CS 6313 STATISTICAL METHODS FOR DATA SCIENCE

Mini Project 6

SEPTEMBER 10, 2021

AFRAA ALSHAMMARI
PRAVALIKA DOMAL

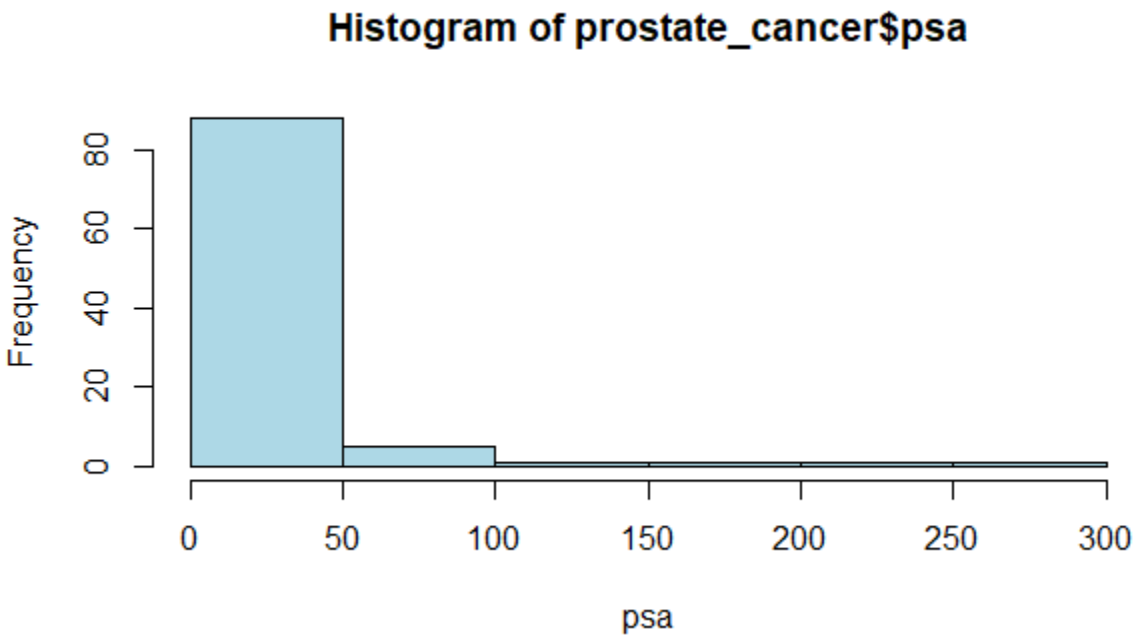
Contribution of each member:

Both worked together to finish and submit this mini project. Starting with learning R and conducted the needed statistics scripts to solve Q1 and Q2. Afraa did the documentation and report the experiments and Pravalika worked on analytical solution scripts for finding accuracy of scripts.

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.

Solution:

a) Histogram of PSA Level



The following conclusions can be drawn from the aforementioned histogram:

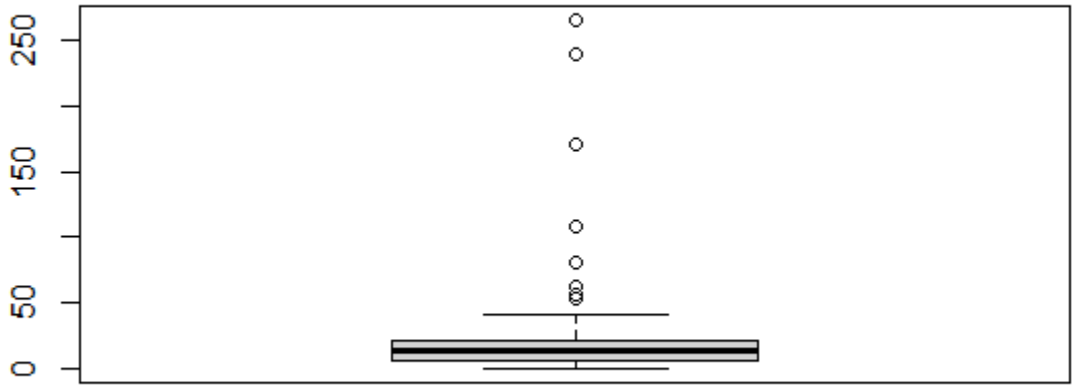
1.Many people have extremely low PSA levels.

2.There is an indirect relationship between PSA levels and the number of people, because when PSA levels rise, the number of people decreases dramatically.

3.The distribution looks to be exponential, which is not the case with the Normal Distribution.

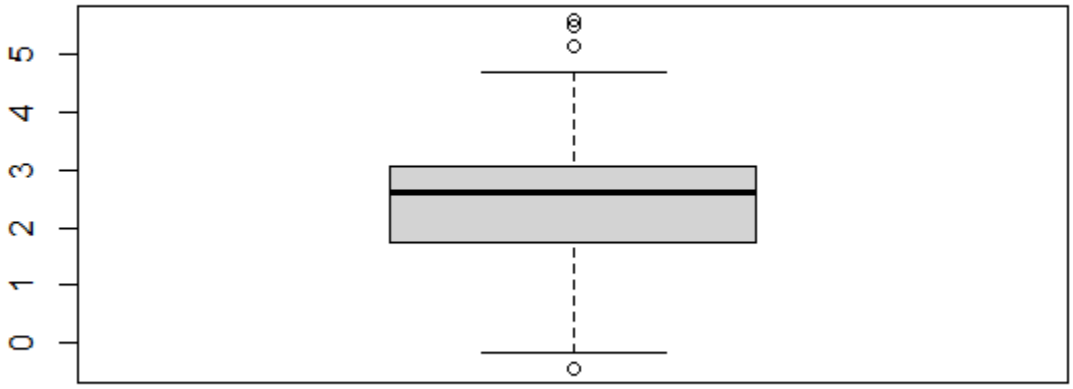
4.We plot the Boxplot because there is no significant evidence for the presence of Outliers in the psa data.

b) Boxplot of Response Variable(PSA)



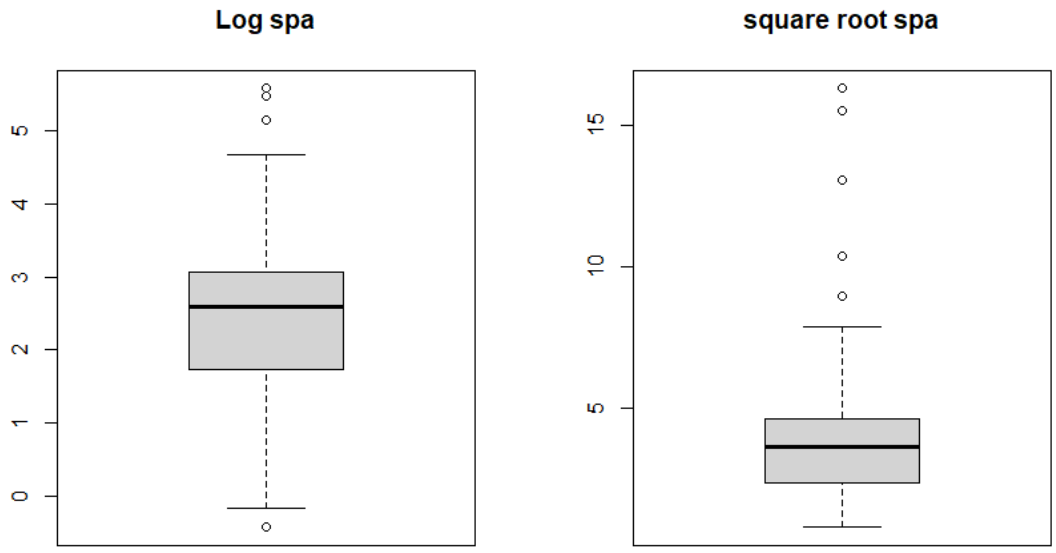
As the Histogram did not provide any significant evidence for the presence of Outlier in the psa data, the Boxplot of Response Variable (PSA) was used. The presence of many outliers in the psa data can be seen in the above boxplot, and no symmetry can be seen. Look at the Boxplot of the Natural logtransformation of the response variable for a more detailed analysis.

c) Boxplot of Natural Log Function PSA



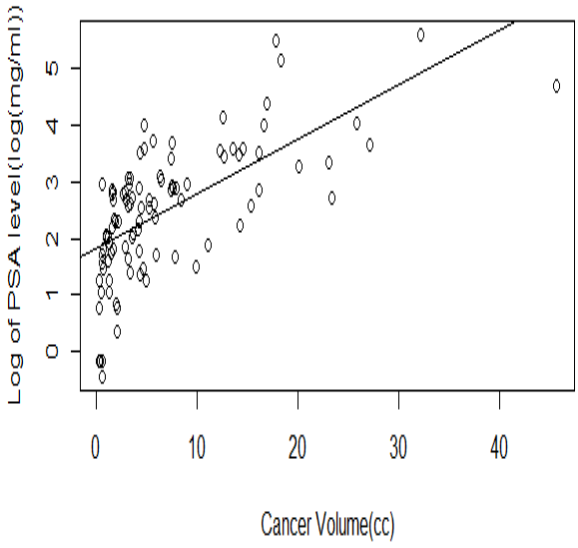
The distribution has now become symmetric, and the number of outliers in the psa data has decreased, as shown in the above Boxplot. We now utilize the transformed answer as our response variable because the distribution is more symmetric here and the presence of outliers is lower than in the normal boxplot.

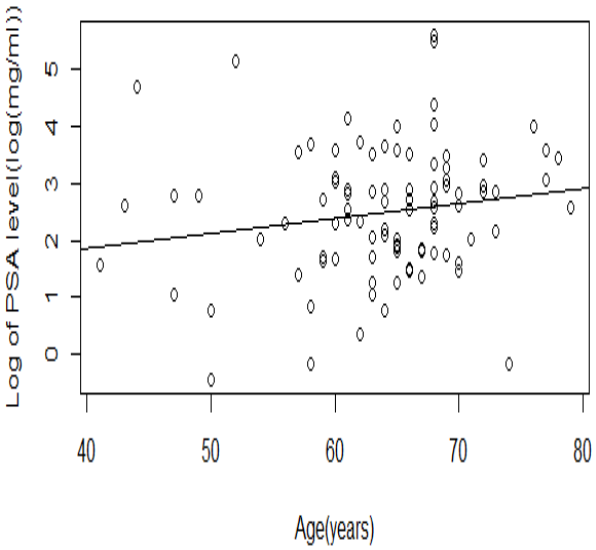
From Graph 2 the log transformation will be chosen as it shows less outliers and its distribution is closer to normal distribution than the square root.

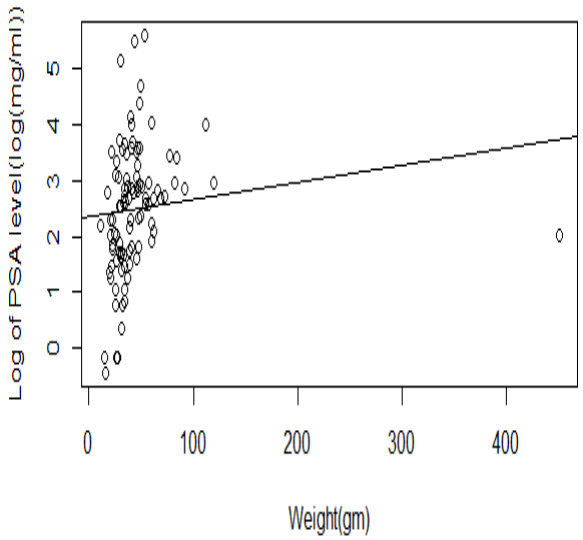


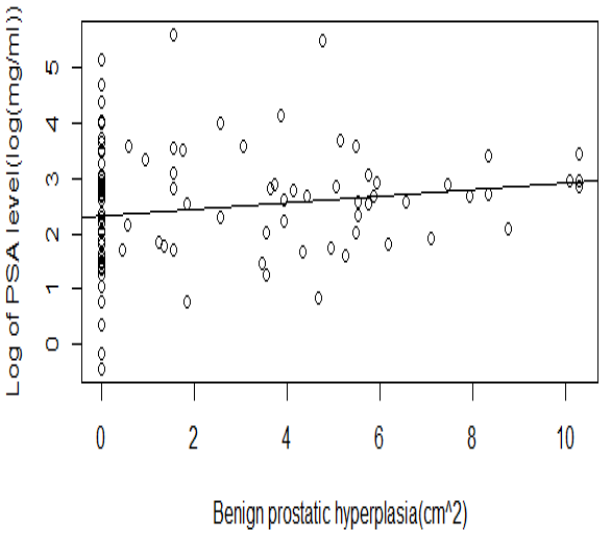
Graph 2 Psa Log and Psa Square Root Boxplots

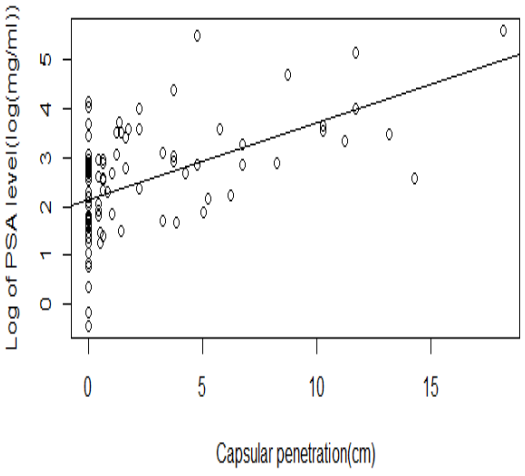
The data contains 2 qualitative predictor vesinv and gleason and the remaining predictors are quantitative. A model will be constructed on the quantitative predictors which will start by plotting the response variable psa with these quantitative predictors. These plots are presented in Graph 3 and it can be concluded that Capspen, benpros and cancervol are high impact factors.











Graph 3 Plot Psa with the Quantitative Predictors

Now 2 models will be built one with the high impact factors and the other with the qualitative. A F test will be used to compare between the two models.

```
#Create model
fit<-lm(formula =  Log_psa ~ (cancervol+weight+age+benpros+capspen))
summary(fit)

##
## Call:
## lm(formula = Log_psa ~ (cancervol + weight + age + benpros +
##     capspen))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.91508 -0.54429  0.06032  0.56605  1.74268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.037961   0.770412   1.347   0.1812
## cancervol   0.088925   0.015093   5.892 6.36e-08 ***
## weight      0.001028   0.001974   0.521   0.6038
## age         0.007634   0.012438   0.614   0.5409
## benpros     0.082325   0.031966   2.575   0.0116 *
## capspen     0.033572   0.031409   1.069   0.2880
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8349 on 91 degrees of freedom
## Multiple R-squared:  0.5036, Adjusted R-squared:  0.4763
## F-statistic: 18.46 on 5 and 91 DF,  p-value: 1.288e-12

fit2<-lm(formula =  Log_psa ~ (cancervol+benpros+capspen))
summary(fit2)

##
## Call:
## lm(formula = Log_psa ~ (cancervol + benpros + capspen))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01425 -0.52582  0.02225  0.54145  1.73694
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.53504    0.13922  11.026  < 2e-16 ***
## cancervol    0.08924    0.01497   5.960 4.48e-08 ***
## benpros      0.09449    0.02816   3.355  0.00115 **
## capspen      0.03544    0.03102   1.143  0.25606
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8289 on 93 degrees of freedom
## Multiple R-squared:  0.4998, Adjusted R-squared:  0.4837
## F-statistic: 30.98 on 3 and 93 DF,  p-value: 5.649e-14
```

```
#Compare 2 fits
anova(fit2,fit)
```

```
## Analysis of Variance Table
##
## Model 1: Log_psa ~ (cancervol + benpros + capspen)
## Model 2: Log_psa ~ (cancervol + weight + age + benpros + capspen)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     93 63.904
## 2     91 63.430  2   0.47464 0.3405 0.7123
```

From the test, it is shown that for βage and βwight is small, and the 3 height factors have a large value. Moreover, the p value is large (0.7123). therefore, the null hypothesis for βage =0 and βwight =0.

Now the stepwise assumption will be conducted to confirm the previous findings.

```
# step wise selection
#Fit 3 Forward
fit3_Forward <-step(lm(Log_psa~1), scope = list(upper=~cancervol+weight
+age+benpros+capspen),direction = "forward")
```

```
## Start:  AIC=28.72
## Log_psa ~ 1
##
##             Df Sum of Sq     RSS      AIC
## + cancervol  1    55.164  72.605 -24.0986
## + capspen    1    34.286  93.482   0.4169
## + age        1     3.688 124.080  27.8831
## + benpros    1     3.166 124.603  28.2911
## <none>                   127.769  28.7246
## + weight     1     1.893 125.876  29.2767
##
## Step:  AIC=-24.1
## Log_psa ~ cancervol
##
##           Df Sum of Sq    RSS     AIC
## + benpros  1    7.8034 64.802 -33.128
## + age      1    2.6615 69.944 -25.721
## + weight   1    1.7901 70.815 -24.520
## <none>                 72.605 -24.099
## + capspen  1    0.9673 71.638 -23.400
##
## Step:  AIC=-33.13
## Log_psa ~ cancervol + benpros
```

```
##
##           Df Sum of Sq    RSS     AIC
## <none>                  64.802 -33.128
## + capspen  1   0.89737 63.904 -32.480
## + age      1   0.39609 64.406 -31.723
## + weight   1   0.20572 64.596 -31.436
```

```
fit3_Forward
```

```
##
## Call:
## lm(formula = Log_psa ~ cancervol + benpros)
##
## Coefficients:
## (Intercept)    cancervol      benpros
##      1.5309       0.1010       0.0949
```

```
#Fit 3 Backward
fit3_Backward <-step(lm(Log_psa~cancervol+weight+age+benpros+capspen),d
irection = "backward")
```

```
## Start:  AIC=-29.2
## Log_psa ~ cancervol + weight + age + benpros + capspen
##
##             Df Sum of Sq    RSS      AIC
## - weight     1    0.1891 63.619 -30.9149
## - age        1    0.2626 63.692 -30.8029
## - capspen    1    0.7963 64.226 -29.9934
## <none>                   63.430 -29.2036
## - benpros    1    4.6231 68.053 -24.3794
## - cancervol  1   24.1971 87.627   0.1424
##
## Step:  AIC=-30.91
## Log_psa ~ cancervol + age + benpros + capspen
##
##             Df Sum of Sq    RSS     AIC
## - age        1    0.2856 63.904 -32.480
## - capspen    1    0.7869 64.406 -31.723
## <none>                   63.619 -30.915
## - benpros    1    5.6465 69.265 -24.667
## - cancervol  1   24.4216 88.040  -1.401
##
## Step:  AIC=-32.48
## Log_psa ~ cancervol + benpros + capspen
##
##             Df Sum of Sq    RSS     AIC
## - capspen    1    0.8974 64.802 -33.128
## <none>                   63.904 -32.480
## - benpros    1    7.7334 71.638 -23.400
## - cancervol  1   24.4110 88.315  -3.098
##
```

```
## Step:  AIC=-33.13
## Log_psa ~ cancervol + benpros
##
##             Df Sum of Sq     RSS     AIC
## <none>                    64.802 -33.128
## - benpros    1     7.803  72.605 -24.099
## - cancervol  1    59.802 124.603  28.291

fit3_Backward

##
## Call:
## lm(formula = Log_psa ~ cancervol + benpros)
##
## Coefficients:
## (Intercept)    cancervol      benpros
##      1.5309       0.1010       0.0949

#Fit 3 both Backward and Forward
fit3_both<-step(lm(Log_psa~1),  scope = list(lower=~1, upper=~cancervol
+weight+age+benpros+capspen),direction = "both")

## Start:  AIC=28.72
## Log_psa ~ 1
##
##             Df Sum of Sq     RSS      AIC
## + cancervol  1    55.164  72.605 -24.0986
## + capspen    1    34.286  93.482   0.4169
## + age        1     3.688 124.080  27.8831
## + benpros    1     3.166 124.603  28.2911
## <none>                   127.769  28.7246
## + weight     1     1.893 125.876  29.2767
##
## Step:  AIC=-24.1
## Log_psa ~ cancervol
##
##             Df Sum of Sq     RSS     AIC
## + benpros    1     7.803  64.802 -33.128
## + age        1     2.662  69.944 -25.721
## + weight     1     1.790  70.815 -24.520
## <none>                    72.605 -24.099
## + capspen    1     0.967  71.638 -23.400
## - cancervol  1    55.164 127.769  28.725
##
## Step:  AIC=-33.13
## Log_psa ~ cancervol + benpros
##
##             Df Sum of Sq     RSS     AIC
## <none>                    64.802 -33.128
## + capspen    1     0.897  63.904 -32.480
## + age        1     0.396  64.406 -31.723
```

```
## + weight     1     0.206  64.596 -31.436
## - benpros    1     7.803  72.605 -24.099
## - cancervol  1    59.802 124.603  28.291

fit3_both

##
## Call:
## lm(formula = Log_psa ~ cancervol + benpros)
##
## Coefficients:
## (Intercept)    cancervol      benpros
##      1.5309       0.1010       0.0949


## we have new formula based on above method

> fit3 <- lm(formula = psalog ~ cancervol + benpros)

> summary(fit3)

Call:

lm(formula = psalog ~ cancervol + benpros)

Residuals:

     Min       1Q   Median       3Q      Max

-2.01672 -0.55101  0.06457  0.56870  1.75415

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  1.53090    0.13940  10.982  < 2e-16 ***

cancervol    0.10105    0.01085   9.314 5.29e-15 ***

benpros      0.09490    0.02821   3.364  0.00111 **

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.8303 on 94 degrees of freedom

Multiple R-squared:  0.4928,  Adjusted R-squared:  0.482

F-statistic: 45.67 on 2 and 94 DF,  p-value: 1.389e-14
```

```
## Compare it with previous quantitative predictors.
> anova(fit3, fit2)
Analysis of Variance Table
Model 1: psalog ~ cancervol + benpros
Model 2: psalog ~ cancervol + capspen + benpros
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     94 64.802
2     93 63.904  1   0.89737 1.3059 0.2561
## p value is large (0.2561)
## Hence null hypothesis is accepted
```

Figure below is Residual graph (fit3)

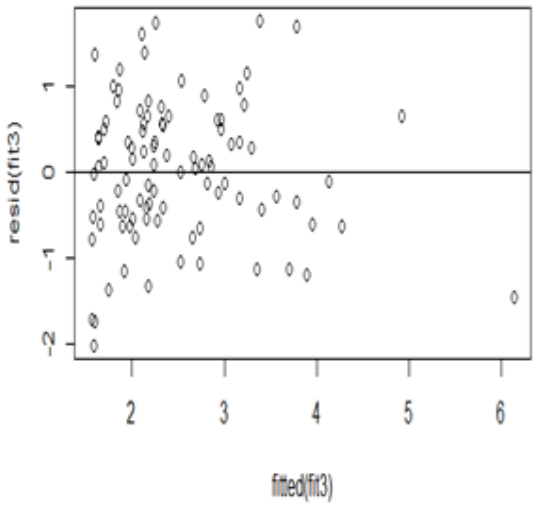


Figure below is Absolute Residual graph (fit3)

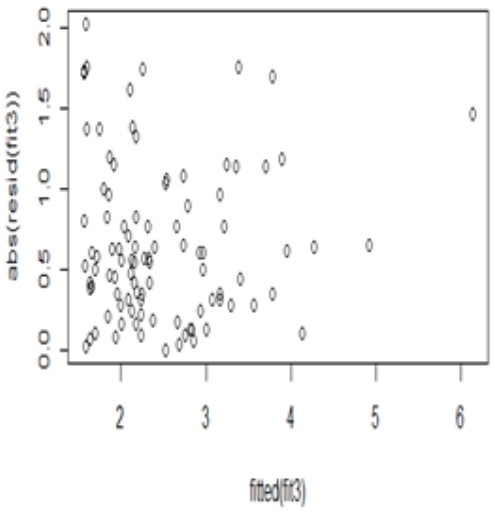


Figure shown below is time series plot for the Model with only Quantitative variables (fit3).

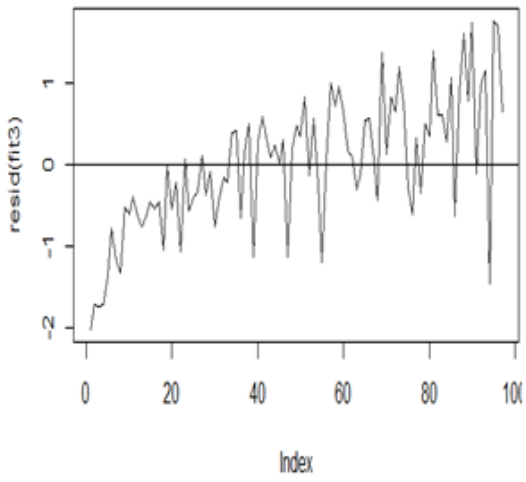
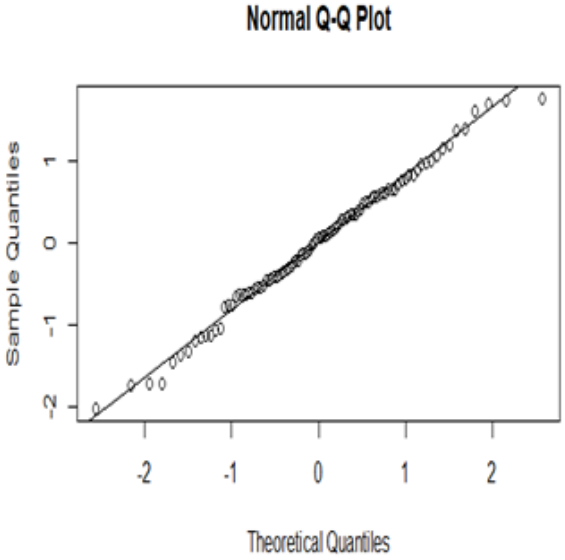


Figure shown below is Normal QQ Plot for Model with Quantitative Variables (fit3).



We now considering the categorical variables.

We add two variables at first to the model.

> fit4 <- update(fit3, . ~ . + factor(vesinv))

> fit5 <- update(fit3, . ~ . + factor(gleason))

Comparing two categorical variables

> summary(fit4)

Call:

lm(formula = psalog ~ cancervol + benpros + factor(vesinv))

Residuals:

Min	1Q	Median	3Q	Max
-1.9867	-0.4996	0.1032	0.5545	1.4993

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.51484	0.13206	11.471	< 2e-16	***
cancervol	0.07618	0.01256	6.067	2.78e-08	***
benpros	0.09971	0.02674	3.729	0.000331	***
factor(vesinv)1	0.82194	0.23858	3.445	0.000858	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7861 on 93 degrees of freedom

Multiple R-squared: 0.5502, Adjusted R-squared: 0.5357

F-statistic: 37.92 on 3 and 93 DF, p-value: 4.247e-16

> anova(fit3, fit4)

Analysis of Variance Table

Model 1: psalog ~ cancervol + benpros

Model 2: psalog ~ cancervol + benpros + factor(vesinv)

  Res.Df RSS Df Sum of Sq F Pr(>F)

1 94 64.802

2 93 57.468 1 7.3339 11.868 0.0008583 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

vesinv and gleason are definitely significant to the model so we add the variables to the formula which results in our final model.

#Finalize the model

>fit6 <- update(fit3, . ~ . + factor(vesinv) + factor(gleason))

> summary(fit6)

Call:

lm(formula = psalog ~ cancervol + benpros + factor(vesinv) +

  factor(gleason))

Residuals:

  Min 1Q Median 3Q Max

-1.85235 -0.45777 0.06741 0.51651 1.53204

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.38817	0.15609	8.894	5.27e-14	***
cancervol	0.06241	0.01367	4.566	1.55e-05	***
benpros	0.09265	0.02627	3.527	0.00066	***
factor(vesinv)1	0.69646	0.23837	2.922	0.00439	**
factor(gleason)7	0.26028	0.18280	1.424	0.15790	
factor(gleason)8	0.70545	0.25712	2.744	0.00732	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7636 on 91 degrees of freedom

Multiple R-squared: 0.5848, Adjusted R-squared: 0.5619

F-statistic: 25.63 on 5 and 91 DF, p-value: 4.722e-16

Figure below shown is Residual graph (fit6).

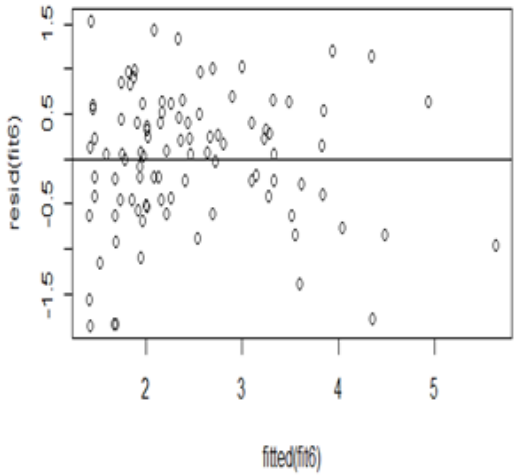


Figure below shown is Absolute Residual graph (fit6).

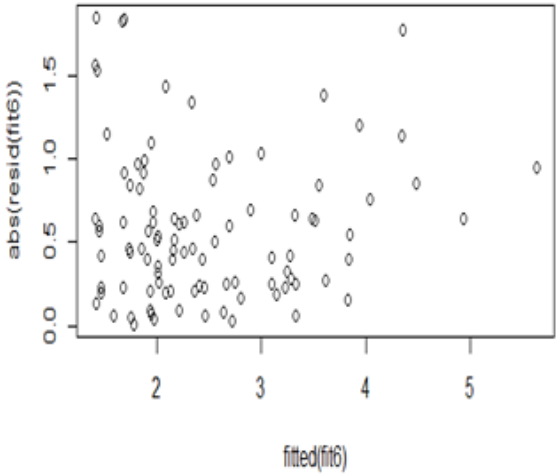


Figure below shown is Time series plot for Final Model (fit6).

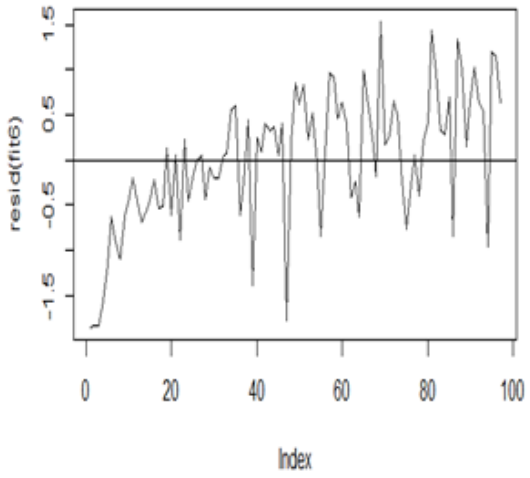
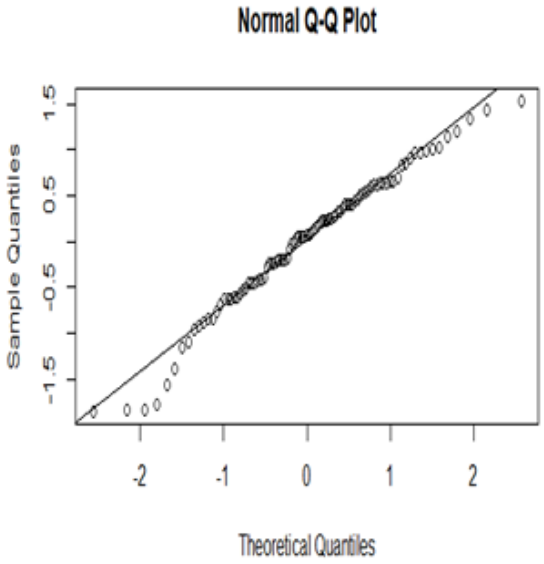


Figure below shown is Normal QQ plot for Final Model (fit6).



The model appears to be realistic for the most part, however there are a few outliers. We can now estimate a patient's PSA level using this model, as long as the quantitative predictors are at their sample means and the qualitative predictors are at their sample modes.

#Predict PSA level for sample mean

```
> pred <- predict(fit6,
+             data.frame(cancervol = mean(cancervol),
+                        benpros   = mean(benpros),
+                        vesinv    = getmode(vesinv),
+                        gleason   = getmode(gleason)))
> exp(pred)
       1
10.17628
```

R code

```
#project 6
#Read prostate_cancer
data<-read.csv("prostate_cancer.csv");
#Attach the data
attach(data)
# Investigating the response variable psa using boxplot
psa<-data [,2];
boxplot(psa);
# Transformation variable of response to log and square root and draw the box plot to
choose the optimal transformation
par(mfrow=c(1,2))
#1. log
Log_psa<-log(psa)
boxplot(Log_psa, main= 'Log spa')
#2.squre root
boxplot(sqrt(psa),main= 'square root spa')


# draw plot bwtween Log_psa and each quantitativevarible
par(mfrow=c(2,3))
#1. with cancervol
plot(cancervol,Log_psa, xlab ="cancer volume (cc)",ylab = "prostate-specific antigen level
(log(mg/ml))", main = "cancervol")
abline(lm(Log_psa~cancervol) )

#2. with weight
plot(weight,Log_psa, xlab ="prostate weight (gm)",ylab = "prostate-specific antigen level
(log(mg/ml))",main = "weight")
abline(lm(Log_psa~weight) )

#3. with age
plot(age,Log_psa, xlab ="years",ylab = "prostate-specific antigen level (log(mg/ml))",main
="age")
abline(lm(Log_psa~age) )

#4. with benpros
plot(benpros, Log_psa, xlab ="Amount of benign prostatic hyperplasia (cm2)", ylab =
"prostate-specific antigen level (log(mg/ml))",main = "benpros")
abline(lm(Log_psa~benpros) )

#5. with capspen
plot(capspen,Log_psa, xlab ="Degree of capsular penetration (cm)",ylab = "prostate-specific
```

```
antigen level (log(mg/ml))", main = "capspen" )
abline(lm(Log_psa~capspen))

#Create model
fit<-lm(formula = Log_psa ~ (cancervol+weight+age+benpros+capspen))
summary(fit)


fit2<-lm(formula = Log_psa ~ (cancervol+benpros+capspen))
summary(fit2)


#Compare 2 fits
anova(fit2,fit)

# Model selected.

fit3 <- lm(formula = Log_psa ~ cancervol + benpros)

summary(fit3)

# Compare the model with the guess one.

anova(fit3, fit2)


# Residual plot of fit3.

plot(fitted(fit3), resid(fit3))

abline(h = 0)


# Plot the absolute residual of fit3

plot(fitted(fit3), abs(resid(fit3)))


# Plot the residuals' time series plot.

plot(resid(fit3), type="l")

abline(h = 0)


# Normal QQ plot of fit3.

qqnorm(resid(fit3))
```

```
qqline(resid(fit3))

# Consider the categorical variables.
fit4 <- update(fit3, . ~ . + factor(vesinv))
fit5 <- update(fit3, . ~ . + factor(gleason))

# Comparing two categorical variables.
summary(fit5)
#the anova comparing fit3 and fit5
anova(fit3, fit5)

summary(fit4)
#the anova comparing fit3 and fit4
anova(fit3, fit4)

# Finalize the model.
fit6 <- update(fit3, . ~ . + factor(vesinv) + factor(gleason))
summary(fit6)

# Residual plot of fit6.
plot(fitted(fit6), resid(fit6))
abline(h = 0)

# Plot the absolute residual of fit3.
plot(fitted(fit6), abs(resid(fit6)))

# Plot the residuals' time series plot.
plot(resid(fit6), type="l")
abline(h = 0)
```

```
# fit6 - Normal QQ plot
qqnorm(resid(fit6))
qqline(resid(fit6))

# Create the function for getting mode
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

#Predict the PSA level for categorical predictors with values at their sample means.
#predictors with the most common label
pred <- predict(fit6,
   data.frame(cancervol = mean(cancervol),
           benpros   = mean(benpros),
           vesinv    = getmode(vesinv),
           gleason   = getmode(gleason)))

#  Response variable is log(psa)
exp(pred)
```