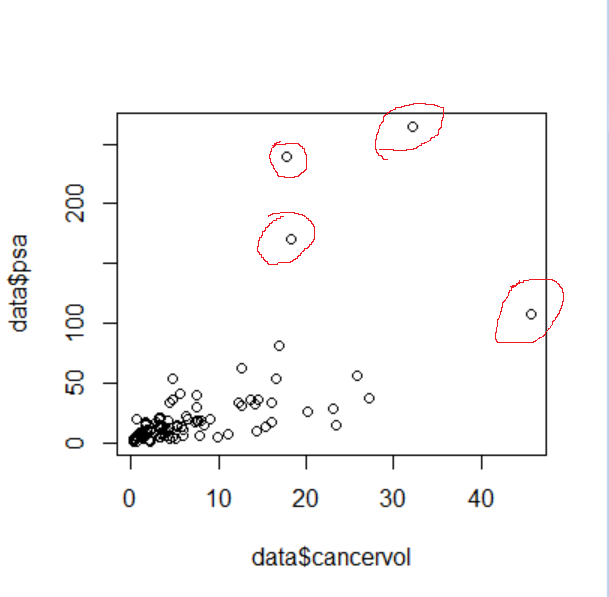
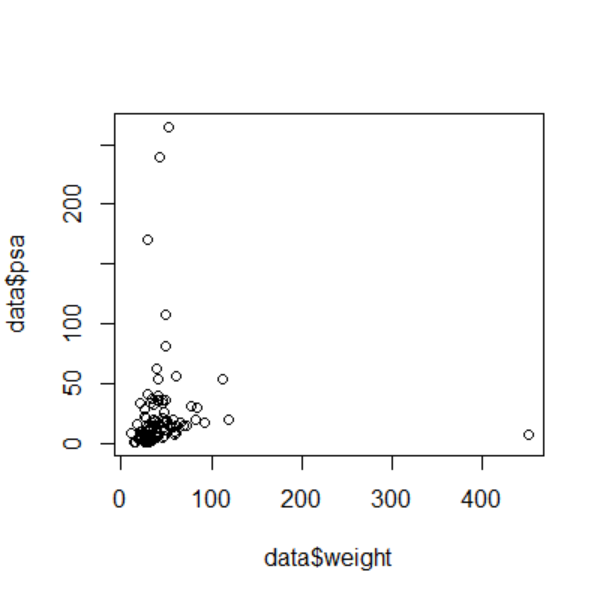
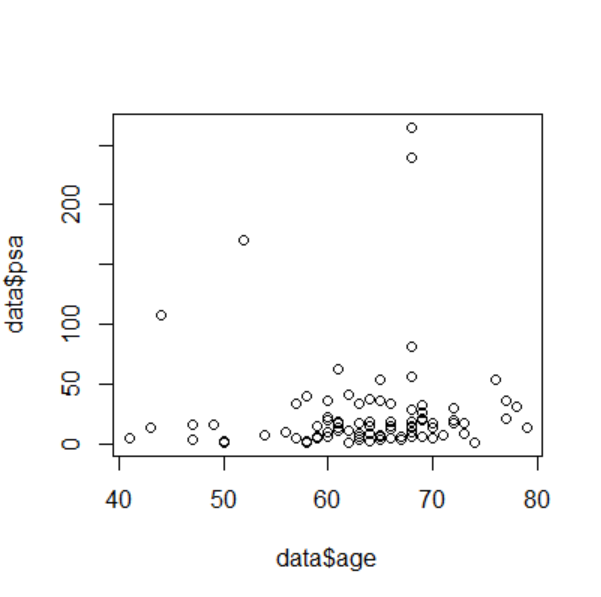
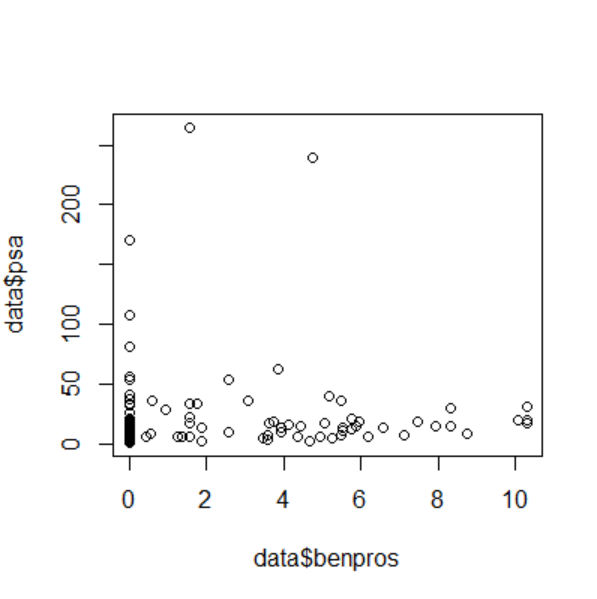
Mini Project #6

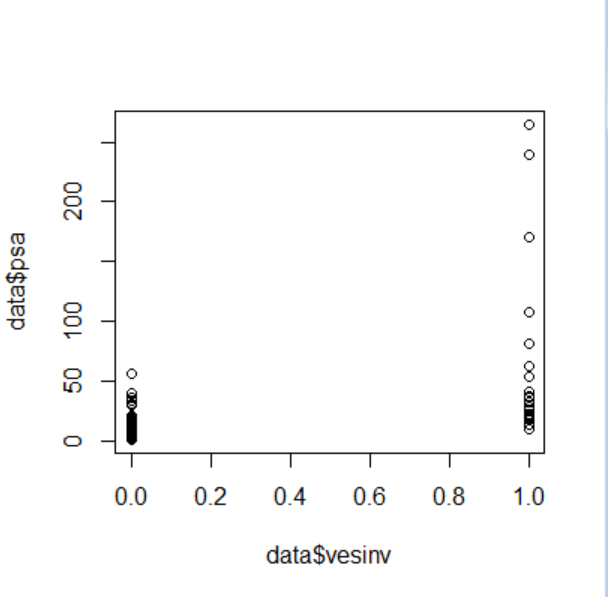
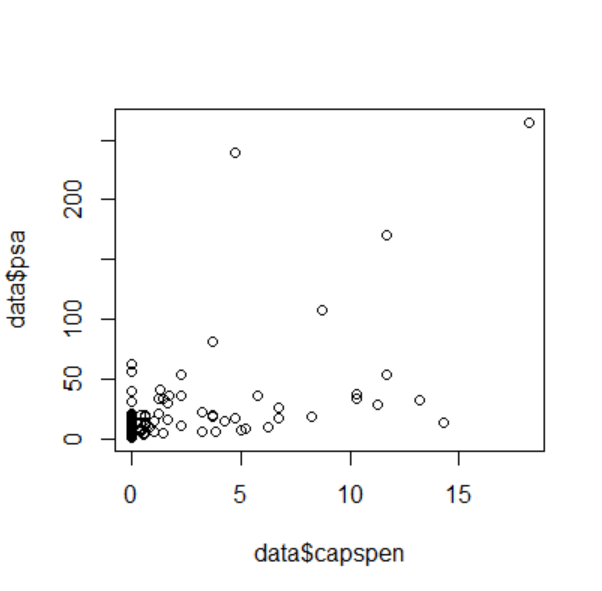
Kainan Xu

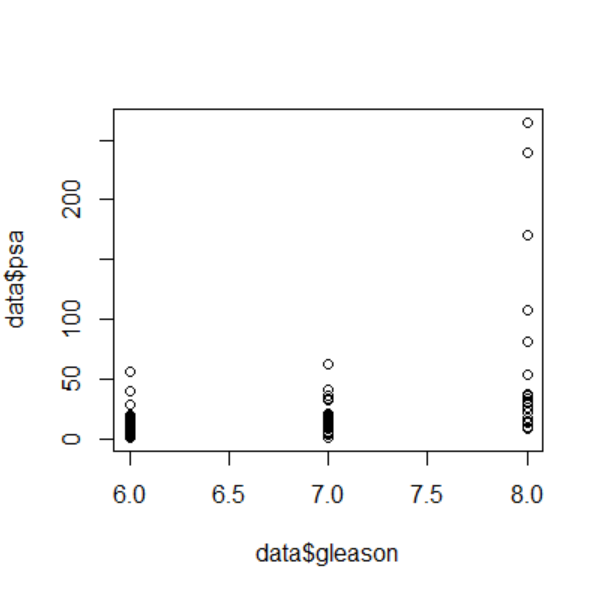
2021229632

1. **Scatterplots of PSA level with other variables are shown below. I choose Cancer Volume (cancervol) as a variable that I think may be used effectively to predict PSA level.**



> data <- read.table("prostate\_cancer.csv", header = T, sep = ",")

> plot(data$cancervol,data$psa)

> plot(data$weight,data$psa)

> plot(data$age,data$psa)

> plot(data$benpros,data$psa)

> plot(data$vesinv,data$psa)

> plot(data$capspen,data$psa)

> plot(data$gleason,data$psa)

1. **Regression diagnostics are shown below. However, the Q-Q plot of residual is not good enough since there are some points do not fall on the line. Then we remedy the model using log of the response and variable.**

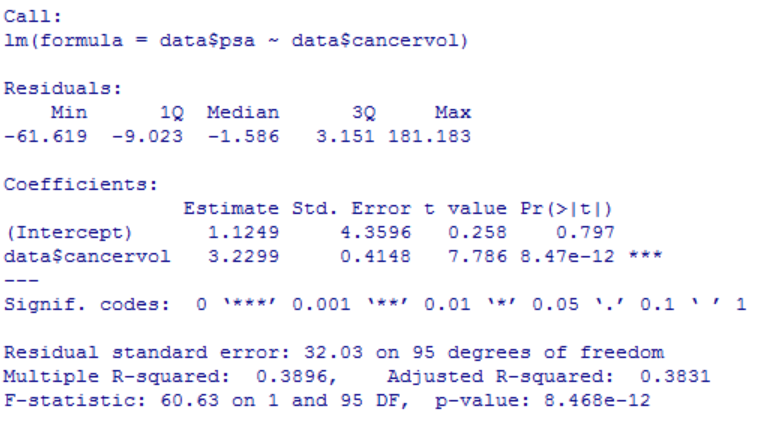
> r = cov(data$cancervol,data$psa)/(sd(data$cancervol)\*sd(data$psa))

> r

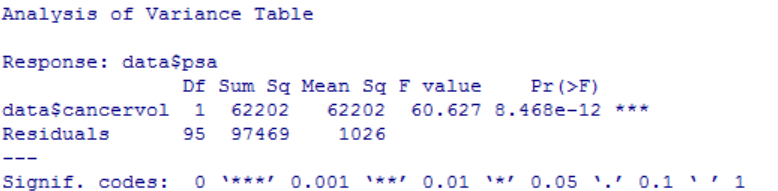
[1] 0.6241506

> fit <- lm(data$psa~data$cancervol)

> summary(fit)



> anova(fit)

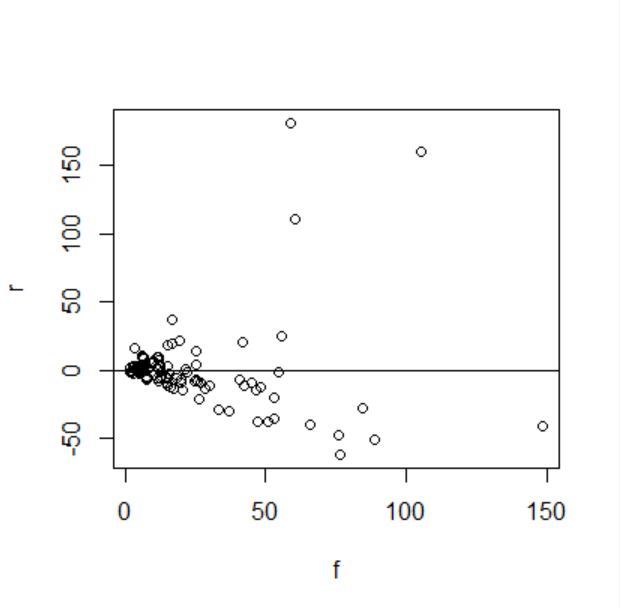


> f = fitted(fit)

> r = resid(fit)

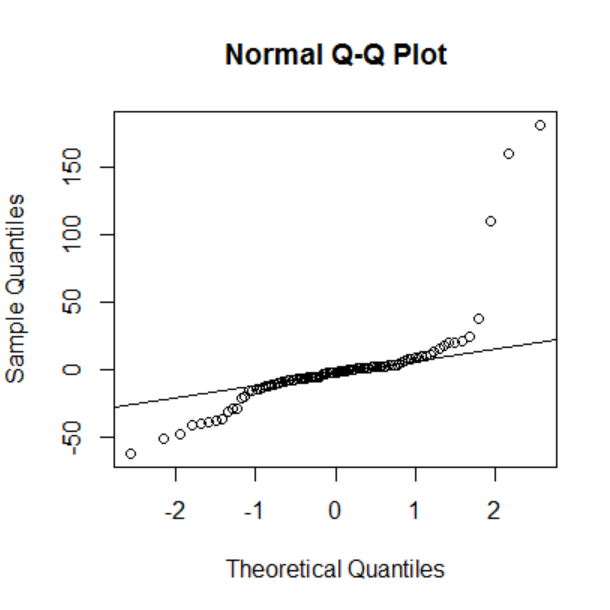
> plot(f,r)

> abline(h = 0)

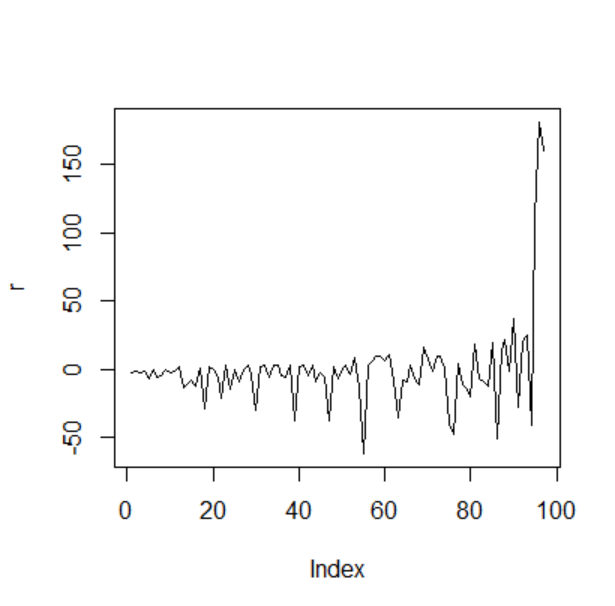


> qqnorm(resid(fit))

> qqline(resid(fit))



> plot(r, type="l")

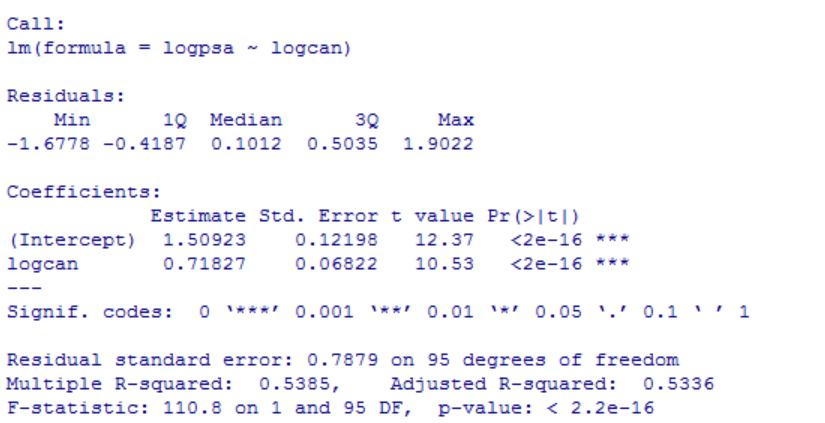


> logpsa = log(data$psa)

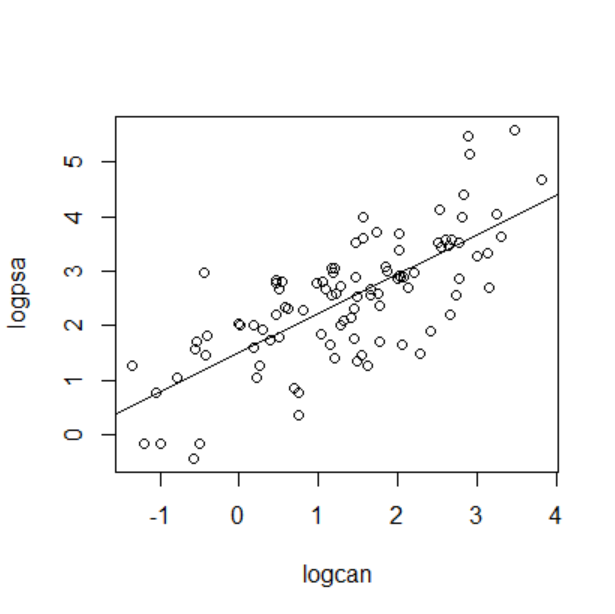
> logcan = log(data$cancervol)

> newfit = lm(logpsa~logcan)

> summary(newfit)



> plot(logcan,logpsa,abline(newfit))

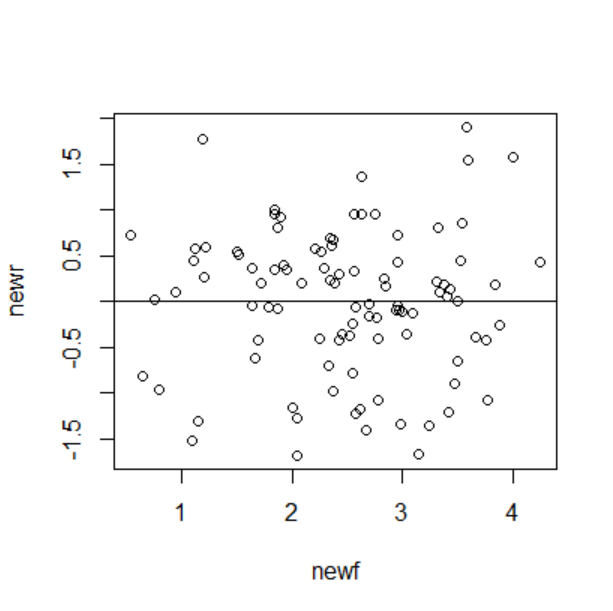


> newf = fitted(newfit)

> newr = resid(newfit)

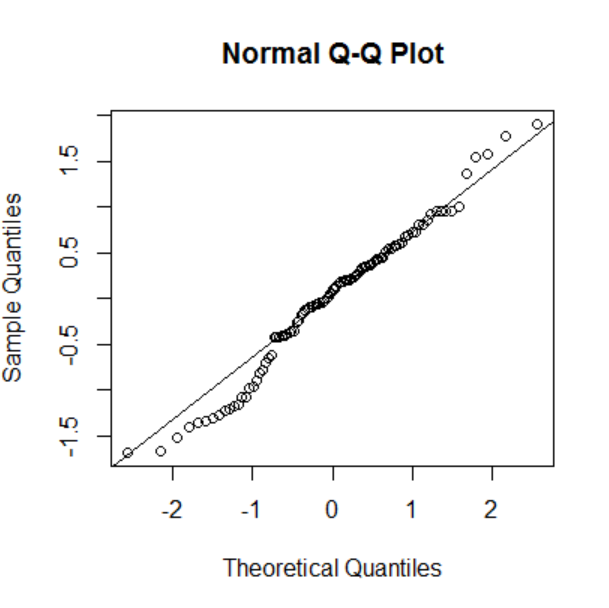
> plot(newf,newr)

> abline(h=0)



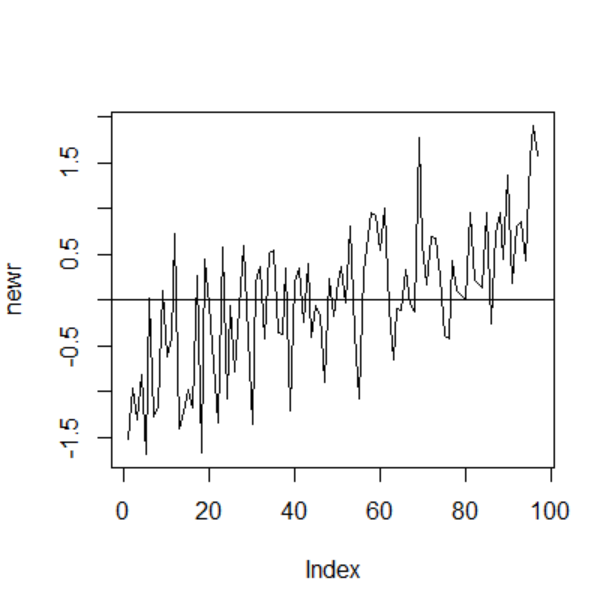
> qqnorm(newr)

> qqline(newr)



> plot(newr,type="l")

> abline(h=0)



**3. When Cancer Volume value is at the median, our new estimated regression line predicts PSA level as 12.81632 mg/ml.**

> p = data.frame(logcan = median(logcan))

> predict(newfit,p)

1

2.550719

> exp(predict(newfit,p))

1

12.81632