Question 1 :

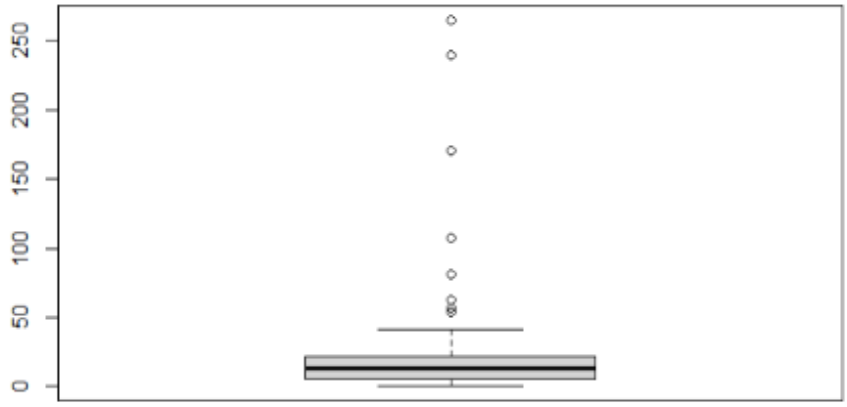
1. Consider the prostate cancer dataset available on eLearning as prostate cancer.csv. It consists of data on 97 men with advanced prostate cancer. A description of the variables is given in Figure 1. We would like to understand how PSA level is related to the other predictors in the dataset. Note that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

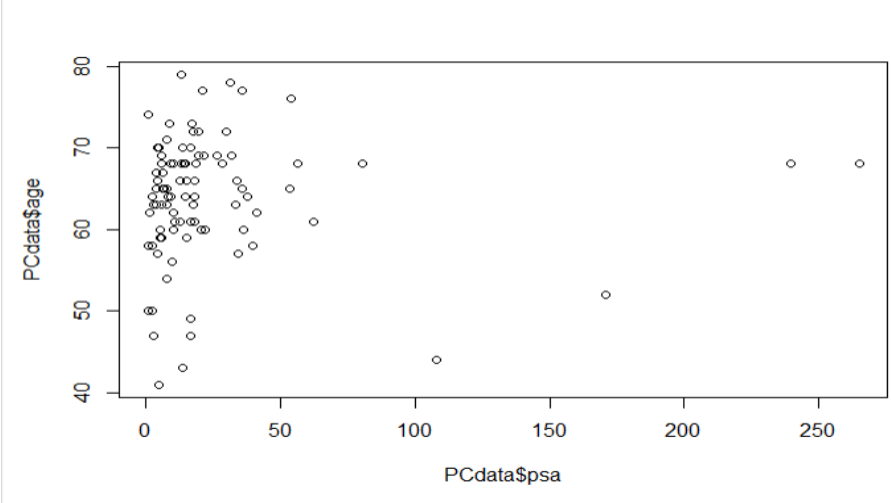
Section 1

Answers to the specific questions asked

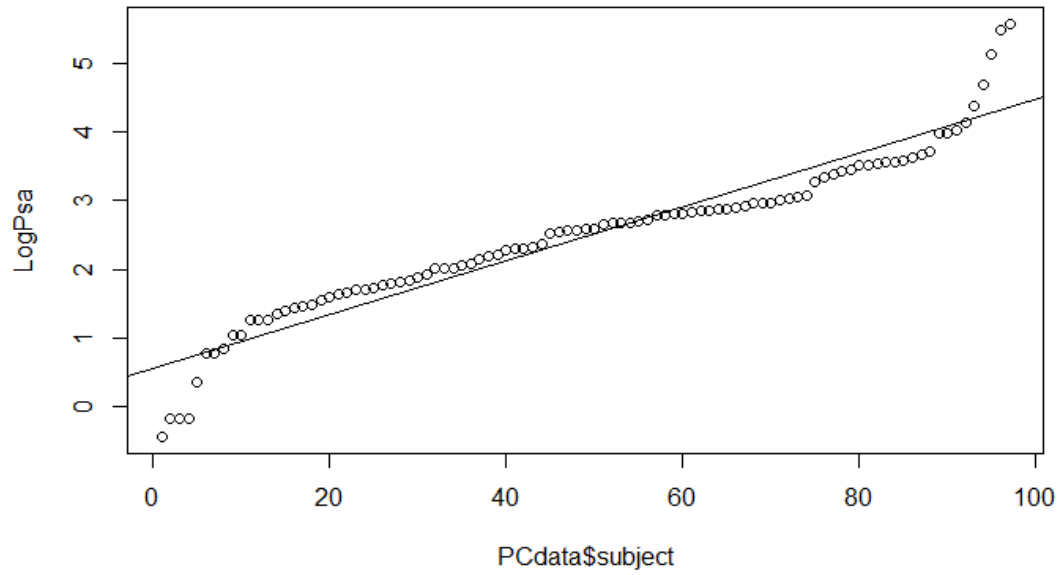
- We read the prostate cancer data set and examined the summary and correlation between the features.
- Compared PSA level with each and every predictor after building a linear model and also found a summary of each.
- Built a model by selecting all the significant predictors which were obtained from the univariate model we developed.
- Altered model by not including some predictors to get significant predictors.
- Compared various model using ANOVA and predicted the value of PSA level and found the best model

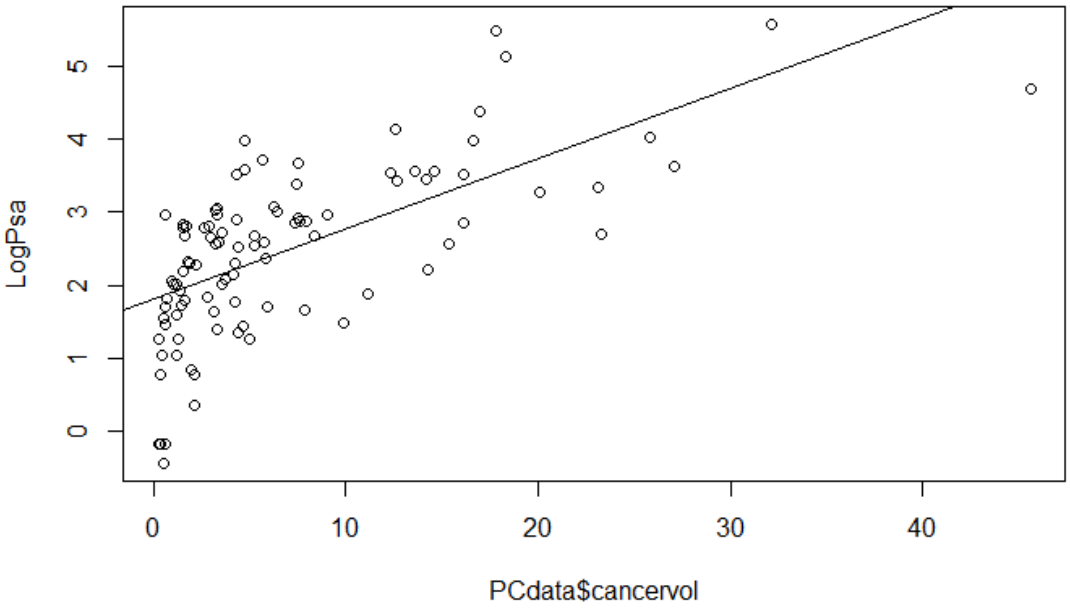
Box Plot of data and PSA

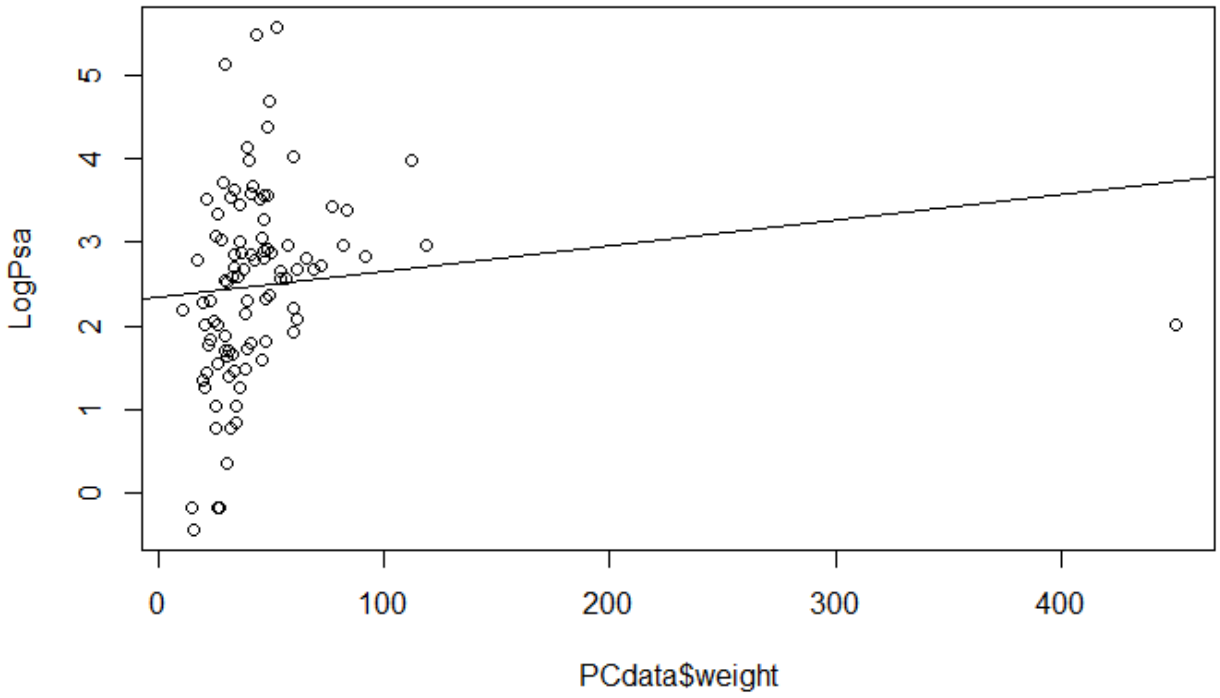


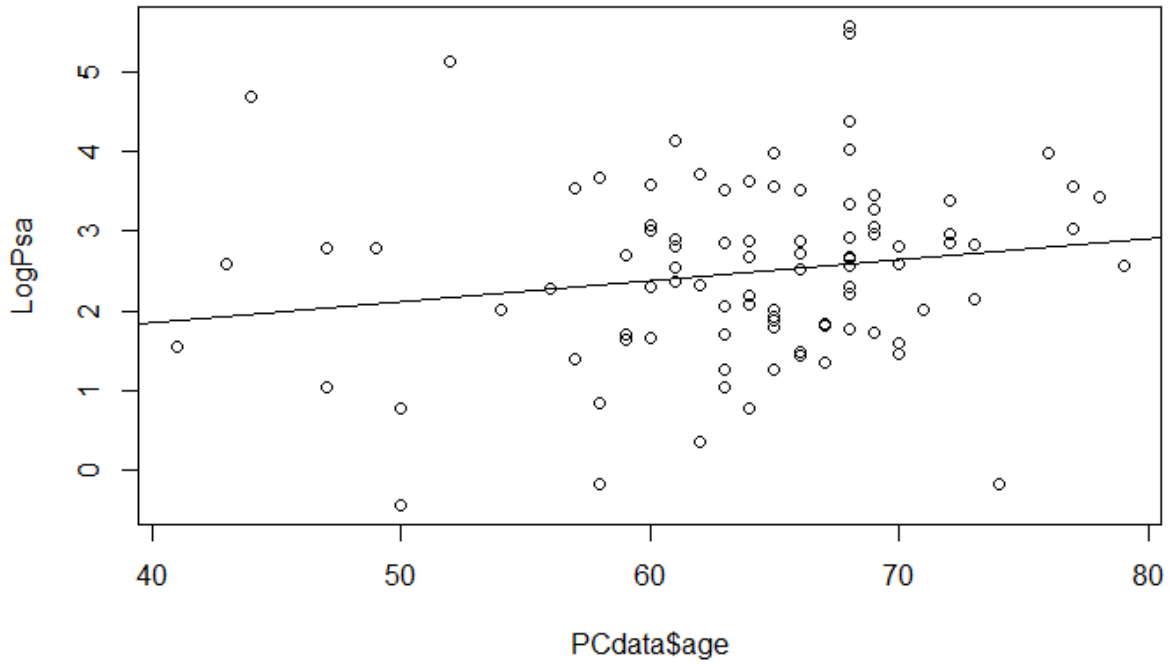


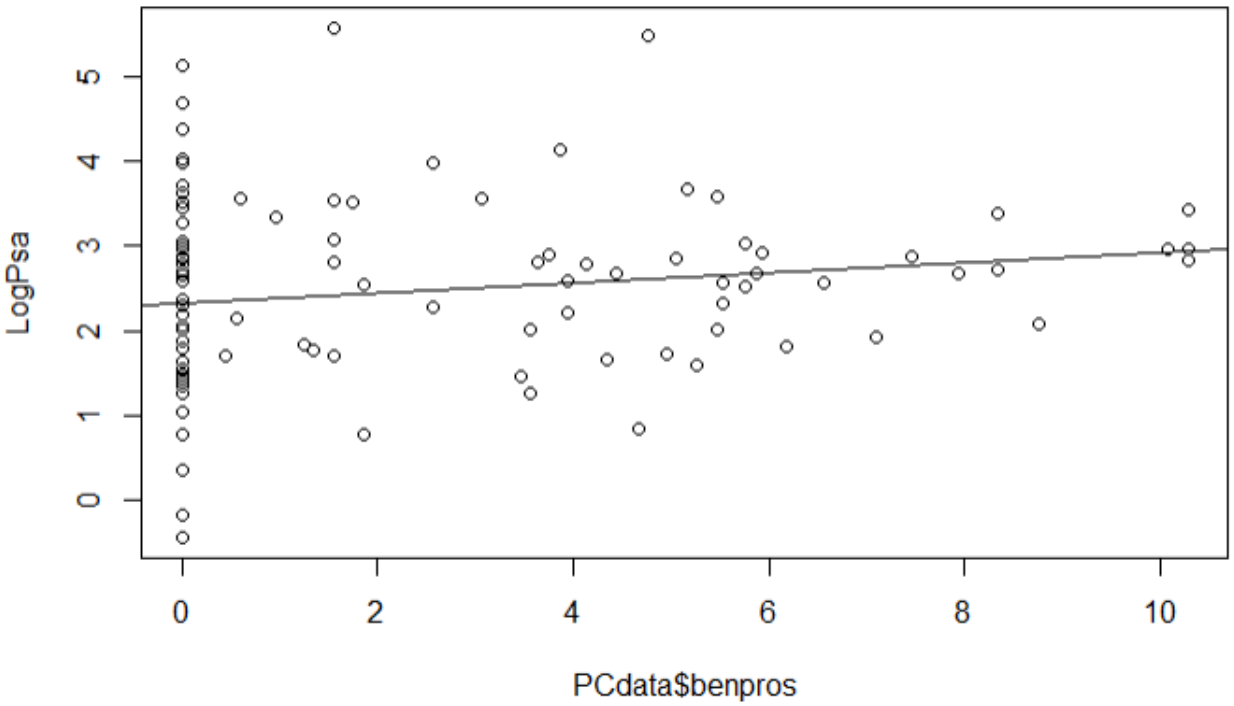
PSA is compared with features like subject, cancervol, weight, age, benpros, vesinv, capspen, gleason after building a linear model.
From the graphs below we can understand the relationship of PSA and predictors. Regression model is built after finding final predictors.

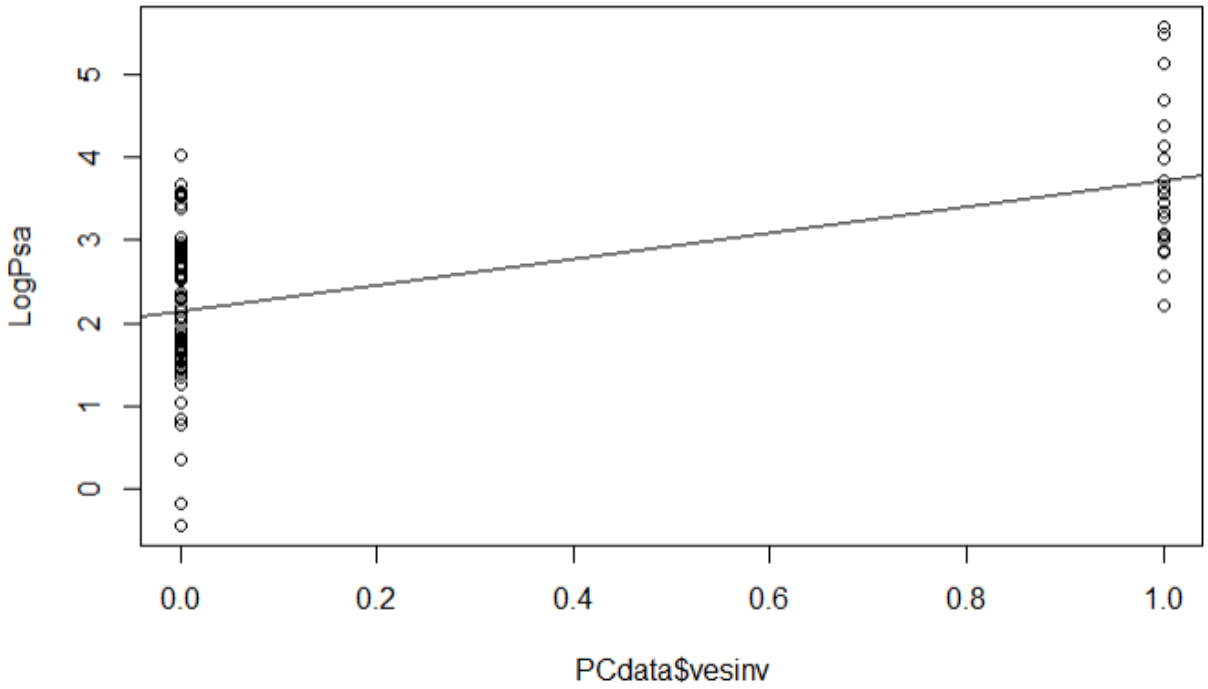

LogPsa
0
1
2
3
4
5
0
20
40
60
80
100
PCdata$subject

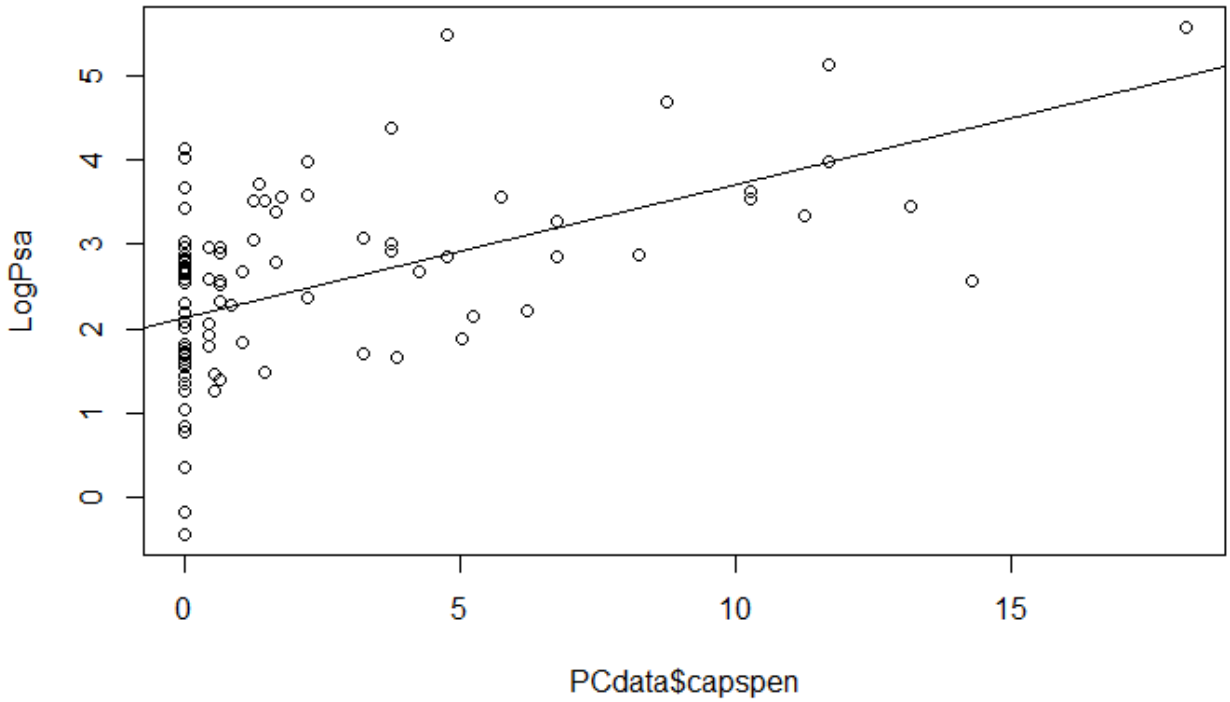

LogPsa
0
1
2
3
4
5
0
10
20
30
40
PCdata$cancervol

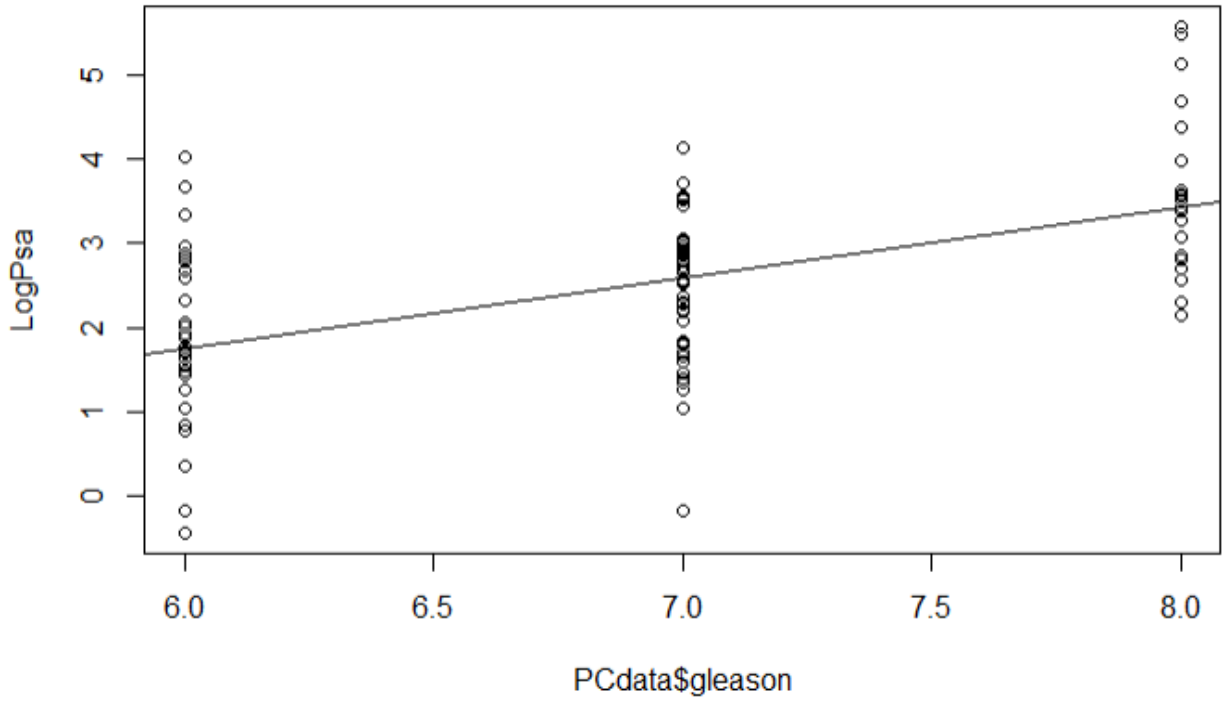

LogPsa
0
1
2
3
4
5
0
100
200
300
400
PCdata$weight

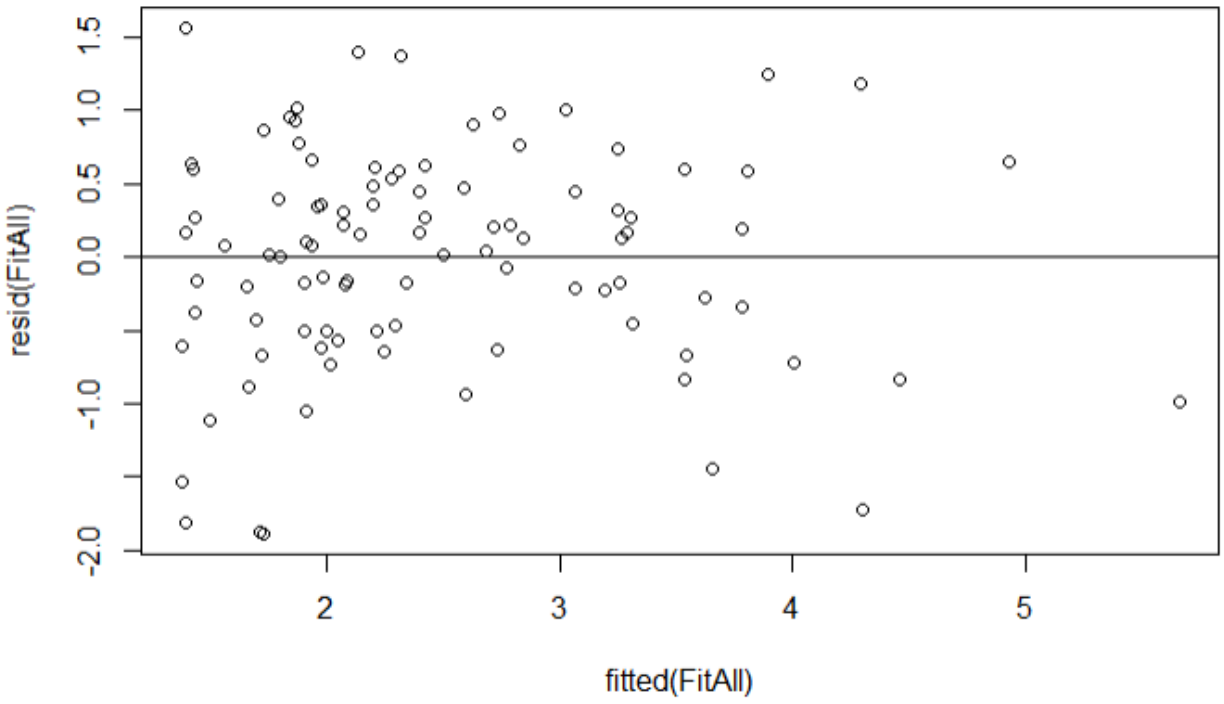

LogPsa
0
1
2
3
4
5
40
50
60
70
80
PCdata$age

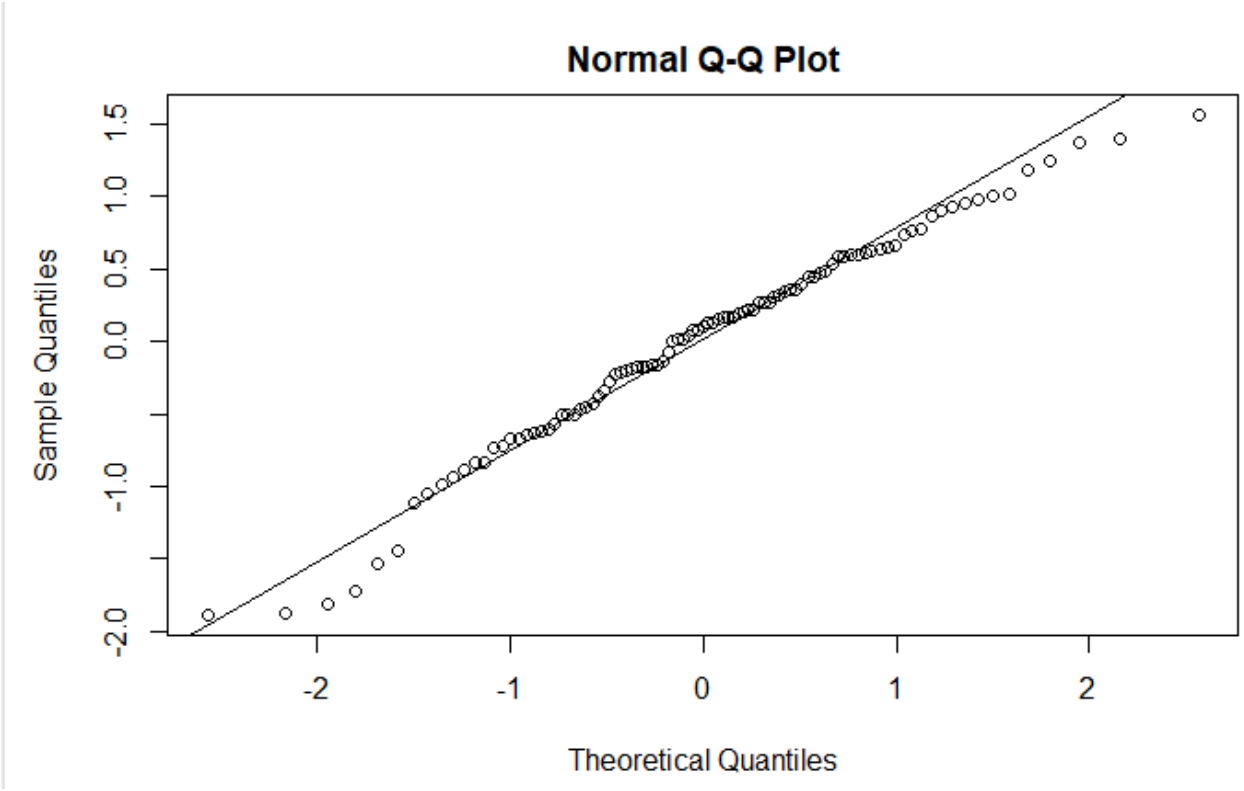

LogPsa
0
1
2
3
4
5
0
2
4
6
8
10
PCdata$benpros


LogPsa
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
PCdata$vesinv


LogPsa
0
1
2
3
4
5
0
5
10
15
PCdata$capspen


LogPsa
0
1
2
3
4
5
6.0
6.5
7.0
7.5
8.0
PCdata$gleason





Residual plot and QQ plot

```r
anova(Fit10, Fit9)
```

```
Analysis of Variance Table

Model 1: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason
Model 2: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason +
    PCdata$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
```

```r
# Capsapen is not a significant predictor as pval is >=0.05
```

```r
# Using all significant predictors
FitAll = lm(LogPsa ~  PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason + PCdata$benpros, data = PCdata )
```

```r
summary(FitAll)
```

```
Call:
lm(formula = LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) +
    PCdata$gleason + PCdata$benpros, data = PCdata)

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
PCdata$cancervol       0.06488    0.01285   5.051 2.22e-06 ***
factor(PCdata$vesinv)1 0.68421    0.23640   2.894 0.004746 **
PCdata$gleason         0.33376    0.12331   2.707 0.008100 **
PCdata$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
```

```r
anova(Fit9, FitAll)
```

```
Analysis of Variance Table

Model 1: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason +
    PCdata$capspen
Model 2: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason +
    PCdata$benpros
  Res.Df    RSS Df Sum of Sq F Pr(>F)
1     92 60.039
2     92 53.229  0    6.8101
```

```r
anova(FitAll, Fit10)
```

```
Analysis of Variance Table

Model 1: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason +
    PCdata$benpros
Model 2: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     92 53.229
2     93 60.340 -1   -7.1115 12.291 0.0007054 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Removed the predictors which are not significant using ANOVA test of different models. Capsepn has p value 0.49 which is greater than 0.05 so we removed that from the model because it is not a significant predictor. FitAll has the lowest standard error among all the Fit.

Predicted the PSA value based on the model and its coefficient standard error

```{r}
x1 = -0.65013
x2 = 0.06488
x3 = 0.68421
x4 = 0.33376
x5 = 0.09136
PredictingAns = exp(x1 + x2 * CancervolMean + x3 * 0 + x4* GleasonMean + x5 * BenprosMean)
PredictingAns
```

```
[1] 10.28357
```

Assumptions:

From the residual graph one can observe that errors are centered around zero with constant variance also they are normally distributed as seen from the QQ plot. The QQ line fits well.

Section 2

Rcode

Read the dataset prostate_cancer

```{r}
PCdata = read.csv("prostate_cancer.csv")
print(PCdata)
```

subject <int>	psa <dbl>	cancervol <dbl>	weight <dbl>	age <int>	benpros <dbl>	vesinv <int>	capspen <dbl>	gleason <int>
1	0.651	0.5599	15.959	50	0.0000	0	0.0000	6
2	0.852	0.3716	27.660	58	0.0000	0	0.0000	7
3	0.852	0.6005	14.732	74	0.0000	0	0.0000	7
4	0.852	0.3012	26.576	58	0.0000	0	0.0000	6
5	1.448	2.1170	30.877	62	0.0000	0	0.0000	6
6	2.160	0.3499	25.280	50	0.0000	0	0.0000	6
7	2.160	2.0959	32.137	64	1.8589	0	0.0000	6
8	2.340	1.9937	34.467	58	4.6646	0	0.0000	6
9	2.858	0.4584	34.467	47	0.0000	0	0.0000	7
10	2.858	1.2461	25.534	63	0.0000	0	0.0000	6

1-10 of 97 rows Previous 1 2 3 4 5 6 ... 10 Next

Summary of the dataset

```{r}
summary(PCdata)
```

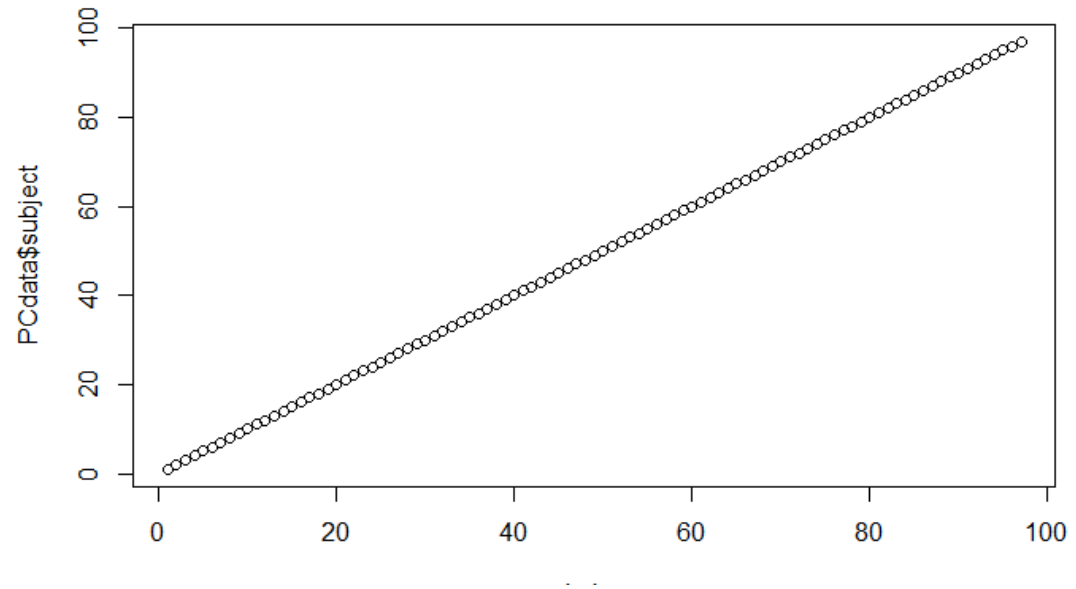
```
    subject          psa             cancervol         weight           age            benpros      
 Min.   :  1   Min.   :  0.651   Min.   : 0.2592   Min.   : 10.70   Min.   :41.00   Min.   : 0.000  
 1st Qu.: 25   1st Qu.:  5.641   1st Qu.: 1.6653   1st Qu.: 29.37   1st Qu.:60.00   1st Qu.: 0.000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

Correlations between the feature variables.

```{r}
cor(PCdata)
```

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

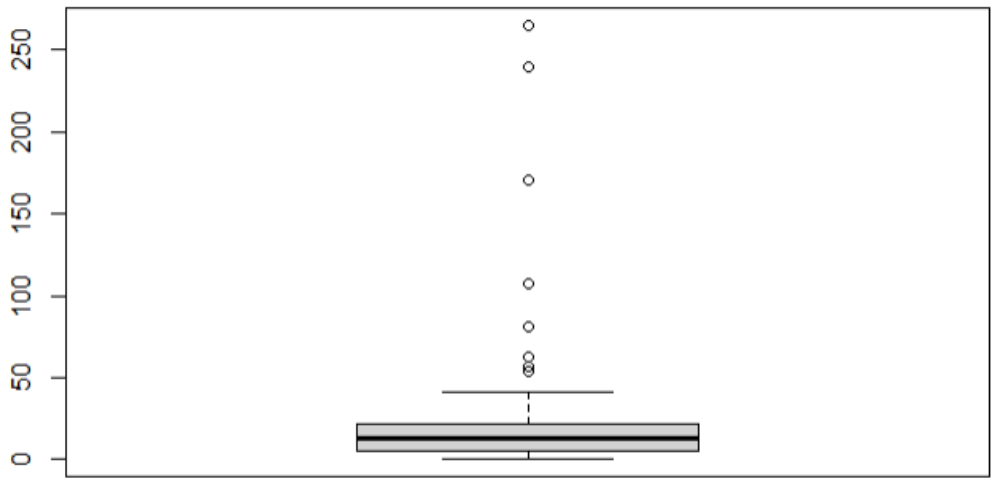
```{r}
plot(PCdata$subject)
```



Taking log to scale the psa and box plot of data and psa

```{r}
# Scaling psa
LogPsa = log(PCdata$psa)
```

```{r}
boxplot(PCdata$psa)
```



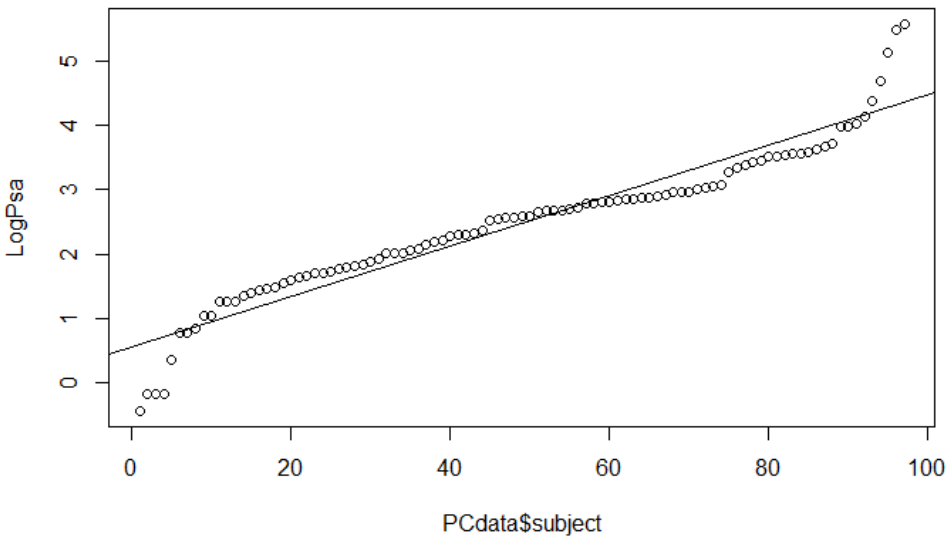
Column names

```{r}
#Column Names
ColNames = colnames(PCdata)
ColNames
```

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

Comparison of psa and subject feature with the help of linear model

```r
#Relation between LogPsa and PCdata and PCdata$subject
plot(PCdata$subject, LogPsa)
Fit1 = lm(LogPsa~PCdata$subject, data = PCdata)
abline(Fit1)
```



Chunk 1 R Markdow

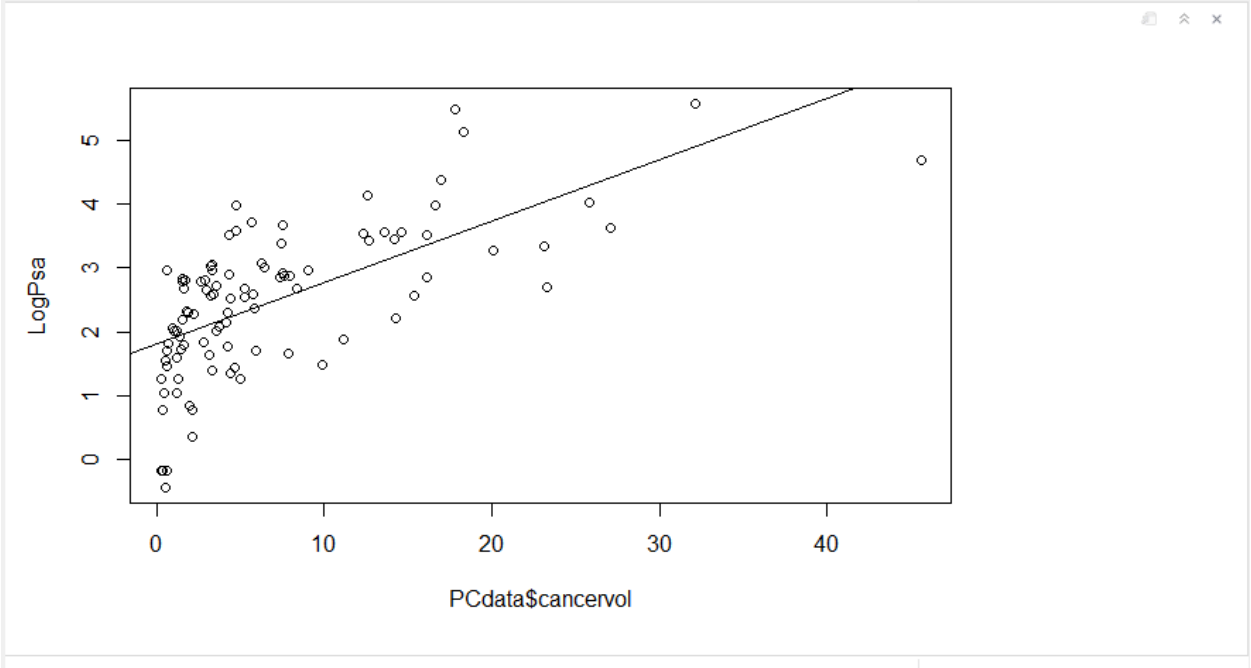
Summary Fit1

```r
summary(Fit1)
```

```
Call:
lm(formula = LogPsa ~ PCdata$subject, data = PCdata)

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

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.554321   0.067966   8.156 1.41e-12 ***
PCdata$subject 0.039272   0.001204  32.610  < 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
```

Comparison of psa and cancervol feature with the help of linear model

```r
#Relation between LogPsa and PCdata and PCdata$cancervol
plot(PCdata$cancervol, LogPsa)
Fit2 = lm(LogPsa~PCdata$cancervol, data = PCdata)
abline(Fit2)
```



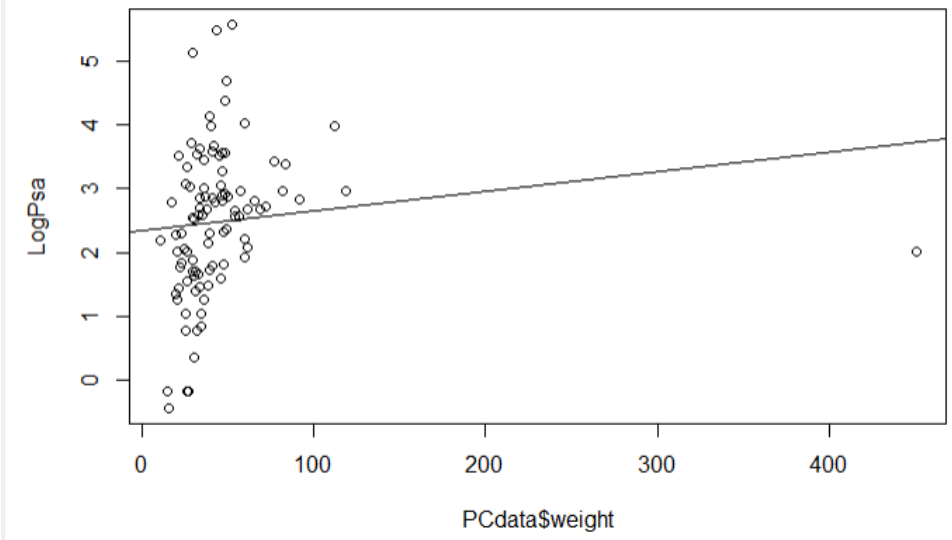
summary Fit2

```r
summary(Fit2)
```

```
Call:
lm(formula = LogPsa ~ PCdata$cancervol, data = PCdata)

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 ***
PCdata$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
```

Comparison of psa and weight feature with the help of linear model

```r
#Relation between LogPsa and PCdata and PCdata$weight
plot(PCdata$weight, LogPsa)
Fit3 = lm(LogPsa~PCdata$weight, data = PCdata)
abline(Fit3)
```



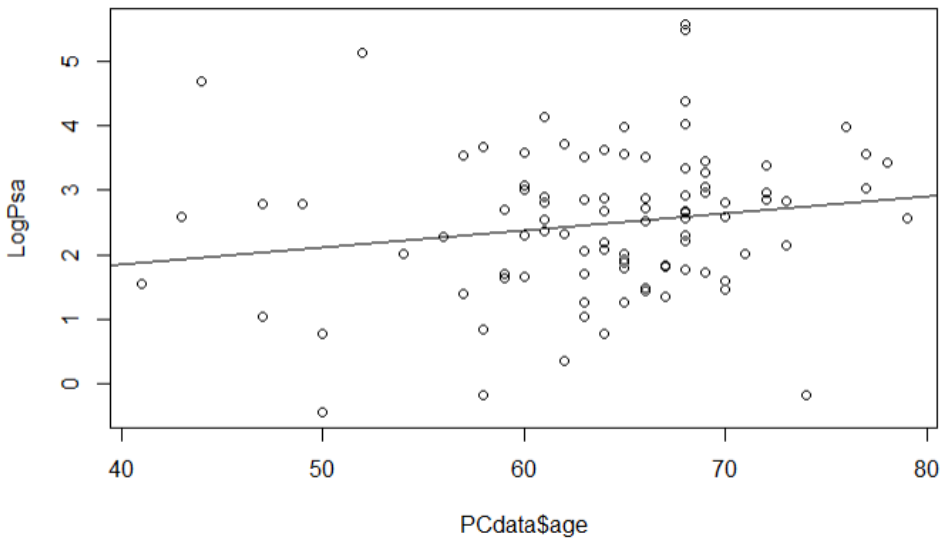
Summary Fit3

```r
summary(Fit3)
```

```
Call:
lm(formula = LogPsa ~ PCdata$weight, data = PCdata)

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.338901   0.165328  14.147   <2e-16 ***
PCdata$weight 0.003072   0.002570   1.195    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
```

Comparison of psa and age feature with the help of linear model

```{r}
#Relation between LogPsa and PCdata and PCdata$age
plot(PCdata$age, LogPsa)
Fit4 = lm(LogPsa~PCdata$age, data = PCdata)
abline(Fit4)
```



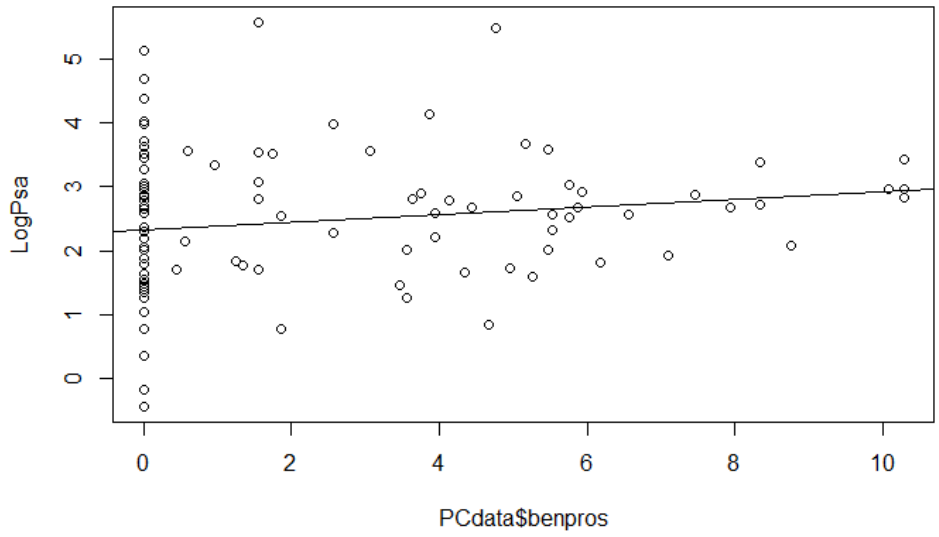
Summary of Fit4

```{r}
summary(Fit4)
```

```
Call:
lm(formula = LogPsa ~ PCdata$age, data = PCdata)

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
PCdata$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
```

Comparison of psa and benpros feature with the help of linear model

```r
#Relation between LogPsa and PCdata and PCdata$benpros
plot(PCdata$benpros, LogPsa)
Fit5 = lm(LogPsa~PCdata$benpros, data = PCdata)
abline(Fit5)
```



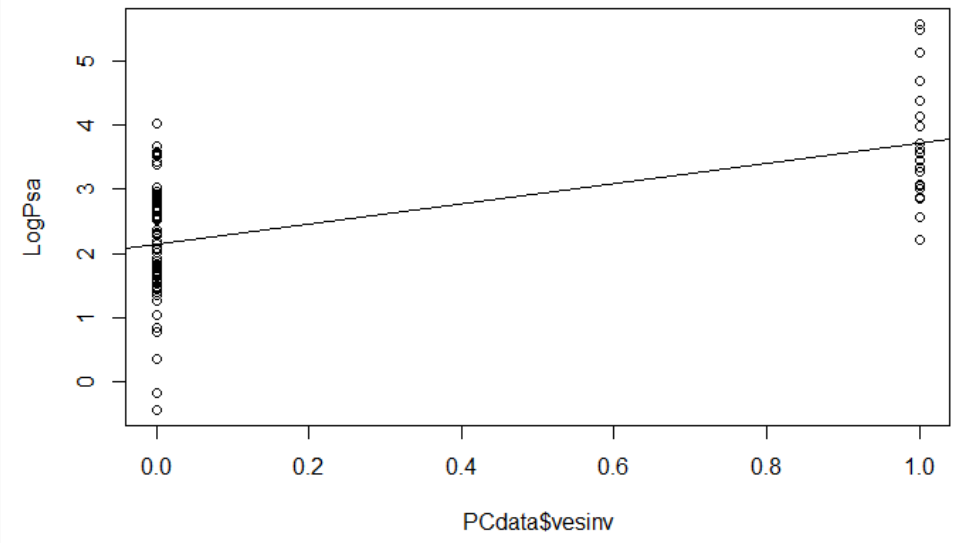
Summary of Fit5

```r
summary(Fit5)
```

```
Call:
lm(formula = LogPsa ~ PCdata$benpros, data = PCdata)

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 ***
PCdata$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
```

Comparison of psa and vesinvt feature with the help of linear model

```r
#Relation between LogPsa and PCdata and PCdata$vesinv
Vesinv = factor(PCdata$vesinv)
plot(PCdata$vesinv, LogPsa)
Fit6 = lm(LogPsa~PCdata$vesinv, data = PCdata)
abline(Fit6)
```



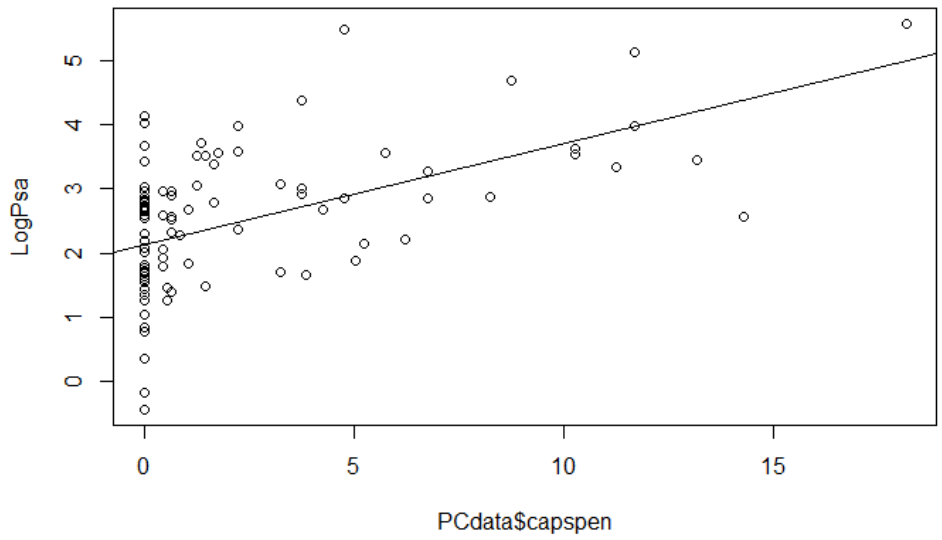
Summary of Fit6

```r
summary(Fit6)
```

```
Call:
lm(formula = LogPsa ~ PCdata$vesinv, data = PCdata)

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 ***
PCdata$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
```

Comparison of psa and capspent feature with the help of linear model

```r
#Relation between LogPsa and PCdata and PCdata$capspen
plot(PCdata$capspen, LogPsa)
Fit7 = lm(LogPsa~PCdata$capspen, data = PCdata)
abline(Fit7)
```



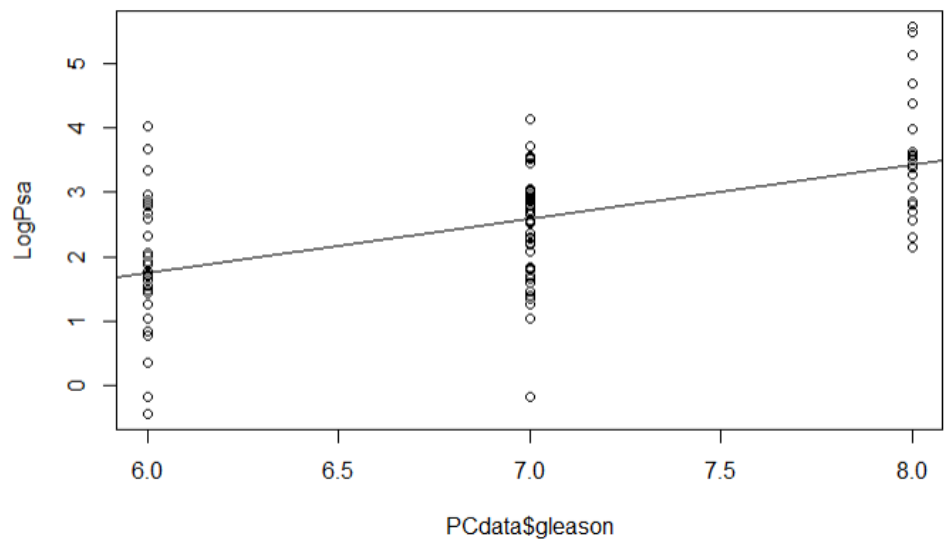
Summary of Fit7

```r
summary(Fit7)
```

```
Call:
lm(formula = LogPsa ~ PCdata$capspen, data = PCdata)

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 ***
PCdata$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
```

Comparison of psa and gleason feature with the help of linear model

```r
#Relation between LogPsa and PCdata and PCdata$gleason
plot(PCdata$gleason, LogPsa)
Fit8 = lm(LogPsa~PCdata$gleason, data = PCdata)
abline(Fit8)
```



Summary of Fit 8

```r
summary(Fit8)
```

```
Call:
lm(formula = LogPsa ~ PCdata$gleason, data = PCdata)

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 ***
PCdata$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
```

```
# One can conclude that carcervol, gleason, vesinv, benepros and capspen are significant and show
considerable relationship with PSA level.
# To predict PSA we use different combinations of above predictors
```{r}
Fit9 = lm(LogPsa ~  PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason + PCdata$capspen, data =
PCdata )

```

```{r}
summary(Fit9)
```
```

```
Call:
lm(formula = LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) +
    PCdata$gleason + PCdata$capspen, data = PCdata)

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
PCdata$cancervol       0.06452    0.01522   4.238 5.35e-05 ***
factor(PCdata$vesinv)1  0.70675    0.28024   2.522  0.01339 *
PCdata$gleason         0.39566    0.13100   3.020  0.00327 **
PCdata$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
```

```
# To  analyze changes in the model by including some of the predictors like PCdata$cancervol,
factor(PCdata$vesinv), PCdata$gleason
```{r}
Fit10 = lm(LogPsa ~  PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason, data = PCdata )
```

```{r}
summary(Fit10)
```
```

```
Call:
lm(formula = LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) +
    PCdata$gleason, data = PCdata)

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
PCdata$cancervol       0.05981    0.01352   4.425 2.62e-05 ***
factor(PCdata$vesinv)1  0.62117    0.24962   2.488   0.0146 *
PCdata$gleason         0.38491    0.12966   2.969   0.0038 **
---
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
```

```r
anova(FitAll, Fit10)
```

```
Analysis of Variance Table

Model 1: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason +
    PCdata$benpros
Model 2: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     92 53.229
2     93 60.340 -1   -7.1115 12.291 0.0007054 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(Fit10, Fit9)
```

```
Analysis of Variance Table

Model 1: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason
Model 2: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason +
    PCdata$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
```

Capsapen has p value = 0.49 which is greater than 0.05. It is not a significant predictor so we can remove it from model

```r
# Capsapen is not a significant predictor as pval is >=0.05
```

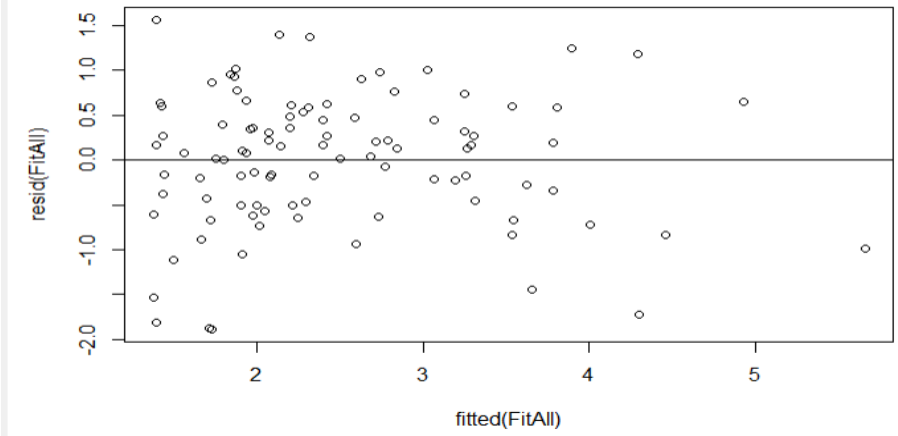
```r
# Using all significant predictors
FitAll = lm(LogPsa ~  PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason + PCdata$benpros, data =
PCdata )
```

```r
summary(FitAll)
```

```
Call:
lm(formula = LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) +
    PCdata$gleason + PCdata$benpros, data = PCdata)

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
PCdata$cancervol       0.06488    0.01285   5.051 2.22e-06 ***
factor(PCdata$vesinv)1  0.68421    0.23640   2.894 0.004746 **
PCdata$gleason         0.33376    0.12331   2.707 0.008100 **
PCdata$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
```

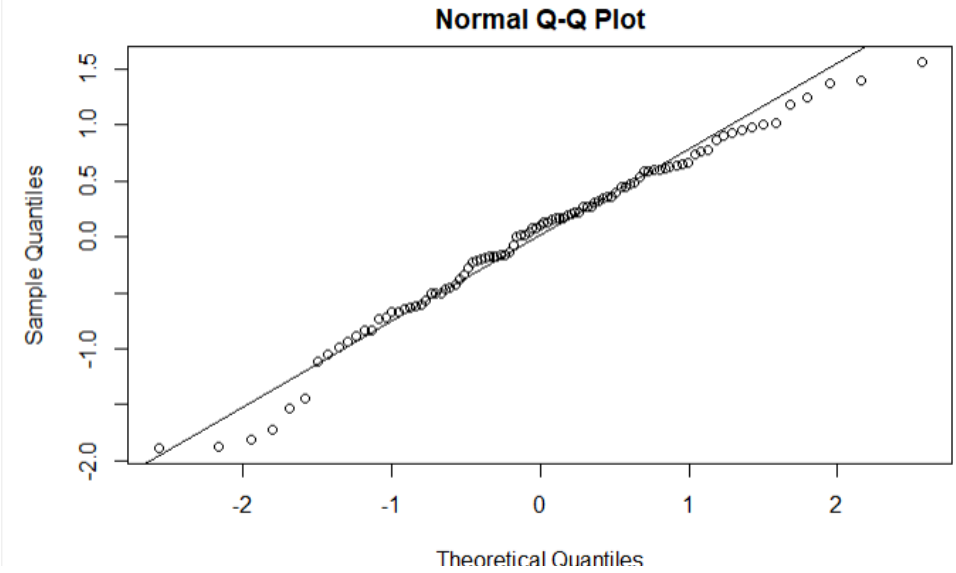
This is the minimum residual standard error obtained so this is the best model.

```r
# Residual Plot
plot(fitted(FitAll), resid(FitAll))
abline(h=0)
```



Errors are centered around zero with constant variance

```r
#QQ- Plot
qqnorm(resid(FitAll))
qqline(resid(FitAll))
```



Errors are normally distributed and also QQ line fits well

```
# Question - 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.

# taking mean of quantitative and qualitative predictors
```{r}
CancervolMean = mean(PCdata$cancervol)
CancervolMean
```
```

```
[1] 6.998682
```

```{r}
GleasonMean = mean(PCdata$gleason)
GleasonMean
```

```
[1] 6.876289
```

```{r}
BenprosMean = mean(PCdata$benpros)
BenprosMean
```

```
[1] 2.534725
```

```{r}
MfVesinv = names(which.max(table(factor(PCdata$vesinv))))
MfVesinv
```

```
[1] "0"
```

Predicting the PSA value based on model and its coefficient error.

```{r}
anova(FitAll, Fit10)
```

```
Analysis of Variance Table

Model 1: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason +
    PCdata$benpros
Model 2: LogPsa ~ PCdata$cancervol + factor(PCdata$vesinv) + PCdata$gleason
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     92 53.229
2     93 60.340 -1   -7.1115 12.291 0.0007054 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```{r}
x1 = -0.65013
x2 = 0.06488
x3 = 0.68421
x4 = 0.33376
x5 = 0.09136
PredictingAns = exp(x1 + x2 * CancervolMean + x3 * 0 + x4* GleasonMean + x5 * BenprosMean)
PredictingAns
```

```
[1] 10.28357
```