STATISTICAL METHODS FOR DATA SCIENCE - MINI PROJECT 6

Names of group members: 1. Venkatesh Sankar Net ID: VXS200014
2. Manneyaa Jayasanker Net ID: MXJ180040

Contributions of each group member:

Venkatesh Sankar:
- Worked on loading dataset
- Worked on plotting correlation matrix
- Worked on Graphical Representations
- Worked on Model Implementation and Comparison

Manneyaa Jayasanker:
- Implemented model using AIC
- Model comparison using AIC
- Model evaluation using Graphical Representation
- Worked on Summary statistics and conclusion

---

1. We plot a correlation matrix in order to find the correlation among the variables. We initially load the sample data set and install the required package called “corrplot” for plotting the correlation matrix.

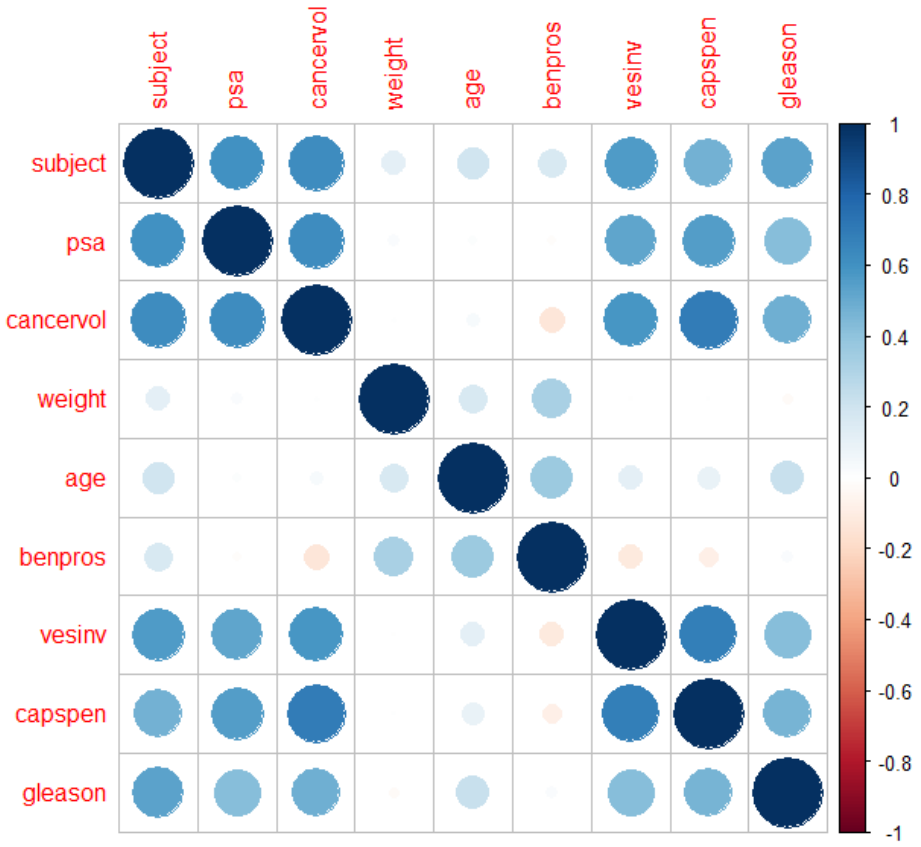
```
> setwd("D:/Academics/Spring 21/Statistical Methods Data Science/MP6/")
> ipdata = read.csv("prostate_cancer.csv")
> install.packages("corrplot")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate ver
sion of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/19254/Documents/R/win-library/4.0'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/corrplot_0.84.zip'
Content type 'application/zip' length 5450182 bytes (5.2 MB)
downloaded 5.2 MB

package 'corrplot' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\19254\AppData\Local\Temp\Rtmps5KcFg\downloaded_packages
> library(corrplot)
corrplot 0.84 loaded
Warning message:
package 'corrplot' was built under R version 4.0.5
> cor.data = cor(ipdata)
> corrplot(cor.data)
>
```

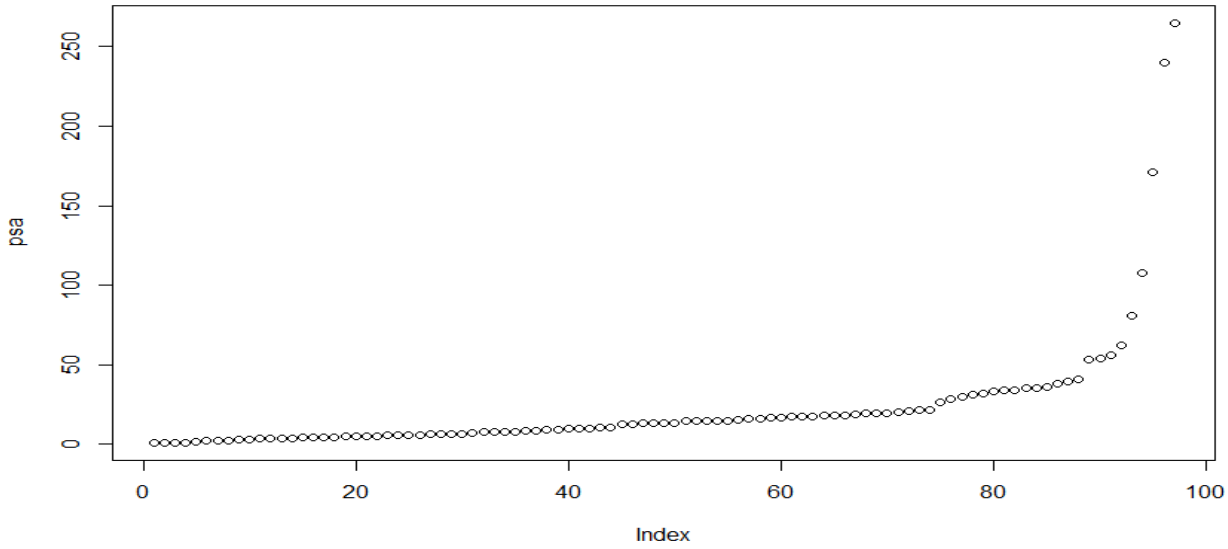
It gives the following output.



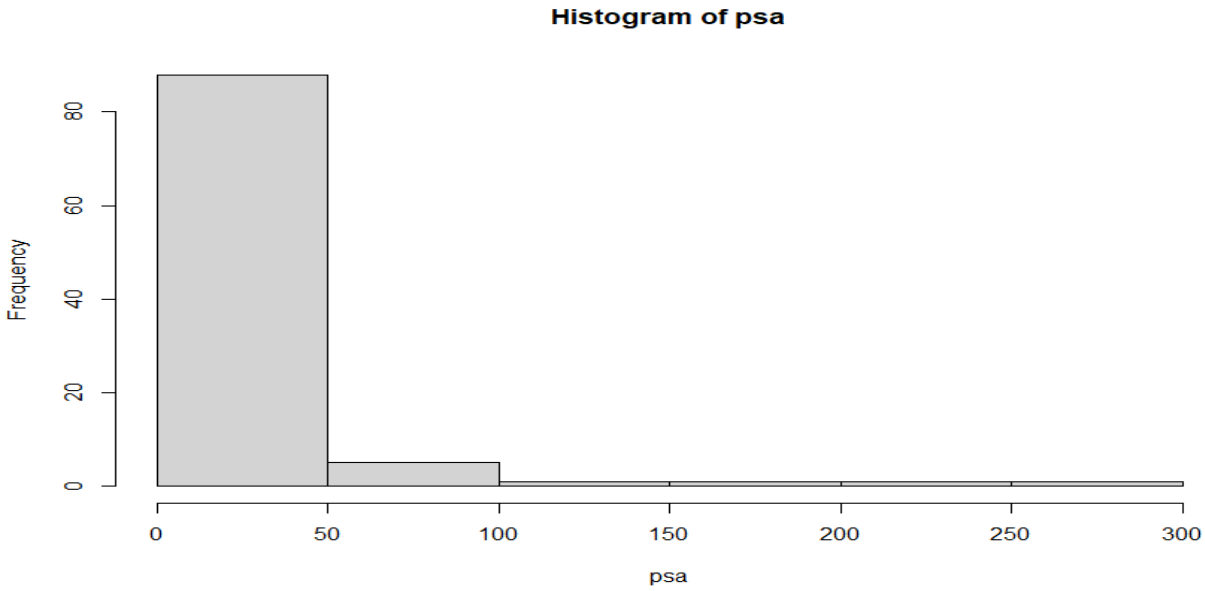
Since, we would like to understand the PSA level column from the dataset, we do graphical plotting of PSA level using plot(), hist() and boxplot() function.

```
> attach(data)
>
> plot(psa) #scattered plot of psa level
> hist(psa)
> boxplot(psa) #boxplot of psa level
>
```

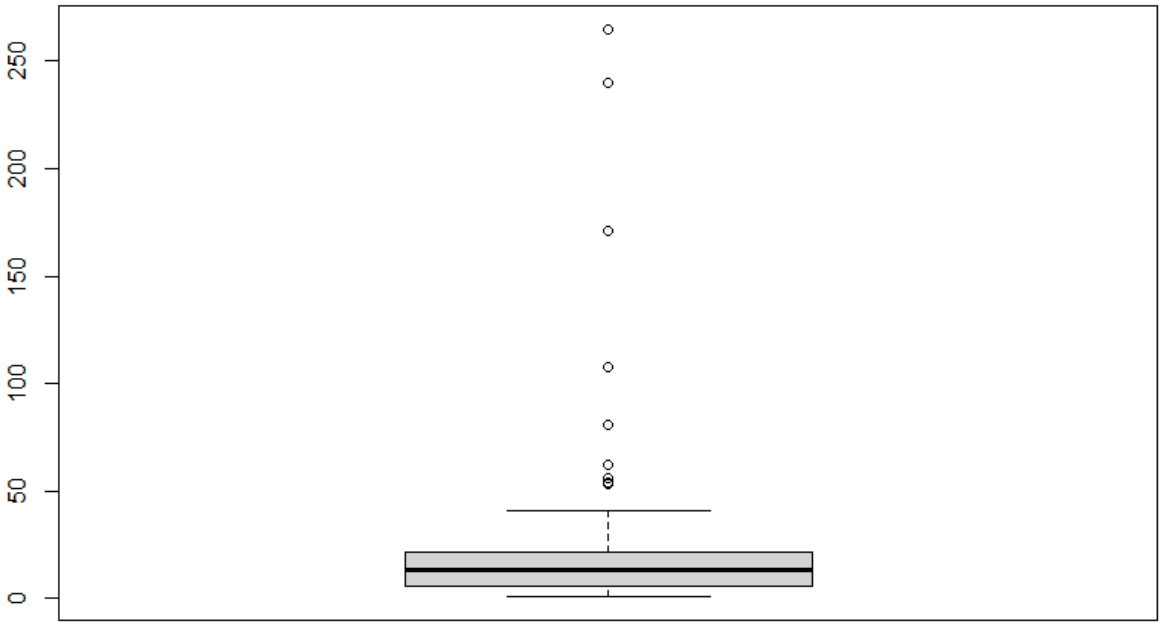
Scattered Plot of PSA



Histogram of PSA



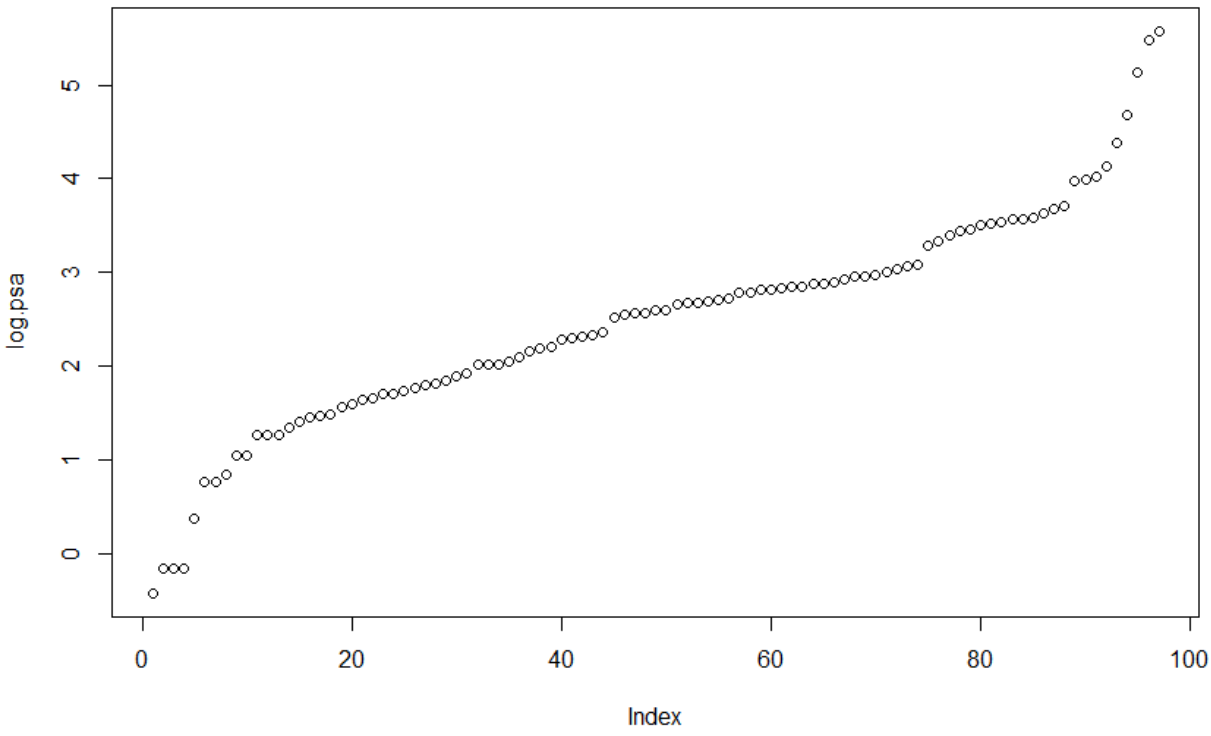
Boxplot of PSA



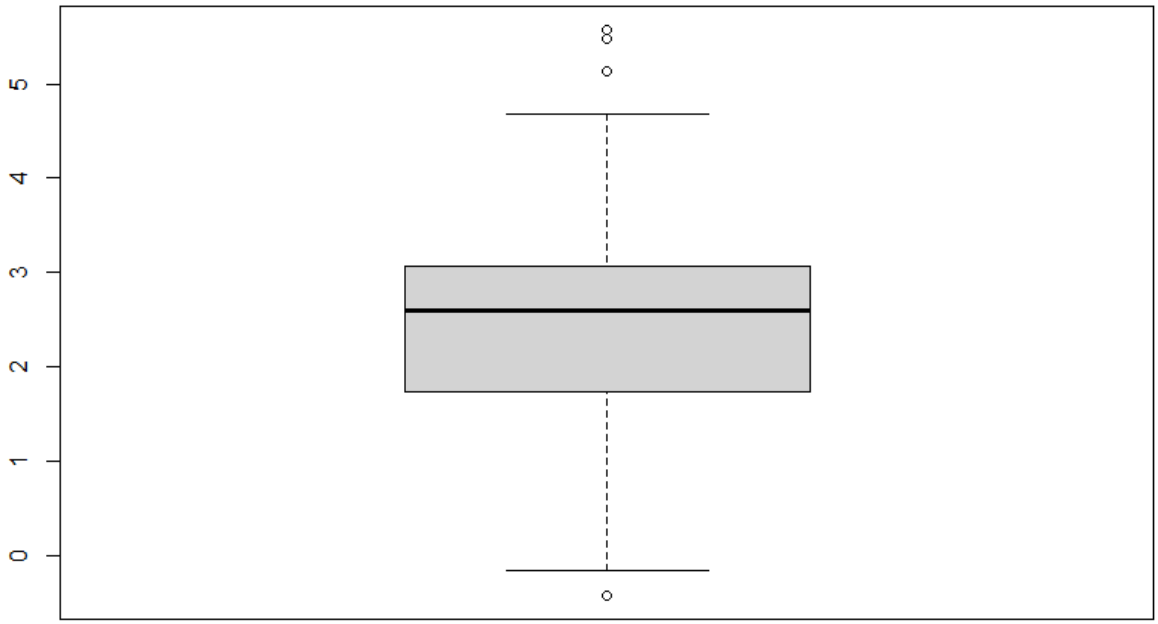
We can infer from the boxplot that there are many outliers. Hence, we do logarithmic transformation to the data to fit out linear model. Then , we do plotting again of the same PSA level.

```
> attach(data)
>
> plot(psa) #scattered plot of psa level
> hist(psa)
> boxplot(psa) #boxplot of psa level
>
```

Scattered Plot of PSA level after applying log transformation



Boxplot of PSA level after applying log transformation



Given that, vesinv is a qualitative variable, we use as.factor() to convert into a factor and preserve the variable label and value attributes.

```
>
> data$vesinv = as.factor(data$vesinv)
>
```

Model 1

Null Hypothesis : H_0 : None of the predictors are useful for predicting response.

Alternative hypothesis : H_1 : Atleast one of the predictor is useful for predicting response.

```
> fit1 = lm(log.psa ~ cancervol + vesinv +capspen + gleason + weight + age + benpros) #linear model #1
> summary(fit1)

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

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

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

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

From model 1, we can infer that cancervol which has *** and vesinv, gleason , benpros which has ** are the significant predictors. Hence, null hypothesis can be rejected.

Model 2

In this model, we consider only the significant predictors.

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

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

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

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

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

From correlation matrix, we can infer that capspen is also important. So for next model, we can update model 2 by including capspen as well.

Model 3

```
>
> fit3 = update(fit2,.~. + capspen) #linear model 3
> summary(fit3)

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

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

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

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

We can infer that the adjusted R-squared decreases from which we can conclude that capspen is not an optimal predictor for the response variable prediction.

Model Comparison

Next we **compare** all the three models using anova()

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

Response: log.psa
          Df Sum Sq Mean Sq F value    Pr(>F)
cancervol  1 55.164  55.164 93.5572 1.522e-15 ***
vesinv     1  6.547   6.547 11.1034  0.001256 **
capspen    1  0.066   0.066  0.1114  0.739372
gleason    1  5.954   5.954 10.0971  0.002042 **
weight     1  2.041   2.041  3.4624  0.066083 .
age        1  0.374   0.374  0.6344  0.427866
benpros    1  5.147   5.147  8.7291  0.004007 **
Residuals 89 52.477   0.590
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> anova(fit2)
Analysis of Variance Table

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

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

Response: log.psa
          Df Sum Sq Mean Sq F value    Pr(>F)
cancervol  1 55.164  55.164 95.0078 8.619e-16 ***
vesinv     1  6.547   6.547 11.2755 0.0011481 **
gleason    1  5.718   5.718  9.8478 0.0022919 **
benpros    1  7.111   7.111 12.2480 0.0007232 ***
capspen    1  0.392   0.392  0.6757 0.4132368
Residuals 91 52.837   0.581
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> anova(fit1, fit2)
Analysis of Variance Table

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

```
> anova(fit2, fit3)
Analysis of Variance Table

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

Model 1: log.psa ~ cancervol + vesinv + capspen + gleason + weight + age +
    benpros
Model 2: log.psa ~ cancervol + vesinv + gleason + benpros + capspen
  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
>
> anova(fit1, fit2, fit3)
Analysis of Variance Table

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

Based on the above comparison of the three models, we can conclude that **Model 2** is the best linear model.

Checking the best model using AIC :

```
fit.full <- fit1 <- lm(psa.log ~ cancervol + vesinv + capspen + gleason + weight + age + benpros)
for.aic <- step(lm(psa.log ~ 1), direction = "forward",scope = formula(fit.full), k = 2, trace = 0)
for.bic <- step(lm(psa.log ~ 1), direction = "forward",scope = formula(fit.full), k = log(32), trace = 0)
back.aic <- step(fit.full, direction = "backward", k = 2, trace = 0)
back.bic <- step(fit.full, direction = "backward", k = log(32), trace = 0)
(Adjusted_R.square <- data.frame("Method"=c("for.aic", "for.bic", "back.aic", "back.bic"),
                                 "Adj.r.square"=c(summary(for.aic)$adj.r.square,
                                                  summary(for.bic)$adj.r.square,summary(back.aic)$adj.r.square,
                                                  summary(back.bic)$adj.r.square)))
```

Output:

```
    Method Adj.r.square
1  for.aic    0.5652831
2  for.bic    0.5652831
3 back.aic    0.5652831
4 back.bic    0.5652831
```

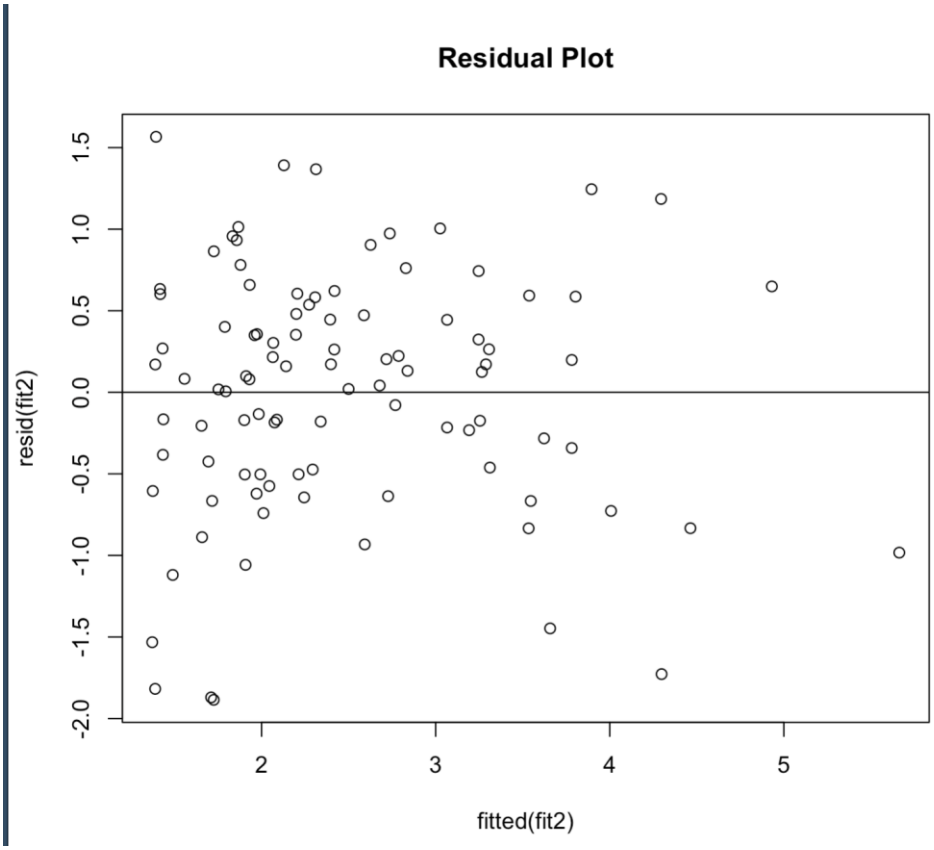
Checking the best model using AIC :

```
> l1 <- glm(fit2)
> l2 <- glm(fit1)
> l3 <- glm(fit3)
> print(l1$aic)
[1] 229.0635
> print(l2$aic)
[1] 233.6828
> print(l3$aic)
[1] 230.346
```

We can see from the above results l1 or fit2 linear model has the lowest aic score telling that it's the best model among all the models.

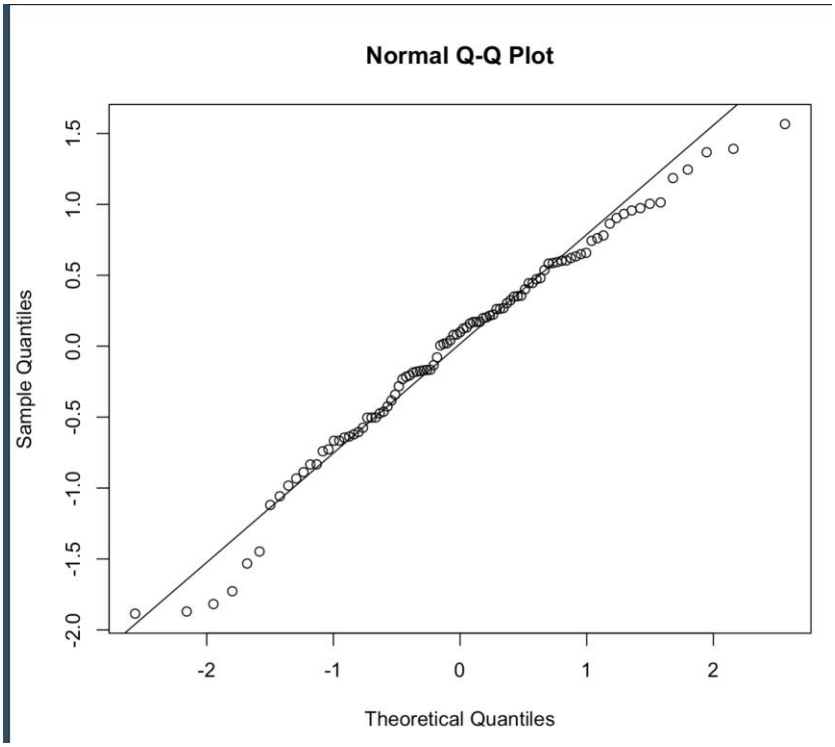
Model Evaluation:

```
> plot(fitted(fit2),resid(fit2), main = "Residual Plot")
> abline(h=0)
```



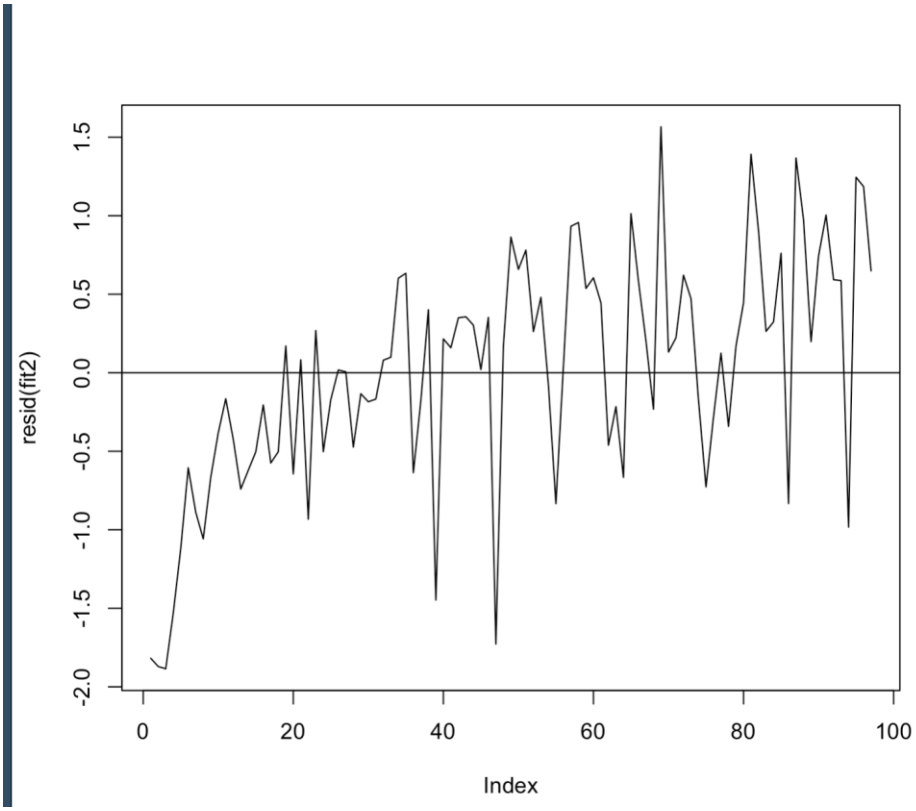
The points are scattered around zero and there is not pattern. So, we can say the errors have mean zero and constant variance.

```
>
> qqnorm(resid(fit2))
> qqline(resid(fit2))
```



Errors are normally distributed .

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



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

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

```
> summary(fit2)

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

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

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

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

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

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

From the above results we can see that gleason value 7 is highest in the data, vesinv value 0 is higher in the data and the mean of cancervol and benpros are 6.998 and 2.534 respectively.

predicted value is equal to: -0.65013 + 6.998682*(0.06488) + 7*(0.33376) + 0.09136*(2.534725) = 2.371837 Thus, the actual value of PSA is exp(2.371837) which is equal to 10.71706