CS-6313 – Statistical Methods for Data Science

Mini Project #6

Group No - 5
Manan Dalal (MUD200000)
Lipi Patel (LDP210000)

Contribution of Team Members:

We both collaborated and solved the questions together for a thorough understanding of functions in R and solved the question concurrently to check for accuracy, debugging and application.

Question 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

Solution

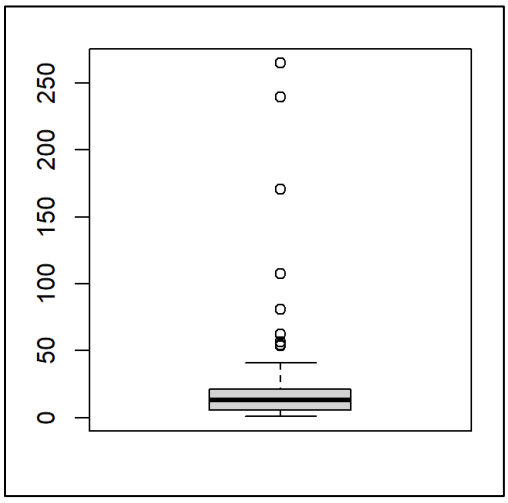
- ⇨ First, we read the csv file containing the prostate cancer data.
- ⇨ Then, we study the different columns of the data and summarize the data.
- ⇨ Then, we analyze the correlation between attributes.
- ⇨ We then create a boxplot to analyze the PSA attribute.
- ⇨ We also transform the PSA attribute using natural logarithm function.
- ⇨ Then, we check the distribution of age and PSA column by plotting a graph.

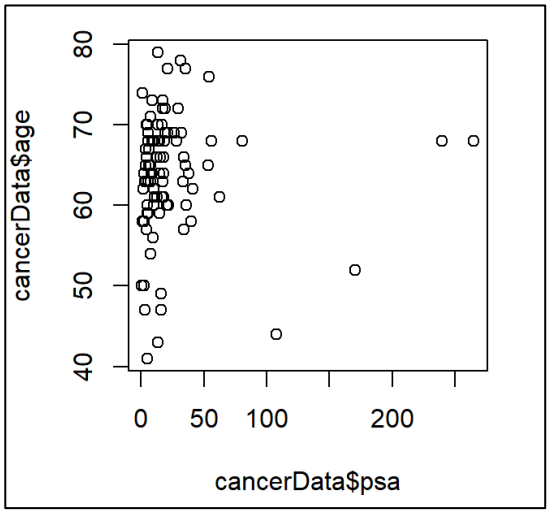
```r
# Reading the file
cancerData = read.csv("E:/MS-CS/Spring 22/CS6313 - SMDS/Mini Projects/6/prostate_cancer.csv")

# Getting names of all columns
names = colnames(cancerData)

names
summary(cancerData)
cor(cancerData)

# Using log to scale PSA column
logPSA = log(cancerData$psa)

# Visualization of the data
boxplot(cancerData$psa)

# Checking the distributions of age and psa
plot(cancerData$psa,cancerData$age)
```

```
> names
[1] "subject"   "psa"       "cancervol" "weight"    "age"       "benpros"   "vesinv"    "capspen"
[9] "gleason"
```

```
> summary(cancerData)
    subject          psa             cancervol         weight           age       
 Min.   : 1   Min.   :  0.651   Min.   : 0.2592   Min.   : 10.70   Min.   :41.00  
 1st Qu.:25   1st Qu.:  5.641   1st Qu.: 1.6653   1st Qu.: 29.37   1st Qu.:60.00  
 Median :49   Median : 13.330   Median : 4.2631   Median : 37.34   Median :65.00  
 Mean   :49   Mean   : 23.730   Mean   : 6.9987   Mean   : 45.49   Mean   :63.87  
 3rd Qu.:73   3rd Qu.: 21.328   3rd Qu.: 8.4149   3rd Qu.: 48.42   3rd Qu.:68.00  
 Max.   :97   Max.   :265.072   Max.   :45.6042   Max.   :450.34   Max.   :79.00  
    benpros          vesinv           capspen          gleason     
 Min.   : 0.000   Min.   :0.0000   Min.   : 0.0000   Min.   :6.000  
 1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:6.000  
 Median : 1.350   Median :0.0000   Median : 0.4493   Median :7.000  
 Mean   : 2.535   Mean   :0.2165   Mean   : 2.2454   Mean   :6.876  
 3rd Qu.: 4.759   3rd Qu.:0.0000   3rd Qu.: 3.2544   3rd Qu.:7.000  
 Max.   :10.278   Max.   :1.0000   Max.   :18.1741   Max.   :8.000  
```

```
> cor(cancerData)
              subject         psa    cancervol       weight        age     benpros
subject    1.0000000  0.60268375  0.620997842  0.113741022 0.19655569  0.16500536
psa        0.6026837  1.00000000  0.624150588  0.026213430 0.01719938 -0.01648649
cancervol  0.6209978  0.62415059  1.000000000  0.005107148 0.03909442 -0.13320943
weight     0.1137410  0.02621343  0.005107148  1.000000000 0.16432371  0.32184875
age        0.1965557  0.01719938  0.039094423  0.164323714 1.00000000  0.36634121
benpros    0.1650054 -0.01648649 -0.133209431  0.321848748 0.36634121  1.00000000
vesinv     0.5667803  0.52861878  0.581741687 -0.002410475 0.11765804 -0.11955319
capspen    0.4767525  0.55079252  0.692896688  0.001578905 0.09955535 -0.08300865
gleason    0.5379241  0.42957975  0.481438397 -0.024206925 0.22585181  0.02682555
               vesinv      capspen     gleason
subject    0.566780347  0.476752459  0.53792405
psa        0.528618785  0.550792517  0.42957975
cancervol  0.581741687  0.692896688  0.48143840
weight    -0.002410475  0.001578905 -0.02420693
age        0.117658038  0.099555351  0.22585181
benpros   -0.119553192 -0.083008649  0.02682555
vesinv     1.000000000  0.680284092  0.42857348
capspen    0.680284092  1.000000000  0.46156590
gleason    0.428573479  0.461565896  1.00000000
```



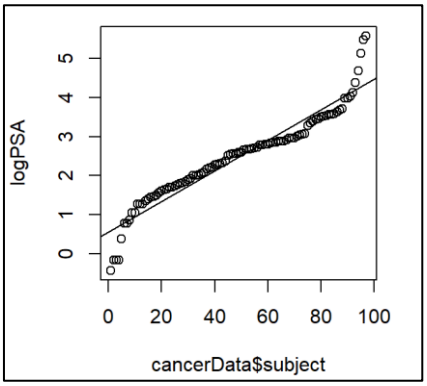


- ⇨ Now, we will compare all the predictors with PSA to find out how they are related.
- ⇨ To do this, we will build univariate models for all the predictors.
- ⇨ Following are the models we built for all the various predictors.

Predictor: ***subject***

```
20  # finding relationship between the predictor 'subject' and PSA
21  plot(cancerData$subject, logPSA)
22  fitSubject <- lm(logPSA ~ cancerData$subject, data = cancerData)
23  abline(fitSubject)
24  summary(fitSubject)
```

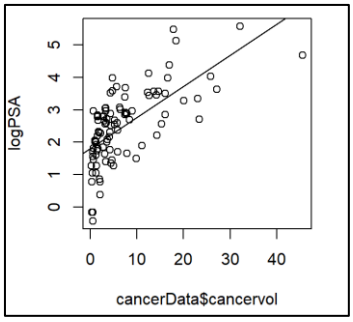
```
lm(formula = logPSA ~ cancerData$subject, data = cancerData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02284 -0.19903  0.07208  0.18334  1.21626

Coefficients:
                    Estimate Std. Error t value
(Intercept)         0.554321   0.067966   8.156
cancerData$subject 0.039272   0.001204  32.610
                    Pr(>|t|)
(Intercept)         1.41e-12 ***
cancerData$subject  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3321 on 95 degrees of freedom
Multiple R-squared:  0.918,     Adjusted R-squared:  0.9171
F-statistic:  1063 on 1 and 95 DF,  p-value: < 2.2e-16
```



Predictor: ***cacervol***

```
# finding relationship between the predictor 'cancervol' and PSA
plot(cancerData$cancervol, logPSA)
fitCancervol <- lm(logPSA ~ cancerData$cancervol, data = cancerData)
abline(fitCancervol)
summary(fitCancervol)
```

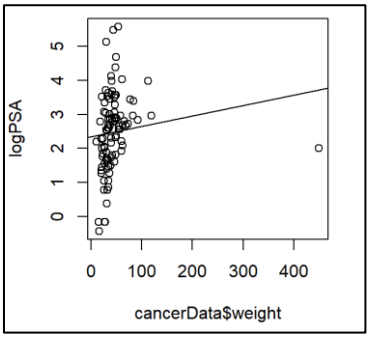
```
lm(formula = logPSA ~ cancerData$cancervol, data = cancerData)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2886 -0.6590  0.1493  0.5769  1.9610

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.80549    0.11899  15.174  < 2e-16 ***
cancerData$cancervol  0.09619    0.01132   8.496 2.69e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8742 on 95 degrees of freedom
Multiple R-squared:  0.4317,    Adjusted R-squared:  0.4258
F-statistic: 72.18 on 1 and 95 DF,  p-value: 2.688e-13
```



Predictor: **weight**

```
# finding relationship between the predictor 'weight' and PSA
plot(cancerData$weight, logPSA)
fitWeight <- lm(logPSA ~ cancerData$weight, data = cancerData)
abline(fitWeight)
summary(fitWeight)
```

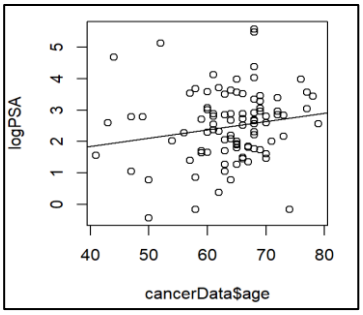
```
lm(formula = logPSA ~ cancerData$weight, data = cancerData)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8172 -0.7291  0.1300  0.6144  3.0783

Coefficients:
                   Estimate Std. Error t value
(Intercept)       2.338901   0.165328  14.147
cancerData$weight 0.003072   0.002570   1.195
                  Pr(>|t|)
(Intercept)         <2e-16 ***
cancerData$weight    0.235
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.151 on 95 degrees of freedom
Multiple R-squared:  0.01482,   Adjusted R-squared:  0.004446
F-statistic: 1.429 on 1 and 95 DF,  p-value: 0.235
```



Predictor: ***age***

```
# finding relationship between the predictor 'age' and PSA
plot(cancerData$age, logPSA)
fitAge <- lm(logPSA ~ cancerData$age, data = cancerData)
abline(fitAge)
summary(fitAge)
```

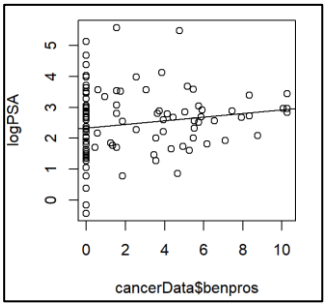
```
lm(formula = logPSA ~ cancerData$age, data = cancerData)

Residuals:
     Min       1Q   Median       3Q      Max
-2.90564 -0.71115  0.07247  0.66617  2.99249

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.79721    1.00729   0.791   0.4307
cancerData$age  0.02633    0.01567   1.680   0.0961 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.143 on 95 degrees of freedom
Multiple R-squared:  0.02887,   Adjusted R-squared:  0.01865
F-statistic: 2.824 on 1 and 95 DF,  p-value: 0.09615
```



Predictor: ***benpros***

```
# finding relationship between the predictor 'benpros' and PSA
plot(cancerData$benpros, logPSA)
fitBenpros <- lm(logPSA ~ cancerData$benpros, data = cancerData)
abline(fitBenpros)
summary(fitBenpros)
```

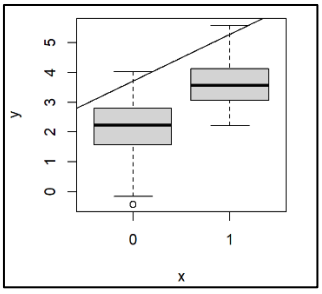
```
lm(formula = logPSA ~ cancerData$benpros, data = cancerData)

Residuals:
     Min       1Q   Median       3Q      Max
-2.75607 -0.76149 -0.01686  0.63318  3.16016

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.32682    0.15191  15.317   <2e-16 ***
cancerData$benpros  0.05991    0.03856   1.554    0.124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.145 on 95 degrees of freedom
Multiple R-squared:  0.02478,   Adjusted R-squared:  0.01451
F-statistic: 2.413 on 1 and 95 DF,  p-value: 0.1236
```



Predictor: ***vesinv***

```
# finding relationship between the predictor 'vesinv' and PSA
vesinv = factor(cancerData$vesinv)
plot(vesinv, logPSA)
fitVesinv <- lm(logPSA ~ vesinv, data = cancerData)
abline(fitVesinv)
summary(fitVesinv)
```

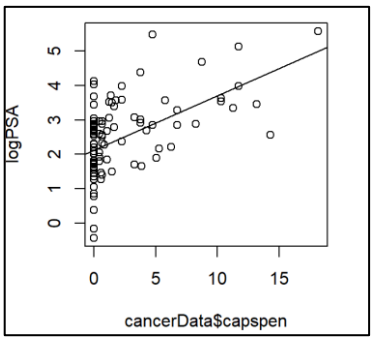
```
lm(formula = logPSA ~ vesinv, data = cancerData)

Residuals:
     Min       1Q   Median       3Q      Max
-2.56623 -0.63526 -0.00524  0.67302  1.89302

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1370     0.1096  19.492  < 2e-16 ***
vesinv        1.5783     0.2356   6.698 1.48e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9558 on 95 degrees of freedom
Multiple R-squared:  0.3208,    Adjusted R-squared:  0.3136
F-statistic: 44.86 on 1 and 95 DF,  p-value: 1.481e-09
```



Predictor: **capspen**

```
50  # finding relationship between the predictor 'capspen' and PSA
51  plot(cancerData$capspen, logPSA)
52  fitCapspen <- lm(logPSA ~ cancerData$capspen, data = cancerData)
53  abline(fitCapspen)
54  summary(fitCapspen)
```

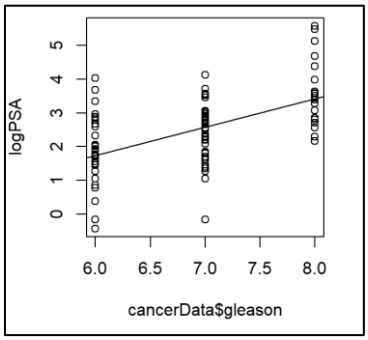
```
lm(formula = logPSA ~ cancerData$capspen, data = cancerData)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5532 -0.6740  0.0071  0.6660  2.6043

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.12399    0.11728  18.110  < 2e-16 ***
cancerData$capspen  0.15796    0.02676   5.903  5.5e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.992 on 95 degrees of freedom
Multiple R-squared:  0.2683,    Adjusted R-squared:  0.2606
F-statistic: 34.84 on 1 and 95 DF,  p-value: 5.503e-08
```



Predictor: **gleason**

```
56  # finding relationship between the predictor 'gleason' and PSA
57  plot(cancerData$gleason, logPSA)
58  fitGleason <- lm(logPSA ~ cancerData$gleason, data = cancerData)
59  abline(fitGleason)
60  summary(fitGleason)
```

```
lm(formula = logPSA ~ cancerData$gleason, data = cancerData)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7428 -0.6134  0.0773  0.4773  2.2881

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3.3026     0.9322  -3.543 0.000616 ***
cancerData$gleason   0.8408     0.1348   6.237 1.23e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9768 on 95 degrees of freedom
Multiple R-squared:  0.2905,    Adjusted R-squared:  0.2831
F-statistic:  38.9 on 1 and 95 DF,  p-value: 1.228e-08
```



- ⇨ Upon Observing the models above, we can see that cancervol, gleason, vesinv, benepros and capspan are significant.
- ⇨ These predictors show an evident linear relation with PSA.
- ⇨ We will now use various combinations of these predictors to predict our PSA.

```
# Creating various models by combining multiple significant predictors
fit1 = lm(logPSA ~ cancerData$cancervol + cancerData$gleason + factor(cancerData$vesinv)
          + cancerData$capspen , data = cancerData)
summary(fit1)
```

```
lm(formula = logPSA ~ cancerData$cancervol + cancerData$gleason +
    factor(cancerData$vesinv) + cancerData$capspen, data = cancerData)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1747 -0.4497  0.1049  0.6215  1.6135

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -0.79386    0.86660  -0.916  0.36203
cancerData$cancervol        0.06452    0.01522   4.238 5.35e-05 ***
cancerData$gleason          0.39566    0.13100   3.020  0.00327 **
factor(cancerData$vesinv)1  0.70675    0.28024   2.522  0.01339 *
cancerData$capspen         -0.02348    0.03455  -0.680  0.49852
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8078 on 92 degrees of freedom
Multiple R-squared:  0.5301,    Adjusted R-squared:  0.5097
F-statistic: 25.95 on 4 and 92 DF,  p-value: 2.075e-14
```

- ⇨ Then, we remove capspen from the model and compare the 2 models to check if capspen is significant or not.

```
# Removing capspen
fit2 = lm(logPSA ~ cancerData$cancervol + cancerData$gleason + factor(cancerData$vesinv),
          data = cancerData)
summary(fit2)

# Comparing the two models
anova(fit2,fit1)
```

```
lm(formula = logPSA ~ cancerData$cancervol + cancerData$gleason +
    factor(cancerData$vesinv), data = cancerData)

Residuals:
     Min       1Q   Median       3Q      Max
-2.16928 -0.44558  0.08431  0.60719  1.64082

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -0.72120    0.85749  -0.841   0.4025
cancerData$cancervol        0.05981    0.01352   4.425 2.62e-05 ***
cancerData$gleason          0.38491    0.12966   2.969   0.0038 **
factor(cancerData$vesinv)1  0.62117    0.24962   2.488   0.0146 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8055 on 93 degrees of freedom
Multiple R-squared:  0.5277,    Adjusted R-squared:  0.5125
F-statistic: 34.64 on 3 and 93 DF,  p-value: 4.022e-15
```

```
Analysis of Variance Table

Model 1: logPSA ~ cancerData$cancervol + cancerData$gleason + factor(cancerData$vesinv)
Model 2: logPSA ~ cancerData$cancervol + cancerData$gleason + factor(cancerData$vesinv) +
    cancerData$capspen
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     93 60.340
2     92 60.039  1   0.30134 0.4617 0.4985
```

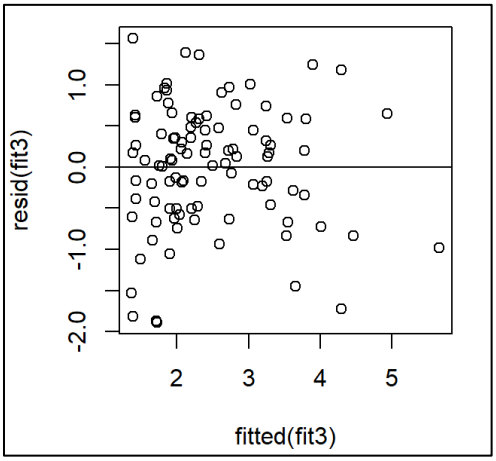
- ⇨ By looking at the above output we can say that capspen is not a significant predictor.
- ⇨ Thus, in the next model we add all other predictors and ignore capspen.

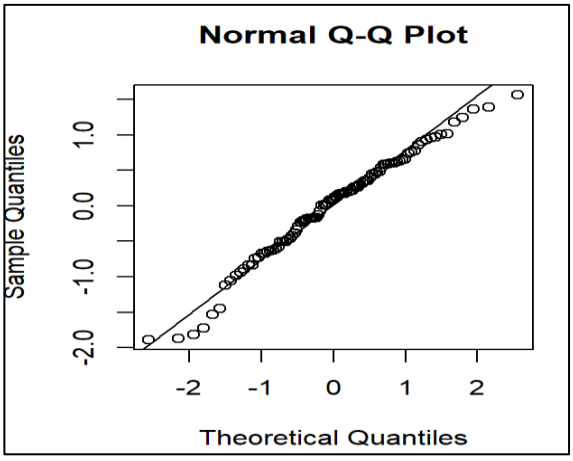
```
82  # Creating model with all predictors
83  fit3 = lm(logPSA ~ cancerData$cancervol + factor(cancerData$vesinv) + cancerData$benpros +
84              cancerData$gleason, data = cancerData )
85  summary(fit3)
```

```
lm(formula = logPSA ~ cancerData$cancervol + factor(cancerData$vesinv) +
    cancerData$benpros + cancerData$gleason, data = cancerData)

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
cancerData$cancervol       0.06488    0.01285   5.051 2.22e-06 ***
factor(cancerData$vesinv)1 0.68421    0.23640   2.894 0.004746 **
cancerData$benpros         0.09136    0.02606   3.506 0.000705 ***
cancerData$gleason         0.33376    0.12331   2.707 0.008100 **
---
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 above summary we can observe minimum residual standard error, thus so far this is the best model.
- ⇨ We will now create a residual and a QQ plot for this model.

```
87  # Creating residual plot for fit3
88  plot(fitted(fit3), resid(fit3))
89  abline(h = 0)
90
91  # Creating QQ plot for fit3
92  qqnorm(resid(fit3))
93  qqline(resid(fit3))
```





Our Assumptions:

• Errors are centered around zero with constant variance, this can be seen from the residual chart thus this is verified.

• Errors are also normally distributed as seen by QQ plot where QQ line fits very well. Thus, this assumption is also verified well.

⇨ Now, we shall use this final model to predict the PSA level of a patient.

⇨ We will consider the sample means of all the quantitative predictors and the highest frequency count for the qualtitative predictor.

```
95  # Using fit3 to predict desired output.
96  # Computing means of all quantitative predictors
97  meanCV = mean(cancerData$cancervol)
98  meanGL = mean(cancerData$gleason)
99  meanBP = mean(cancerData$benpros)
100
101 # Computing frequency count of all qualitative predictors
102 mfVesinv = names(which.max(table(factor(cancerData$vesinv))))
```

```
> meanCV
[1] 6.998682
> meanBP
[1] 2.534725
> meanGL
[1] 6.876289
> mfVesinv
[1] "0"
```

⇨ From the below summary of fit3, we can see the values of beta0, beta1, beta2, beta3, beta4 are -0.65013, 0.06488, 0.68421, 0.09136, 0.33376 respectively.

```
lm(formula = logPSA ~ cancerData$cancervol + factor(cancerData$vesinv) +
    cancerData$benpros + cancerData$gleason, data = cancerData)

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
cancerData$cancervol       0.06488    0.01285   5.051 2.22e-06 ***
factor(cancerData$vesinv)1 0.68421    0.23640   2.894 0.004746 **
cancerData$benpros         0.09136    0.02606   3.506 0.000705 ***
cancerData$gleason         0.33376    0.12331   2.707 0.008100 **
---
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 use these values to predict the PSA value.

```
104 # Computing the PSA value
105 beta0 = -0.65013
106 beta1 = 0.06488
107 beta2 = 0.68421
108 beta3 = 0.09136
109 beta4 = 0.33376
110
111 predAns = exp(beta0 + beta1*meanCV + beta2*0 + beta3*meanBP + beta4*meanGL)
```

```
> predAns
[1] 10.28357
```

⇨ Thus, the final value of PSA level of a patient comes out to be 10.28357 using our best fit model.