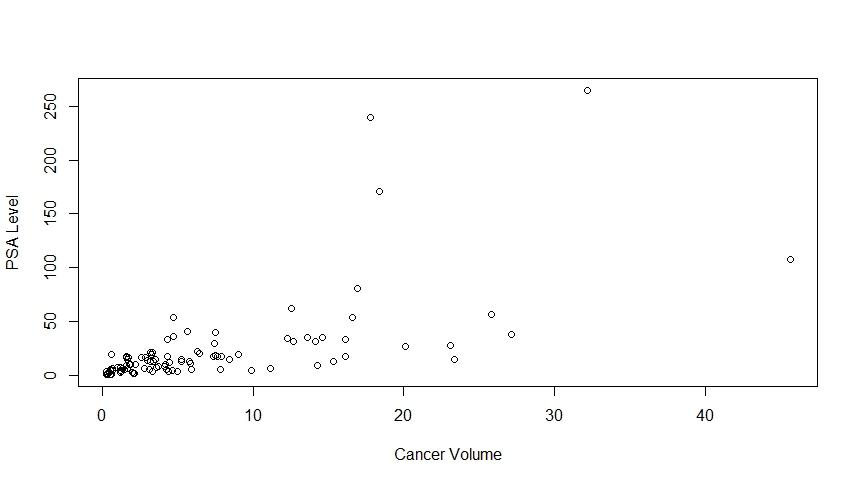
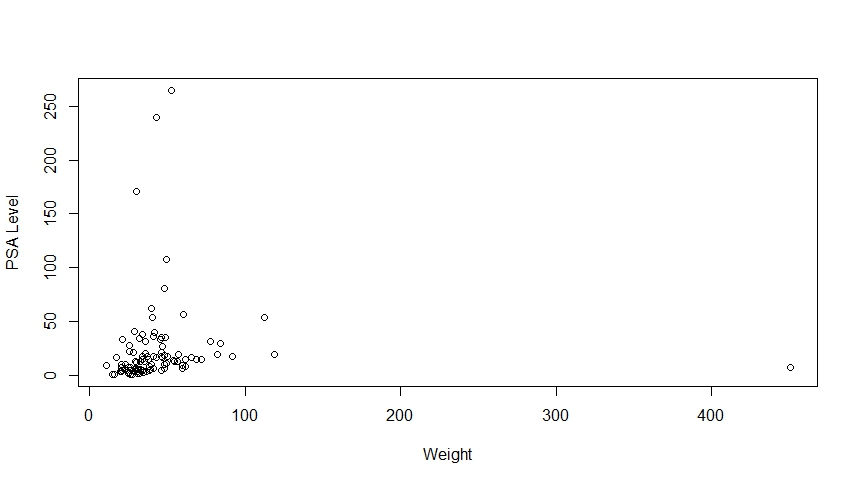
**Mini Project 5**

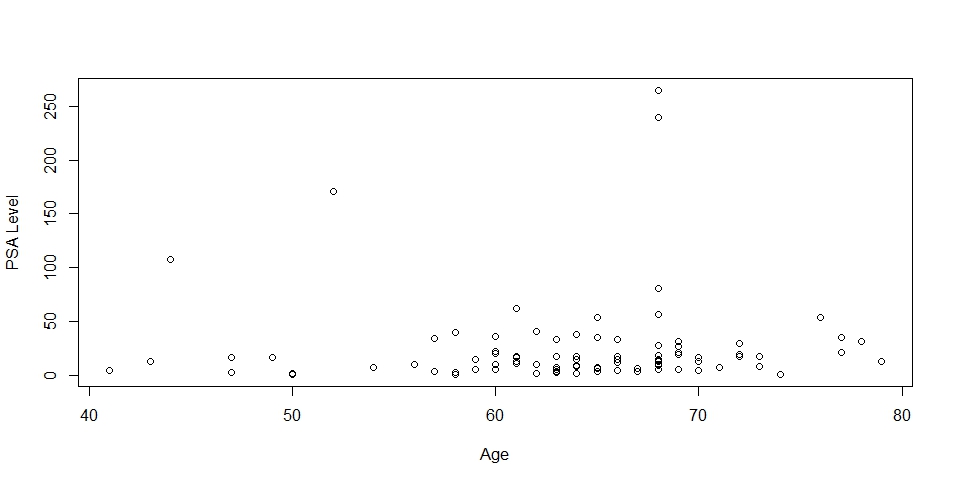
**Section 1**

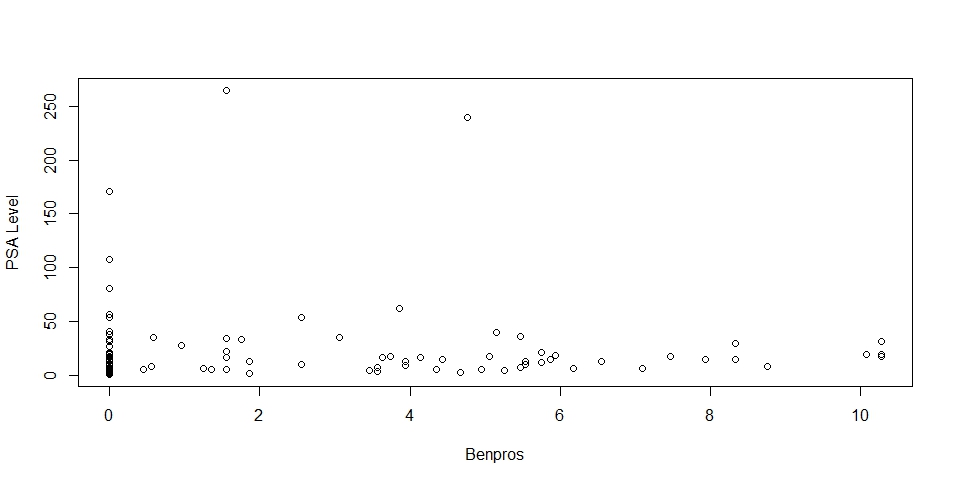
We first read the prostate\_cancer.csv file and saved the data in prostate.

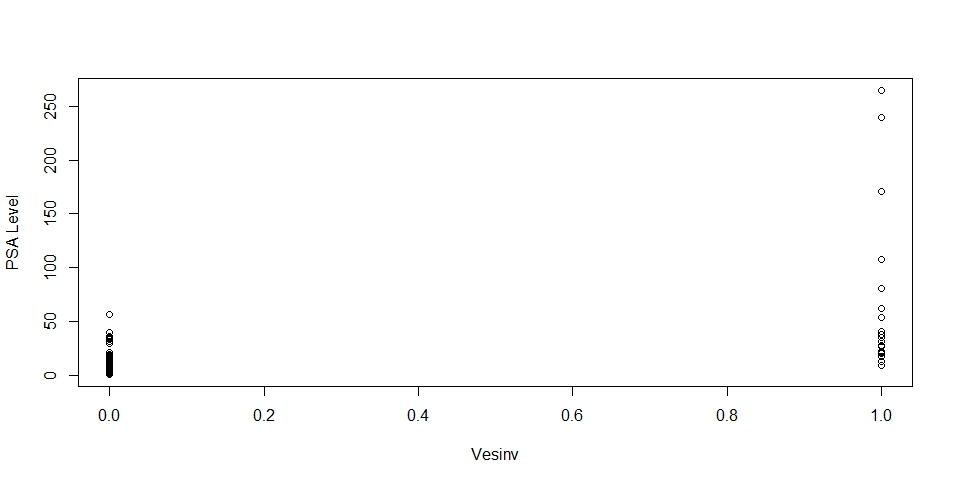
We plotted a scatter plot of all the variables against PSA level

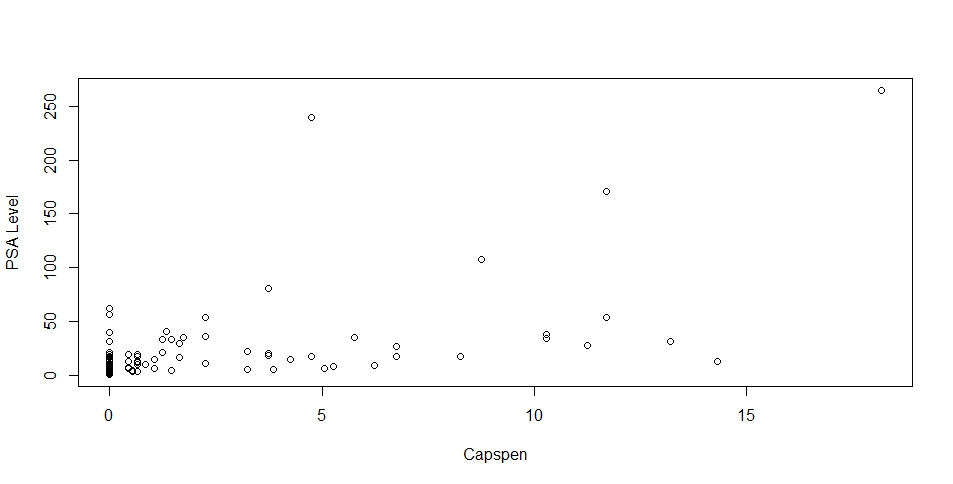


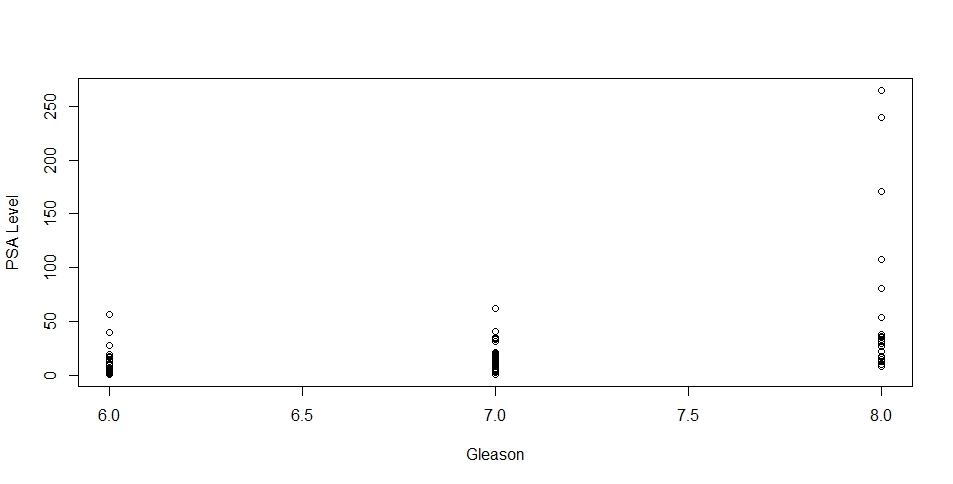












From all the scatter plots for different variables, it is clear that there is a good correlation for PSA level with Cancer volume and weight. We choose the Cancer volume as it has a stronger correlation. So we consider this variable for prediction.

We then also used the lm function to plot psa against all 7 variables (cancervol, wright, age, benpros, vesinv, capspen, and gelason). As the summary shows, the p-value from the F-test is 9.266e-10 with 7 and 89 degrees of feedom.

In **Round 1**, we used Backward elimination algorithm to remove one of the 7 variables using the “update” function. Intriguingly, the “update” function fails to properly process the qualitative variables vesinv and gleason. Thus, for these two updates, we were using the regular lm functions such as:

fit25 <- lm(psa ~ cancervol + weight + age + benpros + capspen + gleason)

fit27 <- lm(psa ~ cancervol + weight + age + benpros + vesinv + capspen)

We subsequently performed ANOVA tests between each of these updated regression models with the original model, and calculated the F-test p-values, which are:

0.0009359 (remove cancervol) \* (\* indicate significant)

0.8787 (remove weight)

0.262 (remove age)

0.2829 (remove benpros)

0.07519 (remove vesinv)

0.4123 (remove capspen)

0.1776 (remove gleason)

Thus, we chose to remove the variable weight.

In **Round 2**, we started with psa ~ cancervol + age + benpros + vesinv + capspen + gleason;

and then removed one of the 6 variables using the “update” function. Again, for the vesinv and gleason, we manually constructed the lm functions as:

fit34 <- lm(psa ~ cancervol + age + benpros + capspen + gleason)

fit36 <- lm(psa ~ cancervol + age + benpros + vesinv + capspen)

We subsequently performed ANOVA tests between each of these updated regression models with the original model, and calculated the F-test p-values, which are:

0.0008356 (remove cancervol) \* (\* indicate significant)

0.2624 (remove age)

0.2402 (remove benpros)

0.07304 (remove vesinv)

0.4103 (remove capspen)

0.1777 (remove gleason)

Thus, we chose to remove the variable capspen.

In **Round 3**, we started with psa ~ cancervol + age + benpros + vesinv + gleason;

and then removed one of the 5 variables using the “update” function. Again, for the vesinv and gleason, we manually constructed the lm functions as:

fit44 <- lm(psa ~ cancervol + age + benpros + gleason)

fit45 <- lm(psa ~ cancervol + age + benpros + vesinv)

We subsequently performed ANOVA tests between each of these updated regression models with the original model, and calculated the F-test p-values, which are:

0.00004072 (remove cancervol) \* (\* indicate significant)

0.2674 (remove age)

0.2326 (remove benpros)

0.01641 (remove vesinv) \*

0.1471 (remove gleason)

Thus, we chose to remove the variable age.

In **Round 4**, we started with psa ~ cancervol + benpros + vesinv + gleason;

and then removed one of the 4 variables using the “update” function. Again, for the vesinv and gleason, we manually constructed the lm functions as:

fit53 <- lm(psa ~ cancervol + benpros + gleason)

fit54 <- lm(psa ~ cancervol + benpros + vesinv)

We subsequently performed ANOVA tests between each of these updated regression models with the original model, and calculated the F-test p-values, which are:

0.00002973 (remove cancervol) \* (\* indicate significant)

0.3987 (remove benpros)

0.02163 (remove vesinv) \*

0.2062 (remove gleason)

Thus, we chose to remove the variable benpros.

In **Round 5**, we started with psa ~ cancervol + vesinv + gleason;

and then removed one of the 3 variables using the “update” function. Again, for the vesinv and gleason, we manually constructed the lm functions as:

fit62 <- lm(psa ~ cancervol + gleason)

fit63 <- lm(psa ~ cancervol + vesinv)

We subsequently performed ANOVA tests between each of these updated regression models with the original model, and calculated the F-test p-values, which are:

0.00003718 (remove cancervol) \* (\* indicate significant)

0.02476 (remove vesinv) \*

0.1693 (remove gleason)

Thus, we chose to remove the variable gleason.

In **Round 6**, we started with psa ~ cancervol + vesinv;

and then removed one of the 2 variables using the “update” function. Again, for the vesinv variable, we manually constructed the lm functions as:

fit72 <- lm(psa ~ cancervol)

We subsequently performed ANOVA tests between each of these updated regression models with the original model, and calculated the F-test p-values, which are:

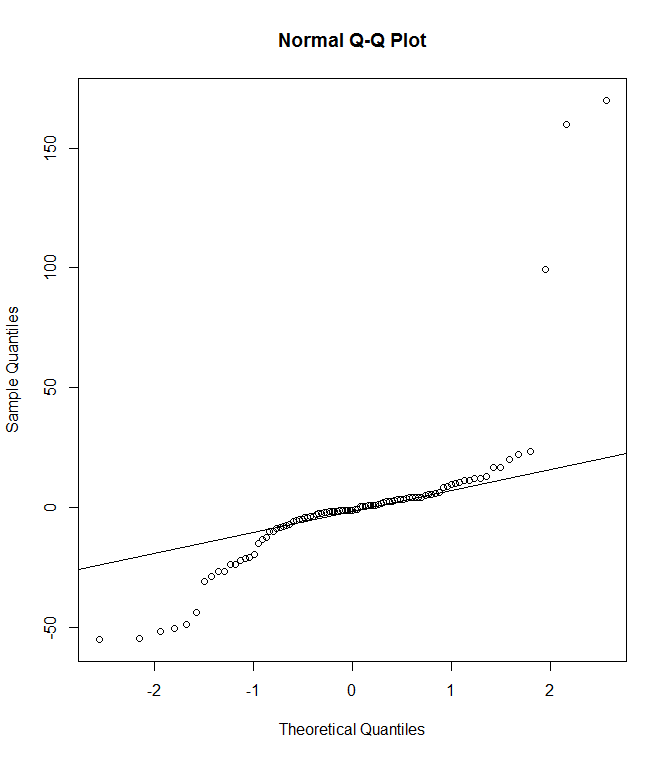
0.00003718 (remove cancervol) \* (\* indicate significant)

0.02476 (remove vesinv) \*

Both are significant, thus we stop the elimination process here, and the final regression model will be: psa ~ cancervol + vesinv.

As the summary shows, this model has a F-test p-value of 3.098e-12 on 2 and 94 degrees of freedom, which is improved from the original model containing all 7 variables.

We then constructed the Normal QQ plot, which showed that the model is reasonable on the middle range, but still have serious deviation at both extremes. It is also important to note that, during this backward elimination process, we constantly resorted to ANOVA tests, which assumes that the error terms follow normal distributions.



Last, we calculated the mean value for the variable cancervol, which is 6.9987. In addition, the mode for the qualitative variable vesinv is calculated as 0.

Our proposed linear model is psa <- 2.477\*cancervol + 24.647\*vesinv +1.060. Thus, the predicted PSA level for a patient whose predictors are at the sample means of the variables will be 18.40.

**Section 2: R code**

# Read the prostate cancer data

prostate <- read.table("C:/Users/yxl121030/My Documents/R/prostate\_cancer.csv", header = T, sep = ",")

# Scatterplot of psa level with other quantitative and qualitative data

plot(prostate$cancervol, prostate$psa,xlab = "Cancer Volume",ylab = "PSA Level")

plot(prostate$weight, prostate$psa,xlab = "Weight",ylab = "PSA Level")

plot(prostate$age, prostate$psa,xlab = "Age",ylab = "PSA Level")

plot(prostate$benpros, prostate$psa,xlab = "Benpros",ylab = "PSA Level")

plot(prostate$vesinv, prostate$psa,xlab = "Vesinv",ylab = "PSA Level")

plot(prostate$capspen, prostate$psa,xlab = "Capspen",ylab = "PSA Level")

plot(prostate$gleason, prostate$psa,xlab = "Gleason",ylab = "PSA Level")

attach(prostate)

# we first plot psa against all 7 variables

fit1 <- lm(psa ~ cancervol + weight + age + benpros + vesinv + capspen + gleason)

summary(fit1)

# Round 1: determine the next to remove

fit21 <- update(fit1, . ~ . - cancervol)

fit22 <- update(fit1, . ~ . - weight)

fit23 <- update(fit1, . ~ . - age)

fit24 <- update(fit1, . ~ . - benpros)

fit25 <- lm(psa ~ cancervol + weight + age + benpros + capspen + gleason)

fit26 <- update(fit1, . ~ . - capspen)

fit27 <- lm(psa ~ cancervol + weight + age + benpros + vesinv + capspen)

anova(fit1, fit21) # F-test value: 0.0009359 (\*)

anova(fit1, fit22) # F-test value: 0.8787

anova(fit1, fit23) # F-test value: 0.262

anova(fit1, fit24) # F-test value: 0.2829

anova(fit1, fit25) # F-test value: 0.07519

anova(fit1, fit26) # F-test value: 0.4123

anova(fit1, fit27) # F-test value: 0.1776

# Round 2: the fit22 yields the largest P-value, remove this "weight" variable, and then determine the next one

fit31 <- update(fit22, . ~ . - cancervol)

fit32 <- update(fit22, . ~ . - age)

fit33 <- update(fit22, . ~ . - benpros)

fit34 <- lm(psa ~ cancervol + age + benpros + capspen + gleason)

fit35 <- update(fit22, . ~ . - capspen)

fit36 <- lm(psa ~ cancervol + age + benpros + vesinv + capspen)

anova(fit22, fit31) # F-test value: 0.0008356 (\*)

anova(fit22, fit32) # F-test value: 0.2624

anova(fit22, fit33) # F-test value: 0.2402

anova(fit22, fit34) # F-test value: 0.07304

anova(fit22, fit35) # F-test value: 0.4103

anova(fit22, fit36) # F-test value: 0.1777

# Round 3: the fit35 yields the largest P-value, remove this "capspen" variable, and then determine the next one

fit41 <- update(fit35, . ~ . -cancervol)

fit42 <- update(fit35, . ~ . -age)

fit43 <- update(fit35, . ~ . -benpros)

fit44 <- lm(psa ~ cancervol + age + benpros + gleason)

fit45 <- lm(psa ~ cancervol + age + benpros + vesinv)

anova(fit35, fit41) # F-test value: 0.00004072 (\*)

anova(fit35, fit42) # F-test value: 0.2674

anova(fit35, fit43) # F-test value: 0.2326

anova(fit35, fit44) # F-test value: 0.01641 (\*)

anova(fit35, fit45) # F-test value: 0.1471

# Round 4: the fit42 yields the largest P-value, remove this "age" variable, and then determine the next one

fit51 <- update(fit42, . ~ . -cancervol)

fit52 <- update(fit42, . ~ . -benpros)

fit53 <- lm(psa ~ cancervol + benpros + gleason)

fit54 <- lm(psa ~ cancervol + benpros + vesinv)

anova(fit42, fit51) # F-test value: 0.00002973 (\*)

anova(fit42, fit52) # F-test value: 0.3987

anova(fit42, fit53) # F-test value: 0.02163 (\*)

anova(fit42, fit54) # F-test value: 0.2062

# Round 5: the fit52 yields the largest P-value, remove this "benpros" variable, and then detrmine the next one

fit61 <- update(fit52, . ~ . -cancervol)

fit62 <- lm(psa ~ cancervol + gleason)

fit63 <- lm(psa ~ cancervol + vesinv)

anova(fit52, fit61) # F-test value: 0.00003718 (\*)

anova(fit52, fit62) # F-test value: 0.02476 (\*)

anova(fit52, fit63) # F-test value: 0.1693

# Round 6: the fit63 yields the largest P-value, remove this "gleason" variable, and then detrmine the next one

fit71 <- update(fit63, . ~ . -cancervol)

fit72 <- lm(psa ~ cancervol)

anova(fit63, fit71) # F-test value: 0.00003718 (\*)

anova(fit63, fit72) # F-test value: 0.02476 (\*)

# Since both two P-values are significant. Thus we stop here. And the final regression will be psa ~ cancervol + vesinv

summary(fit63)

qqnorm(resid(fit63))

qqline(resid(fit63))

# calculate the predicted psa value, the b1 for cancervol is 2.477; the b2 for vesinv is 24.647; the intecept b0 is 1.060

mean1 <- mean(prostate$cancervol)

Mode <- function(x) {

ux <- unique(x)

ux[which.max(tabulate(match(x, ux)))]

}

mean2 <- Mode(prostate$vesinv)

mean1

mean2

predicted\_psa <- 2.477\*mean1 + 24.647\*mean2 + 1.060

predicted\_psa