**MINI PROJECT #6**

**Name:** Omkar Nandkumar Dixit (ond170030)

**Name of group members:** Sushrut Patnaik (sxp175331)

**Contributions of each group members:** Equally contributed for the project

**Section 1:**

**Question:**

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: First, We can look at the relationship between response and each predictor one by one

Reading the data from file

inputData = read.csv(file="prostate\_cancer.csv", header=TRUE, sep = ",")

**Attaching the data so that we are able to access columns via direct names**

attach(inputData)

|  |
| --- |
| head(inputData)  subject psa cancervol weight age benpros vesinv capspen gleason  1 1 0.651 0.5599 15.959 50 0 0 0 6  2 2 0.852 0.3716 27.660 58 0 0 0 7  3 3 0.852 0.6005 14.732 74 0 0 0 7  4 4 0.852 0.3012 26.576 58 0 0 0 6  5 5 1.448 2.1170 30.877 62 0 0 0 6  6 6 2.160 0.3499 25.280 50 0 0 0 6  #using factor function for categorical data  vesinv=as.factor(vesinv)  **Normal QQplot for the response PSA** |
|  |
| |  | | --- | |  |   **As we see there are some outliers, so we will use log transformation.**  logPSA = log(psa)  qqnorm(logPSA)  qqline(logPSA) |

So now we see that the number of outliers is reduced.

We can look at the relationship between response and each predictor one by one

**We will check response wrt cancervol**

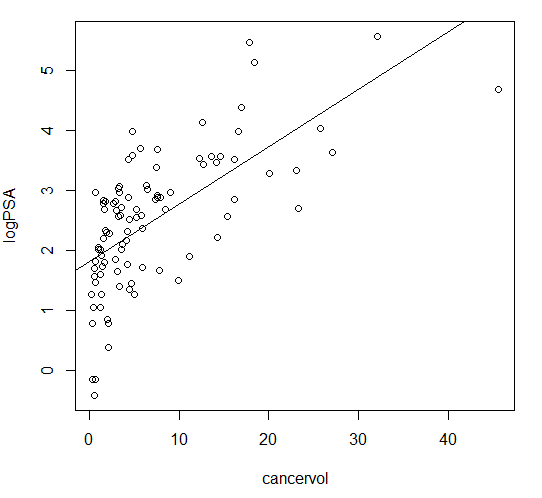
plot(cancervol, logPSA)

#Regress PSA on cancervol

fit1=lm(logPSA~cancervol)

abline(fit1)

#summary(fit1)



**We will check response wrt weight**

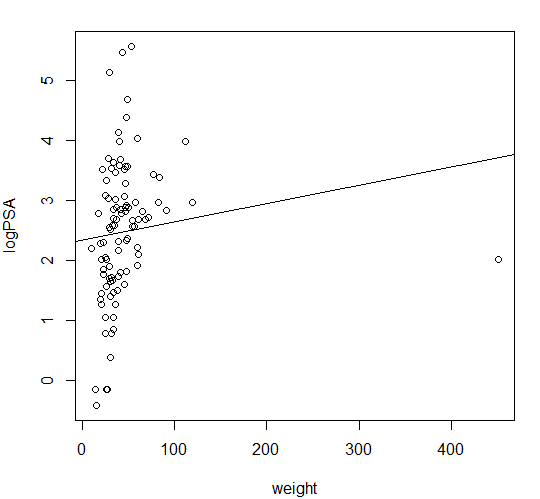
plot(weight, logPSA)

#Regress PSA on weight

fit2=lm(logPSA~weight)

abline(fit2)

#summary(fit2)



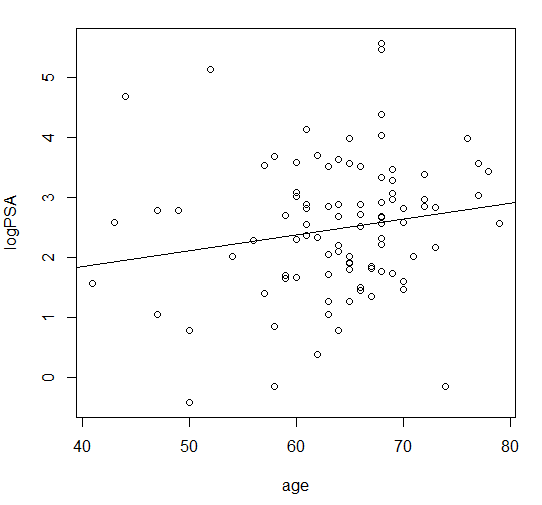
**We will check response wrt age**

plot(age, logPSA)

#Regress PSA on age

fit3=lm(logPSA~age)

abline(fit3)



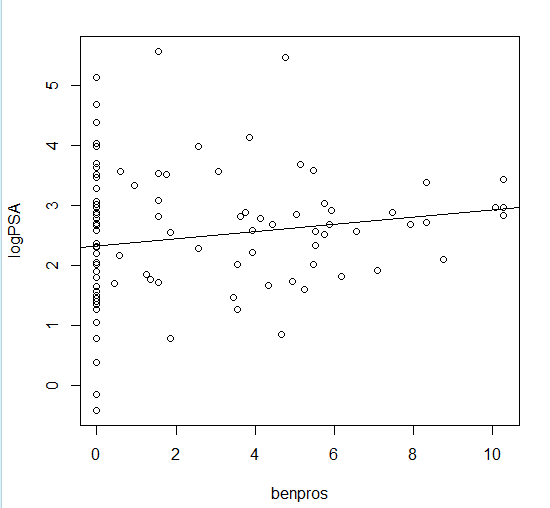
**Checking response wrt benpros**

plot(benpros, logPSA)

#Regress PSA on benpros

fit4=lm(logPSA~benpros)

abline(fit4)



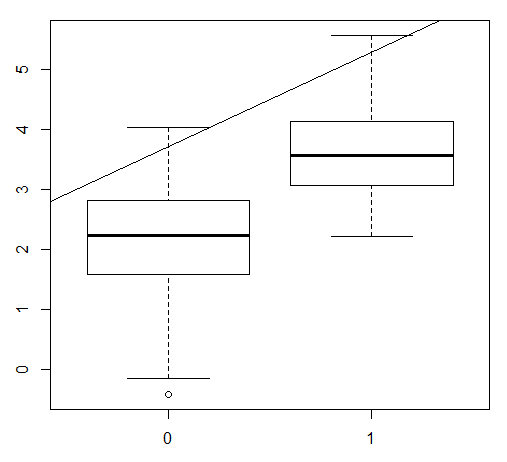
**Checking response wrt vesinv**

plot(vesinv, logPSA)

#Regress PSA on vesinv

fit5=lm(logPSA~vesinv)

abline(fit5)



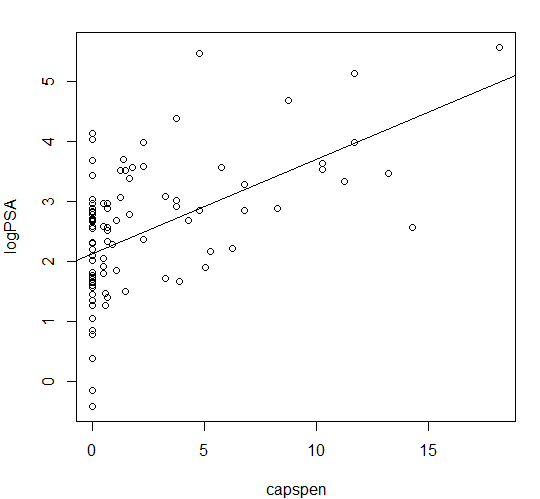
**Checking response wrt capspen**

plot(capspen, logPSA)

#Regress PSA on capspen

fit6=lm(logPSA~capspen)

abline(fit6)



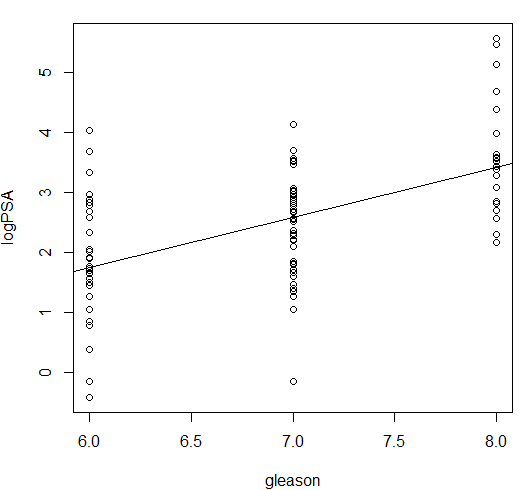
**Checking response wrt gleason**

plot(gleason, logPSA)

#Reress PSA on gleason

fit7=lm(logPSA~gleason)

abline(fit7)



We observe a positive trend in each case. From the predictors given it looks like cancercol, benpros, gleason might be important predictors to predict psa level. So we will build a model with these first.

fit8=lm(logPSA~cancervol+benpros+gleason)

summary(fit8)

Call:

lm(formula = logPSA ~ cancervol + benpros + gleason)

Residuals:

Min 1Q Median 3Q Max

-1.94779 -0.58638 0.05313 0.56068 1.57991

Coefficients:

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

(Intercept) -1.13566 0.82326 -1.379 0.17106

cancervol 0.08202 0.01184 6.926 5.51e-10 \*\*\*

benpros 0.08563 0.02700 3.172 0.00205 \*\*

gleason 0.41058 0.12510 3.282 0.00145 \*\*

---

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

Residual standard error: 0.7902 on 93 degrees of freedom

Multiple R-squared: 0.5455, Adjusted R-squared: 0.5308

F-statistic: 37.2 on 3 and 93 DF, p-value: 6.898e-16

**Adding age and weight to the model as a predictor**

fit9=lm(logPSA~cancervol+benpros+gleason+age+weight)

summary(fit9)

Call:

lm(formula = logPSA ~ cancervol + benpros + gleason + age + weight)

Residuals:

Min 1Q Median 3Q Max

-1.93310 -0.56892 0.02636 0.56067 1.51246

Coefficients:

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

(Intercept) -1.2515789 1.0014027 -1.250 0.2146

cancervol 0.0812940 0.0119709 6.791 1.12e-09 \*\*\*

benpros 0.0776022 0.0304986 2.544 0.0126 \*

gleason 0.4161915 0.1294377 3.215 0.0018 \*\*

age 0.0005454 0.0120864 0.045 0.9641

weight 0.0014928 0.0018885 0.790 0.4313

---

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

Residual standard error: 0.7961 on 91 degrees of freedom

Multiple R-squared: 0.5486, Adjusted R-squared: 0.5238

F-statistic: 22.12 on 5 and 91 DF, p-value: 1.921e-14

We can see that age and weight does not seem important, so we can drop those from the model

fit10=lm(logPSA~cancervol+benpros+gleason+capspen+vesinv)

summary(fit10)

Call:

lm(formula = logPSA ~ cancervol + benpros + gleason + capspen +

vesinv)

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 \*\*\*

benpros 0.09198 0.02612 3.522 0.000672 \*\*\*

gleason 0.34568 0.12437 2.779 0.006617 \*\*

capspen -0.02680 0.03260 -0.822 0.413237

vesinv1 0.78233 0.26520 2.950 0.004041 \*\*

---

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 see that capspen is insignificant, so we remove it.

fit11=lm(logPSA~cancervol+benpros+gleason+vesinv)

> summary(fit11)

Call:

lm(formula = logPSA ~ cancervol + benpros + gleason + vesinv)

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 \*\*\*

benpros 0.09136 0.02606 3.506 0.000705 \*\*\*

gleason 0.33376 0.12331 2.707 0.008100 \*\*

vesinv1 0.68421 0.23640 2.894 0.004746 \*\*

---

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

Now we have built an overfitted model with all Predictors

fit12=lm(logPSA~cancervol+benpros+gleason+vesinv+age+weight+capspen)

summary(fit12)

Call:

lm(formula = logPSA ~ cancervol + benpros + gleason + vesinv +

age + weight + capspen)

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 \*\*\*

benpros 0.087470 0.029605 2.955 0.00401 \*\*

gleason 0.358153 0.127976 2.799 0.00629 \*\*

vesinv1 0.782623 0.268339 2.917 0.00448 \*\*

age -0.002799 0.011724 -0.239 0.81186

weight 0.001380 0.001822 0.757 0.45079

capspen -0.026521 0.032860 -0.807 0.42177

---

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

**Perform a partial F-test to check significance of age, weight, capspen.**

anova(fit11, fit12)

|  |
| --- |
| Analysis of Variance Table  Model 1: logPSA ~ cancervol + benpros + gleason + vesinv  Model 2: logPSA ~ cancervol + benpros + gleason + vesinv + age + weight +  capspen  Res.Df RSS Df Sum of Sq F Pr(>F)  1 92 53.229  2 89 52.477 3 0.75232 0.4253 0.7353 |
|  |
| We don't need any of the three variables, age, weight, capspen. Also, the model in fit11 does not have any non-significant predictors. Therefore, we take this as our preliminary model for the data. However, we need to perform the diagnostics before accepting this model.  Let us compare fit11 with automatic stepwise model selection procedures based on AIC. In the output below '+' means 'add variable' and '-' means 'drop variable.'  fit13.forward <- step(lm(logPSA ~ 1), scope = list(upper = ~cancervol+benpros+gleason+vesinv+age+weight+capspen), direction = "forward")  Start: AIC=28.72  logPSA ~ 1  Df Sum of Sq RSS AIC  + cancervol 1 55.164 72.605 -24.0986  + vesinv 1 40.984 86.785 -6.7944  + gleason 1 37.122 90.647 -2.5707  + 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  logPSA ~ cancervol  Df Sum of Sq RSS AIC  + gleason 1 8.2468 64.358 -33.794  + benpros 1 7.8034 64.802 -33.128  + vesinv 1 6.5468 66.058 -31.265  + 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=-33.79  logPSA ~ cancervol + gleason  Df Sum of Sq RSS AIC  + benpros 1 6.2827 58.075 -41.758  + vesinv 1 4.0178 60.340 -38.047  + weight 1 2.0334 62.325 -34.908  <none> 64.358 -33.794  + age 1 0.9611 63.397 -33.253  + capspen 1 0.1685 64.190 -32.048  Step: AIC=-41.76  logPSA ~ cancervol + gleason + benpros  Df Sum of Sq RSS AIC  + vesinv 1 4.8466 53.229 -48.211  <none> 58.075 -41.758  + weight 1 0.4006 57.675 -40.429  + capspen 1 0.1863 57.889 -40.069  + age 1 0.0059 58.070 -39.768  Step: AIC=-48.21  logPSA ~ cancervol + gleason + benpros + vesinv  Df Sum of Sq RSS AIC  <none> 53.229 -48.211  + capspen 1 0.39230 52.837 -46.928  + weight 1 0.33060 52.898 -46.815  + age 1 0.02497 53.204 -46.256   |  | | --- | |  | |

fit14.backward <- step(lm(logPSA ~ cancervol+benpros+gleason+vesinv+age+weight+capspen), scope = list(lower = ~1), direction = "backward")

Start: AIC=-43.59

logPSA ~ cancervol + benpros + gleason + vesinv + age + weight +

capspen

Df Sum of Sq RSS AIC

- age 1 0.0336 52.510 -45.529

- weight 1 0.3383 52.815 -44.968

- capspen 1 0.3841 52.861 -44.884

<none> 52.477 -43.591

- gleason 1 4.6180 57.095 -37.410

- vesinv 1 5.0155 57.492 -36.737

- benpros 1 5.1469 57.624 -36.516

- cancervol 1 13.2994 65.776 -23.680

Step: AIC=-45.53

logPSA ~ cancervol + benpros + gleason + vesinv + weight + capspen

Df Sum of Sq RSS AIC

- weight 1 0.3264 52.837 -46.928

- capspen 1 0.3881 52.898 -46.815

<none> 52.510 -45.529

- gleason 1 4.6365 57.147 -39.322

- vesinv 1 4.9820 57.492 -38.737

- benpros 1 5.4873 57.998 -37.888

- cancervol 1 13.4654 65.976 -25.386

Step: AIC=-46.93

logPSA ~ cancervol + benpros + gleason + vesinv + capspen

Df Sum of Sq RSS AIC

- capspen 1 0.3923 53.229 -48.211

<none> 52.837 -46.928

- gleason 1 4.4852 57.322 -41.025

- vesinv 1 5.0526 57.889 -40.069

- benpros 1 7.2024 60.039 -36.532

- cancervol 1 13.7311 66.568 -26.520

Step: AIC=-48.21

logPSA ~ cancervol + benpros + gleason + vesinv

Df Sum of Sq RSS AIC

<none> 53.229 -48.211

- gleason 1 4.2389 57.468 -42.778

- vesinv 1 4.8466 58.075 -41.758

- benpros 1 7.1115 60.340 -38.047

- cancervol 1 14.7580 67.987 -26.473

fit15.both <- step(lm(logPSA ~ 1), scope = list(lower = ~1, upper = ~cancervol+benpros+gleason+vesinv+age+weight+capspen), direction = "both")

Start: AIC=28.72

logPSA ~ 1

Df Sum of Sq RSS AIC

+ cancervol 1 55.164 72.605 -24.0986

+ vesinv 1 40.984 86.785 -6.7944

+ gleason 1 37.122 90.647 -2.5707

+ 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

logPSA ~ cancervol

Df Sum of Sq RSS AIC

+ gleason 1 8.247 64.358 -33.794

+ benpros 1 7.803 64.802 -33.128

+ vesinv 1 6.547 66.058 -31.265

+ 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=-33.79

logPSA ~ cancervol + gleason

Df Sum of Sq RSS AIC

+ benpros 1 6.2827 58.075 -41.758

+ vesinv 1 4.0178 60.340 -38.047

+ weight 1 2.0334 62.325 -34.908

<none> 64.358 -33.794

+ age 1 0.9611 63.397 -33.253

+ capspen 1 0.1685 64.190 -32.048

- gleason 1 8.2468 72.605 -24.099

- cancervol 1 26.2887 90.647 -2.571

Step: AIC=-41.76

logPSA ~ cancervol + gleason + benpros

Df Sum of Sq RSS AIC

+ vesinv 1 4.8466 53.229 -48.211

<none> 58.075 -41.758

+ weight 1 0.4006 57.675 -40.429

+ capspen 1 0.1863 57.889 -40.069

+ age 1 0.0059 58.070 -39.768

- benpros 1 6.2827 64.358 -33.794

- gleason 1 6.7262 64.802 -33.128

- cancervol 1 29.9589 88.034 -3.407

Step: AIC=-48.21

logPSA ~ cancervol + gleason + benpros + vesinv

Df Sum of Sq RSS AIC

<none> 53.229 -48.211

+ capspen 1 0.3923 52.837 -46.928

+ weight 1 0.3306 52.898 -46.815

+ age 1 0.0250 53.204 -46.256

- gleason 1 4.2389 57.468 -42.778

- vesinv 1 4.8466 58.075 -41.758

- benpros 1 7.1115 60.340 -38.047

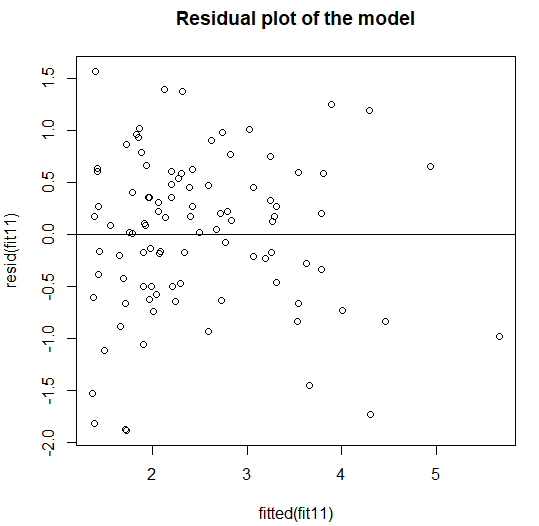
- cancervol 1 14.7580 67.987 -26.473

**Based on AIC we can easily see that the forward backward and both directions give the exact same model as we predicted earlier(fit11).**

**Lets make some Residual Plots and QQPlots**

plot(fitted(fit11), resid(fit11),main="Residual plot of the model")

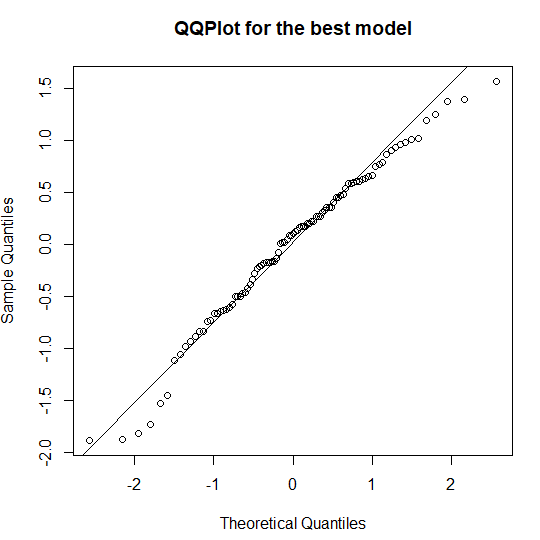
abline(h = 0)



We can see that there is no Trend and no Change in vertical spread which is good.

qqnorm(resid(fit11), main="QQPlot for the best model")

qqline(resid(fit11))



**This preliminary model(fit11) passes the diagnostics. So we can take this as our final model. Therefore fit 11 is our best model**

Now 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 we have performed the following.

summary(vesinv)

0 1

76 21

we see that 0s are 76 and 1s are 21

# calculate all means

> (mean.cancervol = mean(cancervol))

[1] 6.998682

> (mean.benpros = mean(benpros))

[1] 2.534725

> (mean.gleason = mean(gleason))

[1] 6.876289

# here 0 being most frequent in vesinv

psa=-0.65013 + (0.06488 \* mean.cancervol ) + ( 0.09136 \* mean.benpros ) + ( 0.33376 \* mean.gleason ) + ( 0.68421 \* 0 )

[1] 2.330547

exp(psa)

[1] 10.28357

**PSA level for a patient will be 10.28357**

**Section 2:**

setwd("c:/")

# Reading Data from the input file

inputData = read.csv(file="prostate\_cancer.csv", header=TRUE, sep = ",")

#attaching the data so that we are able to access columsn via direct names

attach(inputData)

head(inputData)

#using factor function for categorical data

vesinv=as.factor(vesinv)

par(mfrow=c(1,1))

boxplot(psa)

qqnorm(psa)

qqline(psa)

# as we see there are a some outliers, so we wil transform, as its asked in the

# question to use log transformation if need be, we will use log transform only,

# whereas a sqrt transformer could have also done the job

logPSA = log(psa)

qqnorm(logPSA)

qqline(logPSA)

# So now we see that the number of outliers are reduced

#we will check response wrt cancervol

plot(cancervol, logPSA)

#Regress PSA on cancervol

fit1=lm(logPSA~cancervol)

abline(fit1)

#summary(fit1)

#we will check response wrt weight

plot(weight, logPSA)

#Regress PSA on weight

fit2=lm(logPSA~weight)

abline(fit2)

#summary(fit2)

#we will check response wrt age

plot(age, logPSA)

#Regress PSA on age

fit3=lm(logPSA~age)

abline(fit3)

#checking response wrt benpros

plot(benpros, logPSA)

#Regress PSA on benpros

fit4=lm(logPSA~benpros)

abline(fit4)

#checking response wrt vesinv

plot(vesinv, logPSA)

#Regress PSA on vesinv

fit5=lm(logPSA~vesinv)

abline(fit5)

#checking response wrt capspen

plot(capspen, logPSA)

#Regress PSA on capspen

fit6=lm(logPSA~capspen)

abline(fit6)

#checking response wrt gleason

plot(gleason, logPSA)

#Reress PSA on gleason

fit7=lm(logPSA~gleason)

abline(fit7)

# We observe a positive trend in each case

# from the predictors given it looks like cancercol, benpros, gleason might be important predictors to predict psa level

# so we will build a model with these first

fit8=lm(logPSA~cancervol+benpros+gleason)

summary(fit8)

# adding age and weight to the model as a predictor

fit9=lm(logPSA~cancervol+benpros+gleason+age+weight)

summary(fit9)

#we can see that age and weight does not seem important

#so we can drop those from the model

fit10=lm(logPSA~cancervol+benpros+gleason+capspen+vesinv)

summary(fit10)

#we can see that capspen is insignificant, so we remove it

fit11=lm(logPSA~cancervol+benpros+gleason+vesinv)

summary(fit11)

#built an overfitted model with all Predictors

fit12=lm(logPSA~cancervol+benpros+gleason+vesinv+age+weight+capspen)

summary(fit12)

# Perform a partial F-test to check significance of age, weight, capspen

anova(fit11, fit12)

# We don't need any of the three variables, age, weight, capspen. Also, the model in fit11 does not

# have any non-significant predictors. Therefore, we take this as our

# preliminary model for the data. However, we need to perform the diagnostics

# before accepting this model.

# Let us compare fit11 with automatic stepwise model selection procedures based on AIC

# In the output below '+' means 'add variable' and '-' means 'drop variable.'

fit13.forward <- step(lm(logPSA ~ 1), scope = list(upper = ~cancervol+benpros+gleason+vesinv+age+weight+capspen), direction = "forward")

fit14.backward <- step(lm(logPSA ~ cancervol+benpros+gleason+vesinv+age+weight+capspen), scope = list(lower = ~1), direction = "backward")

fit15.both <- step(lm(logPSA ~ 1), scope = list(lower = ~1, upper = ~cancervol+benpros+gleason+vesinv+age+weight+capspen), direction = "both")

# Based on AIC we can easily see that the forward backward and both directions give the exact same model as we predicted earlier(fit11)

# Lets make some Residual Plots and QQPlots

plot(fitted(fit11), resid(fit11),main="Residual plot of the model")

abline(h = 0)

# No Trend

# No Change in vertical spread ! Good news !

qqnorm(resid(fit11), main="QQPlot for the best model")

qqline(resid(fit11))

# This preliminary model(fit11) passes the diagnostics. So we can take this as our final model.

# Therefore fit 11 is our best model

summary(vesinv)

# we see that 0s are 76 and 1s are 21

# calculate all means

(mean.cancervol = mean(cancervol))

(mean.benpros = mean(benpros))

(mean.gleason = mean(gleason))

# here 0 being most frequent in vesinv

(psa = - 0.65013 + (0.06488 \* mean.cancervol)+(0.09136\*mean.benpros)+(0.33376\*mean.gleason)+(0.68421\*0))

exp(psa)