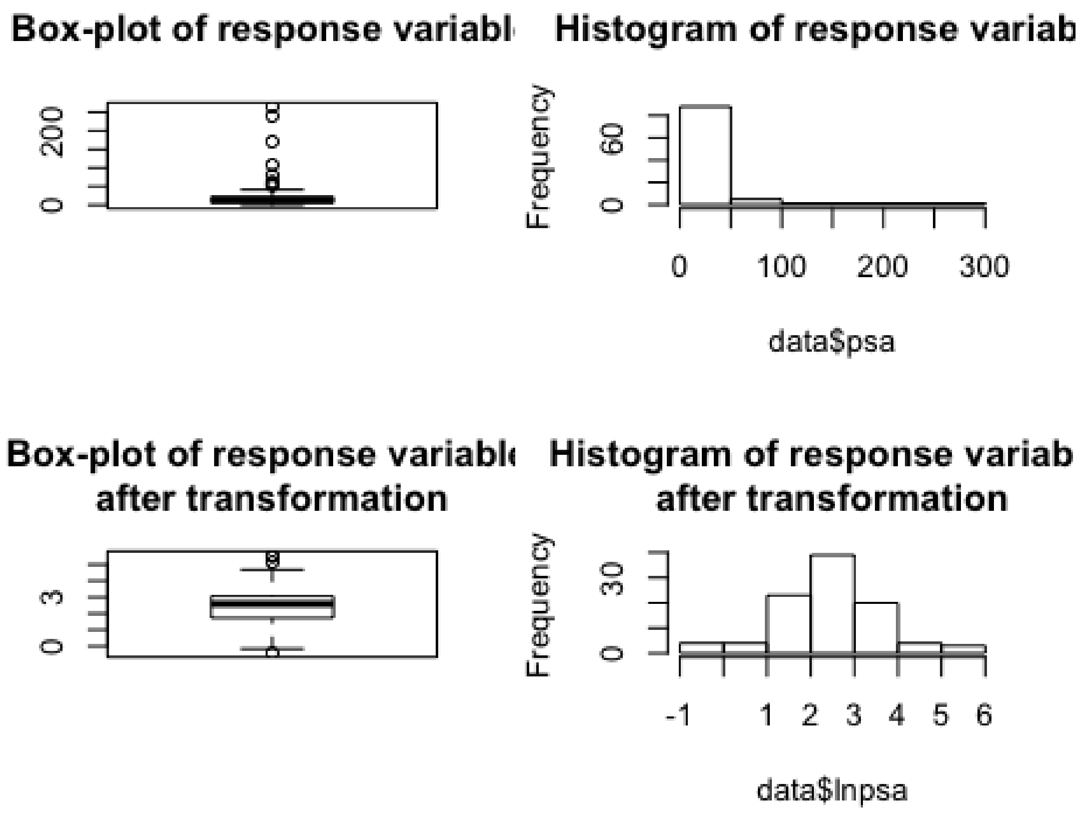
Model-Building

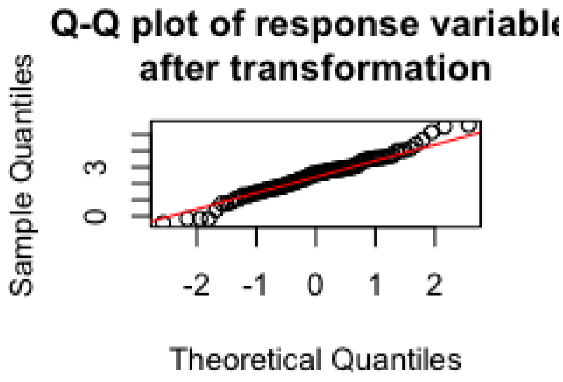
Build a model with correlation analysis:

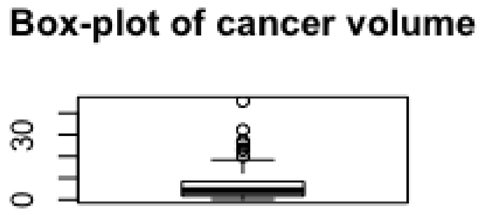
```
file <- "prostate_cancer.csv"
data <- read.csv(file,header = T)

par(mfrow=c(2,2))
#1 explore data and transformation
boxplot(data$psa,main="Box-plot of response variable")
hist(data$psa,main="Histogram of response variable")
data$lnpsa <- log(data$psa)
boxplot(data$lnpsa,main="Box-plot of response variable \nafter
transformation")
hist(data$lnpsa,main="Histogram of response variable \nafter transformation")
```

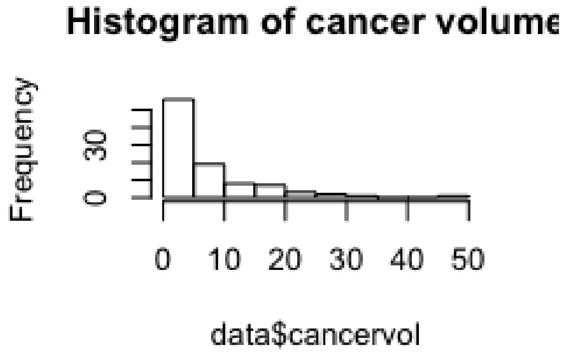


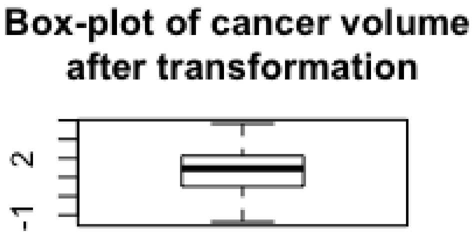
```
qqnorm(data$lnpsa, main="Q-Q plot of response variable \nafter
transformation")
qqline(data$lnpsa,col="red")

boxplot(data$cancervol,main="Box-plot of cancer volume")
```

```
hist(data$cancervol,main="Histogram of cancer volume")
data$lncan <- log(data$cancervol)
boxplot(data$lncan,main="Box-plot of cancer volume \nafter transformation")
```

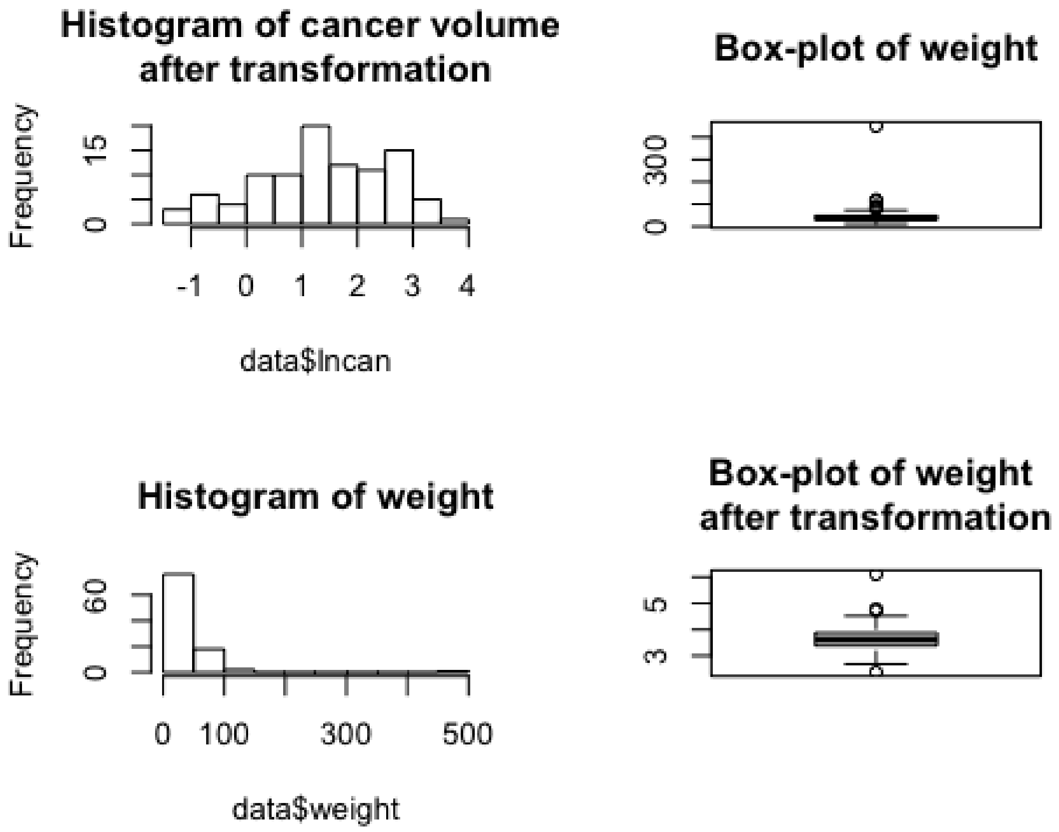




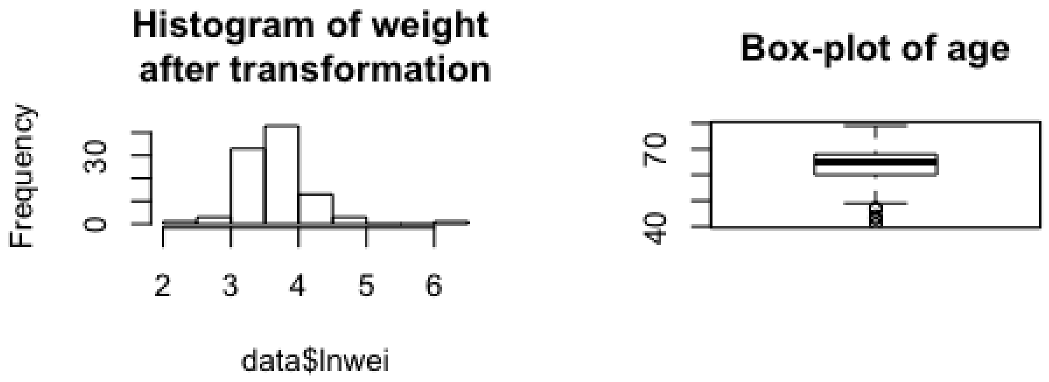


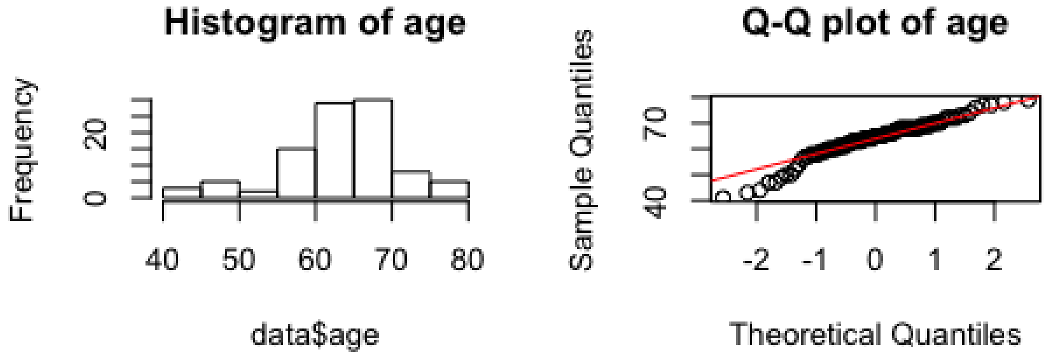


```
hist(data$lncan,main="Histogram of cancer volume \nafter transformation")

boxplot(data$weight,main="Box-plot of weight")
hist(data$weight,main="Histogram of weight")
data$lnwei <- log(data$weight)
boxplot(data$lnwei,main="Box-plot of weight \nafter transformation")
```

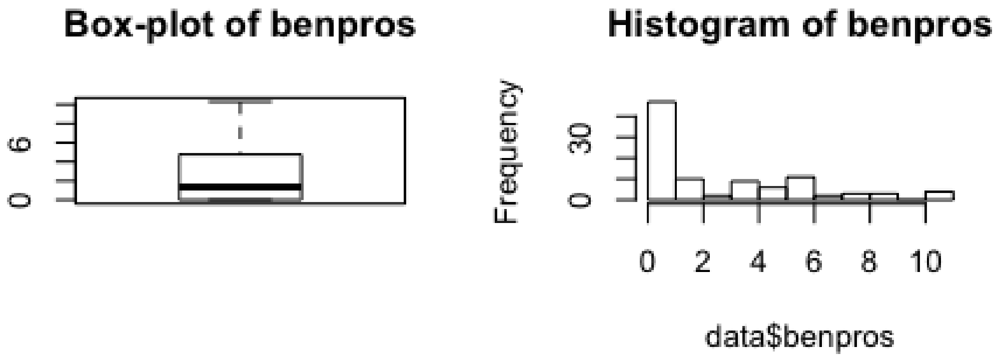


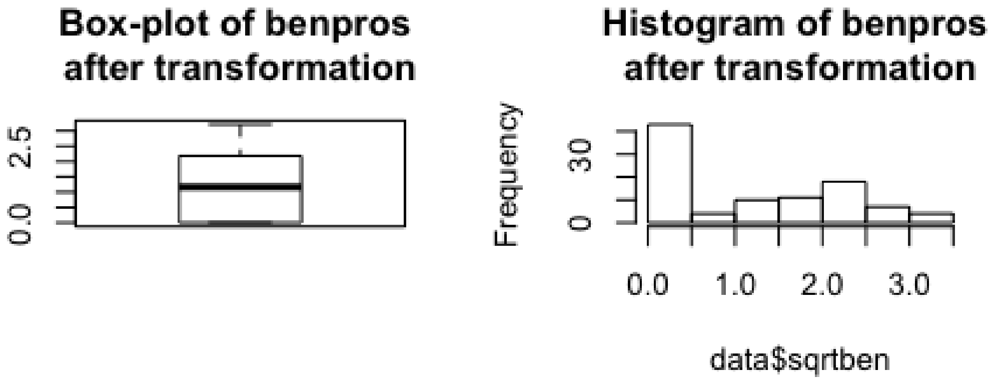
```r
hist(data$lnwei,main="Histogram of weight \nafter transformation")

boxplot(data$age,main="Box-plot of age")
hist(data$age,main="Histogram of age")
qqnorm(data$age, main="Q-Q plot of age")
qqline(data$age,col="red")
```



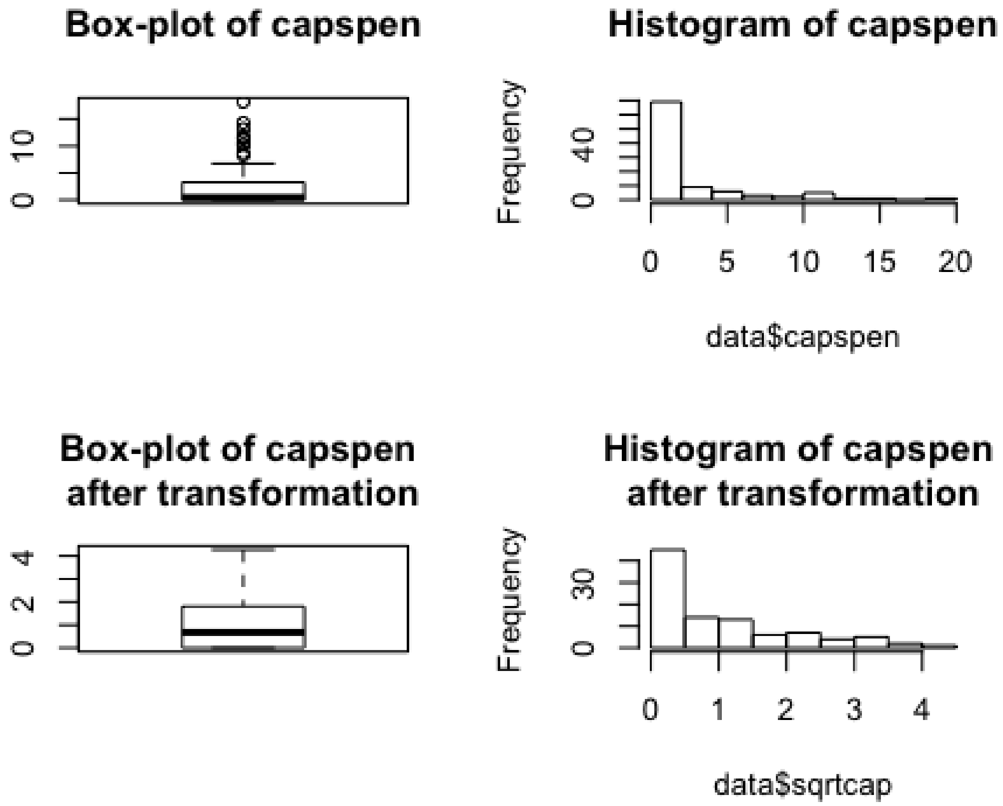


```r
boxplot(data$benpros,main="Box-plot of benpros")
hist(data$benpros,main="Histogram of benpros")
data$sqrtben <- sqrt(data$benpros)
boxplot(data$sqrtben,main="Box-plot of benpros \nafter transformation")
hist(data$sqrtben,main="Histogram of benpros \nafter transformation")
```

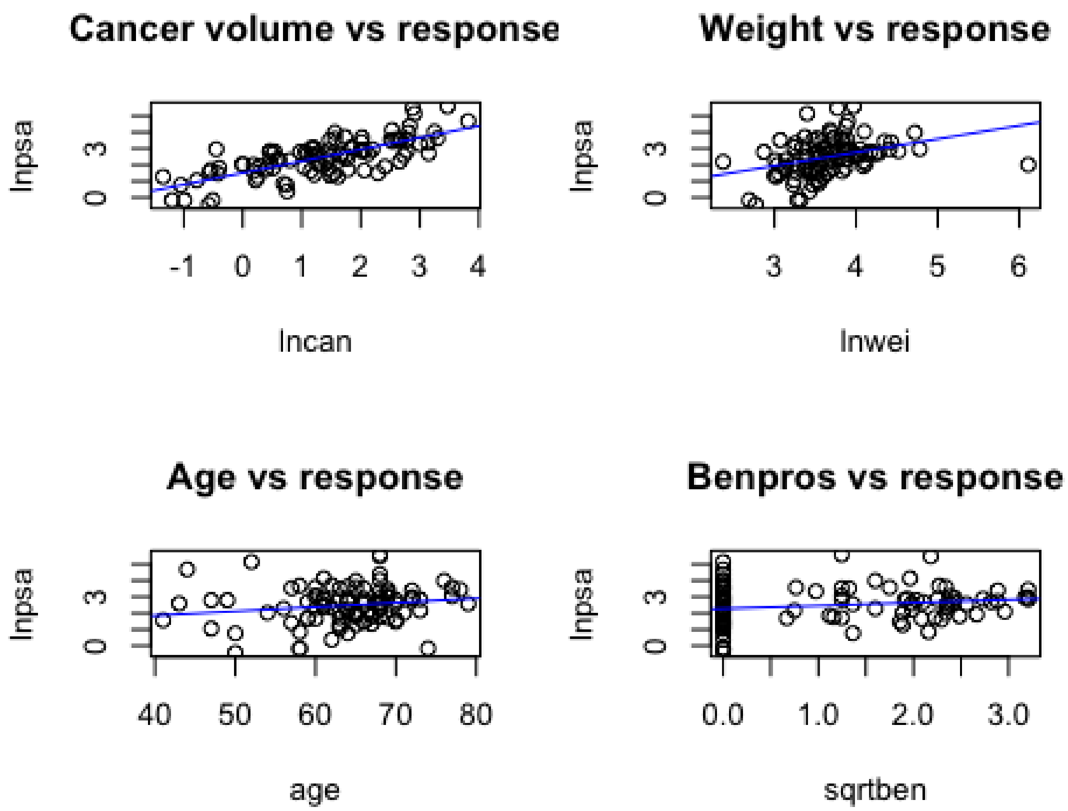




```
data$factves <- factor(data$vesinv)
table(data$vesinv)

##
##  0  1
## 76 21

boxplot(data$capspen,main="Box-plot of capspen")
hist(data$capspen,main="Histogram of capspen")
data$sqrtcap <- sqrt(data$capspen)
boxplot(data$sqrtcap,main="Box-plot of capspen \nafter transformation")
hist(data$sqrtcap,main="Histogram of capspen \nafter transformation")
```



```
table(data$gleason)

##
##  6  7  8
## 33 43 21

attach(data)

#2 explore relationship between individual columns vs response variable
plot(lncan,lnpsa,main="Cancer volume vs response")
abline(lm(lnpsa~lncan), col="blue")
plot(lnwei,lnpsa,main="Weight vs response")
abline(lm(lnpsa~lnwei), col="blue")
plot(age,lnpsa,main="Age vs response")
abline(lm(lnpsa~age), col="blue")
plot(sqrtben,lnpsa,main="Benpros vs response")
abline(lm(lnpsa~sqrtben), col="blue")
```



```
plot(vesinv,lnpsa,main="Vesinv vs response")
abline(lm(lnpsa~vesinv), col="blue")
plot(sqrtcap,lnpsa,main="Capspen vs response")
abline(lm(lnpsa~sqrtcap), col="blue")
plot(gleason,lnpsa,main="Gleason vs response")
abline(lm(lnpsa~gleason), col="blue")


# 3 build initial models:
fit1 <- lm(lnpsa ~ lncan + lnwei + sqrtcap)
anova(fit1)

## Analysis of Variance Table
##
## Response: lnpsa
##           Df Sum Sq Mean Sq F value    Pr(>F)
## lncan      1 68.801  68.801 122.897 < 2.2e-16 ***
## lnwei      1  5.956   5.956  10.639  0.001549 **
## sqrtcap    1  0.948   0.948   1.694  0.196288
## Residuals 93 52.064   0.560
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit2a <- lm(lnpsa ~ lncan + lnwei + sqrtcap + age)
anova(fit2a)
```

```
## Analysis of Variance Table
##
## Response: lnpsa
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## lncan      1 68.801  68.801 122.4760 < 2.2e-16 ***
## lnwei      1  5.956   5.956  10.6026  0.001582 **
## sqrtcap    1  0.948   0.948   1.6882  0.197083
## age        1  0.383   0.383   0.6817  0.411143
## Residuals 92 51.681   0.562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit2b <- lm(lnpsa ~ lncan + lnwei + sqrtcap + factves)
anova(fit2b)
```

```
## Analysis of Variance Table
##
## Response: lnpsa
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## lncan      1 68.801  68.801 132.5263 < 2.2e-16 ***
## lnwei      1  5.956   5.956  11.4726  0.001041 **
## sqrtcap    1  0.948   0.948   1.8267  0.179826
## factves    1  4.302   4.302   8.2871  0.004965 **
## Residuals 92 47.762   0.519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit2c <- lm(lnpsa ~ lncan + lnwei + sqrtcap + gleason)
anova(fit2c)
```

```
## Analysis of Variance Table
##
## Response: lnpsa
##           Df Sum Sq Mean Sq  F value  Pr(>F)
## lncan      1 68.801  68.801 130.1249 < 2e-16 ***
## lnwei      1  5.956   5.956  11.2647 0.00115 **
## sqrtcap    1  0.948   0.948   1.7936 0.18378
## gleason    1  3.421   3.421   6.4699 0.01264 *
## Residuals 92 48.643   0.529
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit3a <- lm(lnpsa ~ lncan + lnwei + factves + gleason)
anova(fit3a)
```

```
## Analysis of Variance Table
##
## Response: lnpsa
##           Df Sum Sq Mean Sq  F value    Pr(>F)    
## lncan      1 68.801  68.801 139.9962 < 2.2e-16 ***
## lnwei      1  5.956   5.956  12.1193 0.0007652 ***
## factves    1  5.194   5.194  10.5696 0.0016071 ** 
## gleason    1  2.605   2.605   5.2999 0.0235824 *  
## Residuals 92 45.213   0.491                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit3b <- lm(lnpsa ~ lncan + lnwei + factves + gleason + sqrtcap + age)
anova(fit3b)
```

```
## Analysis of Variance Table
## 
## Response: lnpsa
##           Df Sum Sq Mean Sq  F value    Pr(>F)    
## lncan      1 68.801  68.801 140.6106 < 2.2e-16 ***
## lnwei      1  5.956   5.956  12.1725 0.0007528 ***
## factves    1  5.194   5.194  10.6160 0.0015824 ** 
## gleason    1  2.605   2.605   5.3232 0.0233378 *  
## sqrtcap    1  0.395   0.395   0.8067 0.3715049    
## age        1  0.781   0.781   1.5971 0.2095851    
## Residuals 90 44.037   0.489                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit3a,fit3b)
```

```
## Analysis of Variance Table
## 
## Model 1: lnpsa ~ lncan + lnwei + factves + gleason
## Model 2: lnpsa ~ lncan + lnwei + factves + gleason + sqrtcap + age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     92 45.213                           
## 2     90 44.037  2    1.1761 1.2019 0.3054
```

```
# 4 model diagonostics using stepwise model selection
fit.forward <- step(lm(lnpsa ~ 1, data = data), scope = list(
  upper = ~lncan + lnwei + factves + gleason + sqrtcap + age + sqrtben),
  direction = "forward")
```

```
## Start:  AIC=28.72
## lnpsa ~ 1
## 
##           Df Sum of Sq     RSS     AIC
## + lncan    1    68.801  58.968 -44.278
## + factves  1    40.984  86.785  -6.794
## + sqrtcap  1    38.679  89.090  -4.251
## + gleason  1    37.122  90.647  -2.571
## + lnwei    1    15.985 111.784  17.760
## + sqrtben  1     4.031 123.738  27.615
## + age      1     3.688 124.080  27.883
## <none>                 127.769  28.725
##
## Step:  AIC=-44.28
## lnpsa ~ lncan
##
##           Df Sum of Sq    RSS     AIC
## + lnwei    1    5.9560 53.012 -52.606
## + factves  1    5.2731 53.695 -51.365
## + gleason  1    4.5889 54.379 -50.137
## + sqrtben  1    3.1654 55.803 -47.630
## <none>                 58.968 -44.278
## + sqrtcap  1    0.8366 58.131 -43.664
## + age      1    0.0031 58.965 -42.283
##
## Step:  AIC=-52.61
## lnpsa ~ lncan + lnwei
##
##           Df Sum of Sq    RSS     AIC
## + factves  1    5.1944 47.818 -60.610
## + gleason  1    4.2337 48.778 -58.680
## <none>                 53.012 -52.606
## + sqrtcap  1    0.9484 52.064 -52.357
## + sqrtben  1    0.5173 52.495 -51.558
## + age      1    0.4121 52.600 -51.363
##
## Step:  AIC=-60.61
## lnpsa ~ lncan + lnwei + factves
##
##           Df Sum of Sq    RSS     AIC
## + gleason  1   2.60463 45.213 -64.042
## + sqrtben  1   1.15018 46.668 -60.971
## <none>                 47.818 -60.610
## + age      1   0.39331 47.424 -59.411
## + sqrtcap  1   0.05616 47.762 -58.724
##
## Step:  AIC=-64.04
## lnpsa ~ lncan + lnwei + factves + gleason
##
##           Df Sum of Sq    RSS     AIC
## + sqrtben  1   1.02520 44.188 -64.267
## <none>                 45.213 -64.042
## + age      1   0.71746 44.496 -63.594
## + sqrtcap  1   0.39470 44.818 -62.893
##
## Step:  AIC=-64.27
## lnpsa ~ lncan + lnwei + factves + gleason + sqrtben
##
##           Df Sum of Sq    RSS     AIC
## + age      1   1.39891 42.789 -65.388
## <none>                 44.188 -64.267
## + sqrtcap  1   0.44007 43.748 -63.238
##
## Step:  AIC=-65.39
## lnpsa ~ lncan + lnwei + factves + gleason + sqrtben + age
##
##           Df Sum of Sq    RSS     AIC
## <none>                 42.789 -65.388
## + sqrtcap  1   0.55801 42.231 -64.661
```

```
fit.backward <- step(lm(lnpsa ~ lncan + lnwei + factves + gleason + sqrtcap +
age + sqrtben,
  data = data), scope = list(lower = ~1), direction = "backward")
```

```
## Start:  AIC=-64.66
## lnpsa ~ lncan + lnwei + factves + gleason + sqrtcap + age + sqrtben
##
##           Df Sum of Sq    RSS     AIC
## - sqrtcap  1    0.5580 42.789 -65.388
## <none>                 42.231 -64.661
## - age      1    1.5169 43.748 -63.238
## - sqrtben  1    1.8060 44.037 -62.599
## - lnwei    1    2.9027 45.134 -60.213
## - gleason  1    3.3852 45.616 -59.182
## - factves  1    4.5804 46.811 -56.673
## - lncan    1   18.9521 61.183 -30.702
##
## Step:  AIC=-65.39
## lnpsa ~ lncan + lnwei + factves + gleason + age + sqrtben
##
##           Df Sum of Sq    RSS     AIC
## <none>                 42.789 -65.388
## - age      1    1.3989 44.188 -64.267
## - sqrtben  1    1.7067 44.496 -63.594
## - gleason  1    2.9291 45.718 -60.965
## - lnwei    1    3.0222 45.811 -60.768
## - factves  1    4.1357 46.925 -58.438
## - lncan    1   19.7174 62.506 -30.626
```

```
fit.both <- step(lm(lnpsa ~ 1, data = data), scope = list(
  lower = ~1, upper = ~lncan + lnwei + factves + gleason + sqrtcap + age +
sqrtben),
  direction = "both")
```

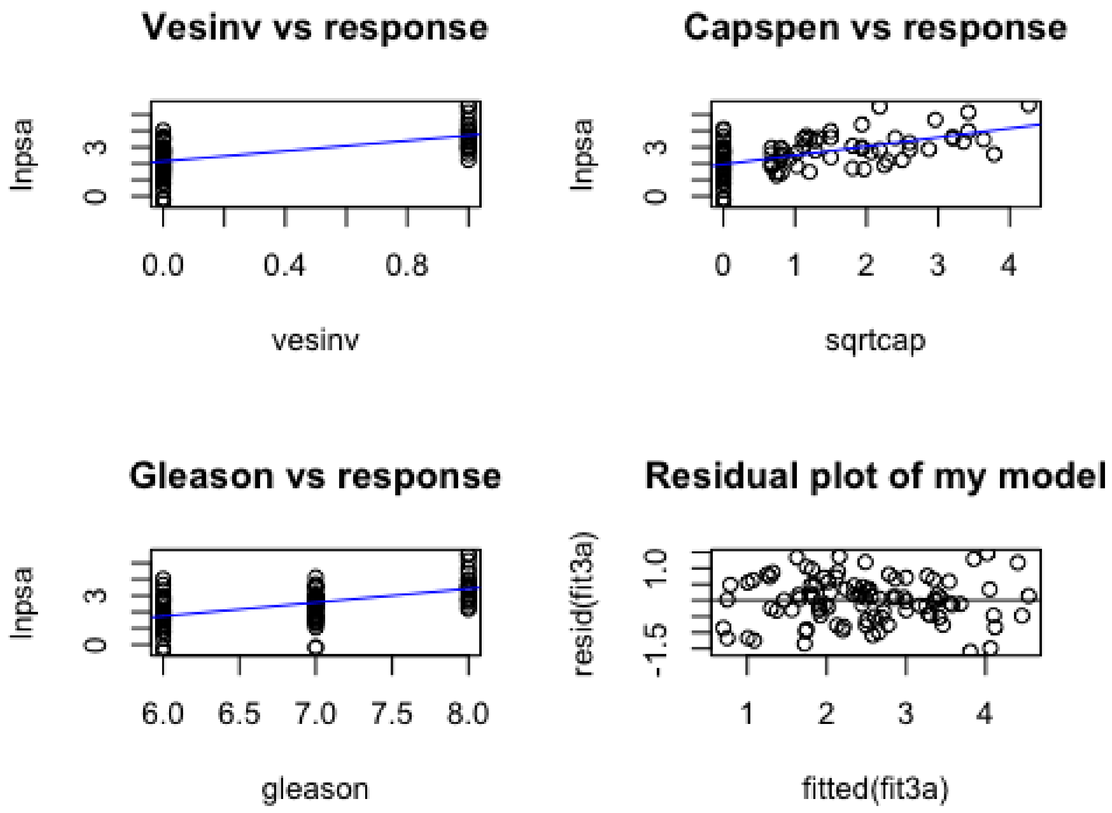
```
## Start:  AIC=28.72
## lnpsa ~ 1
##
##           Df Sum of Sq     RSS     AIC
## + lncan    1    68.801  58.968 -44.278
## + factves  1    40.984  86.785  -6.794
## + sqrtcap  1    38.679  89.090  -4.251
## + gleason  1    37.122  90.647  -2.571
## + lnwei    1    15.985 111.784  17.760
## + sqrtben  1     4.031 123.738  27.615
## + age      1     3.688 124.080  27.883
## <none>                 127.769  28.725
##
## Step:  AIC=-44.28
## lnpsa ~ lncan
##
##           Df Sum of Sq     RSS     AIC
## + lnwei    1     5.956  53.012 -52.606
## + factves  1     5.273  53.695 -51.365
## + gleason  1     4.589  54.379 -50.137
## + sqrtben  1     3.165  55.803 -47.630
## <none>                  58.968 -44.278
## + sqrtcap  1     0.837  58.131 -43.664
## + age      1     0.003  58.965 -42.283
## - lncan    1    68.801 127.769  28.725
##
## Step:  AIC=-52.61
## lnpsa ~ lncan + lnwei
##
##           Df Sum of Sq     RSS     AIC
## + factves  1     5.194  47.818 -60.610
## + gleason  1     4.234  48.778 -58.680
## <none>                  53.012 -52.606
## + sqrtcap  1     0.948  52.064 -52.357
## + sqrtben  1     0.517  52.495 -51.558
## + age      1     0.412  52.600 -51.363
## - lnwei    1     5.956  58.968 -44.278
## - lncan    1    58.772 111.784  17.760
##
## Step:  AIC=-60.61
## lnpsa ~ lncan + lnwei + factves
##
##           Df Sum of Sq    RSS     AIC
## + gleason  1    2.6046 45.213 -64.042
## + sqrtben  1    1.1502 46.668 -60.971
## <none>                 47.818 -60.610
## + age      1    0.3933 47.424 -59.411
## + sqrtcap  1    0.0562 47.762 -58.724
## - factves  1    5.1944 53.012 -52.606
## - lnwei    1    5.8772 53.695 -51.365
## - lncan    1   27.9829 75.801 -17.921
##
## Step:  AIC=-64.04
## lnpsa ~ lncan + lnwei + factves + gleason
##
##           Df Sum of Sq    RSS     AIC
## + sqrtben  1    1.0252 44.188 -64.267
## <none>                 45.213 -64.042
## + age      1    0.7175 44.496 -63.594
## + sqrtcap  1    0.3947 44.818 -62.893
## - gleason  1    2.6046 47.818 -60.610
## - factves  1    3.5653 48.778 -58.680
## - lnwei    1    5.6038 50.817 -54.709
## - lncan    1   18.8940 64.107 -32.173
##
## Step:  AIC=-64.27
## lnpsa ~ lncan + lnwei + factves + gleason + sqrtben
##
##           Df Sum of Sq    RSS     AIC
## + age      1    1.3989 42.789 -65.388
## <none>                 44.188 -64.267
## - sqrtben  1    1.0252 45.213 -64.042
## + sqrtcap  1    0.4401 43.748 -63.238
## - gleason  1    2.4797 46.668 -60.971
## - lnwei    1    2.5838 46.772 -60.755
## - factves  1    4.0873 48.275 -57.686
## - lncan    1   18.8602 63.048 -31.789
##
## Step:  AIC=-65.39
## lnpsa ~ lncan + lnwei + factves + gleason + sqrtben + age
##
##           Df Sum of Sq    RSS     AIC
## <none>                 42.789 -65.388
## + sqrtcap  1    0.5580 42.231 -64.661
## - age      1    1.3989 44.188 -64.267
## - sqrtben  1    1.7067 44.496 -63.594
## - gleason  1    2.9291 45.718 -60.965
## - lnwei    1    3.0222 45.811 -60.768
## - factves  1    4.1357 46.925 -58.438
## - lncan    1   19.7174 62.506 -30.626
```

```
# 5 compare my model against stepwise selection
anova(fit3a, fit.both)
```

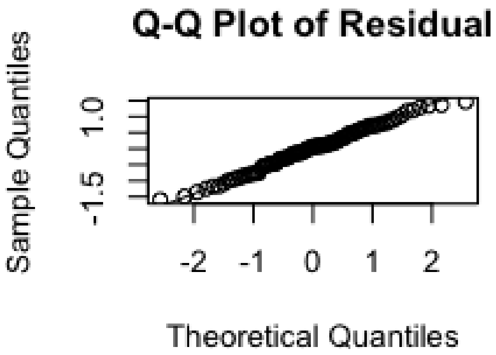
```
## Analysis of Variance Table
##
## Model 1: lnpsa ~ lncan + lnwei + factves + gleason
## Model 2: lnpsa ~ lncan + lnwei + factves + gleason + sqrtben + age
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     92 45.213
## 2     90 42.789  2    2.4241 2.5494 0.08376 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# 6 model evaluation
# residual plot
```

```
plot(fitted(fit3a), resid(fit3a), main="Residual plot of my model")
abline(h = 0)
```



```
# normal QQ plot
qqnorm(resid(fit3a),main="Q-Q Plot of Residual")
qqline(resid(fit3a))

# Time series plot of residuals, ignore because our data is not over time
#plot(resid(fit3a), type="l")
#abline(h=0)



# 7 add new patient data and make prediction
x.new <- data.frame("lncan"=log(mean(cancervol)),
          "lnwei"=log(mean(weight)),
          "factves"=factor(names(sort(table(vesinv),decreasing=TRUE)[1])),
          "gleason"=mean(gleason))

# predict new data
predict(fit3a, newdata=x.new)

##       1
## 2.72739
```


Q-Q Plot of Residual
Sample Quantiles
1.0
-1.5
-2
-1
0
1
2
Theoretical Quantiles