STATISTICAL METHODS FOR DATA SCIENCE Mini-Project 6

Duo Group #23

Members: Hima Sri Tipirineni

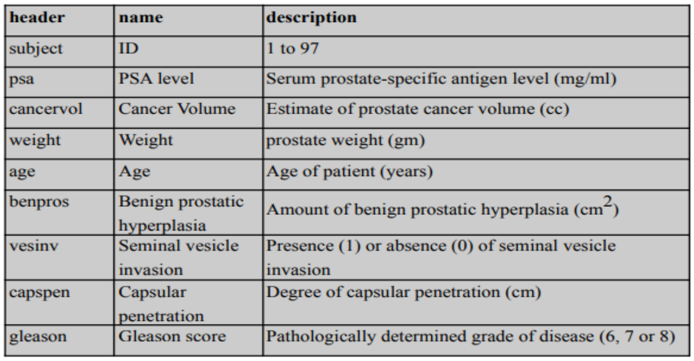
Nithin Pingili

Contribution of each team member:

Hima Sri and Nithin worked together to complete both the questions. Collaborated to learn R and then worked on plotting the scatter plots, qq plots, boxplots and histograms . Also worked on analysing the data and finding the perfect linear model for the data given. Both worked together to answer the question and report all the findings. Hima Sri wrote R code and annotated the code and Nithin worked to check the accuracy of the R code and added the observations.

1. **Consider the prostate cancer dataset available on eLearning as prostate cancer.csv. It consists of data on 97 men with advanced prostate cancer. A description of the variables is given in Figure 1. We would like to understand how PSA level is related to the other predictors in the dataset. Note that vesinv is a qualitative variable. You can treat gleason as a quantitative variable.**

Build a "reasonably good" linear model for these data by taking PSA level as the response variable. Carefully justify all the choices you make in building the model. Be sure to verify the model assumptions. In case a transformation of response is necessary, try the natural log transformation. 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.



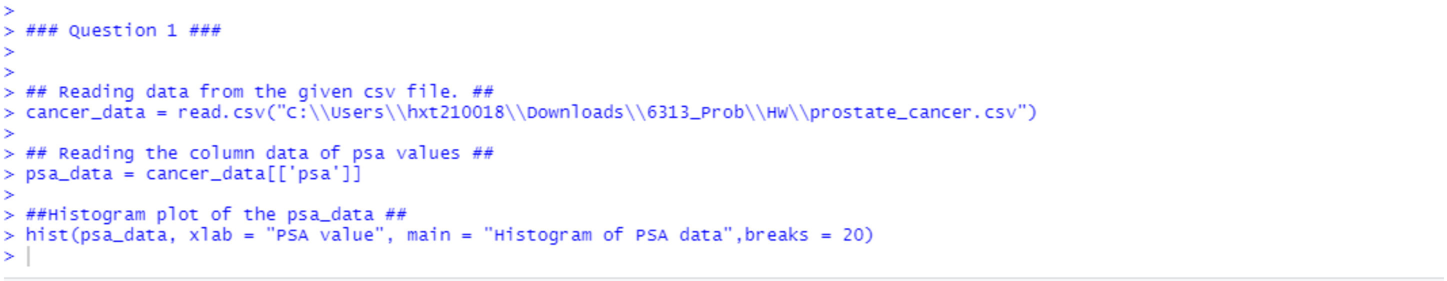
header	name	description
subject	ID	1 to 97
psa	PSA level	Serum prostate-specific antigen level (mg/ml)
cancervol	Cancer Volume	Estimate of prostate cancer volume (cc)
weight	Weight	prostate weight (gm)
age	Age	Age of patient (years)
benpros	Benign prostatic hyperplasia	Amount of benign prostatic hyperplasia (cm^2)
vesinv	Seminal vesicle invasion	Presence (1) or absence (0) of seminal vesicle invasion
capspen	Capsular penetration	Degree of capsular penetration (cm)
gleason	Gleason score	Pathologically determined grade of disease (6, 7 or 8)

Figure 1: List of variables in the prostate cancer data

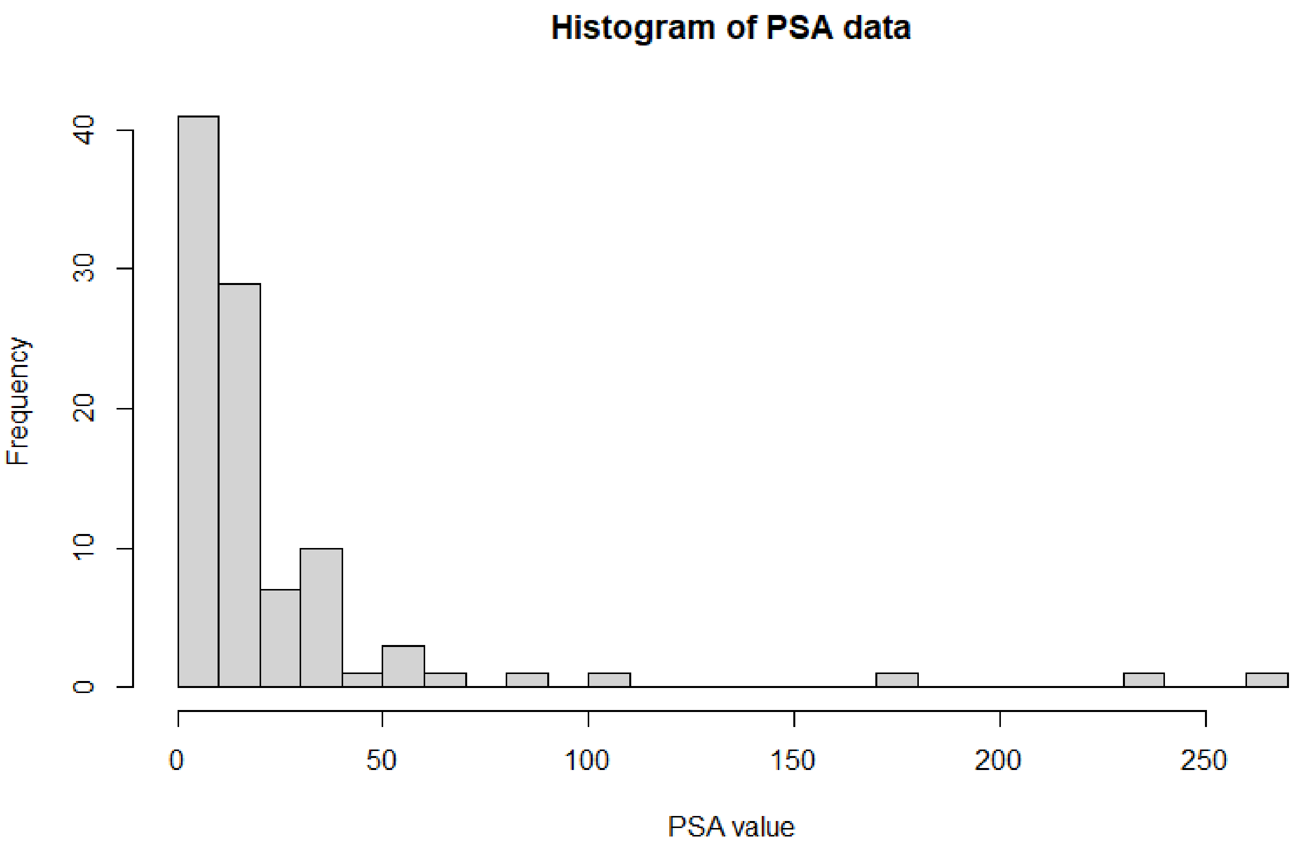
.

Sol :

Performing exploratory analysis on the data which helps in creating an optimal linear model for the given data.

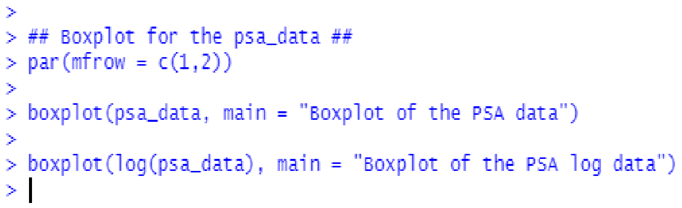


```
>
> ### Question 1 ###
>
>
> ## Reading data from the given csv file. ##
> cancer_data = read.csv("C:\\Users\\hxt210018\\Downloads\\6313_Prob\\HW\\prostate_cancer.csv")
>
> ## Reading the column data of psa values ##
> psa_data = cancer_data[['psa']]
>
> ##Histogram plot of the psa_data ##
> hist(psa_data, xlab = "PSA value", main = "Histogram of PSA data",breaks = 20)
> |
```

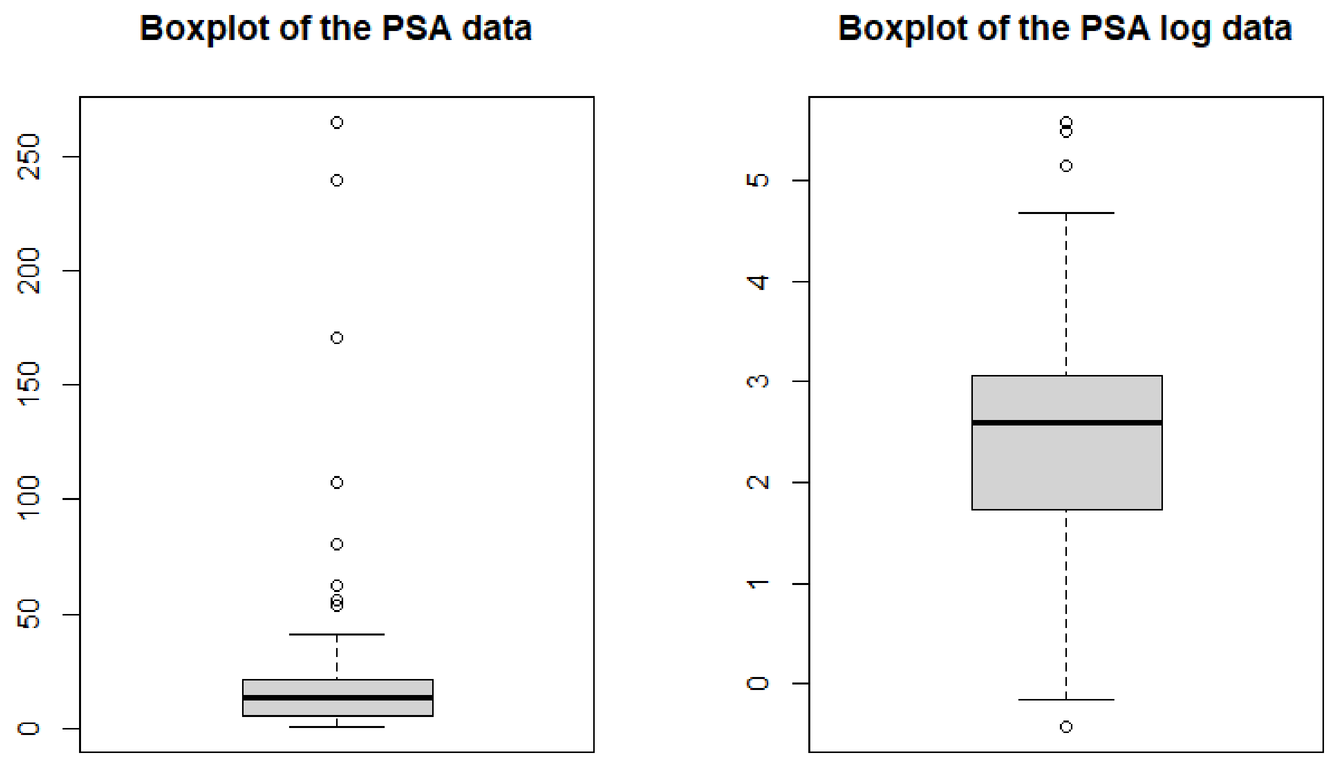


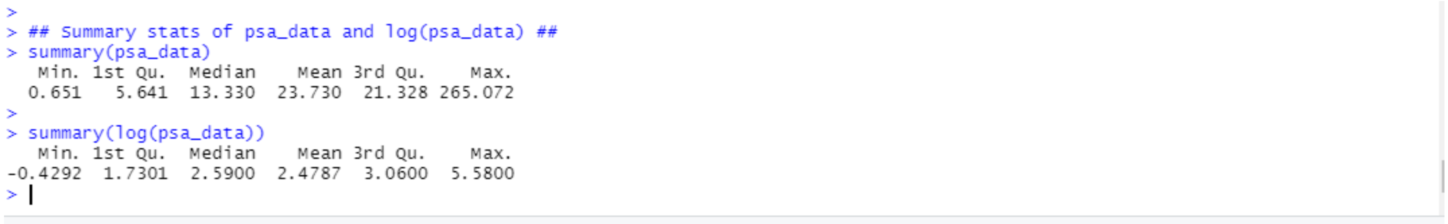
The histogram of PSA data is right skewed. It shows that the population is inversely proportional to the PSA value i.e PSA value decreases as the population increases.

Boxplot analysis :



```
>
> ## Boxplot for the psa_data ##
> par(mfrow = c(1,2))
>
> boxplot(psa_data, main = "Boxplot of the PSA data")
>
> boxplot(log(psa_data), main = "Boxplot of the PSA log data")
> |
```



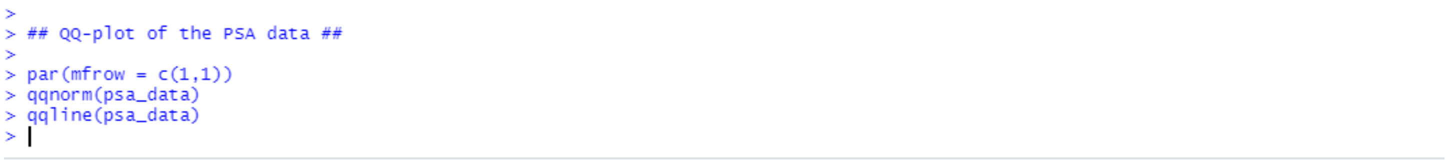


```
>
> ## Summary stats of psa_data and log(psa_data) ##
> summary(psa_data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.651   5.641  13.330  23.730  21.328 265.072 
>
> summary(log(psa_data))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-0.4292  1.7301  2.5900  2.4787  3.0600  5.5800 
> |
```

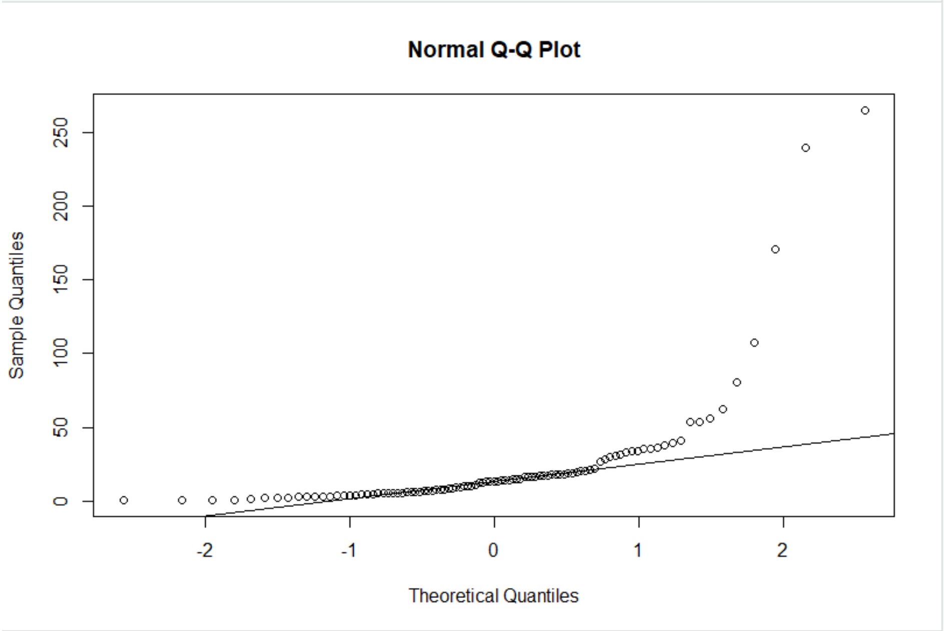
The box plot of PSA data shows the distribution is right skewed because the mean is greater than the median and the whisker is shorter on lower end of the box. Also, the boxplot shows a lot of outliers in the data.

The box plot of natural log transformation of PSA data shows that the distribution is slightly left skewed because the mean is less than the median, the median is closer to Q3 than Q1 and the whisker is shorter on the upper end of the box. This boxplot has less number of outliers compared to the boxplot of the PSA data.

QQ Plot :



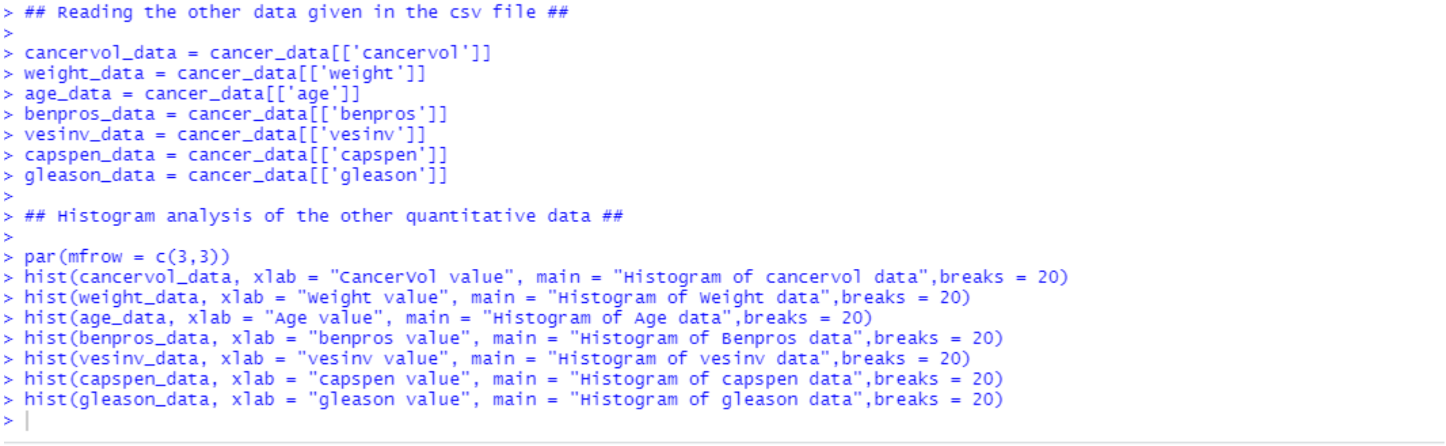
```
>
> ## QQ-plot of the PSA data ##
>
> par(mfrow = c(1,1))
> qqnorm(psa_data)
> qqline(psa_data)
> |
```



In the QQ plot, the points don't fall on a straight line. So, the distribution may not be a good fit for the data. PSA data does not follow Normal distribution.

Analysis of other quantitative data :

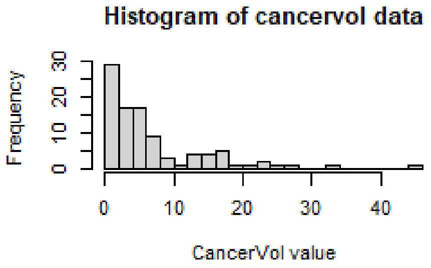
(i)Histogram analysis :

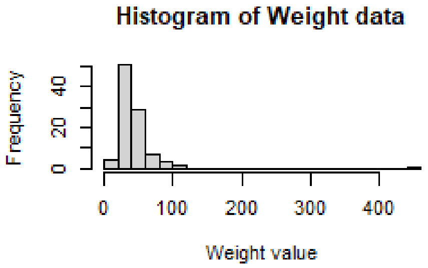


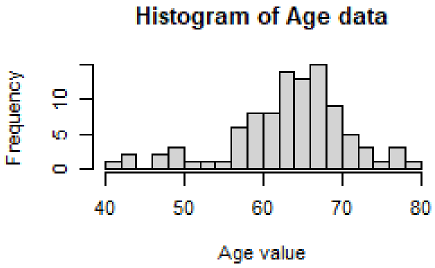
```
> ## Reading the other data given in the csv file ##
>
> cancervol_data = cancer_data[['cancervol']]
> weight_data = cancer_data[['weight']]
> age_data = cancer_data[['age']]
> benpros_data = cancer_data[['benpros']]
> vesinv_data = cancer_data[['vesinv']]
> capspen_data = cancer_data[['capspen']]
> gleason_data = cancer_data[['gleason']]
>
> ## Histogram analysis of the other quantitative data ##
>
> par(mfrow = c(3,3))
> hist(cancervol_data, xlab = "Cancervol value", main = "Histogram of cancervol data",breaks = 20)
> hist(weight_data, xlab = "Weight value", main = "Histogram of Weight data",breaks = 20)
> hist(age_data, xlab = "Age value", main = "Histogram of Age data",breaks = 20)
> hist(benpros_data, xlab = "benpros value", main = "Histogram of Benpros data",breaks = 20)
> hist(vesinv_data, xlab = "vesinv value", main = "Histogram of vesinv data",breaks = 20)
> hist(capspen_data, xlab = "capspen value", main = "Histogram of capspen data",breaks = 20)
> hist(gleason_data, xlab = "gleason value", main = "Histogram of gleason data",breaks = 20)
> |
```

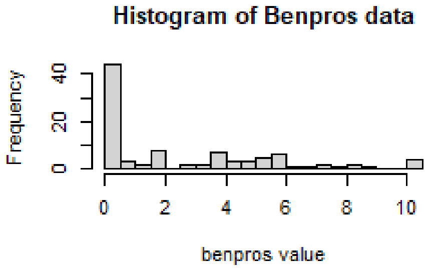
From the below histograms, we can conclude the following :

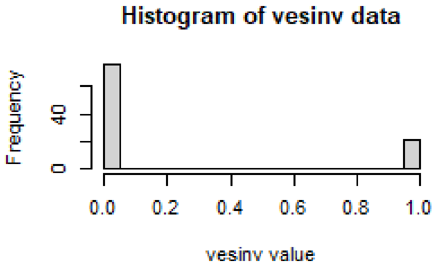
- Cancervol : This distribution is similar to the PSA data distribution. Hence, there might be linear relationship between the cancervol and psa data variables.
- Weight : The weight data distribution has no particular similarity with the psa data. The distribution might be slightly normal or gamma.
- Age : This distribution looks more like a normal distribution.
- Benpros : This distribution is similar to the PSA data and cancervol distribution. Hence , there might be linear relationship between the benpros , cancervol and psa data variables.
- Vesinv : It looks like a Bernoulli variable which takes only 2 values i.e 0 or 1.
- Capspen : This distribution is similar to the PSA data ,cancervol and benspros distribution. Hence , there might be linear relationship between the capspen, benpros , cancervol and psa data variables.
- Gleason : This distribution tells that the gleason variable has only 3 values i.e 6.0, 7.0 and 8.0.

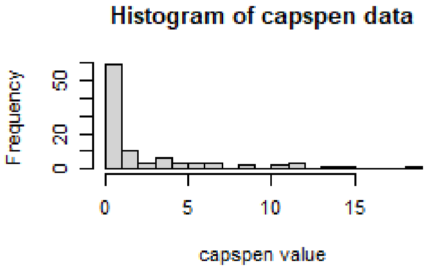


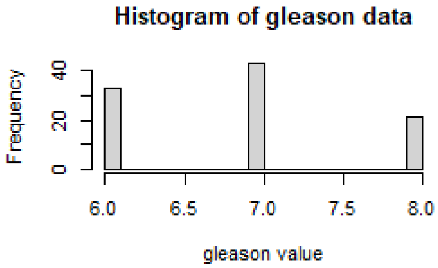






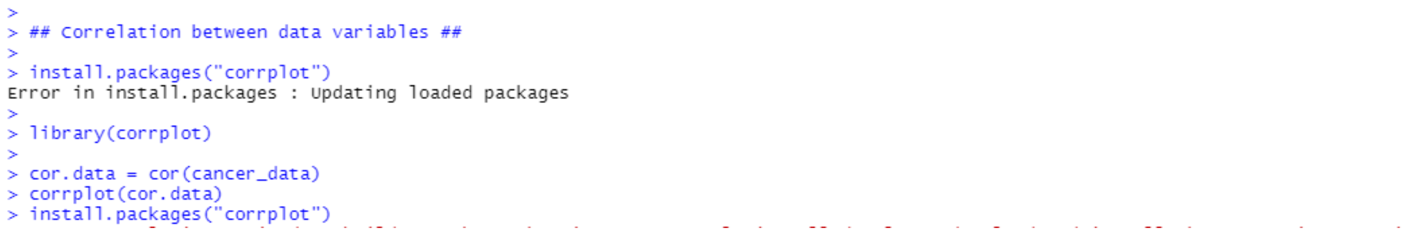




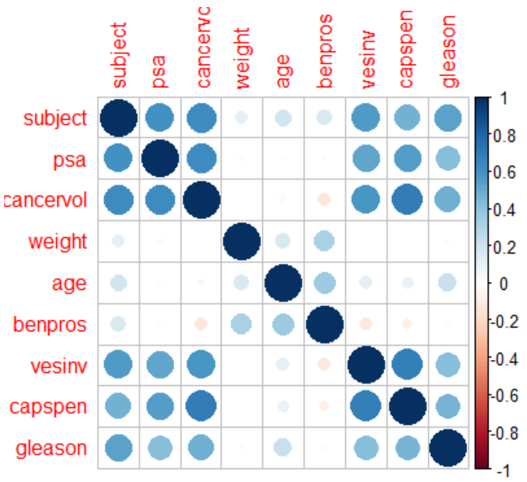


(ii) Correlation between the data variables :

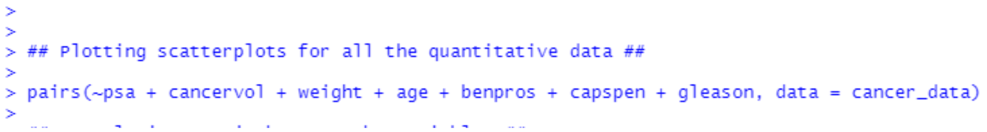
We are plotting a correlation matrix using the corrplot library in order to verify the correlation between the variables.



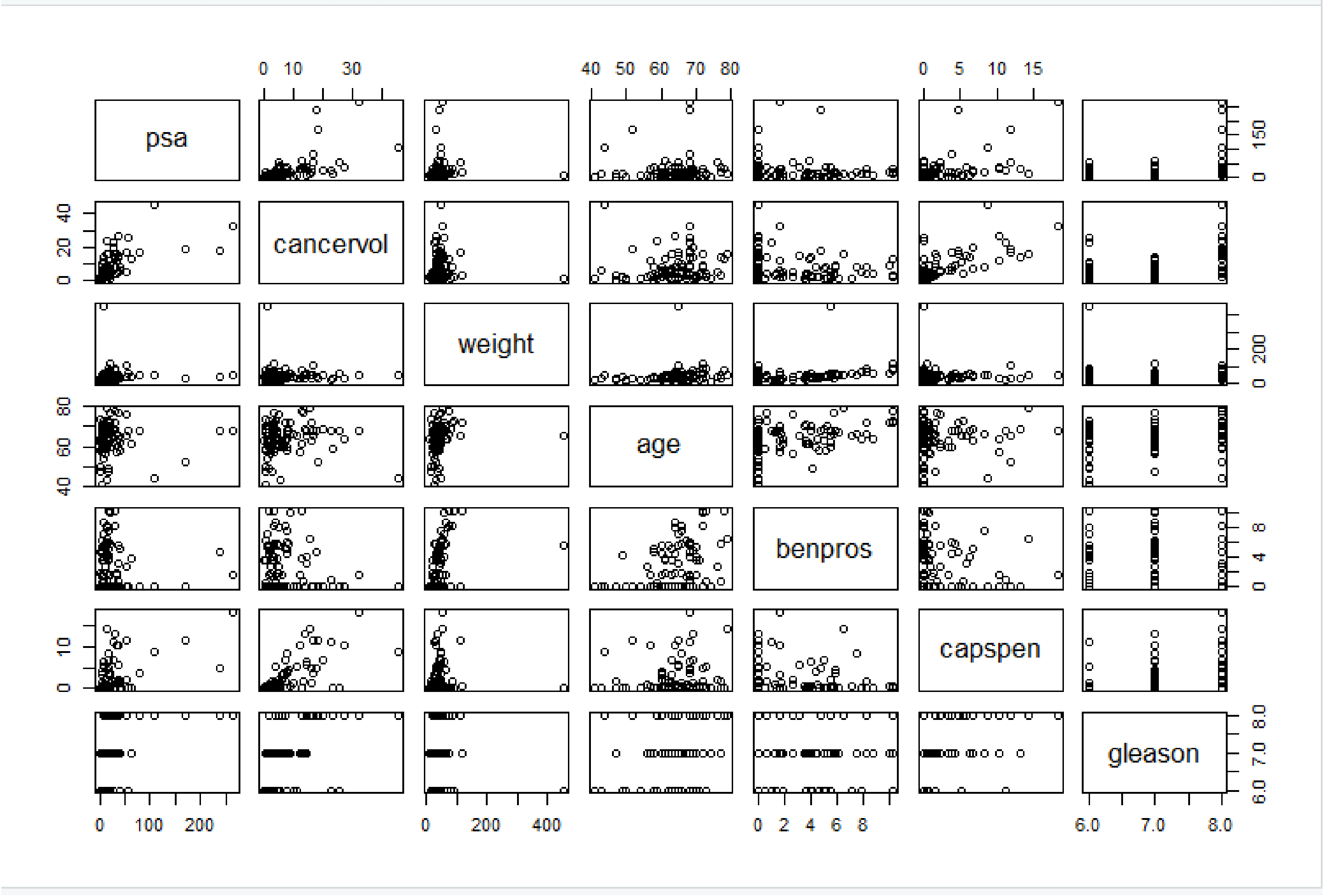
```
>
> ## Correlation between data variables ##
>
> install.packages("corrplot")
Error in install.packages : Updating loaded packages
>
> library(corrplot)
>
> cor.data = cor(cancer_data)
> corrplot(cor.data)
> install.packages("corrplot")
```

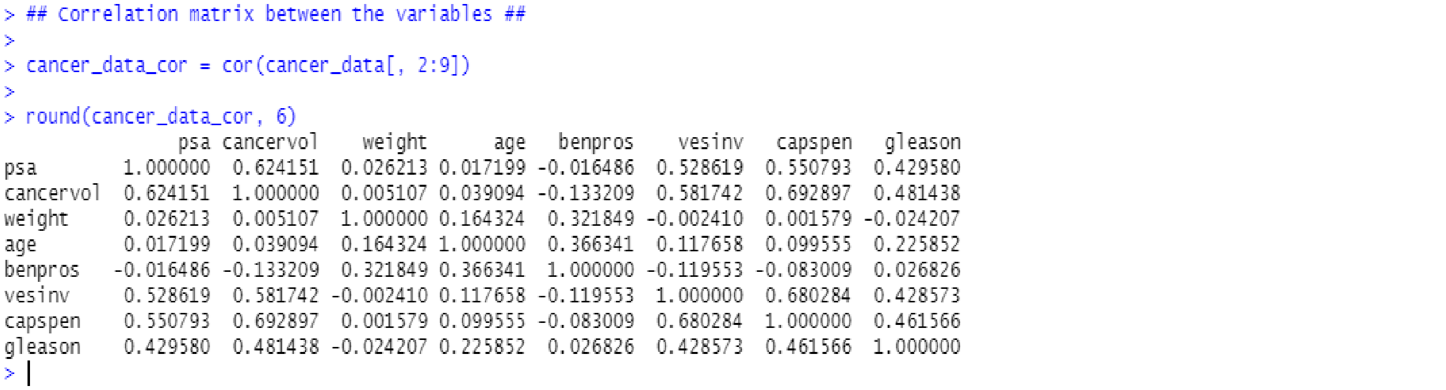


(iii) Plotting scatterplots for the visualization of all the quantitative data :



```
>
>
> ## Plotting scatterplots for all the quantitative data ##
>
> pairs(~psa + cancervol + weight + age + benpros + capspen + gleason, data = cancer_data)
>
```

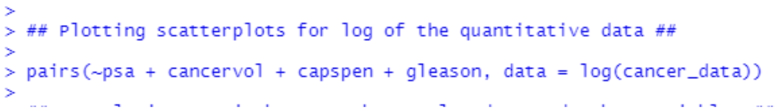




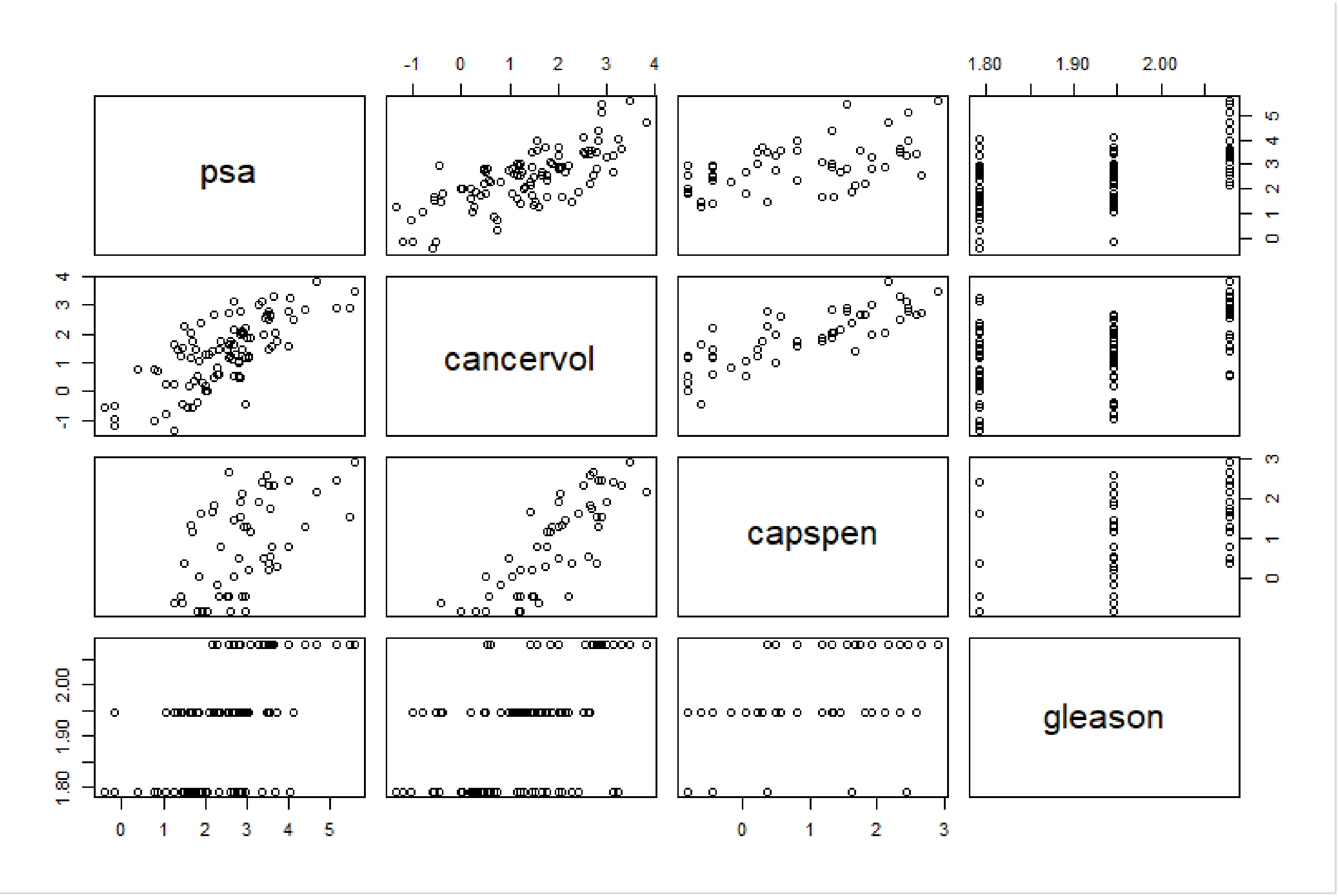
```
> ## Correlation matrix between the variables ##
>
> cancer_data_cor = cor(cancer_data[, 2:9])
>
> round(cancer_data_cor, 6)
               psa cancervol    weight      age   benpros    vesinv   capspen   gleason
psa       1.000000  0.624151  0.026213 0.017199 -0.016486  0.528619  0.550793  0.429580
cancervol  0.624151  1.000000  0.005107 0.039094 -0.133209  0.581742  0.692897  0.481438
weight     0.026213  0.005107  1.000000 0.164324  0.321849 -0.002410  0.001579 -0.024207
age        0.017199  0.039094  0.164324 1.000000  0.366341  0.117658  0.099555  0.225852
benpros   -0.016486 -0.133209  0.321849 0.366341  1.000000 -0.119553 -0.083009  0.026826
vesinv     0.528619  0.581742 -0.002410 0.117658 -0.119553  1.000000  0.680284  0.428573
capspen    0.550793  0.692897  0.001579 0.099555 -0.083009  0.680284  1.000000  0.461566
gleason    0.429580  0.481438 -0.024207 0.225852  0.026826  0.428573  0.461566  1.000000
> |
```

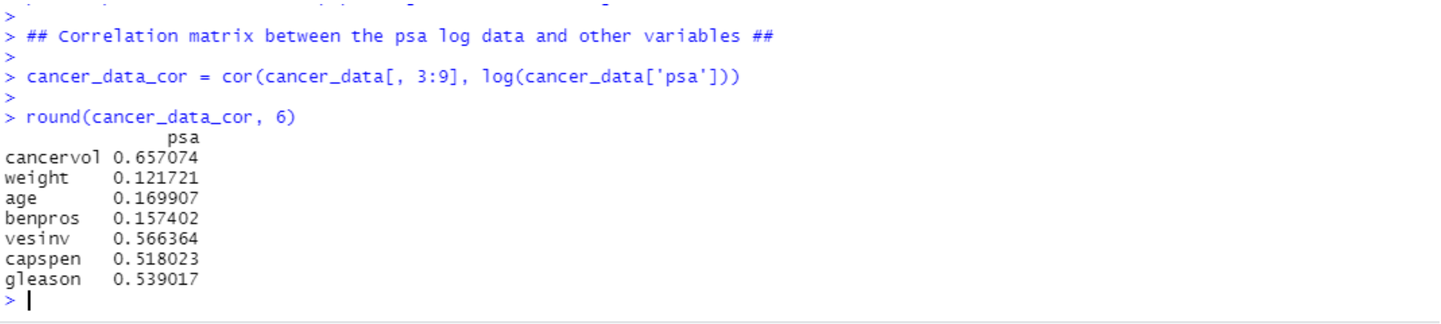
The correlation matrix shows that the correlations between PSA and cancervol, vesinv, capspen are moderate. the correlation values between PSA and weight, age, benpros are very low so there is no relationship between these variables. The correlation between PSA and gleason is weak. So, there is no strong correlation between PSA and the other predictors.

(iv) Scatterplots for the log of the quantitative data :



```
>
> ## Plotting scatterplots for log of the quantitative data ##
>
> pairs(~psa + cancervol + capspen + gleason, data = log(cancer_data))
>
```





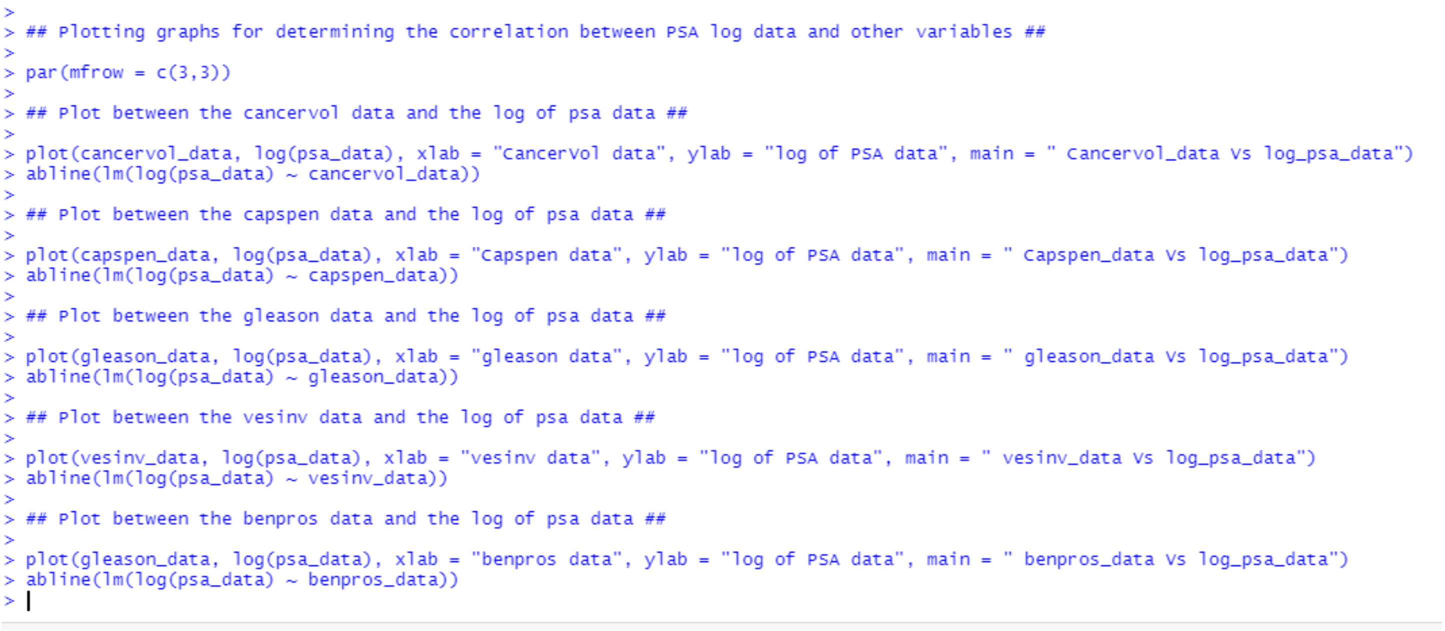
```
>
> ## Correlation matrix between the psa log data and other variables ##
>
> cancer_data_cor = cor(cancer_data[, 3:9], log(cancer_data['psa']))
>
> round(cancer_data_cor, 6)
               psa
cancervol 0.657074
weight    0.121721
age       0.169907
benpros   0.157402
vesinv    0.566364
capspen   0.518023
gleason   0.539017
> |
```

The correlation matrix between natural log transformation of PSA and other variables shows that the correlation values increased for all other variables except capspen. Also, the correlation value between PSA and gleason increased by a significant amount that there is a moderate linear relationship between the two variables.

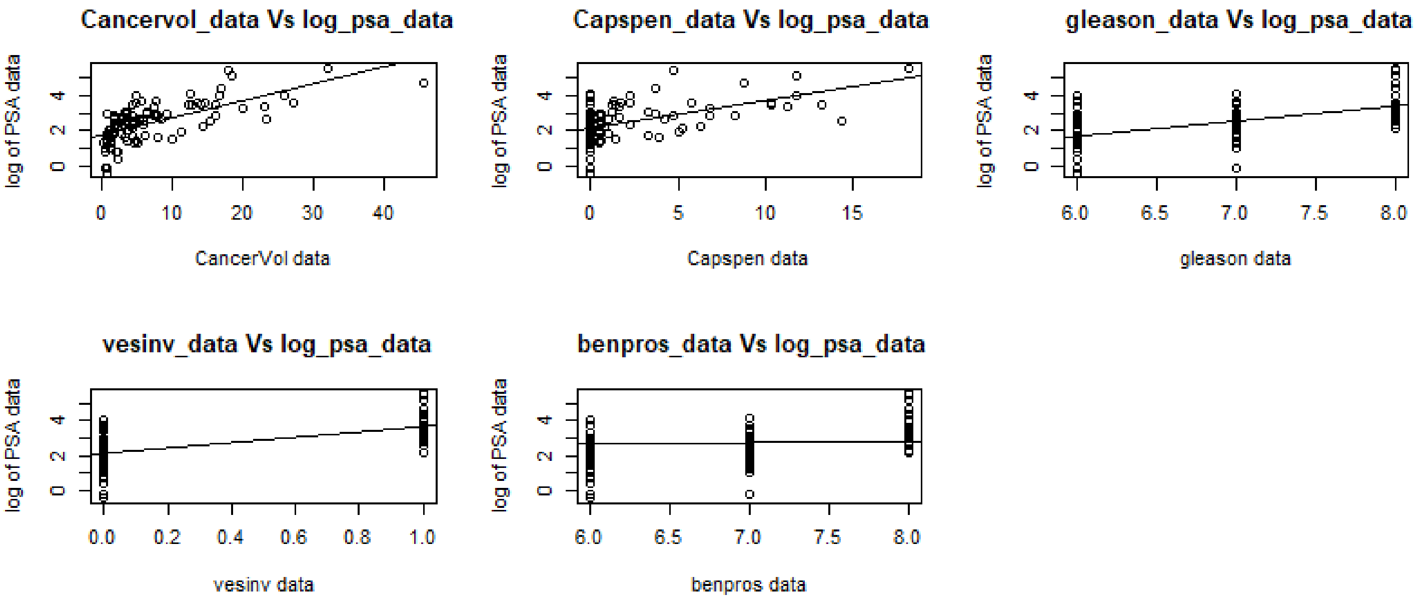
Analysis of variable selection for the linear regression modelling:

We can start with the basic linear model for the log(psa_data) using the cancervol, capspen ,vesinv ,gleason, benpros variables. We can eliminate other variables because the correlation values are less and adding them will not give any value.

Later we will look out for the log(psa_data) Vs the cancervoldata, capspen , benpros and gleason which gives an ideal linear model.



```
>
> ## Plotting graphs for determining the correlation between PSA log data and other variables ##
>
> par(mfrow = c(3,3))
>
> ## Plot between the cancervol data and the log of psa data ##
>
> plot(cancervol_data, log(psa_data), xlab = "CancerVol data", ylab = "log of PSA data", main = " Cancervol_data Vs log_psa_data")
> abline(lm(log(psa_data) ~ cancervol_data))
>
> ## Plot between the capspen data and the log of psa data ##
>
> plot(capspen_data, log(psa_data), xlab = "Capspen data", ylab = "log of PSA data", main = " Capspen_data Vs log_psa_data")
> abline(lm(log(psa_data) ~ capspen_data))
>
> ## Plot between the gleason data and the log of psa data ##
>
> plot(gleason_data, log(psa_data), xlab = "gleason data", ylab = "log of PSA data", main = " gleason_data Vs log_psa_data")
> abline(lm(log(psa_data) ~ gleason_data))
>
> ## Plot between the vesinv data and the log of psa data ##
>
> plot(vesinv_data, log(psa_data), xlab = "vesinv data", ylab = "log of PSA data", main = " vesinv_data Vs log_psa_data")
> abline(lm(log(psa_data) ~ vesinv_data))
>
> ## Plot between the benpros data and the log of psa data ##
>
> plot(gleason_data, log(psa_data), xlab = "benpros data", ylab = "log of PSA data", main = " benpros_data Vs log_psa_data")
> abline(lm(log(psa_data) ~ benpros_data))
> |
```



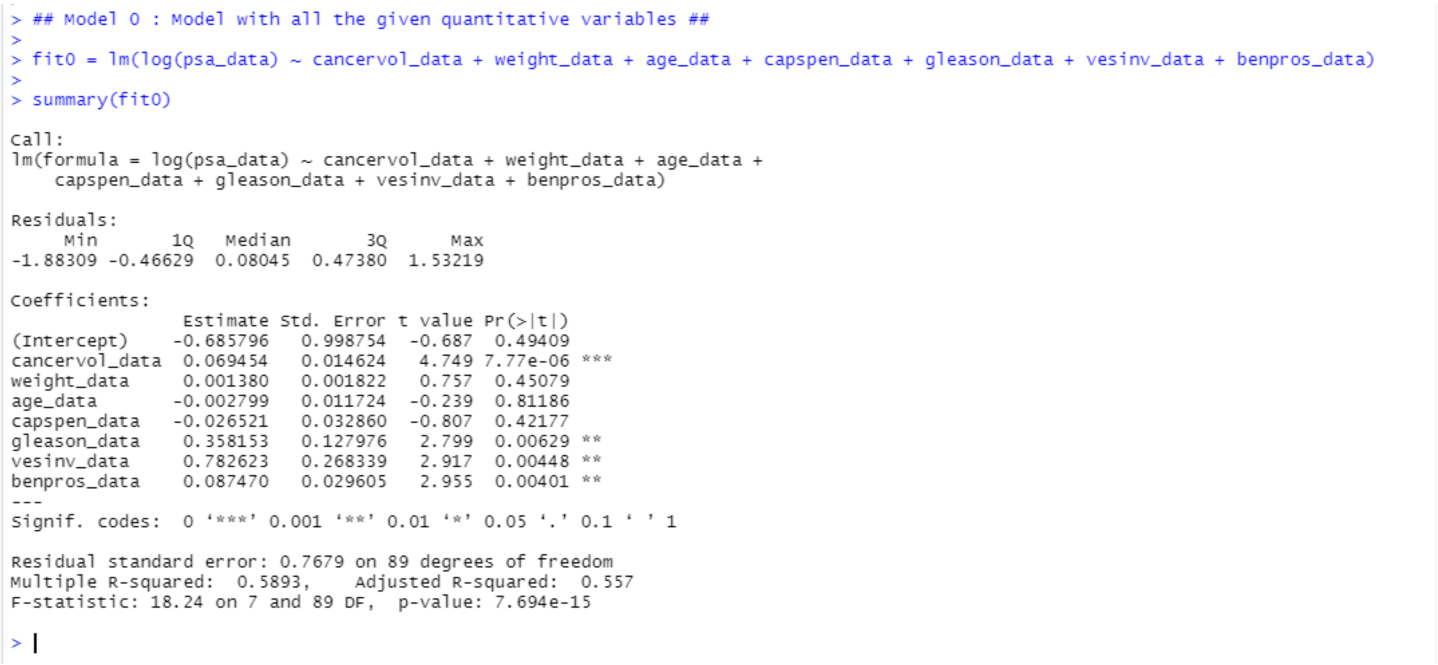
Linear Regression Modelling :

Model 0 :

Predictors used : cancervoldata, weight, age, capspendata, gleasondata ,vesinv, benspros data.

Null Hypothesis: None of the predictors help predict the PSA level.

Alternative Hypothesis: At least one of the predictors helps predict the PSA level.



```
> ## Model 0 : Model with all the given quantitative variables ##
>
> fit0 = lm(log(psa_data) ~ cancervol_data + weight_data + age_data + capspen_data + gleason_data + vesinv_data + benpros_data)
>
> summary(fit0)

Call:
lm(formula = log(psa_data) ~ cancervol_data + weight_data + age_data +
    capspen_data + gleason_data + vesinv_data + benpros_data)

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_data  0.069454   0.014624   4.749 7.77e-06 ***
weight_data     0.001380   0.001822   0.757  0.45079
age_data       -0.002799   0.011724  -0.239  0.81186
capspen_data   -0.026521   0.032860  -0.807  0.42177
gleason_data    0.358153   0.127976   2.799  0.00629 **
vesinv_data     0.782623   0.268339   2.917  0.00448 **
benpros_data    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

> |
```

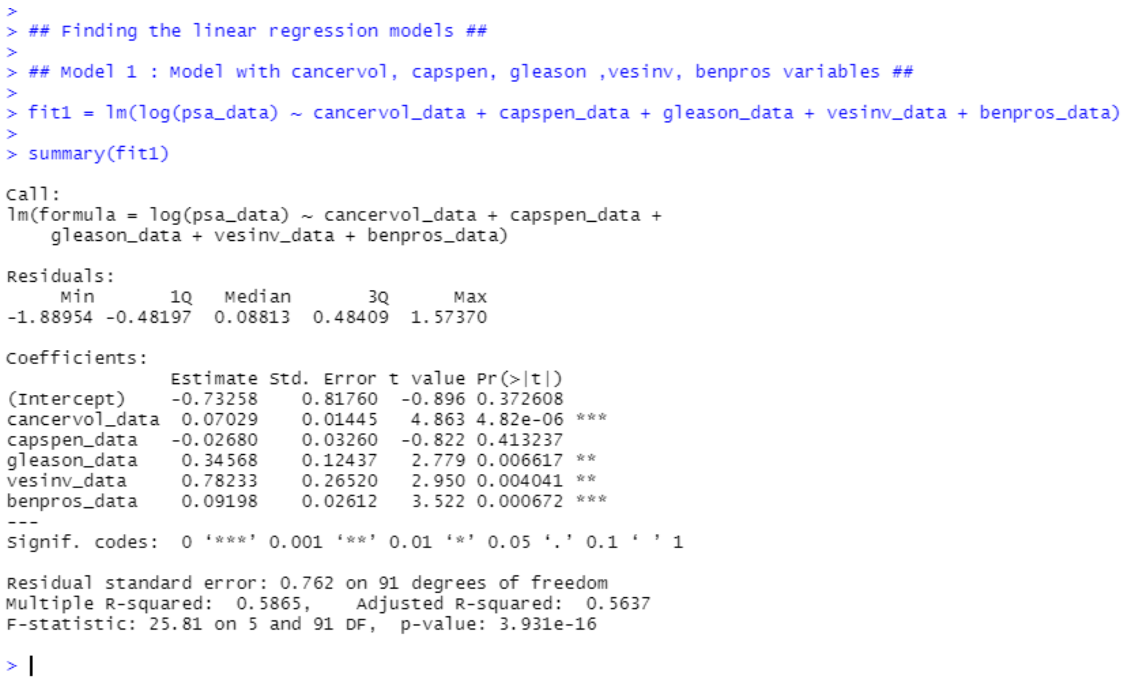
We have seen that the p-value obtained after the t test for the predictors weight_data, age_data , capspen_data is > 0.05. Hence, they are not significant predictors. Since capspen is coming out be in an important variable in the correlation plot, we add that in the next model and eliminate weight and age.

Model 1 :

Predictors used : cancervoldata, capspendata, gleasondata ,vesinv, benspros data.

Null Hypothesis: None of the predictors help predict the PSA level.

Alternative Hypothesis: At least one of the predictors helps predict the PSA level.



```
>
> ## Finding the linear regression models ##
>
> ## Model 1 : Model with cancervol, capspen, gleason ,vesinv, benpros variables ##
>
> fit1 = lm(log(psa_data) ~ cancervol_data + capspen_data + gleason_data + vesinv_data + benpros_data)
>
> summary(fit1)

Call:
lm(formula = log(psa_data) ~ cancervol_data + capspen_data +
    gleason_data + vesinv_data + benpros_data)

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_data  0.07029    0.01445   4.863 4.82e-06 ***
capspen_data   -0.02680    0.03260  -0.822 0.413237
gleason_data    0.34568    0.12437   2.779 0.006617 **
vesinv_data     0.78233    0.26520   2.950 0.004041 **
benpros_data    0.09198    0.02612   3.522 0.000672 ***
---
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

> |
```

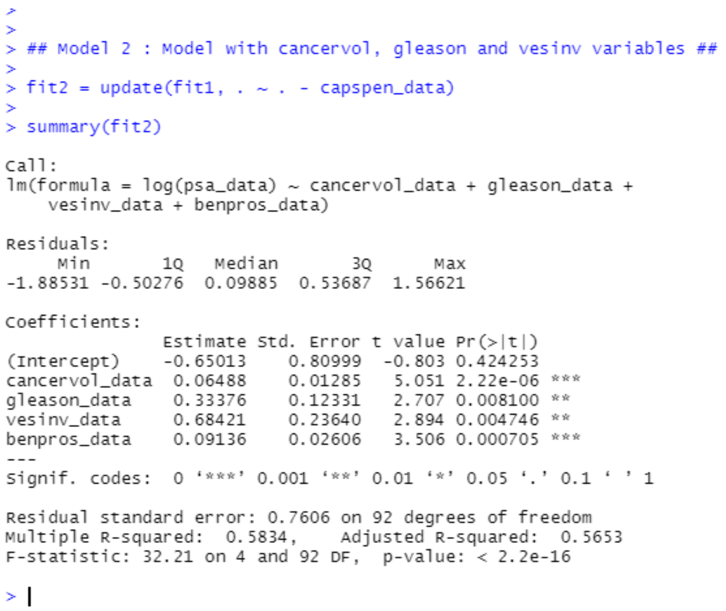
The above results shows that cancervol, gleason, vesinv and benpros are significant predictors of the PSA level because the p-value for these variables is less than 0.05. Therefore, we reject the null hypothesis. We have seen that the p-value obtained after the t test for capspen_data is greater than 0.05. Hence, capspen is not significant predictor. Eliminating capspen in the next model.

Model 2 :

Predictors used : cancervoldata, gleasondata ,vesinv, benspros data.

Null Hypothesis: None of the predictors help predict the PSA level.

Alternative Hypothesis: At least one of the predictors helps predict the PSA level.



```
>
>
> ## Model 2 : Model with cancervol, gleason and vesinv variables ##
>
> fit2 = update(fit1, . ~ . - capspen_data)
>
> summary(fit2)

Call:
lm(formula = log(psa_data) ~ cancervol_data + gleason_data +
    vesinv_data + benpros_data)

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_data  0.06488    0.01285   5.051 2.22e-06 ***
gleason_data    0.33376    0.12331   2.707 0.008100 **
vesinv_data     0.68421    0.23640   2.894 0.004746 **
benpros_data    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

> |
```

The above results shows that cancervol, gleason, benspros and vesinv are significant predictors of the PSA level because the p-value obtained after the t test for these variables is less than 0.05. Therefore, we reject the null hypothesis.

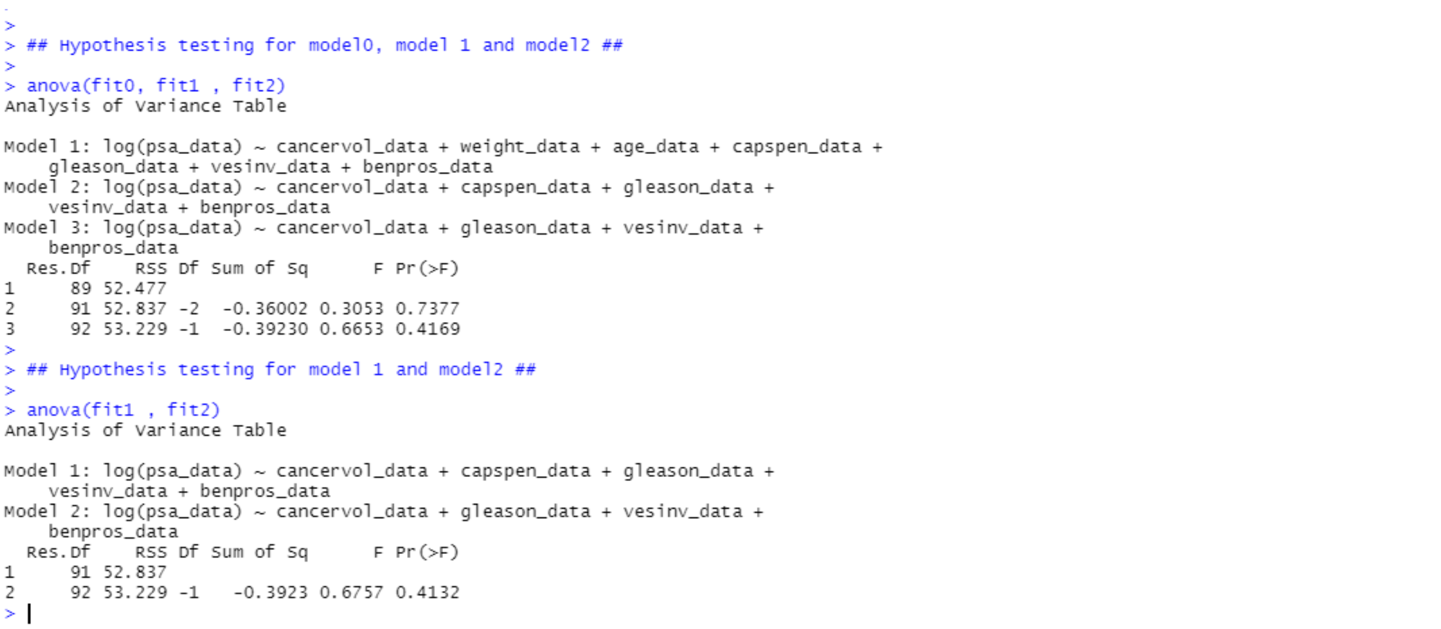
Also, the adjusted r-squared value in model 2 is greater than the value in model 1. This shows that capspen is a bad predictor of the PSA level. Therefore, capspen is a useless predictor.

ANOVA test:

We can perform the hypothesis testing between the model 1 and model 2 in order to make a decision about the final model.

Null hypothesis : The predictor variable capspen_data is useless i.e $ß_{capspen\_data} = 0$

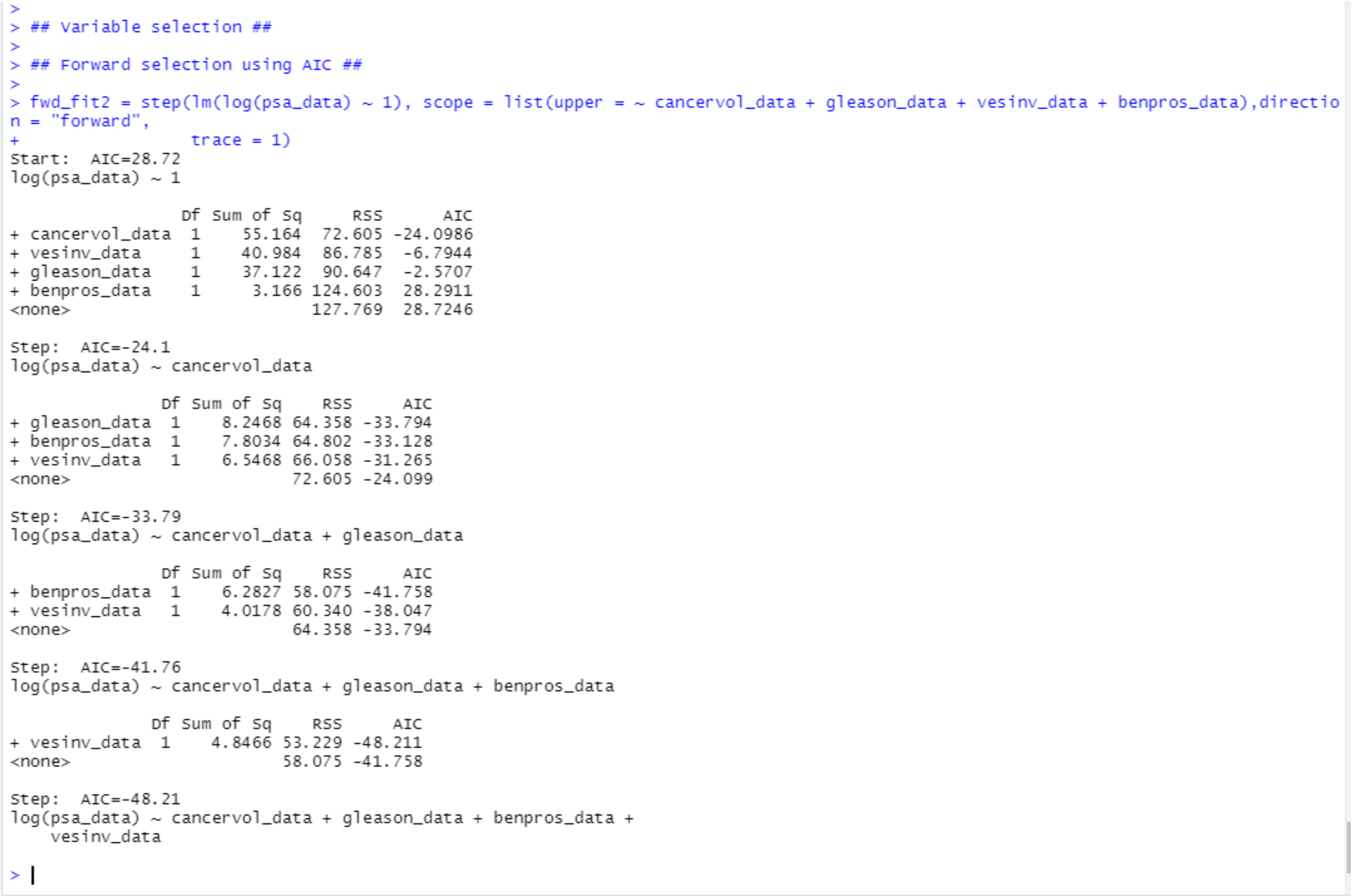
Alternative hypothesis : The predictor variable capspen_data is significant i.e $ß_{capspen\_data} \neq 0$



```
>
> ## Hypothesis testing for model0, model 1 and model2 ##
>
> anova(fit0, fit1 , fit2)
Analysis of Variance Table

Model 1: log(psa_data) ~ cancervol_data + weight_data + age_data + capspen_data +
    gleason_data + vesinv_data + benpros_data
Model 2: log(psa_data) ~ cancervol_data + capspen_data + gleason_data +
    vesinv_data + benpros_data
Model 3: log(psa_data) ~ cancervol_data + gleason_data + vesinv_data +
    benpros_data
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     89 52.477
2     91 52.837 -2  -0.36002 0.3053 0.7377
3     92 53.229 -1  -0.39230 0.6653 0.4169
>
> ## Hypothesis testing for model 1 and model2 ##
>
> anova(fit1 , fit2)
Analysis of Variance Table

Model 1: log(psa_data) ~ cancervol_data + capspen_data + gleason_data +
    vesinv_data + benpros_data
Model 2: log(psa_data) ~ cancervol_data + gleason_data + vesinv_data +
    benpros_data
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     91 52.837
2     92 53.229 -1   -0.3923 0.6757 0.4132
> |
```

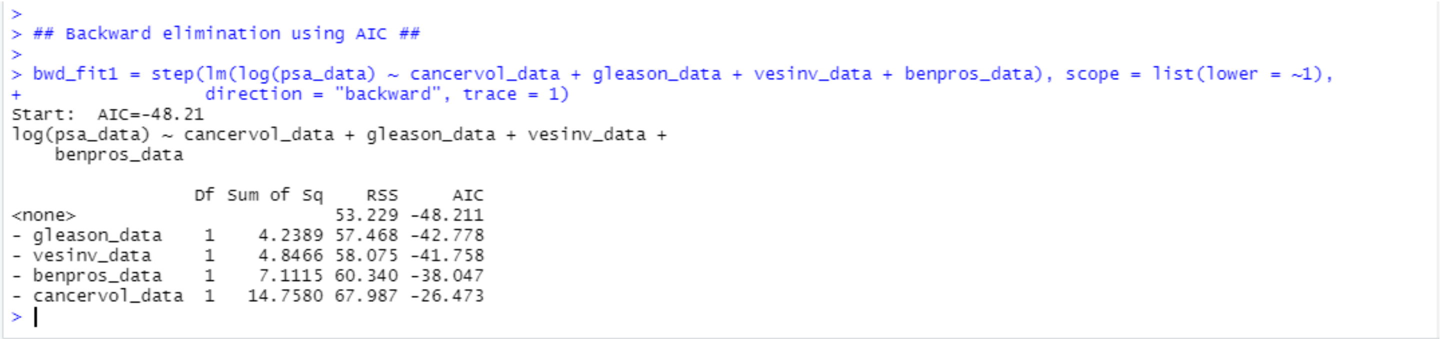
Since the f-statistic is high, i.e > 0.05; null hypothesis is accepted. Therefore the capspen_data is not a significant predictor.

Forward selection using AIC :



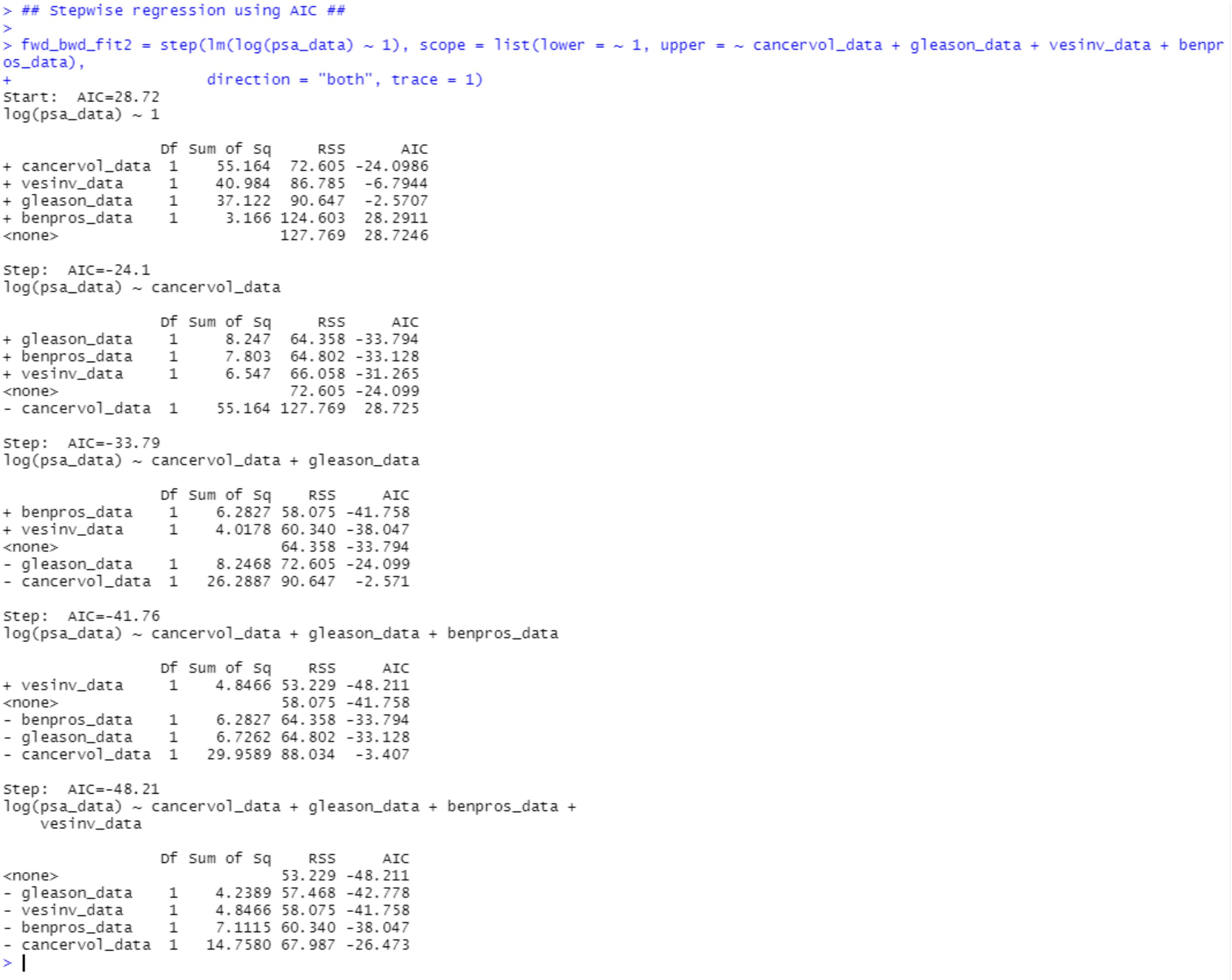
```
>
> ## Variable selection ##
>
> ## Forward selection using AIC ##
>
> fwd_fit2 = step(lm(log(psa_data) ~ 1), scope = list(upper = ~ cancervol_data + gleason_data + vesinv_data + benpros_data),directio
n = "forward",
+                 trace = 1)
Start:  AIC=28.72
log(psa_data) ~ 1

                 Df Sum of Sq     RSS      AIC
+ cancervol_data  1    55.164  72.605 -24.0986
+ vesinv_data     1    40.984  86.785  -6.7944
+ gleason_data    1    37.122  90.647  -2.5707
+ benpros_data    1     3.166 124.603  28.2911
<none>                        127.769  28.7246

Step:  AIC=-24.1
log(psa_data) ~ cancervol_data

               Df Sum of Sq    RSS     AIC
+ gleason_data  1    8.2468 64.358 -33.794
+ benpros_data  1    7.8034 64.802 -33.128
+ vesinv_data   1    6.5468 66.058 -31.265
<none>                      72.605 -24.099

Step:  AIC=-33.79
log(psa_data) ~ cancervol_data + gleason_data

               Df Sum of Sq    RSS     AIC
+ benpros_data  1    6.2827 58.075 -41.758
+ vesinv_data   1    4.0178 60.340 -38.047
<none>                      64.358 -33.794

Step:  AIC=-41.76
log(psa_data) ~ cancervol_data + gleason_data + benpros_data

              Df Sum of Sq    RSS     AIC
+ vesinv_data  1    4.8466 53.229 -48.211
<none>                     58.075 -41.758

Step:  AIC=-48.21
log(psa_data) ~ cancervol_data + gleason_data + benpros_data +
    vesinv_data

> |
```

Backward elimination using AIC :

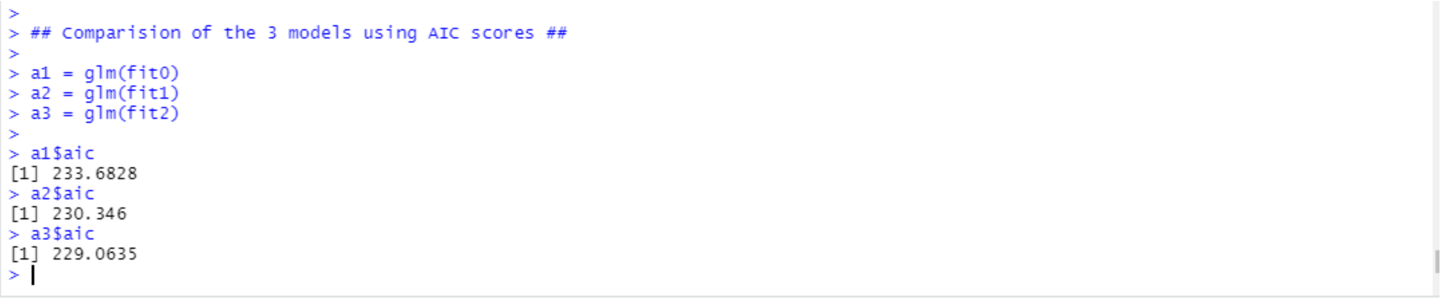


```
>
> ## Backward elimination using AIC ##
>
> bwd_fit1 = step(lm(log(psa_data) ~ cancervol_data + gleason_data + vesinv_data + benpros_data), scope = list(lower = ~1),
+                  direction = "backward", trace = 1)
Start:  AIC=-48.21
log(psa_data) ~ cancervol_data + gleason_data + vesinv_data +
    benpros_data

                 Df Sum of Sq    RSS     AIC
<none>                        53.229 -48.211
- gleason_data    1    4.2389 57.468 -42.778
- vesinv_data     1    4.8466 58.075 -41.758
- benpros_data    1    7.1115 60.340 -38.047
- cancervol_data  1   14.7580 67.987 -26.473
>
```

Stepwise Regression using AIC :



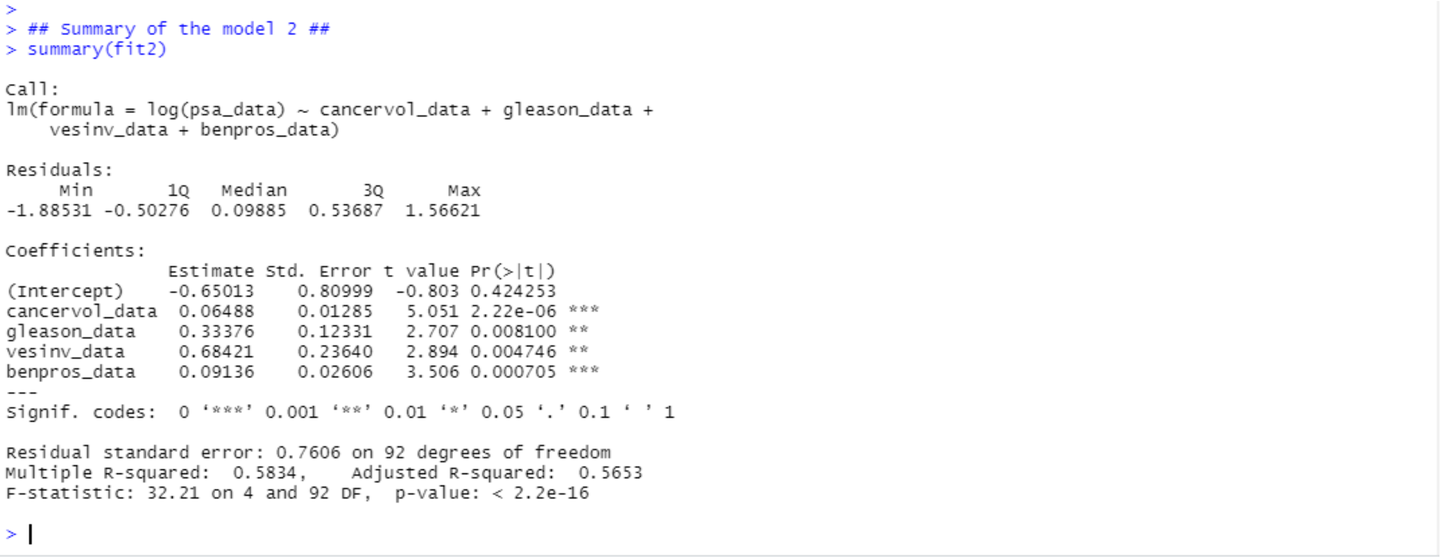
```
> ## Stepwise regression using AIC ##
>
> fwd_bwd_fit2 = step(lm(log(psa_data) ~ 1), scope = list(lower = ~ 1, upper = ~ cancervol_data + gleason_data + vesinv_data + benpr
os_data),
+                      direction = "both", trace = 1)
Start:  AIC=28.72
log(psa_data) ~ 1

                 Df Sum of Sq     RSS      AIC
+ cancervol_data  1    55.164  72.605 -24.0986
+ vesinv_data     1    40.984  86.785  -6.7944
+ gleason_data    1    37.122  90.647  -2.5707
+ benpros_data    1     3.166 124.603  28.2911
<none>                        127.769  28.7246

Step:  AIC=-24.1
log(psa_data) ~ cancervol_data

                 Df Sum of Sq     RSS     AIC
+ gleason_data    1     8.247  64.358 -33.794
+ benpros_data    1     7.803  64.802 -33.128
+ vesinv_data     1     6.547  66.058 -31.265
<none>                         72.605 -24.099
- cancervol_data  1    55.164 127.769  28.725

Step:  AIC=-33.79
log(psa_data) ~ cancervol_data + gleason_data

                 Df Sum of Sq    RSS     AIC
+ benpros_data    1    6.2827 58.075 -41.758
+ vesinv_data     1    4.0178 60.340 -38.047
<none>                        64.358 -33.794
- gleason_data    1    8.2468 72.605 -24.099
- cancervol_data  1   26.2887 90.647  -2.571

Step:  AIC=-41.76
log(psa_data) ~ cancervol_data + gleason_data + benpros_data

                 Df Sum of Sq    RSS     AIC
+ vesinv_data     1    4.8466 53.229 -48.211
<none>                        58.075 -41.758
- benpros_data    1    6.2827 64.358 -33.794
- gleason_data    1    6.7262 64.802 -33.128
- cancervol_data  1   29.9589 88.034  -3.407

Step:  AIC=-48.21
log(psa_data) ~ cancervol_data + gleason_data + benpros_data +
    vesinv_data

                 Df Sum of Sq    RSS     AIC
<none>                        53.229 -48.211
- gleason_data    1    4.2389 57.468 -42.778
- vesinv_data     1    4.8466 58.075 -41.758
- benpros_data    1    7.1115 60.340 -38.047
- cancervol_data  1   14.7580 67.987 -26.473
>
```



```
>
> ## Comparision of the 3 models using AIC scores ##
>
> a1 = glm(fit0)
> a2 = glm(fit1)
> a3 = glm(fit2)
>
> a1$aic
[1] 233.6828
> a2$aic
[1] 230.346
> a3$aic
[1] 229.0635
>
```

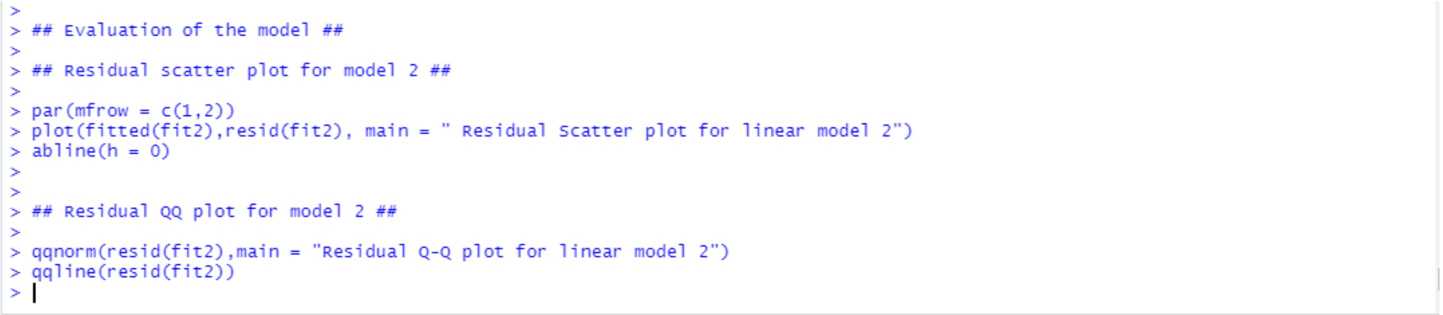
From the above results, a3 or fit2 linear model has the lowest AIC score. Therefore, we could tell us that the fit2 is the best model among all.

Summary :

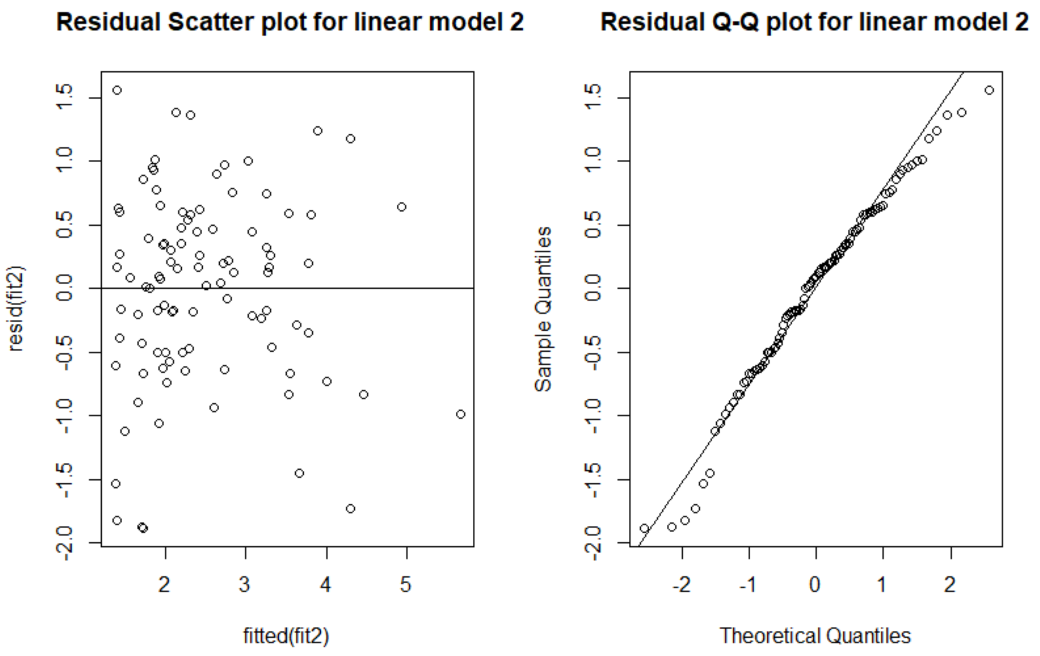


```
>
> ## Summary of the model 2 ##
> summary(fit2)

Call:
lm(formula = log(psa_data) ~ cancervol_data + gleason_data +
    vesinv_data + benpros_data)

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_data  0.06488    0.01285   5.051 2.22e-06 ***
gleason_data    0.33376    0.12331   2.707 0.008100 **
vesinv_data     0.68421    0.23640   2.894 0.004746 **
benpros_data    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 evaluation :

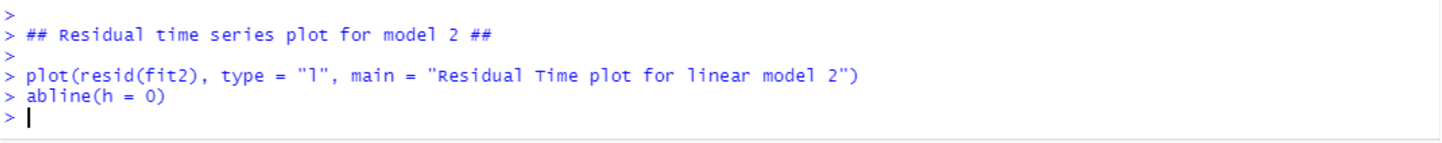


```
>
> ## Evaluation of the model ##
>
> ## Residual scatter plot for model 2 ##
>
> par(mfrow = c(1,2))
> plot(fitted(fit2),resid(fit2), main = " Residual Scatter plot for linear model 2")
> abline(h = 0)
>
>
> ## Residual QQ plot for model 2 ##
>
> qqnorm(resid(fit2),main = "Residual Q-Q plot for linear model 2")
> qqline(resid(fit2))
> |
```

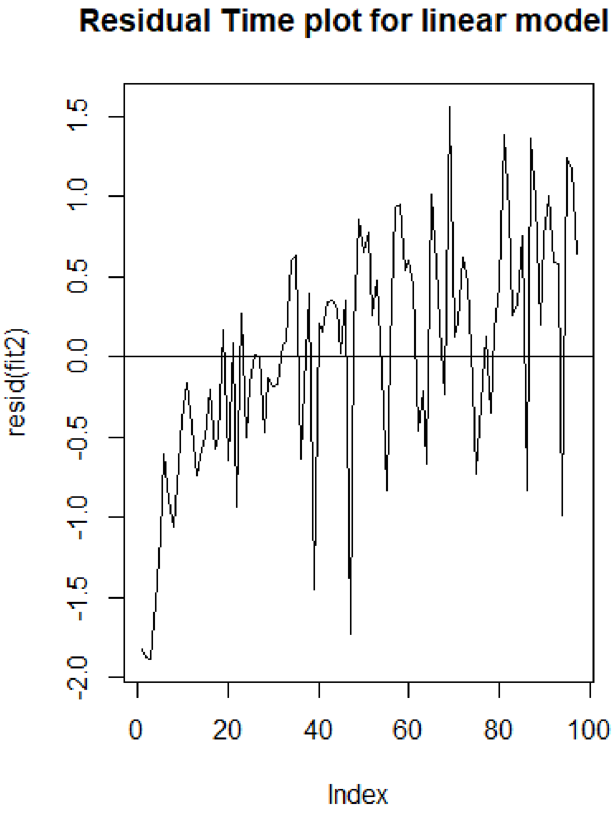


We assumed that residual errors have mean zero and constant variance. The residual scatterplot for linear model 2 shows that the points are scattered around zero. Also, there is no pattern. This verifies that the errors have mean zero and constant variance. This means the standard deviation is constant indicating the linear model is a good estimate.

We assumed that the residual errors are normally distributed. To validate this assumption, we plotted the QQ plot of fitted model. The QQ plot shows that the data is almost normally distributed.



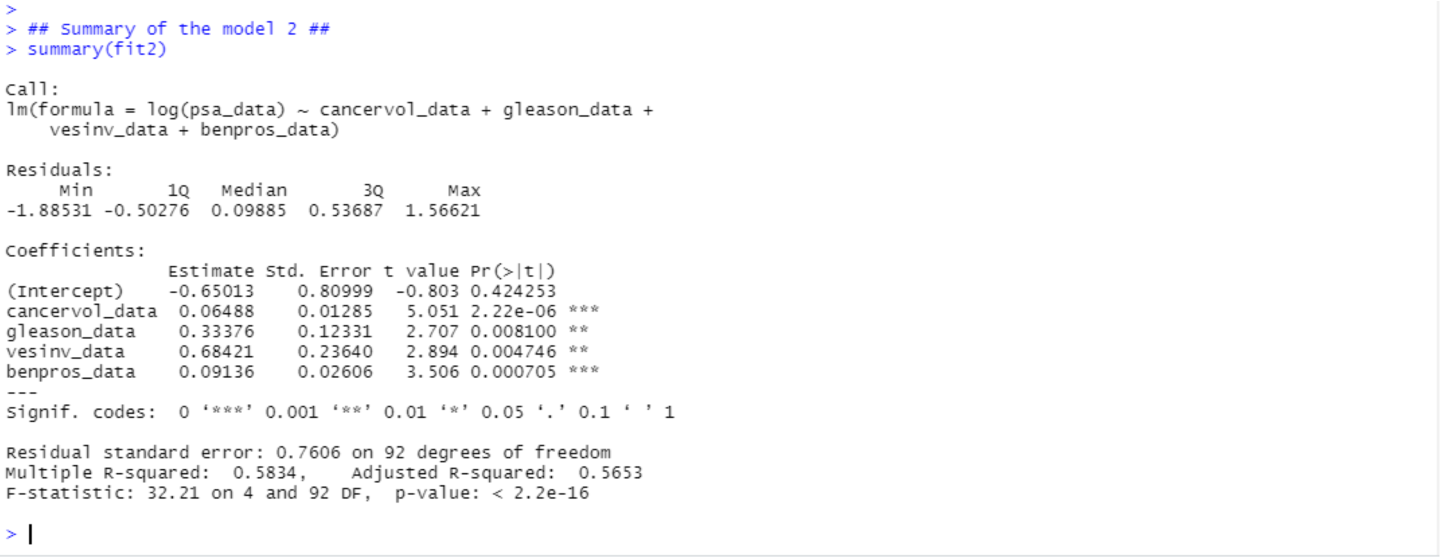
```
>
> ## Residual time series plot for model 2 ##
>
> plot(resid(fit2), type = "l", main = "Residual Time plot for linear model 2")
> abline(h = 0)
>
```



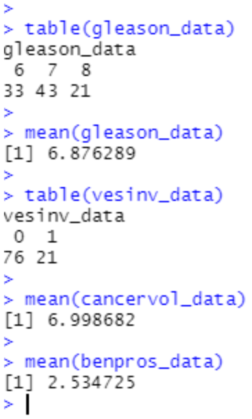
We assumed residual errors are independent. The time series plot doesn't have any dependence, which verifies independence assumption.

We 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.

Predict PSA with the model lm (formula = y ~ cancervolData + vesinvData + gleasonData + benprosData)



```
>
> ## Summary of the model 2 ##
> summary(fit2)

Call:
lm(formula = log(psa_data) ~ cancervol_data + gleason_data +
    vesinv_data + benpros_data)

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_data  0.06488    0.01285   5.051 2.22e-06 ***
gleason_data    0.33376    0.12331   2.707 0.008100 **
vesinv_data     0.68421    0.23640   2.894 0.004746 **
benpros_data    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_data)
gleason_data
 6  7  8
33 43 21
>
> mean(gleason_data)
[1] 6.876289
>
> table(vesinv_data)
vesinv_data
 0  1
76 21
>
> mean(cancervol_data)
[1] 6.998682
>
> mean(benpros_data)
[1] 2.534725
>
```

From the above results:

Means for cancervol ,benpros and gleason_data are 6.998682 ,2.534725 and 6.876289 respectively

Most frequent categories for gleason and vesinv are 7 and 0 respectively.

Predicting the PSA level using linear model 2:

- -0.65013 + 6.998682*(0.06488) + 6.876289 *(0.33376) + 0 * (0.68421) + 2.534725 * (0.09136)
  = -0.65013 + 0.45407 + 2.29503 + 0 + 0.231572
  = 2.330542

Predicted PSA level :

Y = log(PSA_value) = 2.3305

Psa_level = exp(2.37183) = 10.28308

Hence, the actual value of PSA level is 10.28308

R code :

```
### Question 1 ###
## Reading data from the given csv file. ##
cancer_data = read.csv("C:\\Users\\hxt210018\\Downloads\\6313_Prob\\HW\\prostate_cancer.csv")
## Reading the column data of psa values ##
psa_data = cancer_data[['psa']]
##Histogram plot of the psa_data ##
hist(psa_data, xlab = "PSA value", main = "Histogram of PSA data",breaks = 20)
## Boxplot for the psa_data ##
par(mfrow = c(1,2))
boxplot(psa_data, main = "Boxplot of the PSA data")
boxplot(log(psa_data), main = "Boxplot of the PSA log data")
## Summary stats of psa_data and log(psa_data) ##
summary(psa_data)
summary(log(psa_data))
## QQ-plot of the PSA data ##
par(mfrow = c(1,1))
qqnorm(psa_data)
qqline(psa_data)
## Reading the other data given in the csv file ##
cancervol_data = cancer_data[['cancervol']]
weight_data = cancer_data[['weight']]
age_data = cancer_data[['age']]
benpros_data = cancer_data[['benpros']]
vesinv_data = cancer_data[['vesinv']]
capspen_data = cancer_data[['capspen']]
gleason_data = cancer_data[['gleason']]
## Histogram analysis of the other quantitative data ##
par(mfrow = c(3,3))
hist(cancervol_data, xlab = "CancerVol value", main = "Histogram of cancervol data",breaks = 20)
hist(weight_data, xlab = "Weight value", main = "Histogram of Weight data",breaks = 20)
hist(age_data, xlab = "Age value", main = "Histogram of Age data",breaks = 20)
hist(benpros_data, xlab = "benpros value", main = "Histogram of Benpros data",breaks = 20)
hist(vesinv_data, xlab = "vesinv value", main = "Histogram of vesinv data",breaks = 20)

hist(capspen_data, xlab = "capspen value", main = "Histogram of capspen data",breaks = 20)

hist(gleason_data, xlab = "gleason value", main = "Histogram of gleason data",breaks = 20)

## Correlation between data variables ##

install.packages("corrplot")

library(corrplot)

cor.data = cor(cancer_data)

corrplot(cor.data)

## Plotting scatterplots for all the quantitative data ##

pairs(~psa + cancervol + weight + age + benpros + capspen + gleason, data = cancer_data)

## Correlation matrix between the variables ##

cancer_data_cor = cor(cancer_data[, 2:9])

round(cancer_data_cor, 6)

## Plotting scatterplots for log of the quantitative data ##

pairs(~psa + cancervol + capspen + gleason, data = log(cancer_data))

## Correlation matrix between the psa log data and other variables ##

cancer_data_cor = cor(cancer_data[, 3:9], log(cancer_data['psa']))

round(cancer_data_cor, 6)

## Plotting graphs for determining the correlation between PSA log data and other variables ##

par(mfrow = c(3,3))

## Plot between the cancervol data and the log of psa data ##

plot(cancervol_data, log(psa_data), xlab = "CancerVol data", ylab = "log of PSA data", main = " Cancervol_data Vs log_psa_data")

abline(lm(log(psa_data) ~ cancervol_data))

## Plot between the capspen data and the log of psa data ##

plot(capspen_data, log(psa_data), xlab = "Capspen data", ylab = "log of PSA data", main = " Capspen_data Vs log_psa_data")

abline(lm(log(psa_data) ~ capspen_data))

## Plot between the gleason data and the log of psa data ##

plot(gleason_data, log(psa_data), xlab = "gleason data", ylab = "log of PSA data", main = " gleason_data Vs log_psa_data")

abline(lm(log(psa_data) ~ gleason_data))

## Plot between the vesinv data and the log of psa data ##
```

```
plot(vesinv_data, log(psa_data), xlab = "vesinv data", ylab = "log of PSA data", main = " vesinv_data Vs log_psa_data")

abline(lm(log(psa_data) ~ vesinv_data))

## Plot between the benpros data and the log of psa data ##

plot(gleason_data, log(psa_data), xlab = "benpros data", ylab = "log of PSA data", main = " benpros_data Vs log_psa_data")

abline(lm(log(psa_data) ~ benpros_data))

## Finding the linear regression models ##

## Model 0 : Model with all the given quantitative variables ##

fit0 = lm(log(psa_data) ~ cancervol_data + weight_data + age_data + capspen_data + gleason_data + vesinv_data + benpros_data)

summary(fit0)

## Model 1 : Model with cancervol, capspen, gleason ,vesinv, benpros variables ##

fit1 = lm(log(psa_data) ~ cancervol_data + capspen_data + gleason_data + vesinv_data + benpros_data)

summary(fit1)

## Model 2 : Model with cancervol, gleason and vesinv variables ##

fit2 = update(fit1, . ~ . - capspen_data)

summary(fit2)

## Hypothesis testing for model0, model 1 and model2 ##

anova(fit0, fit1 , fit2)

## Hypothesis testing for model 1 and model2 ##

anova(fit1 , fit2)

## Variable selection ##

## Forward selection using AIC ##

fwd_fit2 = step(lm(log(psa_data) ~ 1), scope = list(upper = ~ cancervol_data + gleason_data + vesinv_data + benpros_data),direction = "forward",trace = 1)

## Backward elimination using AIC ##

bwd_fit1 = step(lm(log(psa_data) ~ cancervol_data + gleason_data + vesinv_data + benpros_data), scope = list(lower = ~1), direction = "backward", trace = 1)

## Stepwise regression using AIC ##

fwd_bwd_fit2 = step(lm(log(psa_data) ~ 1), scope = list(lower = ~ 1, upper = ~ cancervol_data + gleason_data + vesinv_data + benpros_data), direction = "both", trace = 1)

## Comparision of the 3 models using AIC scores ##

a1 = glm(fit0)

a2 = glm(fit1)
```

```
a3 = glm(fit2)
a1$aic
a2$aic
a3$aic
## Summary of the model 2 ##
summary(fit2)
## Evaluation of the model ##
## Residual scatter plot for model 2 ##
par(mfrow = c(1,2))
plot(fitted(fit2),resid(fit2), main = " Residual Scatter plot for linear model 2")
abline(h = 0)
## Residual QQ plot for model 2 ##
qqnorm(resid(fit2),main = "Residual Q-Q plot for linear model 2")
qqline(resid(fit2))
## Residual time series plot for model 2 ##
plot(resid(fit2), type = "l", main = "Residual Time plot for linear model 2")
abline(h = 0)
table(gleason_data)
table(vesinv_data)
mean(cancervol_data)
mean(benpros_data)
```