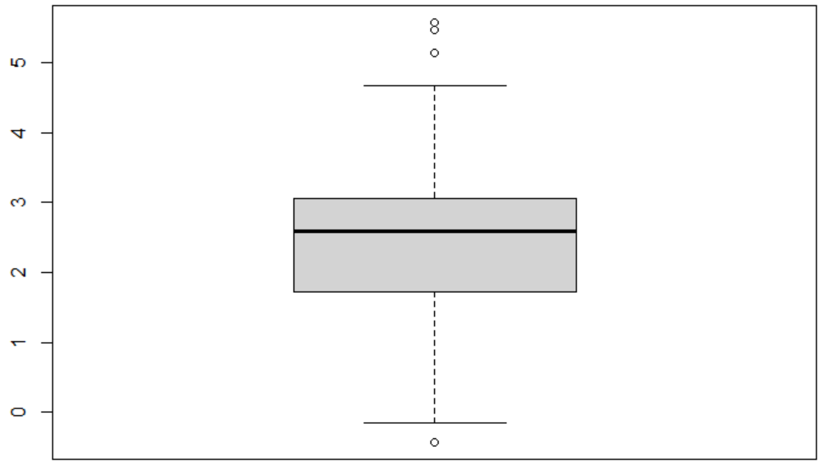
From the box plot we can see there are many outliers in the data. We need some kind of transformation to the data to fit our linear model. Here we are applying logarithmic transformation.

We apply Logarithmic Transformation for the psa and

Plot the scattered plot



Plotting the boxplot for the logarithmically transformed data:



As vesinv is a qualitative variable we use as_factor converts a variable into a factor and preserves the value and variable label attributes.

cancer_data$vesinv <- as.factor(cancer_data$vesinv)

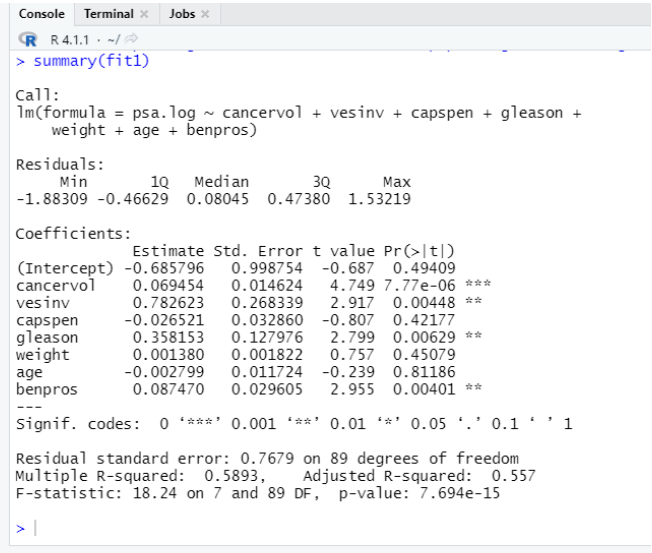
Fitting linear models:

Model 1

Null hypothesis -> H0 : None of the predictors are useful for predicting response.

Alternate hypothesis -> H1: Atleast one of the predictors is useful for predicting the response.

summary(fit1)



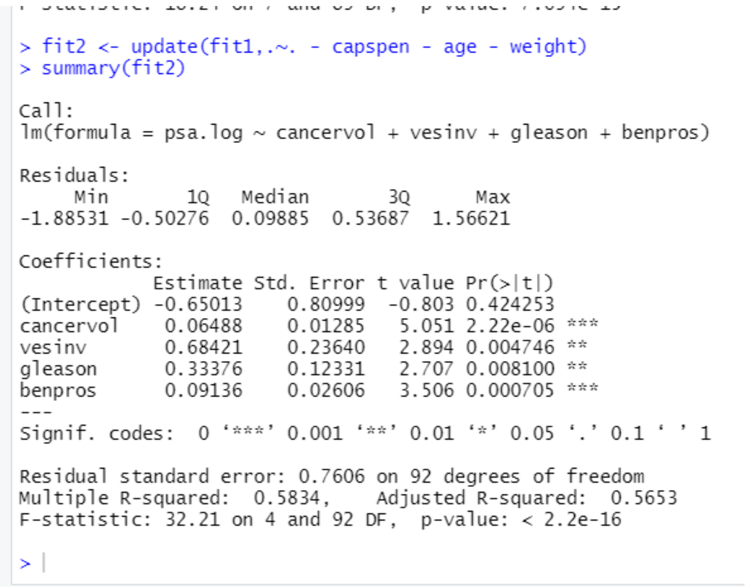
```
> summary(fit1)

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 results we can see that the cancervol which is ***, vesinv, gleason, benpros which has ** are the significant predictors. Hence, we reject the null hypothesis.

Model 2 : Reduced model

For the above hypothesis now we will only consider the significant predictors.

summary(fit2)

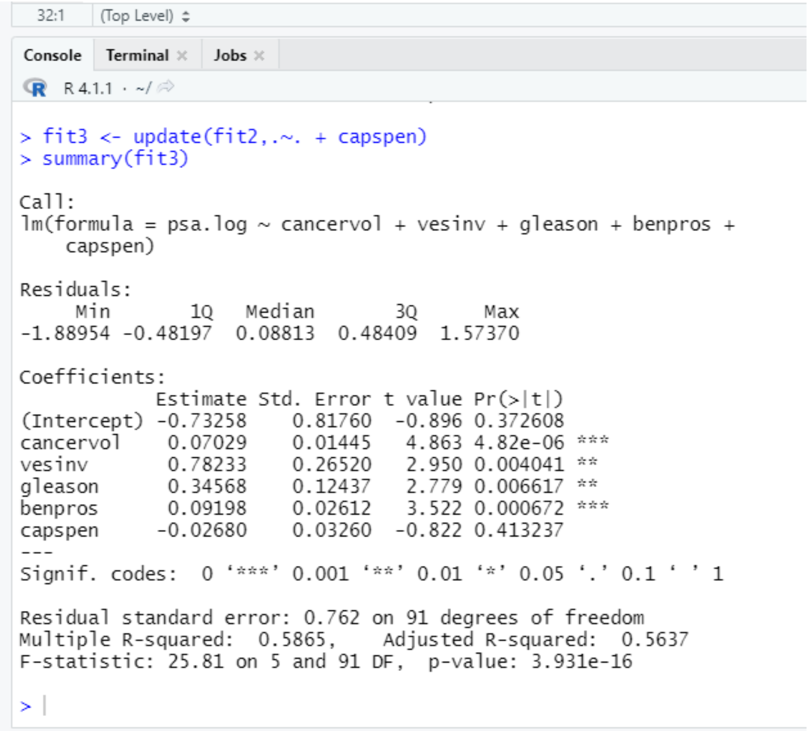


```
> fit2 <- update(fit1,.~. - capspen - age - weight)
> summary(fit2)

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

> 
```

From the Correlation matrix we know that capspen is also important so for the model 3 we can add capspen for model 2.

Model 3:

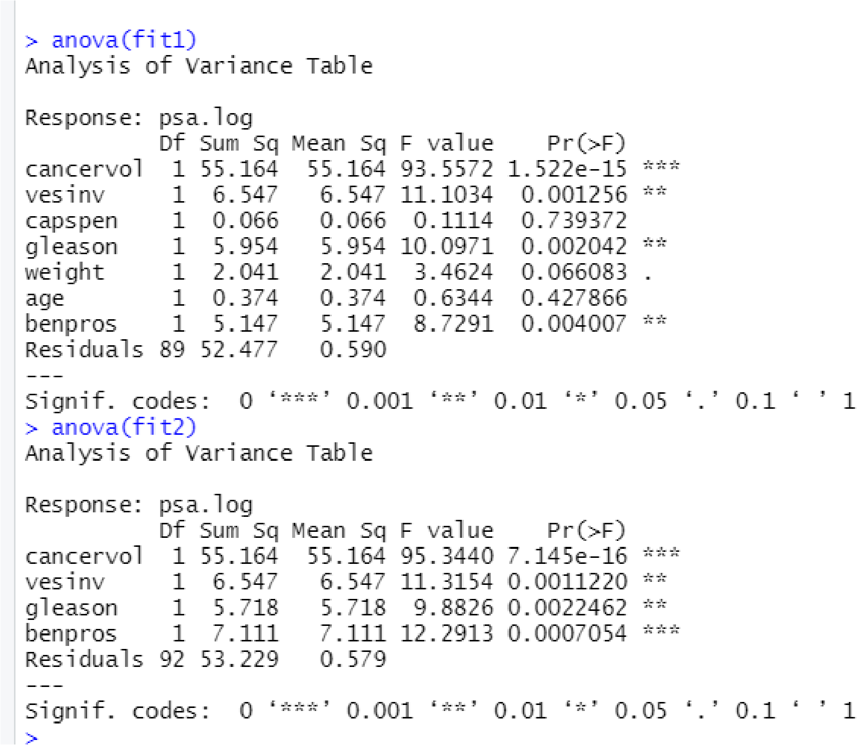
Summary(fit3)



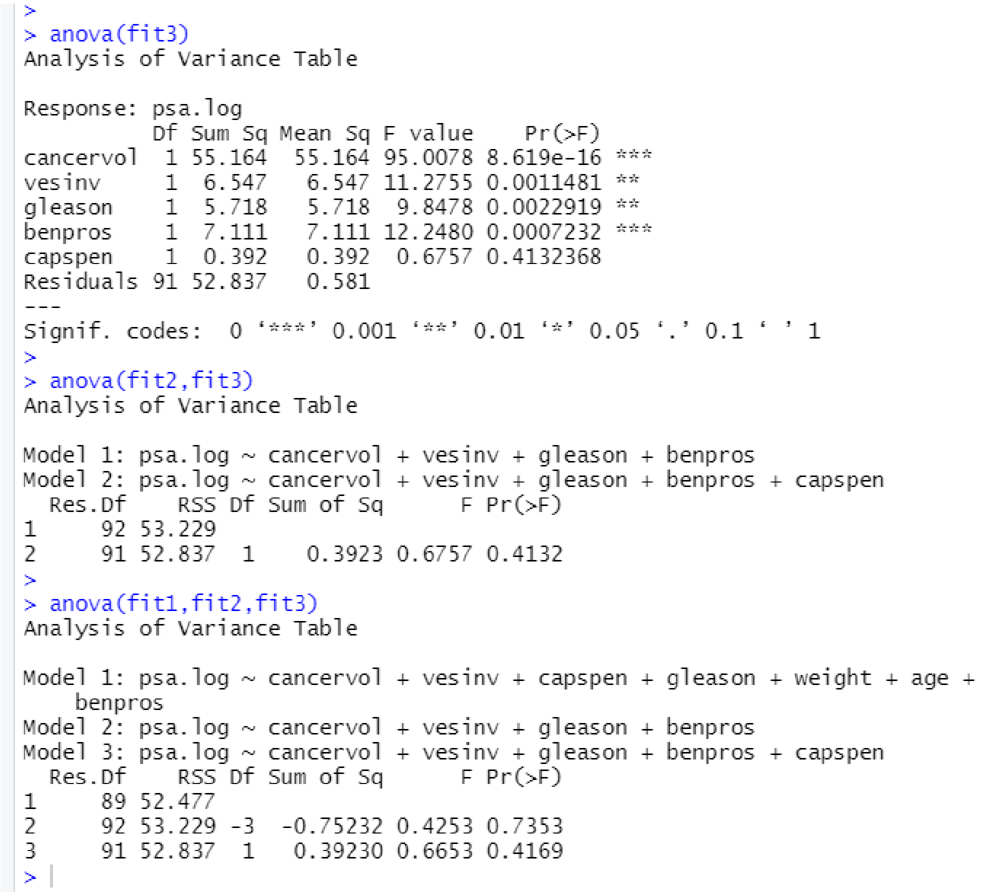
```
> fit3 <- update(fit2,.~. + capspen)
> summary(fit3)

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

> 
```

we can see that the adjusted R-squared value decreases telling that capspen is not an optimal predictor for predicting the response variable.

Comparing all the three models



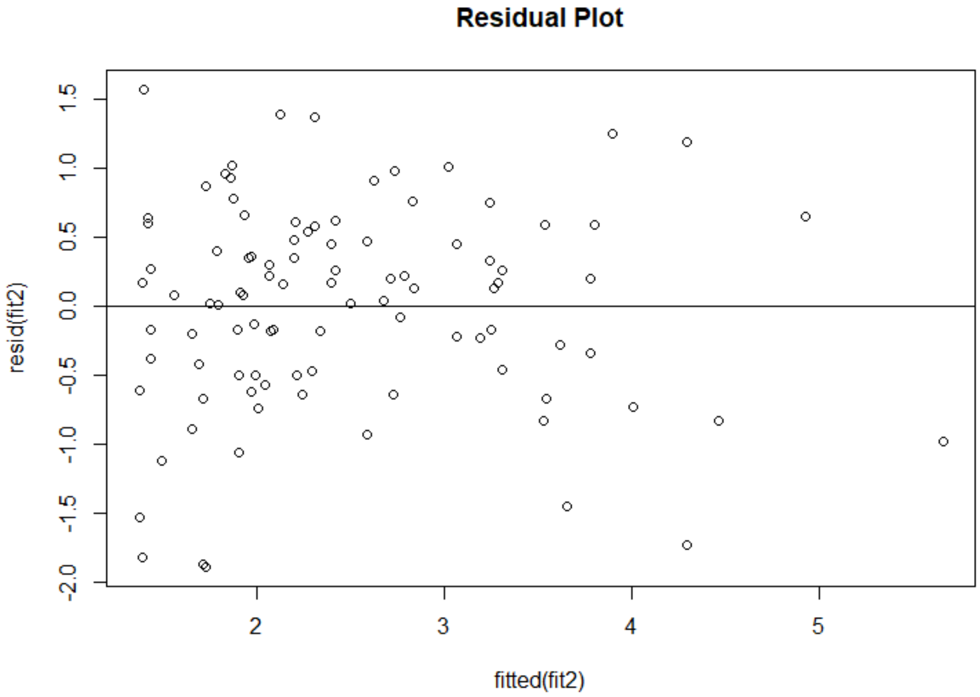
```
> anova(fit1)
Analysis of Variance Table

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
> 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
>
```



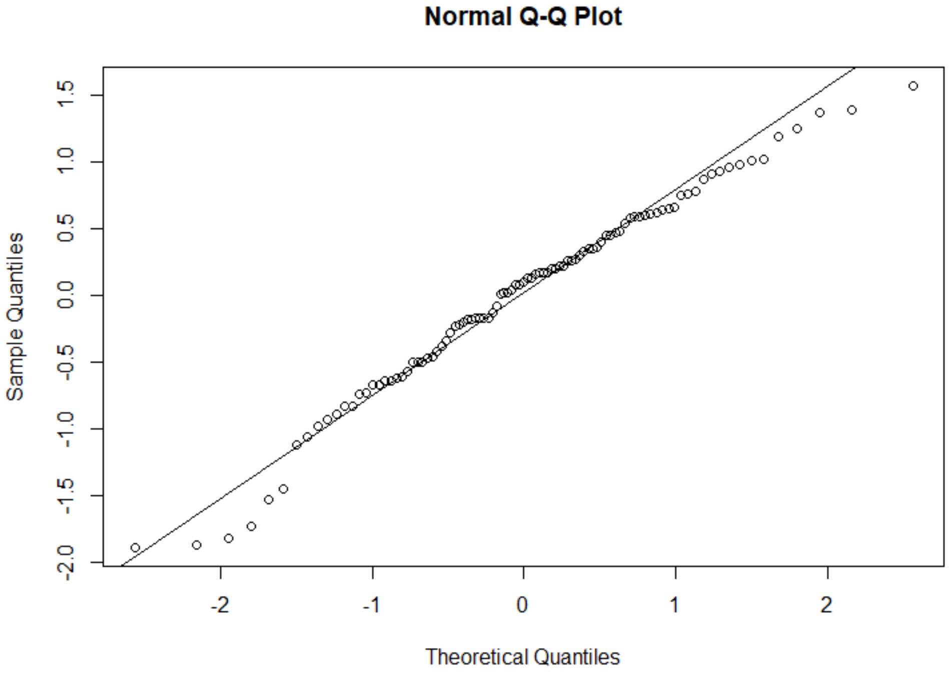
```
>
> anova(fit3)
Analysis of Variance Table

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
>
> 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
>
> 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
> 
```

From the above results we can say that model 2 is the best linear model.

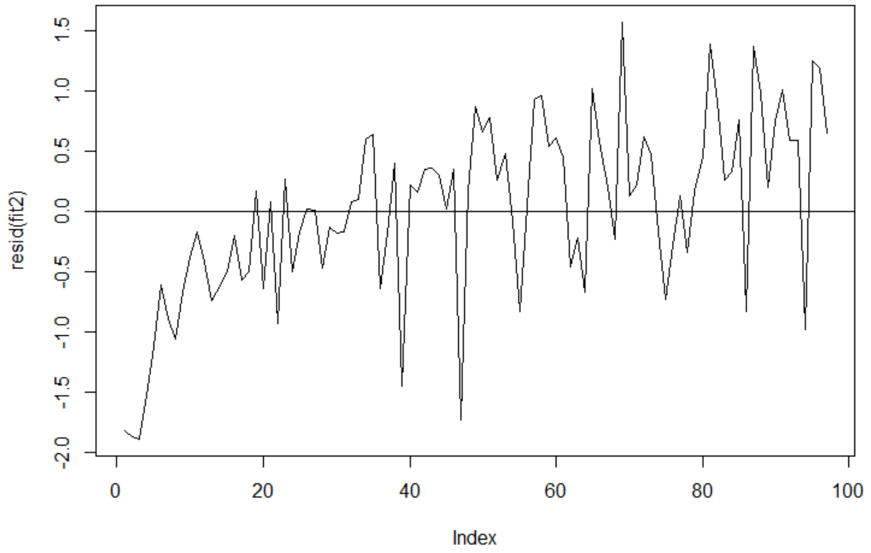
Model Evaluation



The points are scattered around zero and there is not pattern. So, we can say the errors have mean zero and constant variance.

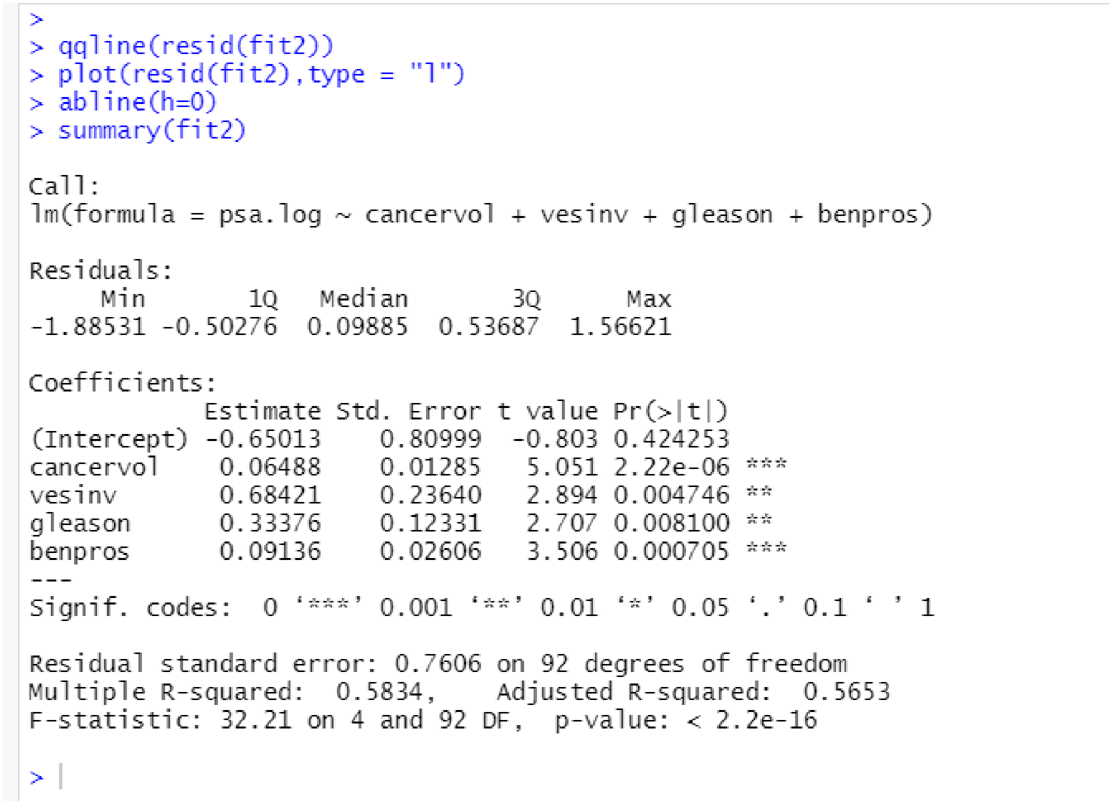


Errors are normally distributed



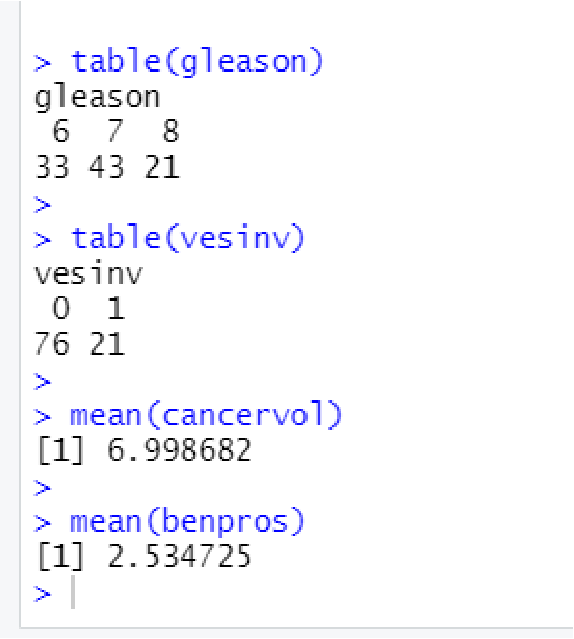
Use the final model to predict the PSA level for a patient whose quantitative predictors are at the sample means of the variables and qualitative predictors are at the most frequent category.

lm(formula = y ~ cancervol + vesinv + gleason + benpros)



```
>
> qqline(resid(fit2))
> plot(resid(fit2),type = "l")
> abline(h=0)
> 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

> |
```

Predict PSA with the model lm(formula = y ~ cancervol + vesinv + gleason + benpros)



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

Observation:

From the above results we can see that gleason value 7 is being dominated in the data, vesinv value 0 is

being dominated in the data and the mean of cancervol and benpros are 6.998 and 2.534 respectively.

predicted value is equal to:

-0.65013 + 6.998682*(0.06488) + 7*(0.33376) + 0.09136*(2.534725) = 2.371837

Thus, the actual value of PSA is exp(2.371837) which is equal to 10.71706

Rcode:

```
install.packages("corrplot")

library("corrplot")

prostate_cancer_data= read.csv("D:\\classes\\Statistics\\Assignments\\prostate_cancer.csv")

cor.data <- cor(prostate_cancer_data)

corrplot(cor.data)

attach(prostate_cancer_data)

plot(psa)

hist(psa)

boxplot(psa)

psa.log = log(psa)

plot(psa.log)

boxplot(psa.log)
```

```
prostate_cancer_data$vesinv <- as.factor(prostate_cancer_data$vesinv)

fit1 <- lm(psa.log ~ cancervol + vesinv + capspen + gleason + weight + age + benpros)
summary(fit1)
fit2 <- update(fit1,.~. - capspen - age - weight)
summary(fit2)
fit3 <- update(fit2,.~. + capspen)
summary(fit3)
anova(fit1)
anova(fit2)
anova(fit3)
anova(fit2,fit3)
anova(fit1,fit2,fit3)
plot(fitted(fit2),resid(fit2), main = "Residual Plot")
abline(h=0)
qqnorm(resid(fit2))
qqline(resid(fit2))
plot(resid(fit2),type = "l")
abline(h=0)
summary(fit2)
table(gleason)
table(vesinv)
mean(cancervol)
mean(benpros)
```