CS6313.001 Statistical Methods for Data Science

Mini Project # 6

Name: Ning Xin Di Jin

NetID: nxx170130 dxj170930

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Section 1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#read the table

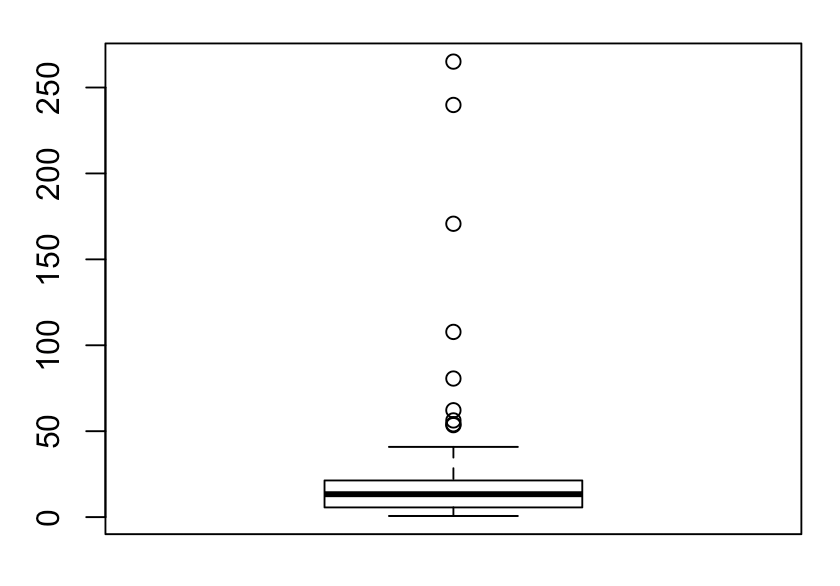
cancer <- read.table("/Users/dijin/Desktop/统计/minipro/6/prostate\_cancer.csv", sep=",", header=T)

#11111111111

#see the connect between each parameter and result

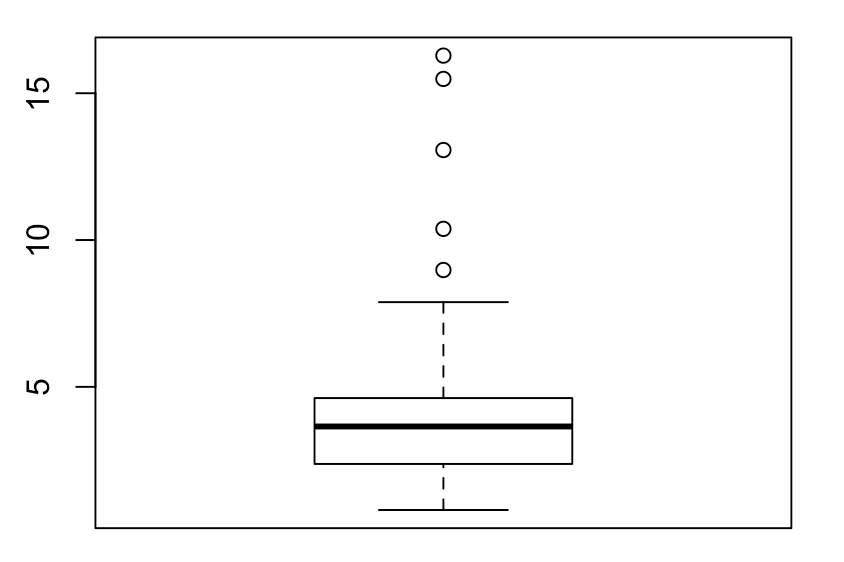
First find a good method to transform response

boxplot(cancer$psa)



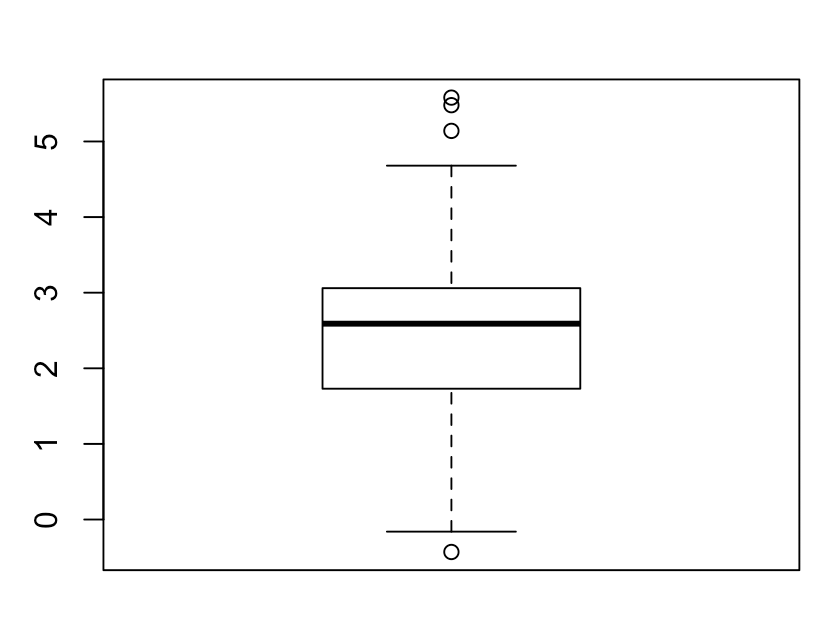
#try sqrt to get better distribution

boxplot(sqrt(cancer$psa)) # not so good



#try nature log to get better distribution

boxplot(log(cancer$psa)) #what has been used in following code



#seems nature log is a better choice

y<-log(cancer$psa)

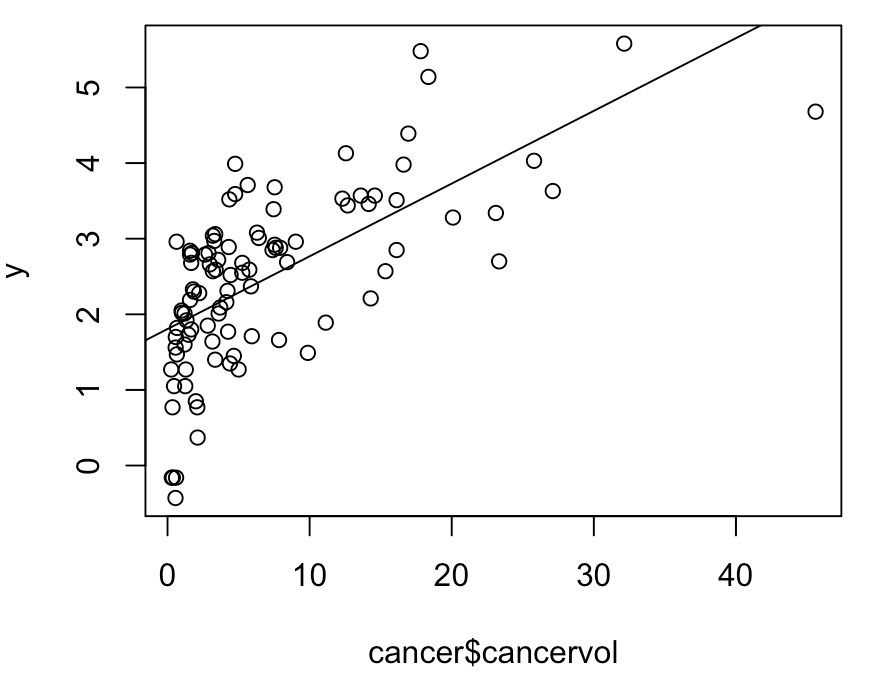
#222222222222

#First, let's look at the relationship between response and each predictor one by one

plot(cancer$cancervol,y)

fit1<-lm(y ~ cancervol,data=cancer)

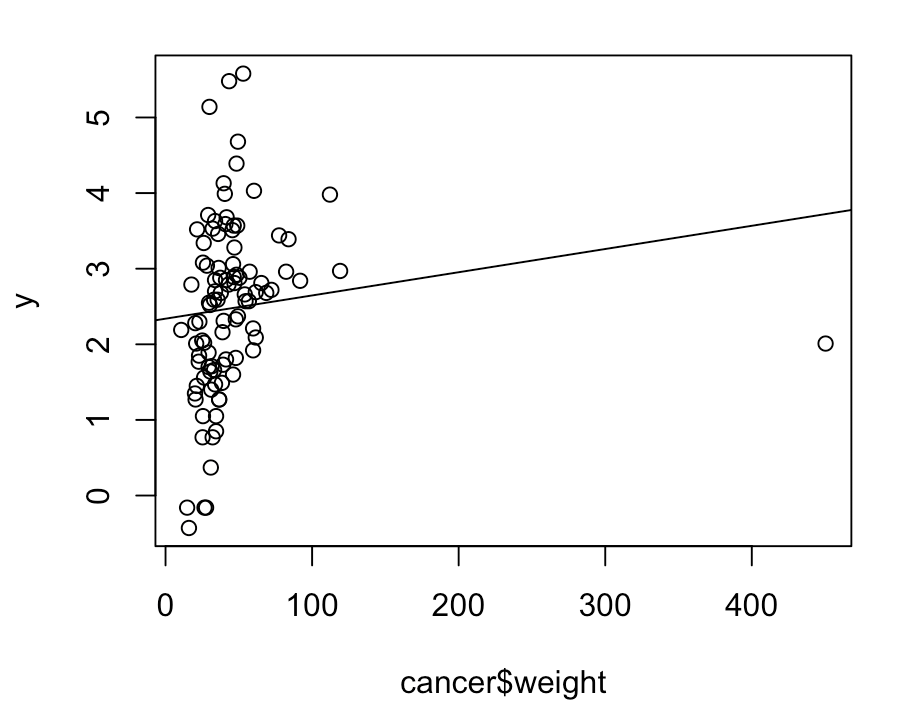
abline(fit1)



plot(cancer$weight,y)

fit2<-lm(y ~ weight,data=cancer)

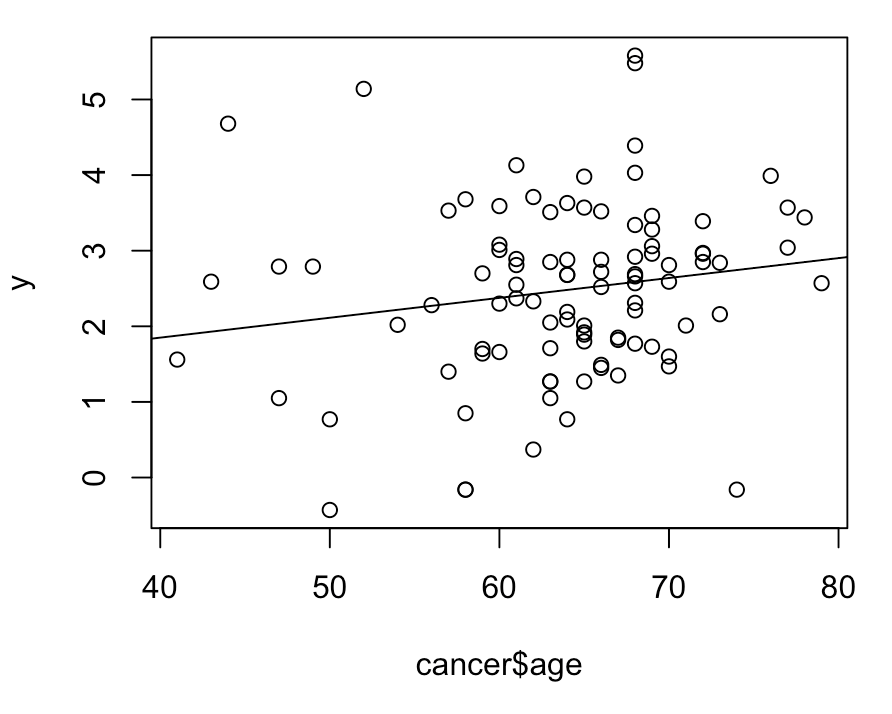
abline(fit2)



plot(cancer$age,y)

fit3<-lm(y ~ age,data=cancer)

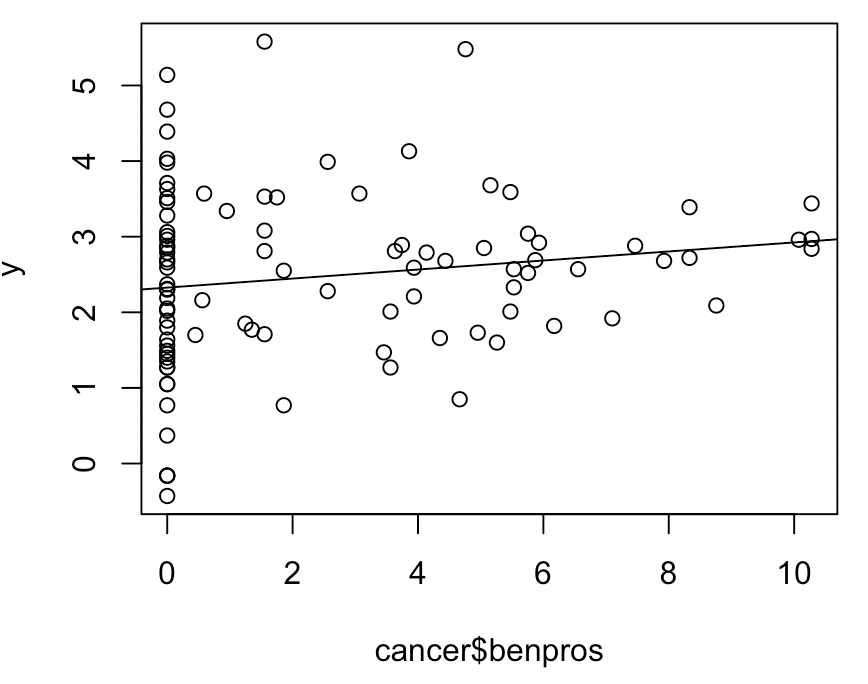
abline(fit3)



plot(cancer$benpros,y)

fit4<-lm(y ~ benpros,data=cancer)

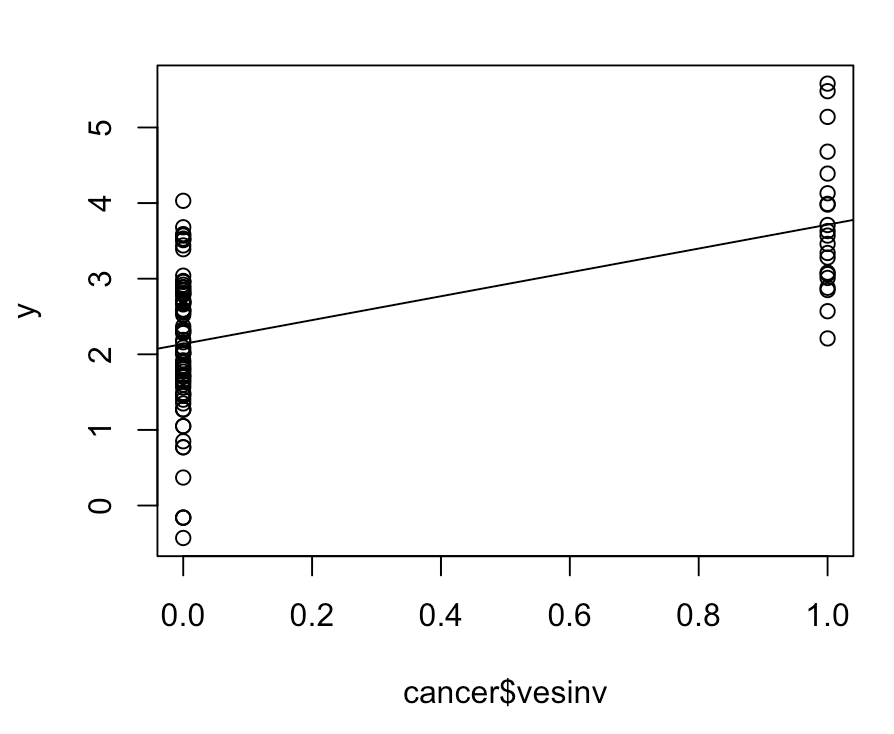
abline(fit4)



plot(cancer$vesinv,y)

fit5<-lm(y ~ factor(vesinv),data=cancer)

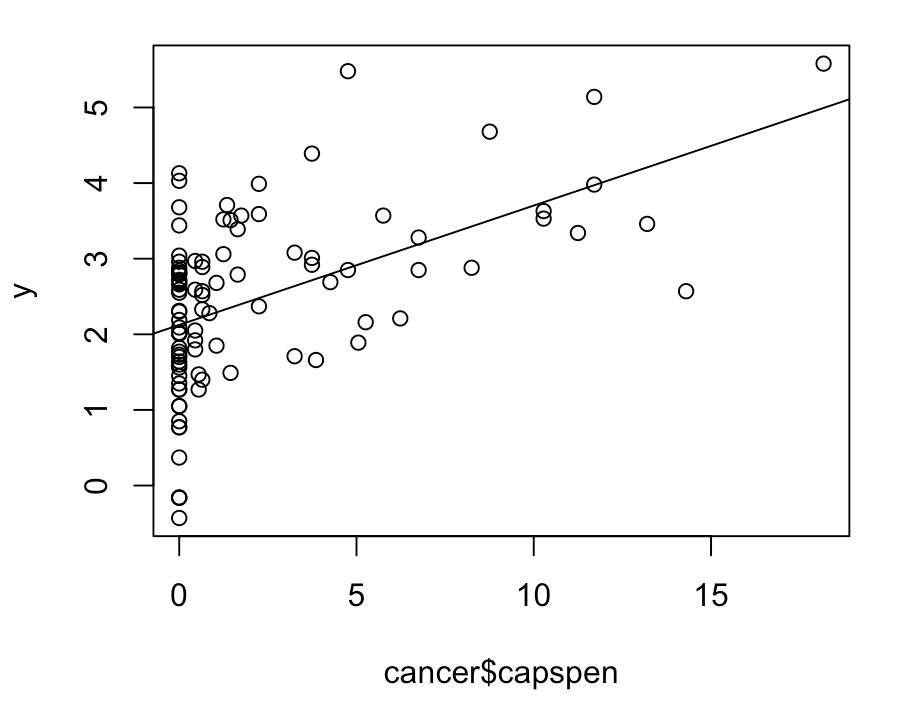
abline(fit5)



plot(cancer$capspen,y)

fit6<-lm(y ~ capspen,data=cancer)

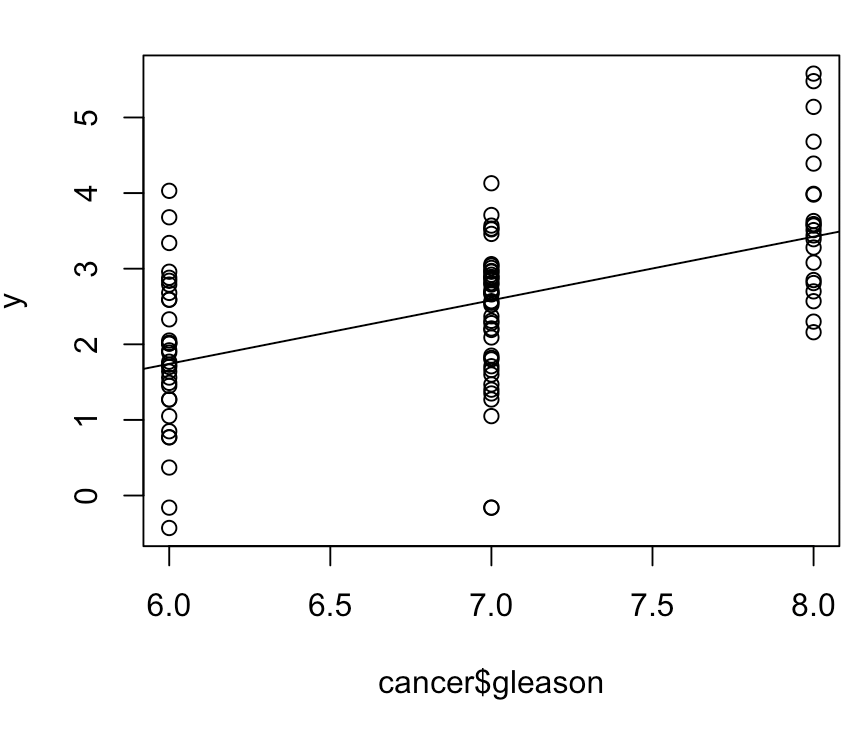
abline(fit6)



plot(cancer$gleason,y)

fit7<-lm(y ~ gleason,data=cancer)

abline(fit7)



# We see a significant positive trend in each case

#3333333333

#We put every parameter into test

fit8<-lm(y ~ cancervol+weight+age+benpros+factor(vesinv)+capspen+gleason,data=cancer)

anova(fit8)

# Response: y

# Df Sum Sq Mean Sq F value Pr(>F)

# cancervol 1 55.164 55.164 93.5572 1.522e-15 \*\*\*

# weight 1 1.790 1.790 3.0360 0.0848889 .

# age 1 2.048 2.048 3.4735 0.0656577 .

# benpros 1 4.541 4.541 7.7014 0.0067245 \*\*

# factor(vesinv) 1 6.983 6.983 11.8438 0.0008832 \*\*\*

# capspen 1 0.148 0.148 0.2510 0.6176363

# gleason 1 4.618 4.618 7.8322 0.0062918 \*\*

# Residuals 89 52.477 0.590

# We can find that cancervol and vesinv is quite important, so we start from them

fit9<-lm(y ~ cancervol+factor(vesinv),data=cancer)

anova(fit9)

# Response: y

# Df Sum Sq Mean Sq F value Pr(>F)

# cancervol 1 55.164 55.164 78.4972 4.877e-14 \*\*\*

# factor(vesinv) 1 6.547 6.547 9.3161 0.002953 \*\*

# Residuals 94 66.058 0.703

# then add weight

fit10<-lm(y ~ cancervol+factor(vesinv)+weight,data=cancer)

anova(fit9,fit10)

# Model 1: y ~ cancervol + factor(vesinv)

# Model 2: y ~ cancervol + factor(vesinv) + weight

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

# 1 94 66.058

# 2 93 64.222 1 1.8358 2.6584 0.1064

# from the output pr is not lower enough

# we assume weight to be non-significant predictors

#then add age

fit11<-lm(y ~ cancervol+factor(vesinv)+age,data=cancer)

anova(fit9,fit11)

# Model 1: y ~ cancervol + factor(vesinv)

# Model 2: y ~ cancervol + factor(vesinv) + age

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

# 1 94 66.058

# 2 93 64.258 1 1.8004 2.6057 0.1099

# from the output pr is not lower enough

# we assume age to be non-significant predictors

#then add benpros

fit12<-lm(y ~ cancervol+factor(vesinv)+benpros,data=cancer)

anova(fit9,fit12)

# Model 1: y ~ cancervol + factor(vesinv)

# Model 2: y ~ cancervol + factor(vesinv) + benpros

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

# 1 94 66.058

# 2 93 57.468 1 8.5905 13.902 0.0003308 \*\*\*

#according to pr value, we consider benpros is significant, model12 is better

#add capspen to model12

fit13<-lm(y ~ cancervol+factor(vesinv)+benpros+capspen,data=cancer)

anova(fit12,fit13)

# Model 1: y ~ cancervol + factor(vesinv) + benpros

# Model 2: y ~ cancervol + factor(vesinv) + benpros + capspen

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

# 1 93 57.468

# 2 92 57.322 1 0.14596 0.2343 0.6295

# from the output pr is not lower enough

# we assume capspen to be non-significant predictors

#add gleason to model 12

fit14<-lm(y ~ cancervol+factor(vesinv)+benpros+gleason,data=cancer)

anova(fit12,fit14)

#according to pr value, we consider gleason is significant, model14 is better

#4444444 use auto model method (forward,backward,both) to compare

forward <- step(lm(y ~ 1, data = cancer),

scope = list(upper = ~ cancervol+weight+age+benpros+factor(vesinv)+capspen+gleason),

direction = "forward")

# Start: AIC=28.72

# y ~ 1

#

# Df Sum of Sq RSS AIC

# + cancervol 1 55.164 72.605 -24.0986

# + factor(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

# y ~ cancervol

#

# Df Sum of Sq RSS AIC

# + gleason 1 8.2468 64.358 -33.794

# + benpros 1 7.8034 64.802 -33.128

# + factor(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

# y ~ cancervol + gleason

#

# Df Sum of Sq RSS AIC

# + benpros 1 6.2827 58.075 -41.758

# + factor(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

# y ~ cancervol + gleason + benpros

#

# Df Sum of Sq RSS AIC

# + factor(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

# y ~ cancervol + gleason + benpros + factor(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

backward <- step(lm(y ~ cancervol+weight+age+benpros+factor(vesinv)+capspen+gleason, data = cancer),

scope = list(lower = ~1), direction = "backward")

# Start: AIC=-43.59

# y ~ cancervol + weight + age + benpros + factor(vesinv) + capspen +

# gleason

#

# 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

# - factor(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

# y ~ cancervol + weight + benpros + factor(vesinv) + capspen +

# gleason

#

# 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

# - factor(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

# y ~ cancervol + benpros + factor(vesinv) + capspen + gleason

#

# 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

# - factor(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

# y ~ cancervol + benpros + factor(vesinv) + gleason

#

# Df Sum of Sq RSS AIC

# <none> 53.229 -48.211

# - gleason 1 4.2389 57.468 -42.778

# - factor(vesinv) 1 4.8466 58.075 -41.758

# - benpros 1 7.1115 60.340 -38.047

# - cancervol 1 14.7580 67.987 -26.473

# >

both <- step(lm(y ~ 1, data = cancer),

scope = list(lower = ~1,upper= ~ cancervol+weight+age+benpros+factor(vesinv)+capspen+gleason), direction = "both")

# Start: AIC=28.72

# y ~ 1

#

# Df Sum of Sq RSS AIC

# + cancervol 1 55.164 72.605 -24.0986

# + factor(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

# y ~ cancervol

#

# Df Sum of Sq RSS AIC

# + gleason 1 8.247 64.358 -33.794

# + benpros 1 7.803 64.802 -33.128

# + factor(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

# y ~ cancervol + gleason

#

# Df Sum of Sq RSS AIC

# + benpros 1 6.2827 58.075 -41.758

# + factor(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

# y ~ cancervol + gleason + benpros

#

# Df Sum of Sq RSS AIC

# + factor(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

# y ~ cancervol + gleason + benpros + factor(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

# - factor(vesinv) 1 4.8466 58.075 -41.758

# - benpros 1 7.1115 60.340 -38.047

# - cancervol 1 14.7580 67.987 -26.473

#they all give the same result with my mode 14

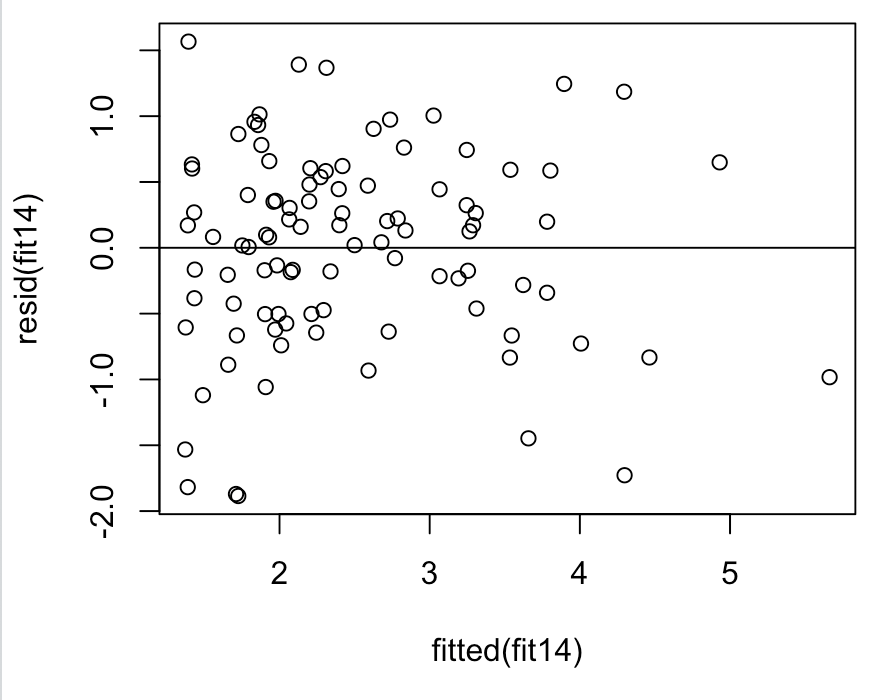
#5555555

#give the model diagnostics

# residual plot

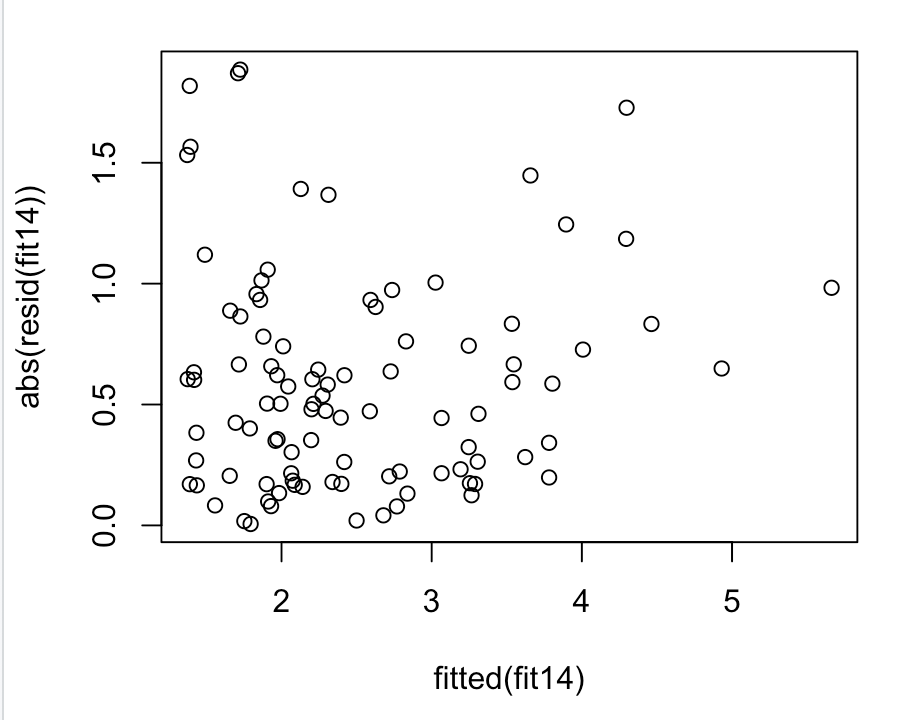
plot(fitted(fit14), resid(fit14))

abline(h = 0)



# plot of absolute residuals

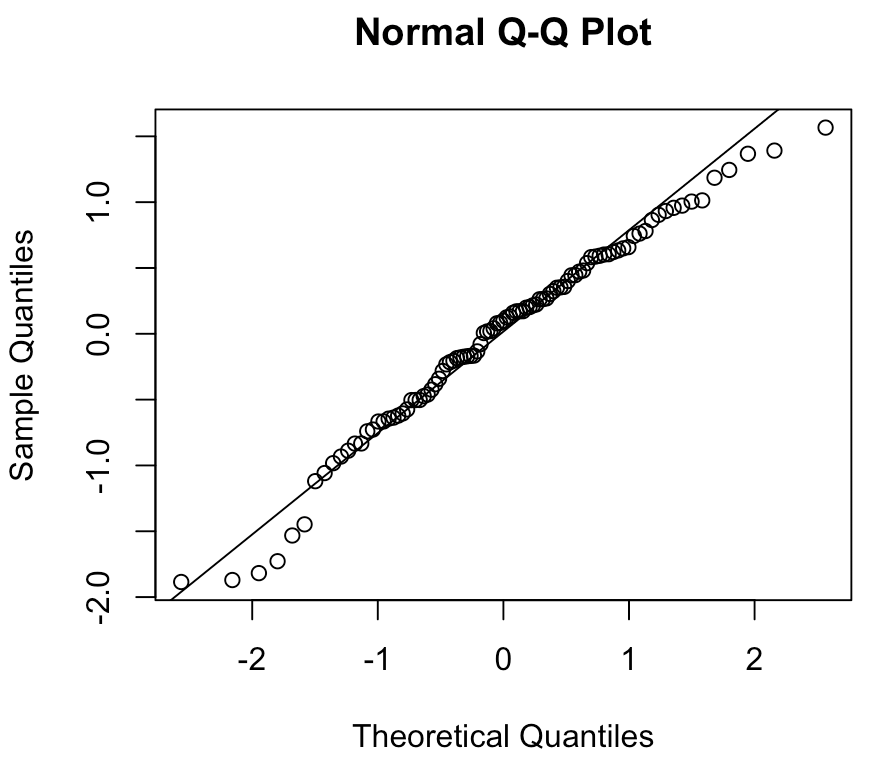
plot(fitted(fit14), abs(resid(fit14)))



# normal QQ plot

qqnorm(resid(fit14))

qqline(resid(fit14))



# This preliminary model passes the diagnostics. So we can take this

# as our final model.

summary（fit14）

Call:

lm(formula = y ~ cancervol + factor(vesinv) + benpros + gleason,

data = cancer)

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 \*\*\*

factor(vesinv) 0.68421 0.23640 2.894 0.004746 \*\*

benpros 0.09136 0.02606 3.506 0.000705 \*\*\*

gleason 0.33376 0.12331 2.707 0.008100 \*\*

---

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

Since the p-value of cancervol, factor, benpros and gleason are all smaller than 0.05. So we reject H0 which cancervol, factor, benpros and gleason are useless. In conclusion, cancervol, factor, benpros and gleason are useful.

#666666

#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.

table(cancer$vesinv) #find the most freaquent

input<-data.frame(cancervol=mean(cancer$cancervol),weight=mean(cancer$weight),age=mean(cancer$age),benpros=mean(cancer$benpros),vesinv=0,capspen=mean(cancer$capspen),gleason=mean(cancer$gleason))

predict(fit14,newdata = input)

final mode:fit14:

Call:

lm(formula = y ~ cancervol + factor(vesinv) + benpros + gleason,

data = cancer)

Coefficients:

(Intercept) cancervol factor(vesinv)1 benpros gleason

-0.65013 0.06488 0.68421 0.09136 0.33376

# > predict(fit14,newdata = input)

# 1

# 2.330541

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Section 2

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#read the table

cancer <- read.table("/Users/dijin/Desktop/??????/minipro/6/prostate\_cancer.csv", sep=",", header=T)

#11111111111see the connect between each parameter and result

boxplot(cancer$psa)

#try sqrt to get better distribution

boxplot(sqrt(cancer$psa)) # not so good

#try nature log to get better distribution

boxplot(log(cancer$psa)) #what has been used in following code

#seems nature log is a better choice

y<-log(cancer$psa)

#222222222222 First, let's look at the relationship between response and each predictor one by one

plot(cancer$cancervol,y)

fit1<-lm(y ~ cancervol,data=cancer)

abline(fit1)

plot(cancer$weight,y)

fit2<-lm(y ~ weight,data=cancer)

abline(fit2)

plot(cancer$age,y)

fit3<-lm(y ~ age,data=cancer)

abline(fit3)

plot(cancer$benpros,y)

fit4<-lm(y ~ benpros,data=cancer)

abline(fit4)

plot(cancer$vesinv,y)

fit5<-lm(y ~ factor(vesinv),data=cancer)

abline(fit5)

plot(cancer$capspen,y)

fit6<-lm(y ~ capspen,data=cancer)

abline(fit6)

plot(cancer$gleason,y)

fit7<-lm(y ~ gleason,data=cancer)

abline(fit7)

# We see a significant positive trend in each case

#3333333333 We put every parameter into test

fit8<-lm(y ~ cancervol+weight+age+benpros+factor(vesinv)+capspen+gleason,data=cancer)

anova(fit8)

# Response: y

# Df Sum Sq Mean Sq F value Pr(>F)

# cancervol 1 55.164 55.164 93.5572 1.522e-15 \*\*\*

# weight 1 1.790 1.790 3.0360 0.0848889 .

# age 1 2.048 2.048 3.4735 0.0656577 .

# benpros 1 4.541 4.541 7.7014 0.0067245 \*\*

# factor(vesinv) 1 6.983 6.983 11.8438 0.0008832 \*\*\*

# capspen 1 0.148 0.148 0.2510 0.6176363

# gleason 1 4.618 4.618 7.8322 0.0062918 \*\*

# Residuals 89 52.477 0.590

# We can find that cancervol and vesinv is quite important, so we start from them

fit9<-lm(y ~ cancervol+factor(vesinv),data=cancer)

anova(fit9)

# Response: y

# Df Sum Sq Mean Sq F value Pr(>F)

# cancervol 1 55.164 55.164 78.4972 4.877e-14 \*\*\*

# factor(vesinv) 1 6.547 6.547 9.3161 0.002953 \*\*

# Residuals 94 66.058 0.703

# then add weight

fit10<-lm(y ~ cancervol+factor(vesinv)+weight,data=cancer)

anova(fit9,fit10)

# Model 1: y ~ cancervol + factor(vesinv)

# Model 2: y ~ cancervol + factor(vesinv) + weight

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

# 1 94 66.058

# 2 93 64.222 1 1.8358 2.6584 0.1064

# from the output pr is not lower enough

# we assume weight to be non-significant predictors

#then add age

fit11<-lm(y ~ cancervol+factor(vesinv)+age,data=cancer)

anova(fit9,fit11)

# Model 1: y ~ cancervol + factor(vesinv)

# Model 2: y ~ cancervol + factor(vesinv) + age

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

# 1 94 66.058

# 2 93 64.258 1 1.8004 2.6057 0.1099

# from the output pr is not lower enough

# we assume age to be non-significant predictors

#then add benpros

fit12<-lm(y ~ cancervol+factor(vesinv)+benpros,data=cancer)

anova(fit9,fit12)

# Model 1: y ~ cancervol + factor(vesinv)

# Model 2: y ~ cancervol + factor(vesinv) + benpros

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

# 1 94 66.058

# 2 93 57.468 1 8.5905 13.902 0.0003308 \*\*\*

#according to pr value, we consider benpros is significant, model12 is better

#add capspen to model12

fit13<-lm(y ~ cancervol+factor(vesinv)+benpros+capspen,data=cancer)

anova(fit12,fit13)

# Model 1: y ~ cancervol + factor(vesinv) + benpros

# Model 2: y ~ cancervol + factor(vesinv) + benpros + capspen

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

# 1 93 57.468

# 2 92 57.322 1 0.14596 0.2343 0.6295

# from the output pr is not lower enough

# we assume capspen to be non-significant predictors

#add gleason to model 12

fit14<-lm(y ~ cancervol+factor(vesinv)+benpros+gleason,data=cancer)

anova(fit12,fit14)

#according to pr value, we consider gleason is significant, model14 is better

#4444444 use auto model method (forward,backward,both) to compare

forward <- step(lm(y ~ 1, data = cancer),

scope = list(upper = ~ cancervol+weight+age+benpros+factor(vesinv)+capspen+gleason),

direction = "forward")

# Start: AIC=28.72

# y ~ 1

#

# Df Sum of Sq RSS AIC

# + cancervol 1 55.164 72.605 -24.0986

# + factor(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

# y ~ cancervol

#

# Df Sum of Sq RSS AIC

# + gleason 1 8.2468 64.358 -33.794

# + benpros 1 7.8034 64.802 -33.128

# + factor(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

# y ~ cancervol + gleason

#

# Df Sum of Sq RSS AIC

# + benpros 1 6.2827 58.075 -41.758

# + factor(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

# y ~ cancervol + gleason + benpros

#

# Df Sum of Sq RSS AIC

# + factor(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

# y ~ cancervol + gleason + benpros + factor(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

backward <- step(lm(y ~ cancervol+weight+age+benpros+factor(vesinv)+capspen+gleason, data = cancer),

scope = list(lower = ~1), direction = "backward")

# Start: AIC=-43.59

# y ~ cancervol + weight + age + benpros + factor(vesinv) + capspen +

# gleason

#

# 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

# - factor(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

# y ~ cancervol + weight + benpros + factor(vesinv) + capspen +

# gleason

#

# 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

# - factor(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

# y ~ cancervol + benpros + factor(vesinv) + capspen + gleason

#

# 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

# - factor(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

# y ~ cancervol + benpros + factor(vesinv) + gleason

#

# Df Sum of Sq RSS AIC

# <none> 53.229 -48.211

# - gleason 1 4.2389 57.468 -42.778

# - factor(vesinv) 1 4.8466 58.075 -41.758

# - benpros 1 7.1115 60.340 -38.047

# - cancervol 1 14.7580 67.987 -26.473

# >

both <- step(lm(y ~ 1, data = cancer),

scope = list(lower = ~1,upper= ~ cancervol+weight+age+benpros+factor(vesinv)+capspen+gleason), direction = "both")

# Start: AIC=28.72

# y ~ 1

#

# Df Sum of Sq RSS AIC

# + cancervol 1 55.164 72.605 -24.0986

# + factor(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

# y ~ cancervol

#

# Df Sum of Sq RSS AIC

# + gleason 1 8.247 64.358 -33.794

# + benpros 1 7.803 64.802 -33.128

# + factor(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

# y ~ cancervol + gleason

#

# Df Sum of Sq RSS AIC

# + benpros 1 6.2827 58.075 -41.758

# + factor(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

# y ~ cancervol + gleason + benpros

#

# Df Sum of Sq RSS AIC

# + factor(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

# y ~ cancervol + gleason + benpros + factor(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

# - factor(vesinv) 1 4.8466 58.075 -41.758

# - benpros 1 7.1115 60.340 -38.047

# - cancervol 1 14.7580 67.987 -26.473

#they all give the same result with my mode 14

#5555555 give the model diagnostics

# residual plot

plot(fitted(fit14), resid(fit14))

abline(h = 0)

# plot of absolute residuals

plot(fitted(fit14), abs(resid(fit14)))

# normal QQ plot

qqnorm(resid(fit14))

qqline(resid(fit14))

# This preliminary model passes the diagnostics. So we can take this

# as our final model.

# Call:

# lm(formula = y ~ cancervol + factor(vesinv) + benpros + gleason,

# data = cancer)

#

# Coefficients:

# (Intercept) cancervol factor(vesinv)1 benpros gleason

# -0.65013 0.06488 0.68421 0.09136 0.33376

#666666 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.

fit14

table(cancer$vesinv) #find the most freaquent

input<-data.frame(cancervol=mean(cancer$cancervol),weight=mean(cancer$weight),age=mean(cancer$age),benpros=mean(cancer$benpros),vesinv=0,capspen=mean(cancer$capspen),gleason=mean(cancer$gleason))

predict(fit14,newdata = input)

# > predict(fit14,newdata = input)

# 1

# 2.330541