CS 6363.001 Statistical Methods for Data Science (Spring 2017)

Mini Project #5

April 20, 2017

Group Members:

NAME	netID
Ashwath Santhanam	axs161730
Haripriyaa U Manian	hum160030

Contribution.
Both of us discussed and completed the entire project together.

- ➢ psa is the response variable.
- ➢ all other variables are predictors.
- ➢ subject ID is not a predictor.
- ➢ Quantitative variables are: cancervol, weight, age, benpros, capspen
- ➢ vesinv is a qualitative variable with no ordering.
- ➢ gleason is a qualitative variable with ordering.

The simplest linear regression model looks like this: $y = \beta0 + \beta1x1 + \beta2x2 + \cdot \cdot \cdot + \beta pxp + \varepsilon$ (1)

The regression model comes with two main assumptions:
1. Linear relationship between response and predictor.
2. $\varepsilon \sim N(0, \sigma^2)$

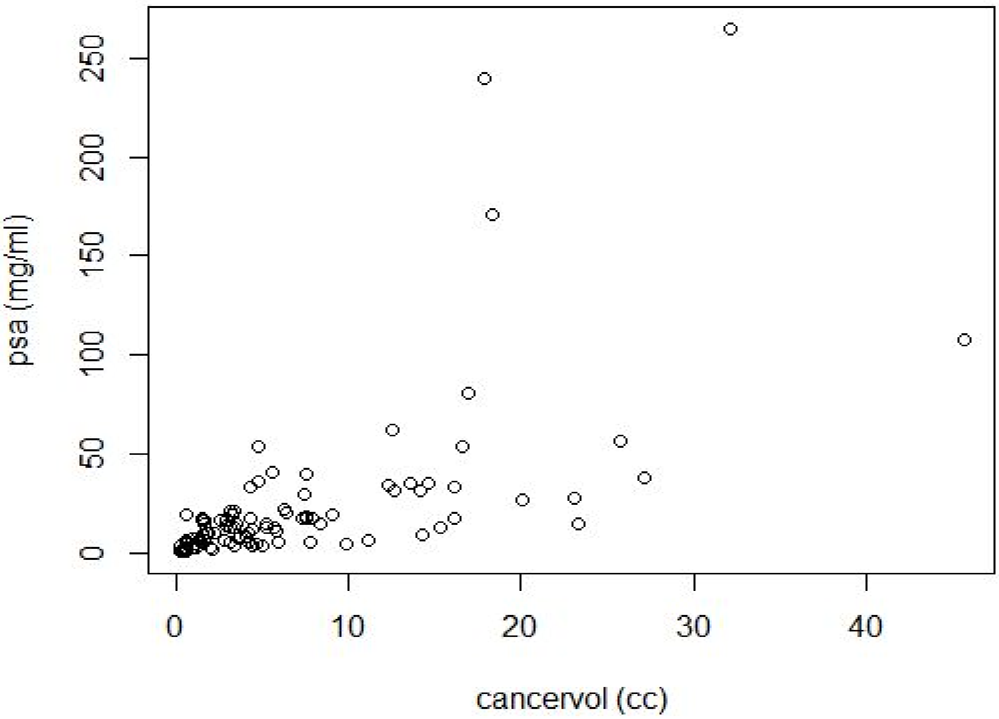
-**UNIVARIATE ANALYSIS:**

#We begin by doing a univariate analysis(one predictor at a time) and see each variable's performance.

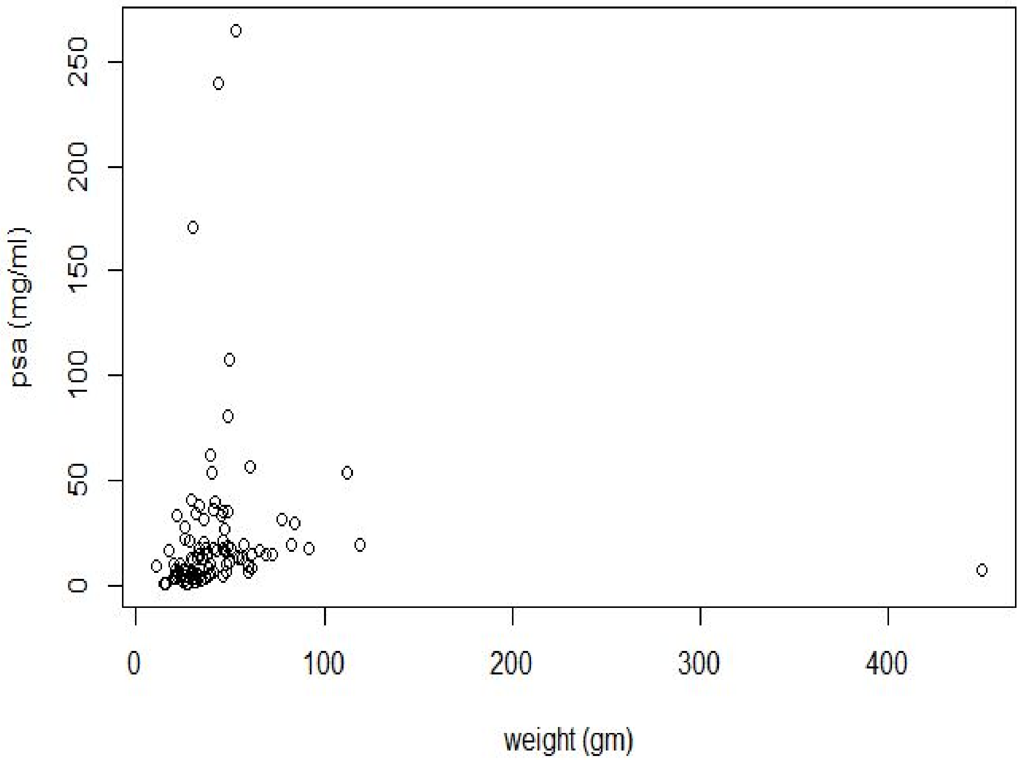
```
#Import the data set
prostate1 <- read.table(file="prostate_cancer.csv", sep=",", header=T)
```

```
#Attach the data set values
attach(prostate1)
```

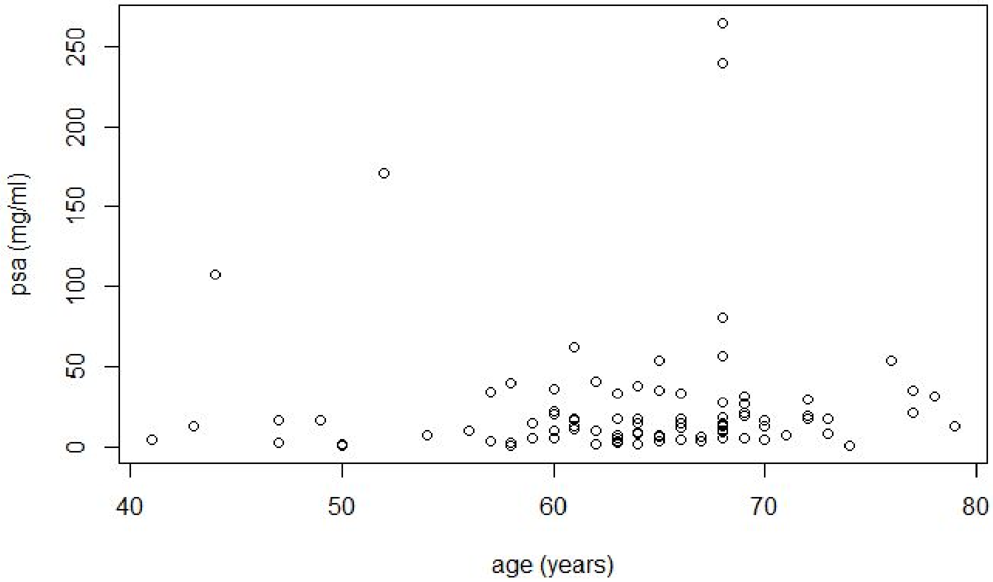
```
#Scatterplot - psa vs cancervol
plot(cancervol, psa, xlab="cancervol (cc)", ylab="psa (mg/ml)")
```



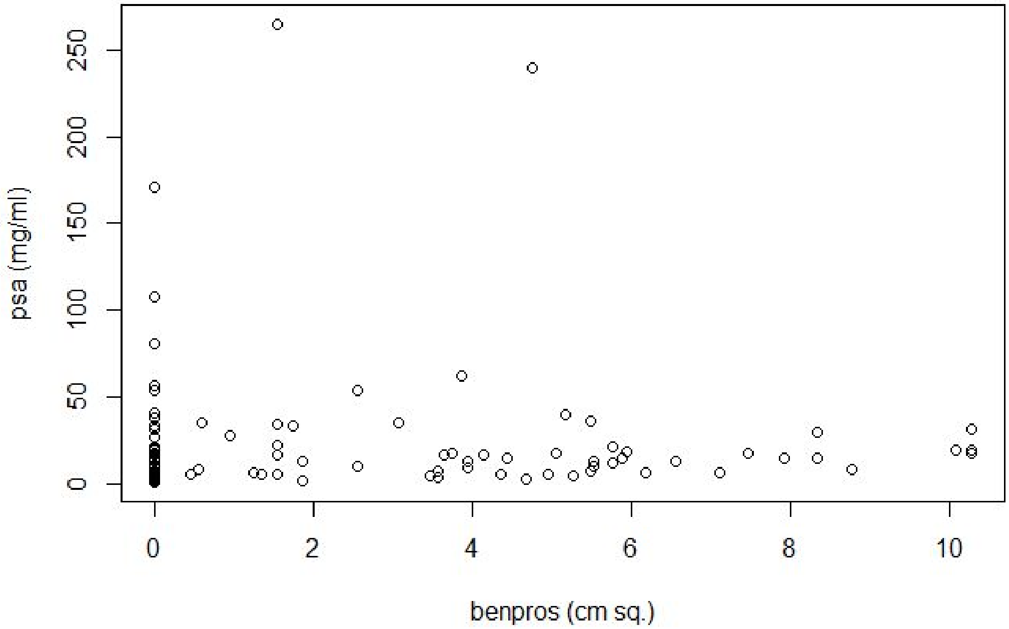
```
#Scatterplot - psa vs weight
plot(weight, psa, xlab="weight (gm)", ylab="psa (mg/ml)")
```



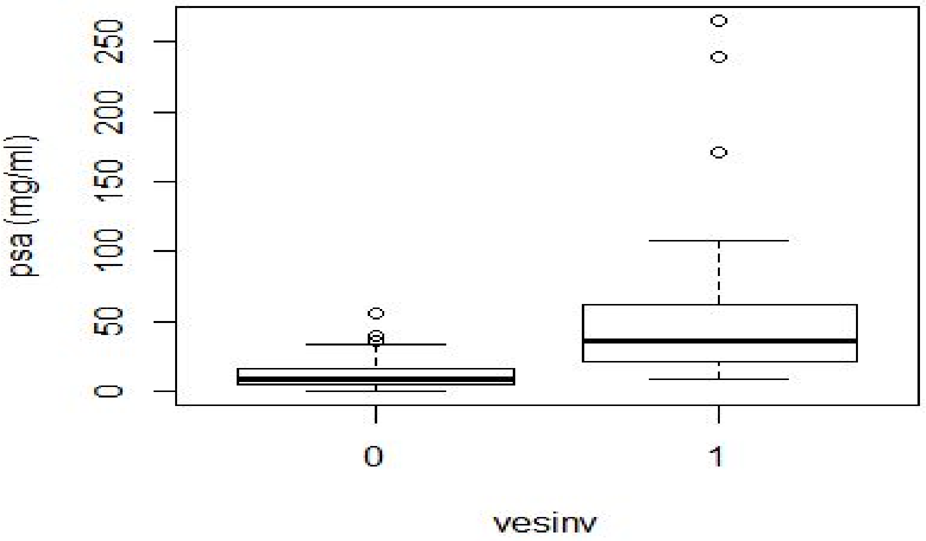
```
#Scatterplot - psa vs age
plot(age, psa, xlab="age (years)", ylab="psa (mg/ml)")
```



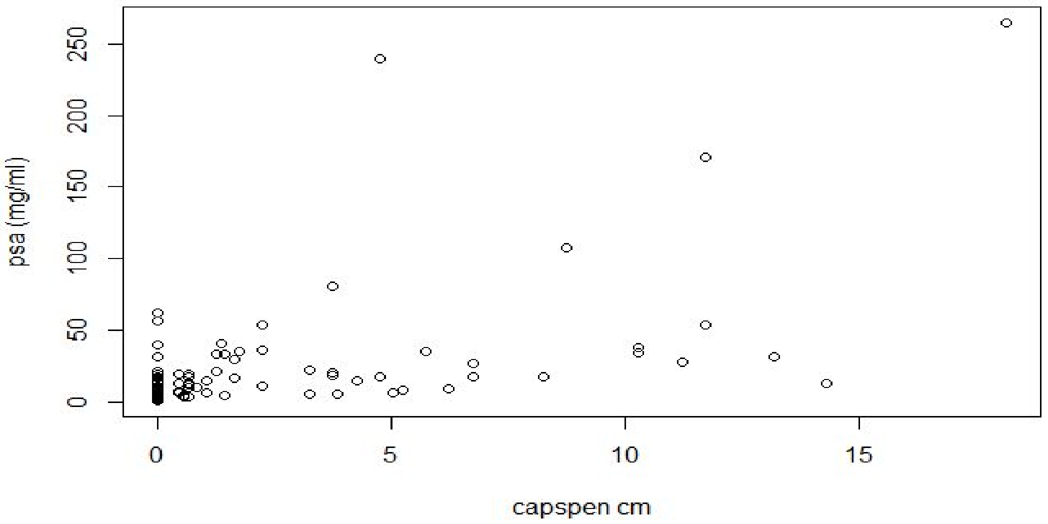
```
#Scatterplot - psa vs benpros
plot(benpros, psa, xlab="benpros (cm sq.)", ylab="psa (mg/ml)")
```



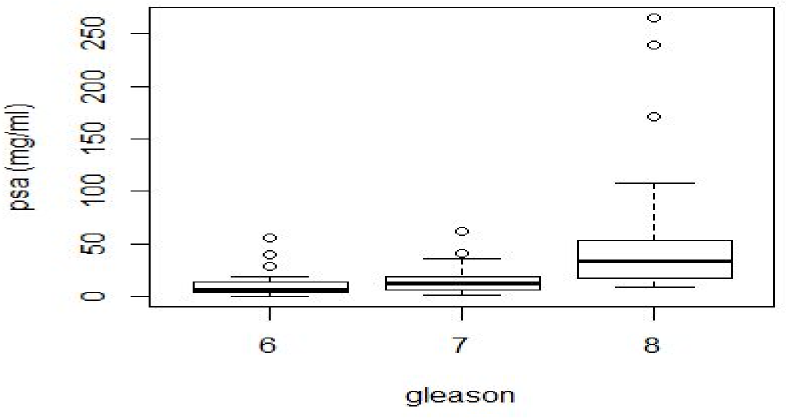
```
#Scatterplot - psa vs vesinv
plot(factor(vesinv), psa, xlab="vesinv ", ylab="psa (mg/ml)")
```



```
#Scatterplot - psa vs capspen
plot(capspen, psa, xlab="capspen cm", ylab="psa (mg/ml)")
```



```
#Scatterplot - psa vs gleason
plot(factor(gleason), psa, xlab="gleason", ylab="psa (mg/ml)")
```



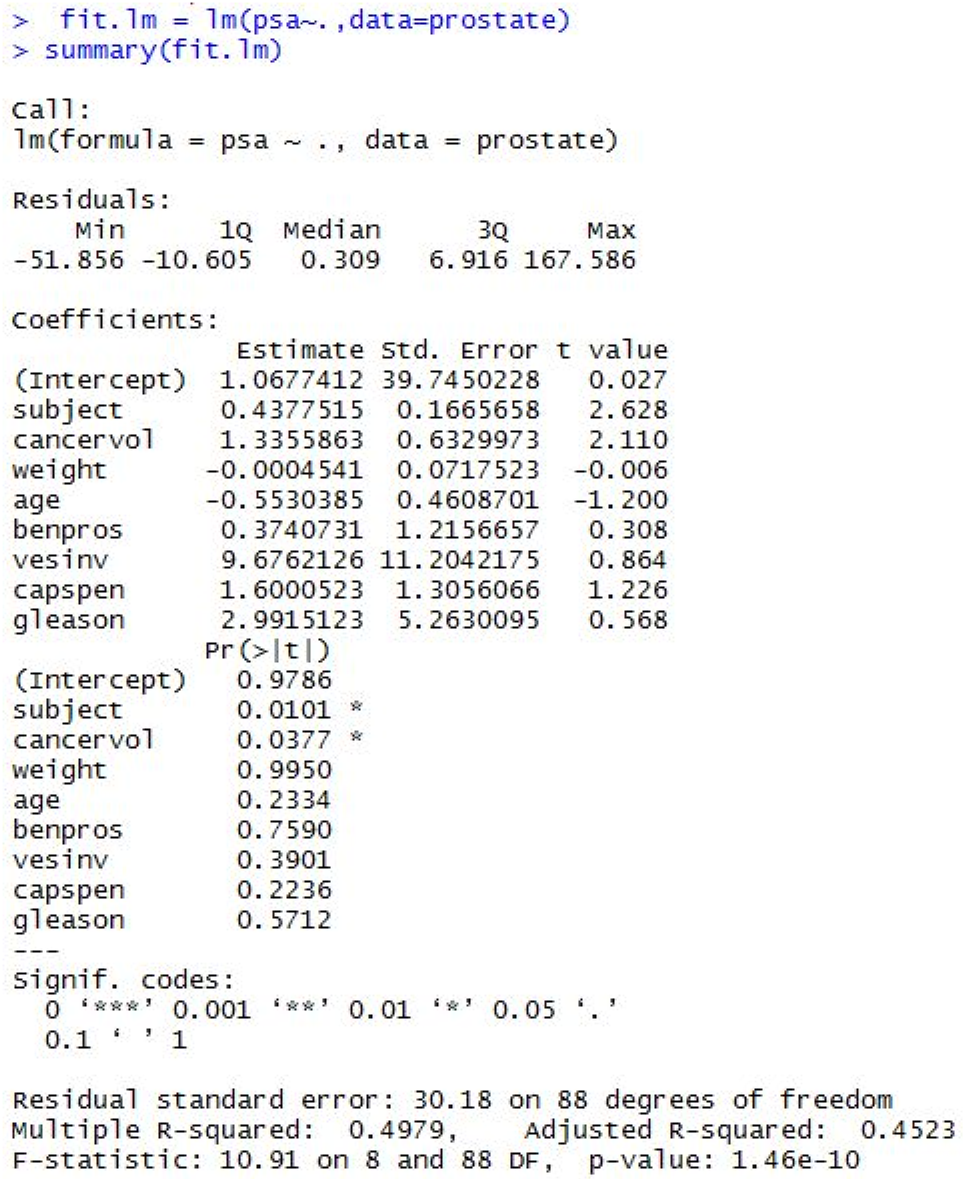
```
#to generate correlations between psa and cancervol
 cor(psa,cancervol)
[1] 0.6241506

#similarly, we find the correlation between psa and other variables

>cor(psa,weight)
[1] 0.02621343

>cor(psa,age)
[1] 0.01719938

>cor(psa,benpros)
[1] -0.01648649

>cor(psa,capspen)
[1] 0.5507925
```

- We can see the correlation is positive and moderately strong for the quantitative variable cancervol.
- So, from the univariate analysis, our predictor for the data is cancervol as the correlation between the PSA level and cancevol is a linear positive relationship from the scatter plot.

MODEL BUILDING:
-Multivariate Analysis:

```
#A sensible solution to picking a good model is to Minimize AIC.

#We wish to find which ones (if any) of the explanatory variables are important.
#We'll start off with including every variable as-is
#To do this, we first fit the full linear model:
```

```
>  fit.lm = lm(psa~.,data=prostate1)
> summary(fit.lm)
```



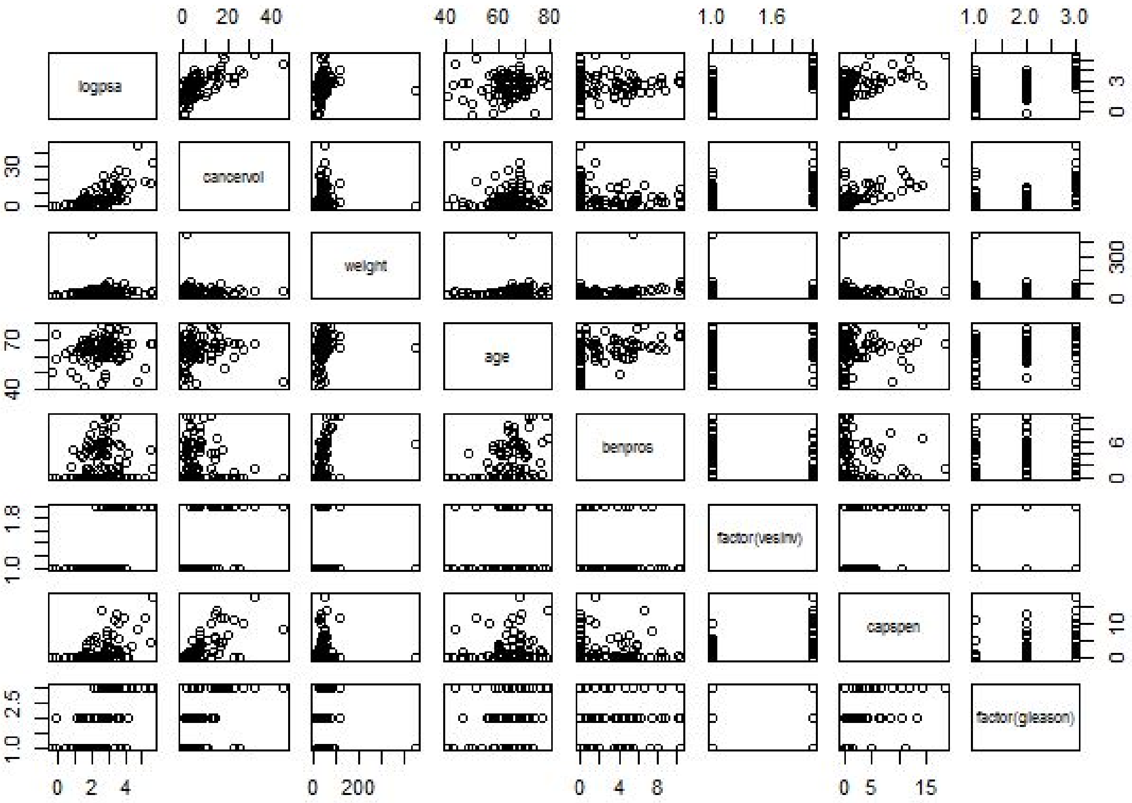
```
>  fit.lm = lm(psa~.,data=prostate)
> summary(fit.lm)

Call:
lm(formula = psa ~ ., data = prostate)

Residuals:
    Min      1Q  Median      3Q     Max 
-51.856 -10.605   0.309   6.916 167.586 

Coefficients:
              Estimate Std. Error t value
(Intercept)  1.0677412 39.7450228   0.027
subject      0.4377515  0.1665658   2.628
cancervol    1.3355863  0.6329973   2.110
weight      -0.0004541  0.0717523  -0.006
age         -0.5530385  0.4608701  -1.200
benpros      0.3740731  1.2156657   0.308
vesinv       9.6762126 11.2042175   0.864
capspen      1.6000523  1.3056066   1.226
gleason      2.9915123  5.2630095   0.568
            Pr(>|t|)  
(Intercept)   0.9786  
subject       0.0101 *
cancervol     0.0377 *
weight        0.9950  
age           0.2334  
benpros       0.7590  
vesinv        0.3901  
capspen       0.2236  
gleason       0.5712  
---
Signif. codes:  
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
  0.1 ‘ ’ 1

Residual standard error: 30.18 on 88 degrees of freedom
Multiple R-squared:  0.4979,    Adjusted R-squared:  0.4523 
F-statistic: 10.91 on 8 and 88 DF,  p-value: 1.46e-10
```

Observation:

Here, some of the predictors are unimportant.
(But there are too many parameters)

Transformations:

Often it is a good idea to transform the data to make it fit the assumptions of the model. Most commonly this means transforming the data to make the relationships more linear. In this case, let’s try a log transformation on the dependent variable psa



```
> logpsa <- log(psa)
> psa = NULL
> pairs(logpsa ~ cancervol + weight + age + benpros + factor(vesinv) + capspen + factor(gleason), data = prostate1)
```

#Let's fit the model again

```
> transformedfit <- lm(logpsa ~ ., data = prostate1)

> summary(transformedfit)

Call:
lm(formula = logpsa ~ ., data = prostate)

Residuals:
    Min       1Q   Median       3Q      Max
-1.11781 -0.09685  0.04960  0.17660  0.24607

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.6574253  0.3490985   1.883    0.063 .
subject      0.0334434  0.0015193  22.012  < 2e-16 ***
psa          0.0063728  0.0009363   6.806 1.22e-09 ***
cancervol    0.0032788  0.0056988   0.575    0.567
weight       0.0004085  0.0006302   0.648    0.519
age         -0.0005843  0.0040810  -0.143    0.886
benpros      0.0085855  0.0106834   0.804    0.424
vesinv      -0.1012339  0.0988275  -1.024    0.309
capspen      0.0047744  0.0115651   0.413    0.681
```

```
gleason      0.0024009  0.0463120   0.052    0.959
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2651 on 87 degrees of freedom
Multiple R-squared:  0.9521,            Adjusted R-squared:  0.9472
F-statistic: 192.3 on 9 and 87 DF,  p-value: < 2.2e-16
```

Observation:

The adjusted R-squared increased from 0.4523 to 0.9472, which is a huge increase. Based on these results, the log transformation of the dependent variables is definitely a good idea.

#But let us also try and see a Transformation using sqrt:

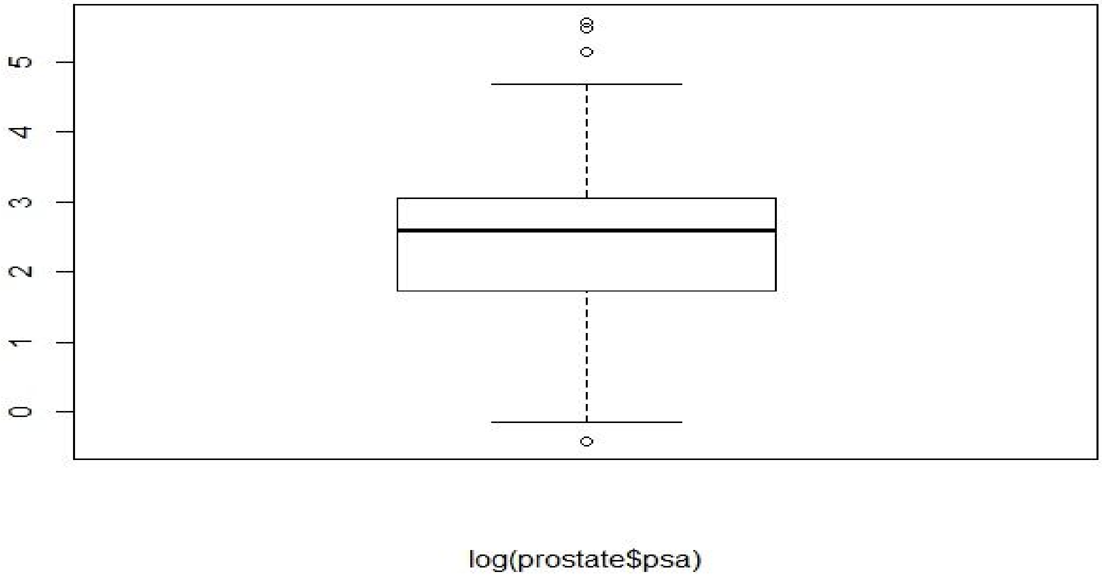
```
> sqrtpsa <-sqrt(prostate$psa)
> prostate$psa <- NULL
> pairs(sqrtpsa ~ cancervol + weight + age + benpros + factor(vesinv) + capspen + factor(gleason), data = prostate1)
> sqrtfit = lm(sqrtpsa ~ ., data = prostate1)

> summary(sqrtfit)

Call:
lm(formula = sqrtpsa ~ ., data = prostate)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4953 -0.5348  0.0312  0.3088  7.2946

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.2813319  1.7877952   0.717   0.4755
subject      0.0576572  0.0074924   7.695 1.93e-11 ***
cancervol    0.0719255  0.0284732   2.526   0.0133 *
weight       0.0002907  0.0032275   0.090   0.9284
age         -0.0295024  0.0207307  -1.423   0.1582
benpros      0.0143057  0.0546826   0.262   0.7942
vesinv       0.5118460  0.5039838   1.016   0.3126
capspen      0.0551802  0.0587283   0.940   0.3500
gleason      0.1557187  0.2367387   0.658   0.5124
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.358 on 88 degrees of freedom
Multiple R-squared:  0.7638,            Adjusted R-squared:  0.7423
F-statistic: 35.56 on 8 and 88 DF,  p-value: < 2.2e-16
```

Observation:

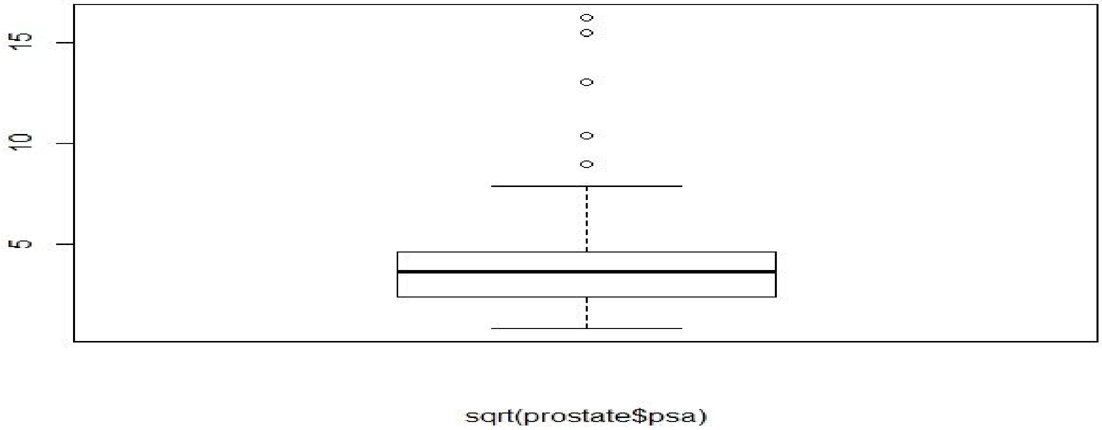
The adjusted R-squared increased from 0.4523 to 0.7423, which is a reasonable increase.

#Let’s see how the boxplot of the log and sqrt transformation look like:

```
> boxplot(log(prostate1$psa),xlab="log(prostate1$psa)")
```



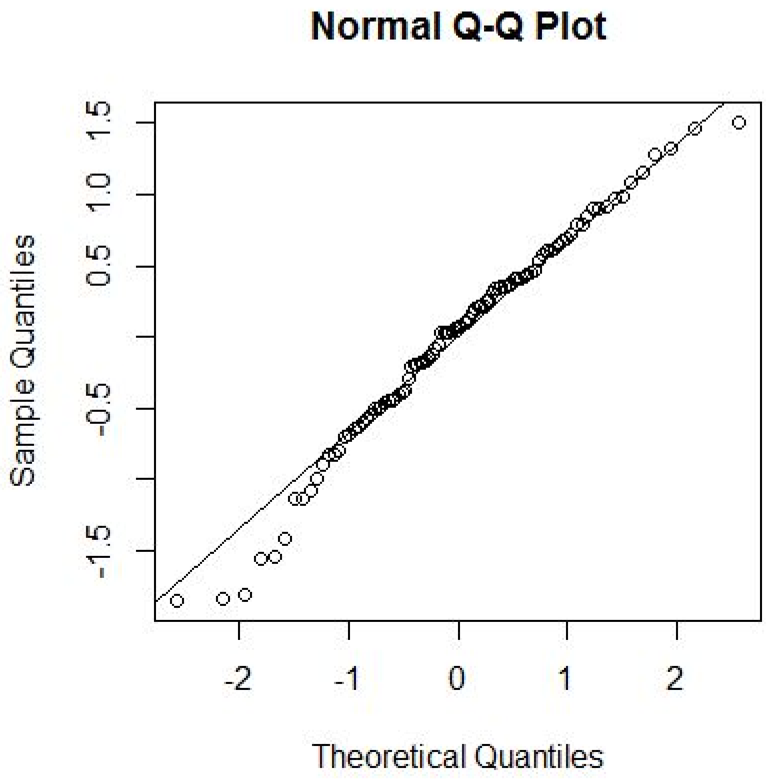
```
> boxplot(sqrt(prostate1$psa),xlab="sqrt(prostate1$psa)")
```



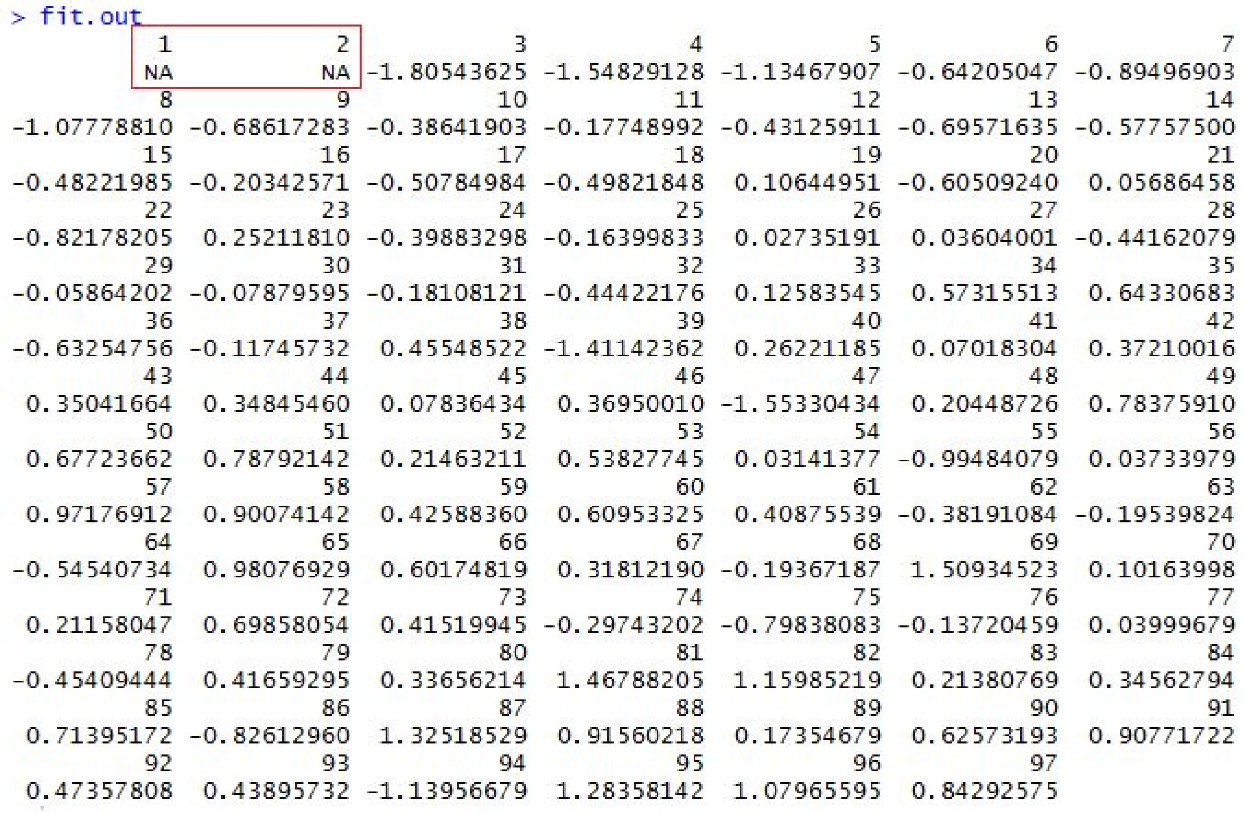
Observations of Box Plots:
We see that sqrt has more outliers than log transformation. So we can choose log as a better pick for transformation.

```
y = log(prostate1$psa)

#check for the residulas in the fit
 fit.check <- lm(y ~ cancervol + weight + age + benpros + factor(vesinv) + capspen + factor(gleason), data =
prostate1)

 #residuals in fit
 fit.res <- resid(fit.check)
```

```
#qqplot for residuals
 qqnorm(fit.res)
 qqline(fit.res)
```



```
#find the outliers using IQR values
 qnt = quantile(fit.res, probs=c(.25,.75))
 >qnt
      25%        75%
-0.4540944  0.4554852
  H = 1.5 * IQR(fit.res)
> H
>[1] 1.364369
 fit.out <- fit.res
 fit.out[fit.res < (qnt[1] - H)] <- NA
 fit.out[fit.res > (qnt[2] + H)] <- NA
```

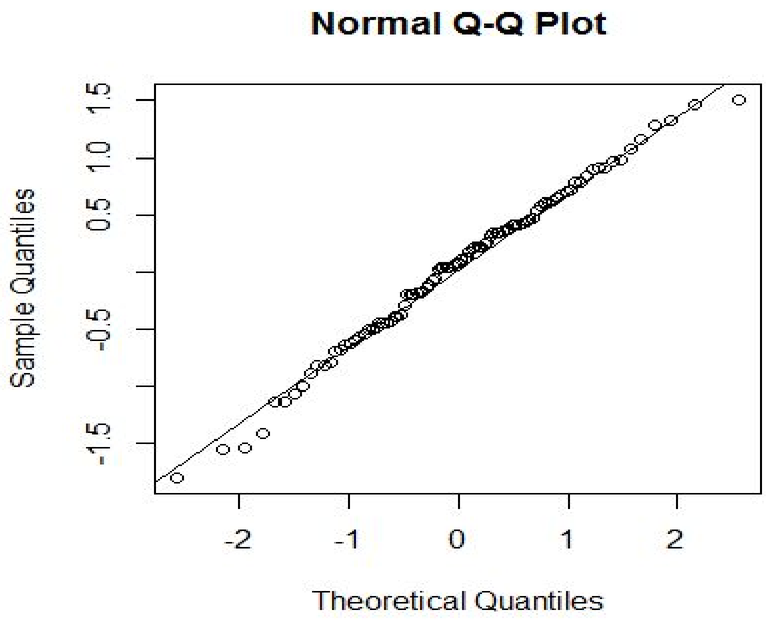


```
> fit.out
          1           2           3           4           5           6           7 
         NA          NA -1.80543625 -1.54829128 -1.13467907 -0.64205047 -0.89496903 
          8           9          10          11          12          13          14 
-1.07778810 -0.68617283 -0.38641903 -0.17748992 -0.43125911 -0.69571635 -0.57757500 
         15          16          17          18          19          20          21 
-0.48221985 -0.20342571 -0.50784984 -0.49821848  0.10644951 -0.60509240  0.05686458 
         22          23          24          25          26          27          28 
-0.82178205  0.25211810 -0.39883298 -0.16399833  0.02735191  0.03604001 -0.44162079 
         29          30          31          32          33          34          35 
-0.05864202 -0.07879595 -0.18108121 -0.44422176  0.12583545  0.57315513  0.64330683 
         36          37          38          39          40          41          42 
-0.63254756 -0.11745732  0.45548522 -1.41142362  0.26221185  0.07018304  0.37210016 
         43          44          45          46          47          48          49 
 0.35041664  0.34845460  0.07836434  0.36950010 -1.55330434  0.20448726  0.78375910 
         50          51          52          53          54          55          56 
 0.67723662  0.78792142  0.21463211  0.53827745  0.03141377 -0.99484079  0.03733979 
         57          58          59          60          61          62          63 
 0.97176912  0.90074142  0.42588360  0.60953325  0.40875539 -0.38191084 -0.19539824 
         64          65          66          67          68          69          70 
-0.54540734  0.98076929  0.60174819  0.31812190 -0.19367187  1.50934523  0.10163998 
         71          72          73          74          75          76          77 
 0.21158047  0.69858054  0.41519945 -0.29743202 -0.79838083 -0.13720459  0.03999679 
         78          79          80          81          82          83          84 
-0.45409444  0.41659295  0.33656214  1.46788205  1.15985219  0.21380769  0.34562794 
         85          86          87          88          89          90          91 
 0.71395172 -0.82612960  1.32518529  0.91560218  0.17354679  0.62573193  0.90771722 
         92          93          94          95          96          97 
 0.47357808  0.43895732 -1.13956679  1.28358142  1.07965595  0.84292575 
```

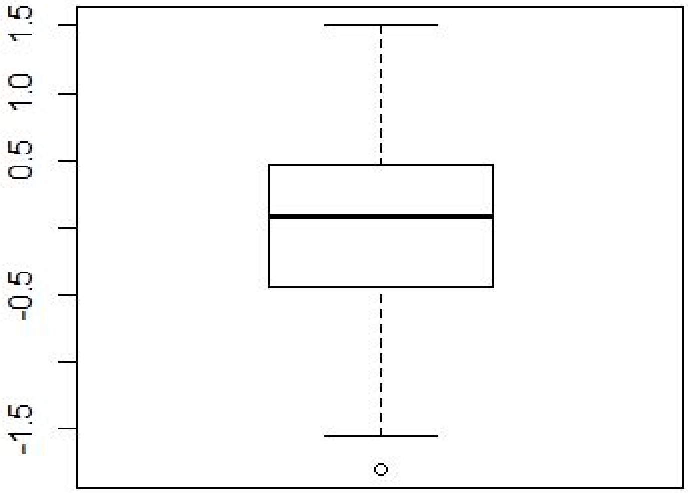
#We observe that the first 2 values have come out as NA, means they are the outlier values.

qqnorm(fit.out)

qqline(fit.out)



```
boxplot(fit.out)
```



```
##############now removing the outliers from the data(first 2 values)
prostate = tail(prostate1, -2)

##set y
y = log(prostate$psa)
```

Note:
prostate1 was our initial dataset.
prostate is now the new dataset after the removal of residuals.

```
###############check for the best fit -manually
#full
fit.full <- lm(y ~ cancervol + weight + age + benpros + factor(vesinv) + capspen + factor(gleason), data =
prostate)

> summary(fit.full)
#
# Call:
#   lm(formula = y ~ cancervol + weight + age + benpros + factor(vesinv) +
#       capspen + factor(gleason), data = prostate)
#
# Residuals:
#   Min       1Q   Median       3Q      Max
# -1.83835 -0.47139  0.05644  0.45003  1.46694
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)       1.991266   0.692637   2.875  0.00509 **
#   cancervol         0.062206   0.014474   4.298 4.53e-05 ***
#   weight            0.001123   0.001734   0.647  0.51907
# age             -0.009134   0.011191  -0.816  0.41665
# benpros          0.082172   0.028201   2.914  0.00455 **
#   factor(vesinv)1   0.791629   0.253819   3.119  0.00247 **
#   capspen         -0.025558   0.030948  -0.826  0.41118
# factor(gleason)7  0.308906   0.179597   1.720  0.08903 .
# factor(gleason)8  0.768287   0.250041   3.073  0.00284 **
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 0.7231 on 86 degrees of freedom
# Multiple R-squared:  0.5986,         Adjusted R-squared:  0.5613
# F-statistic: 16.03 on 8 and 86 DF,  p-value: 3.066e-14
```

```
###########try1---removing weight from fit.full as it has high p-value

fit.try1 <- lm(y ~ cancervol +  age + benpros + factor(vesinv) + capspen + factor(gleason), data = prostate)

> summary(fit.try1)

Call:
lm(formula = y ~ cancervol + age + benpros + factor(vesinv) +
    capspen + factor(gleason), data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.85269 -0.48381  0.05273  0.44920  1.51187

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.011339   0.689630   2.917  0.00450 **
cancervol         0.062319   0.014425   4.320 4.12e-05 ***
age              -0.008738   0.011137  -0.785  0.43480
benpros           0.087693   0.026791   3.273  0.00153 **
factor(vesinv)1   0.797177   0.252826   3.153  0.00222 **
capspen          -0.025681   0.030844  -0.833  0.40735
factor(gleason)7  0.290387   0.176711   1.643  0.10393
factor(gleason)8  0.760915   0.248946   3.057  0.00297 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7207 on 87 degrees of freedom
Multiple R-squared:  0.5966,           Adjusted R-squared:  0.5642
F-statistic: 18.38 on 7 and 87 DF,  p-value: 8.333e-15



anova(fit.try1,fit.full)
# Analysis of Variance Table
#
# Model 1: y ~ cancervol + age + benpros + factor(vesinv) + capspen + factor(gleason)
# Model 2: y ~ cancervol + weight + age + benpros + factor(vesinv) + capspen +
#   factor(gleason)
# Res.Df    RSS Df Sum of Sq       F Pr(>F)
# 1     87 45.187
# 2     86 44.967  1   0.21918 0.4192 0.5191
```

# concluding fit.try1 as a better pick because there is an increase in F-statistic and p-value is > 0.05, accept Null Hypothesis, ie. fit.try2 is as good as fit.try1,we find weight as insignificant predictor,so we proceed with fit.try2.

```
###########try2---removing age as it has the highest pvalue in fit.try1
```

```
fit.try2 <- lm(y ~ cancervol + benpros + factor(vesinv) + capspen + factor(gleason), data = prostate)

> summary(fit.try2)

Call:
lm(formula = y ~ cancervol + benpros + factor(vesinv) + capspen +
    factor(gleason), data = prostate)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9566 -0.4941  0.0405  0.4521  1.4363

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.48341    0.15096   9.827 8.16e-16 ***
cancervol         0.06325    0.01434   4.409 2.92e-05 ***
benpros           0.08025    0.02500   3.210  0.00185 **
factor(vesinv)1   0.77965    0.25129   3.103  0.00258 **
capspen          -0.02613    0.03077  -0.849  0.39800
factor(gleason)7  0.27506    0.17524   1.570  0.12010
factor(gleason)8  0.72626    0.24446   2.971  0.00383 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7191 on 88 degrees of freedom
Multiple R-squared:  0.5938,            Adjusted R-squared:  0.5661
F-statistic: 21.44 on 6 and 88 DF,  p-value: 2.294e-15



anova(fit.try2,fit.try1)

# > anova(fit.try2,fit.try1)
# Analysis of Variance Table
#
# Model 1: y ~ cancervol + benpros + factor(vesinv) + capspen + factor(gleason)
# Model 2: y ~ cancervol + age + benpros + factor(vesinv) + capspen + factor(gleason)
# Res.Df    RSS Df Sum of Sq      F Pr(>F)
# 1     88 45.506
# 2     87 45.187  1   0.31976 0.6157 0.4348

#concluding fit.try2 as better because p-value > 0.05 and F-statistic of fit.try2 has increased.
```

```
######try3---removing capspen from try2

fit.try3 <- lm(y ~ cancervol + benpros + factor(vesinv) + factor(gleason), data = prostate)
> summary(fit.try3)

Call:
lm(formula = y ~ cancervol + benpros + factor(vesinv) + factor(gleason),
    data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95245 -0.51569  0.04707  0.46982  1.42865

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.49439    0.15017   9.951 4.03e-16 ***
cancervol         0.05797    0.01291   4.492 2.11e-05 ***
```

```
benpros          0.07964    0.02495   3.192  0.00196 **
factor(vesinv)1   0.68403    0.22431   3.050  0.00302 **
factor(gleason)7  0.26308    0.17440   1.508  0.13497
factor(gleason)8  0.70283    0.24252   2.898  0.00473 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.718 on 89 degrees of freedom
Multiple R-squared:  0.5905,           Adjusted R-squared:  0.5674
F-statistic: 25.66 on 5 and 89 DF,  p-value: 6.039e-16
```

```
anova(fit.try3,fit.try2)
#
# > anova(fit.try3,fit.try2)
# Analysis of Variance Table
#
# Model 1: y ~ cancervol + benpros + factor(vesinv) + factor(gleason)
# Model 2: y ~ cancervol + benpros + factor(vesinv) + capspen + factor(gleason)
# Res.Df    RSS Df Sum of Sq      F Pr(>F)
# 1     89 45.879
# 2     88 45.506  1   0.37304 0.7214  0.398
```

#Concluding fit.try3 as good because p-value > 0.05, we accept null hypothesis that try 2 and try 3 are same and F-statistic of fit.try2 has increased.

#########try4-------removing factor(gleason) from try3

```
fit.try4 <- lm(y ~ cancervol + benpros + factor(vesinv), data = prostate)
> summary(fit.try4)

Call:
lm(formula = y ~ cancervol + benpros + factor(vesinv), data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.82574 -0.53983  0.07311  0.54884  1.43414

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.62256    0.12833  12.644  < 2e-16 ***
cancervol        0.07161    0.01193   6.001 3.94e-08 ***
benpros          0.08672    0.02553   3.397 0.001013 **
factor(vesinv)1  0.80978    0.22552   3.591 0.000534 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.743 on 91 degrees of freedom
Multiple R-squared:  0.5516,           Adjusted R-squared:  0.5368
F-statistic: 37.32 on 3 and 91 DF,  p-value: 8.118e-16
```

```
anova(fit.try4,fit.try3)
#
# > anova(fit.try4,fit.try3)
# Analysis of Variance Table
#
# Model 1: y ~ cancervol + benpros + factor(vesinv)
# Model 2: y ~ cancervol + benpros + factor(vesinv) + factor(gleason)
# Res.Df    RSS Df Sum of Sq      F  Pr(>F)
# 1     91 50.230
# 2     89 45.879  2    4.3505 4.2197 0.01775 *
#   ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#gleason is a significant variable and we can omit it. Moreover, pval < 0.05, reject null hyp, which means try4 and try3 are not same. So we proceed with fit.try3.

Variable Selection: (Other conventional methods)

There are many methods for selecting model variables. We will use the following ones:

- ***Forward selection based on AIC***
- ***Backward elimination based on AIC***
- ***Stepwise Selection - Both forward/backward***

1. Forward selection based on AIC

1. Find the best possible one predictor model**
2. Keep that predictor and find the best possible two predictor model given the original predictor is selected.
3. Keep both those predictors and find the best possible three predictor model given the first two selected.
...
4. Keep going until adding another predictor no longer reduces AIC

**Assumption:
assume an intercept is included and always retained

```
#################### proceeding with AIC
#################### Forward selection based on AIC
```

```
fit.full.forward <- step(lm(y ~ 1, data = prostate),
                 scope = list(upper = ~cancervol + weight + age + benpros + factor(vesinv) + capspen +
factor(gleason)),
                 direction = "forward")
```

```
#
# Start:  AIC=17.66
# y ~ 1
#
# Df Sum of Sq     RSS     AIC
# + cancervol        1    48.996  63.029 -34.977
# + factor(vesinv)   1    37.423  74.603 -18.962
# + factor(gleason)  2    36.524  75.502 -15.823
# + capspen          1    30.616  81.409 -10.667
# <none>                         112.025  17.661
# + benpros          1     1.705 110.321  18.204
# + age              1     1.358 110.667  18.502
# + weight           1     1.158 110.867  18.674
#
# Step:  AIC=-34.98
# y ~ cancervol
#
# Df Sum of Sq    RSS     AIC
# + factor(gleason)  2    7.8988 55.130 -43.697
# + factor(vesinv)   1    6.4303 56.599 -43.200
# + benpros          1    5.6823 57.347 -41.952
# <none>                         63.029 -34.977
# + weight           1    1.2197 61.809 -34.833
# + age              1    1.1049 61.924 -34.657
# + capspen          1    0.9432 62.086 -34.409
```

```
#
# Step:  AIC=-43.7
# y ~ cancervol + factor(gleason)
#
# Df Sum of Sq    RSS     AIC
# + benpros        1    4.4568 50.673 -49.705
# + factor(vesinv)  1    3.9998 51.130 -48.852
# + weight         1    1.4575 53.673 -44.242
# <none>                       55.130 -43.697
# + age            1    0.1718 54.958 -41.994
# + capspen        1    0.1666 54.964 -41.985
#
# Step:  AIC=-49.71
# y ~ cancervol + factor(gleason) + benpros
#
# Df Sum of Sq    RSS     AIC
# + factor(vesinv)  1    4.7940 45.879 -57.147
# <none>                       50.673 -49.705
# + weight         1    0.2793 50.394 -48.230
# + capspen        1    0.1891 50.484 -48.060
# + age            1    0.1136 50.560 -47.918
#
# Step:  AIC=-57.15
# y ~ cancervol + factor(gleason) + benpros + factor(vesinv)
#
# Df Sum of Sq    RSS     AIC
# <none>                45.879 -57.147
# + capspen  1   0.37304 45.506 -55.922
# + age      1   0.33275 45.547 -55.838
# + weight   1   0.19336 45.686 -55.548
```

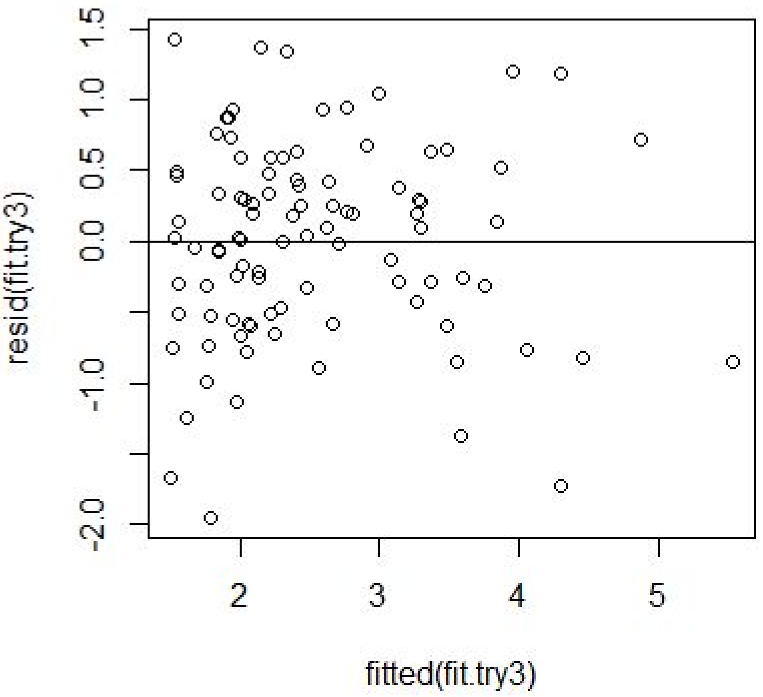
############ Backward elimination based on AIC

```
fit.full.backward <- step(lm(y ~ cancervol + weight + age + benpros + factor(vesinv) + capspen +
factor(gleason), data = prostate),
                  scope = list(lower = ~1), direction = "backward")
#
# Start:  AIC=-53.05
# y ~ cancervol + weight + age + benpros + factor(vesinv) + capspen +
#   factor(gleason)
#
# Df Sum of Sq    RSS     AIC
# - weight          1    0.2192 45.187 -54.592
# - age             1    0.3483 45.316 -54.321
# - capspen         1    0.3566 45.324 -54.304
# <none>                        44.967 -53.054
# - factor(gleason)  2    4.9839 49.951 -47.069
# - benpros         1    4.4394 49.407 -46.110
# - factor(vesinv)   1    5.0862 50.054 -44.874
# - cancervol       1    9.6576 54.625 -36.572
#
# Step:  AIC=-54.59
# y ~ cancervol + age + benpros + factor(vesinv) + capspen + factor(gleason)
#
# Df Sum of Sq    RSS     AIC
# - age             1    0.3198 45.506 -55.922
# - capspen         1    0.3601 45.547 -55.838
# <none>                        45.187 -54.592
# - factor(gleason)  2    4.8786 50.065 -48.852
# - factor(vesinv)   1    5.1637 50.350 -46.313
# - benpros          1    5.5647 50.751 -45.559
# - cancervol        1    9.6942 54.881 -38.128
#
# Step:  AIC=-55.92
# y ~ cancervol + benpros + factor(vesinv) + capspen + factor(gleason)
#
# Df Sum of Sq    RSS     AIC
# - capspen          1    0.3730 45.879 -57.147
# <none>                         45.506 -55.922
# - factor(gleason)  2    4.5887 50.095 -50.796
# - factor(vesinv)   1    4.9780 50.484 -48.060
# - benpros          1    5.3280 50.834 -47.404
# - cancervol        1   10.0544 55.561 -38.958
#
# Step:  AIC=-57.15
# y ~ cancervol + benpros + factor(vesinv) + factor(gleason)
#
# Df Sum of Sq    RSS     AIC
# <none>                         45.879 -57.147
# - factor(gleason)  2    4.3505 50.230 -52.540
# - factor(vesinv)   1    4.7940 50.673 -49.705
# - benpros          1    5.2510 51.130 -48.852
# - cancervol        1   10.4003 56.280 -39.737
```

############### Both forward/backward

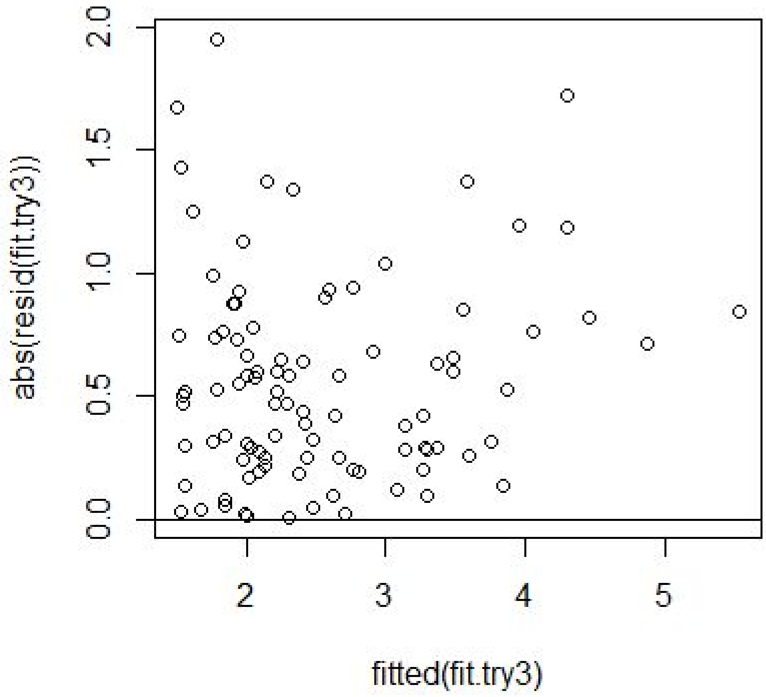
```
fit.full.both <- step(lm(y ~ 1, data = prostate),
                scope = list(lower = ~1, upper = ~cancervol + weight + age + benpros + factor(vesinv) + capspen +
factor(gleason)),
                direction = "both")
#
# Start:  AIC=17.66
# y ~ 1
#
# Df Sum of Sq     RSS     AIC
# + cancervol        1    48.996  63.029 -34.977
# + factor(vesinv)   1    37.423  74.603 -18.962
# + factor(gleason)  2    36.524  75.502 -15.823
# + capspen          1    30.616  81.409 -10.667
# <none>                         112.025  17.661
# + benpros          1     1.705 110.321  18.204
# + age              1     1.358 110.667  18.502
# + weight           1     1.158 110.867  18.674
#
# Step:  AIC=-34.98
# y ~ cancervol
#
# Df Sum of Sq     RSS     AIC
# + factor(gleason)  2     7.899  55.130 -43.697
# + factor(vesinv)   1     6.430  56.599 -43.200
# + benpros          1     5.682  57.347 -41.952
# <none>                          63.029 -34.977
# + weight           1     1.220  61.809 -34.833
# + age              1     1.105  61.924 -34.657
# + capspen          1     0.943  62.086 -34.409
# - cancervol        1    48.996 112.025  17.661
#
# Step:  AIC=-43.7
# y ~ cancervol + factor(gleason)
```

```
#
# Df Sum of Sq    RSS     AIC
# + benpros          1    4.4568 50.673 -49.705
# + factor(vesinv)   1    3.9998 51.130 -48.852
# + weight           1    1.4575 53.673 -44.242
# <none>                         55.130 -43.697
# + age              1    0.1718 54.958 -41.994
# + capspen          1    0.1666 54.964 -41.985
# - factor(gleason)  2    7.8988 63.029 -34.977
# - cancervol        1   20.3718 75.502 -15.823
#
# Step:  AIC=-49.71
# y ~ cancervol + factor(gleason) + benpros
#
# Df Sum of Sq    RSS     AIC
# + factor(vesinv)   1    4.7940 45.879 -57.147
# <none>                         50.673 -49.705
# + weight           1    0.2793 50.394 -48.230
# + capspen          1    0.1891 50.484 -48.060
# + age              1    0.1136 50.560 -47.918
# - benpros          1    4.4568 55.130 -43.697
# - factor(gleason)  2    6.6734 57.347 -41.952
# - cancervol        1   22.7560 73.429 -16.467
#
# Step:  AIC=-57.15
# y ~ cancervol + factor(gleason) + benpros + factor(vesinv)
#
# Df Sum of Sq    RSS     AIC
# <none>                         45.879 -57.147
# + capspen          1    0.3730 45.506 -55.922
# + age              1    0.3327 45.547 -55.838
# + weight           1    0.1934 45.686 -55.548
# - factor(gleason)  2    4.3505 50.230 -52.540
# - factor(vesinv)   1    4.7940 50.673 -49.705
# - benpros          1    5.2510 51.130 -48.852
# - cancervol        1   10.4003 56.280 -39.737

###########Here are the results for AIC
#for step forward------y ~ cancervol + factor(gleason) + benpros + factor(vesinv)
#for step backward----y ~ cancervol + benpros + factor(vesinv) + factor(gleason)
#for step both-------- y ~ cancervol + factor(gleason) + benpros + factor(vesinv)

#In all the three AIC, we get the model with these four variables to be the best, which we also concluded for
fit.try3. So we choose fit.try3 as the best fit.

#So our final model will contain: cancervol + factor(gleason) + benpros + factor(vesinv)

#residual plot
plot(fitted(fit.try3), resid(fit.try3))
abline(h = 0)
```

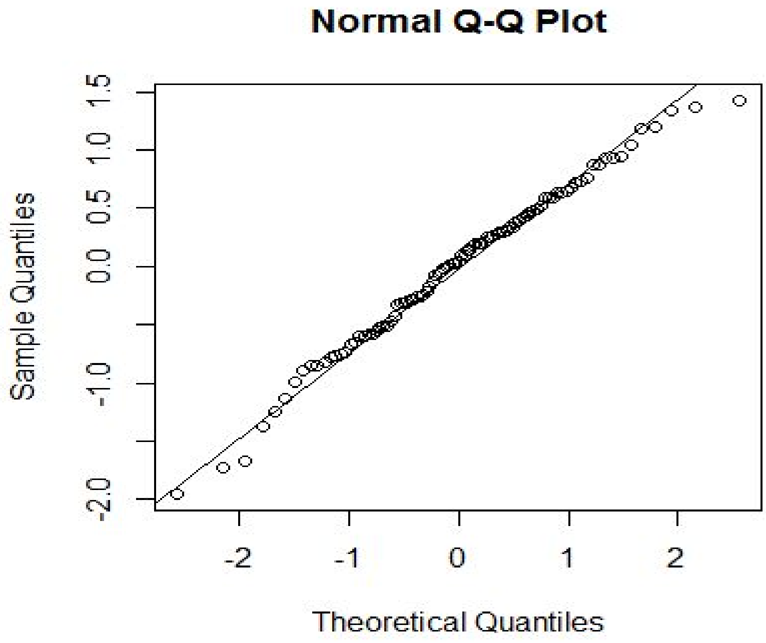


# plot of absolute residuals

```
plot(fitted(fit.try3), abs(resid(fit.try3)))
abline(h=0)
```



```
# normal QQ plot
qqnorm(resid(fit.try3))
qqline(resid(fit.try3))
```



#finding the modes of categorical variables vesinv and gleason

```
mode.vesinv = table(factor(vesinv))
new.vesinv = names(mode.vesinv)[mode.vesinv==max(mode.vesinv)]

mode.gleason = table(factor(gleason))
new.gleason = names(mode.gleason)[mode.gleason==max(mode.gleason)]
------------------------------------------------------------------------------------------
```

#we check the class type of each variable in prostate

```
sapply(prostate, class)
  subject       psa cancervol    weight       age   benpros    vesinv   capspen   gleason 
"integer" "numeric" "numeric" "numeric" "integer" "numeric" "integer" "numeric" "integer" 

------------------------------------------------------------------------------------------
```

#create x.new to use in predict function
#we use means for cancervol and benpros and
#modes for vesinv and gleason

```
x.new=data.frame(cancervol=mean(cancervol),benpros=mean(benpros),vesinv=new.vesinv,gleason=new.gleason)

> x.new
  cancervol  benpros vesinv gleason
1  7.136218 2.588087      0       7
```

#predict

```
> ans = predict(fit.try3,newdata=x.new)
> ans
      1 
2.37726 
```

#now, log(psa)=2.37726
#We need psa. So, psa = $e^{2.37726}$

```
> exp(ans)
       1 
10.77533 
```