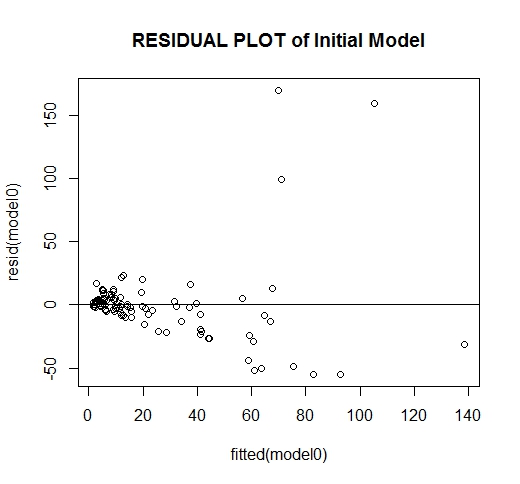
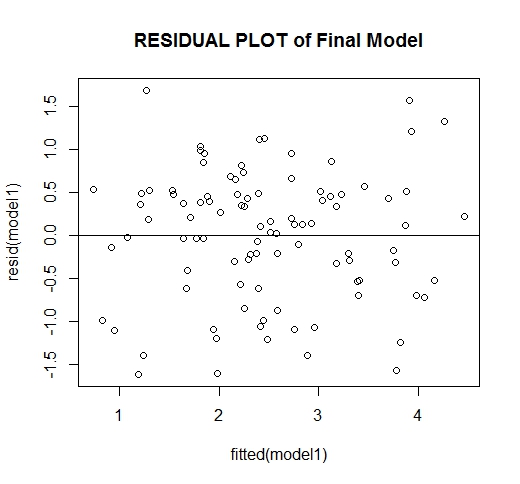
Project 6

1. Here we have chosen the predictors as ‘Cancer Volume’ and ‘Seminal Vesicle Invasion’ to predict the ‘PSA level’. We **consider ‘Cancer Volume’ from Project5**. Thereafter we start adding the 2nd predictor from the given set of predictors. Calculating ANOVA from the list of predictors gives us pval= 0.00308 for ‘Seminal Vesicle Invasion’ which is the best among the given list of predictors.

Initial model comparing the variables without using any remedy situations (of applying log to variables) will give us this value.



Thereafter, when we apply the logarithmic value to the predictor ‘Cancer Volume’ and final result ‘PSA’, we get an appropriate regression plot as shown below:

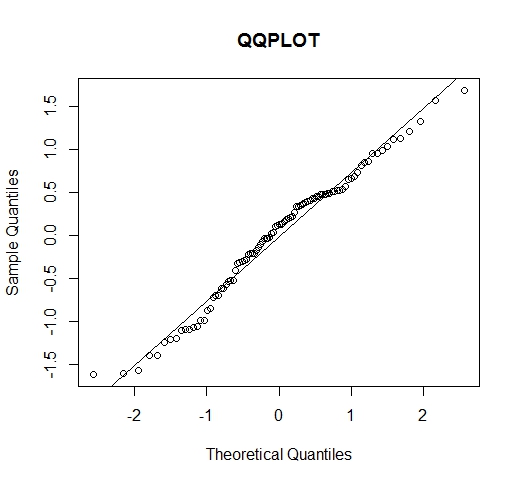


Comment on the fit of the final model using appropriate tests and statistics

We have to show that the model is Normal, Time-Series Independent and Residuals are independent (not increasing)

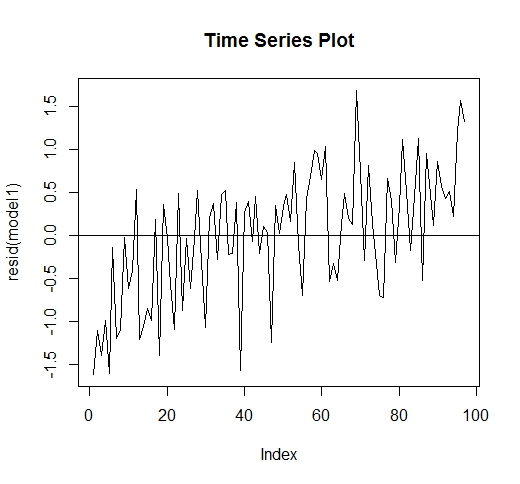
The residual plot is shown above which shows residuals are independent

We have done QQ plot to look for normality



This plot shows us that the model is Normal as all the values are almost near the normal line.

Thereafter we plot the Time-series plot to show the residuals are independent



We get the final statistics as

Residual standard error: 0.7558 on 94 degrees of freedom

Multiple R-squared: 0.5797, Adjusted R-squared: 0.5708

F-statistic: 64.84 on 2 and 94 DF, p-value: < 2.2e-16

We get the p-value very low which proves this model is valid.

1. final model to predict the PSA level for a patient whose predictor variables are at the sample medians of the variables

We get the sample median values of the predictors ‘Cancer Volume’ and ‘Seminal Vesicle Invasion’ as 1.449997 0

The Final PSA value by using the predictor values is 2.392514

**Annotated R-Code**

data.frame=read.csv("prostate\_cancer.csv")

z1<-log(data.frame$psa)

z2<-log(data.frame$cancervol)

z3<-(data.frame$vesinv)

z4<- data.frame$gleason

z5<- data.frame$weight

z6<- data.frame$age

z7<- data.frame$benpros

z8<- data.frame$capspen

y1<- data.frame$psa

y2<- data.frame$cancervol

#Initial Model

model0<- lm(y1~y2+factor(z3))

#Final Models after applying logarithm

model1<-lm(z1 ~z2 + factor(z3))

model2<-lm(z1 ~z2)

model3<-lm(z1~z2+factor(z4))

model4<- lm(z1~z2+ z5)

model5<- lm(z1~z2+ z6)

model6<- lm(z1~z2+ z7)

model7<- lm(z1~z2+ z8)

summary(model1)

#gives us the statistical values

anova(model2,model1)

#pval= 0.00308 \*\*

anova(model2,model3)

#pval= 0.01236 \*

anova(model2,model4)

#pval= 0.08084 .

anova(model2,model5)

#pval= 0.9444

anova(model2,model6)

#pval= 0.04446 \*

anova(model2,model7)

#pval = 0.2429

#from the above models, using the ANOVA test, we conclude that model1 which takes vesinv as 2nd predictor variable is best

#Check for errors whether they have mean zero constant variance,

#normality and independence of residual error

plot(fitted(model0),resid(model0), main="RESIDUAL PLOT of Initial Model")

abline(h=0)

#testing final obtained model with logarithm

plot(fitted(model1),resid(model1), main="RESIDUAL PLOT of Final Model")

abline(h=0)

qqnorm(resid(model1),main="QQPLOT")

qqline(resid(model1))

plot(resid(model1), type="l",main="Time Series Plot")

abline(h=0)

#predicting y-value

model1updated= update(model1,z1~.)

x.new<-data.frame(z2=median(z2),z3=median(z3))

x.new

predict(model1updated,x.new)