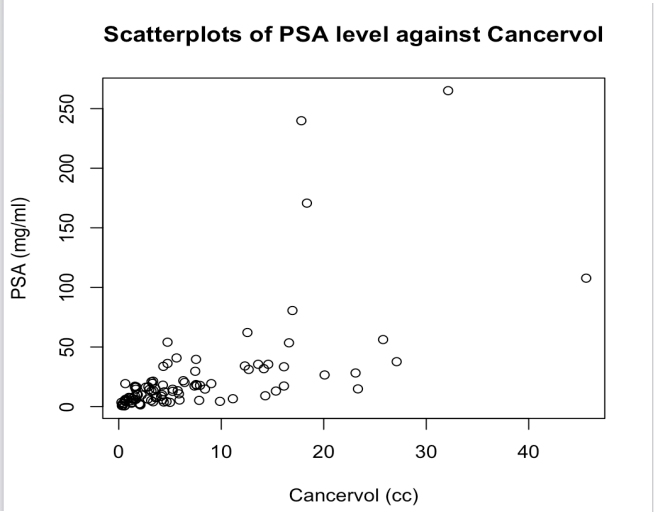
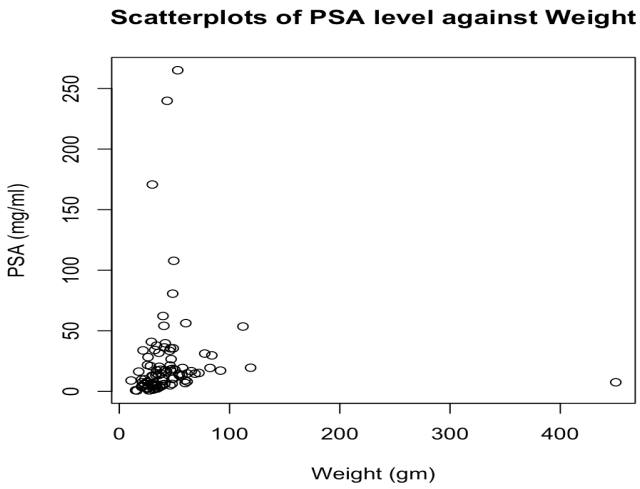
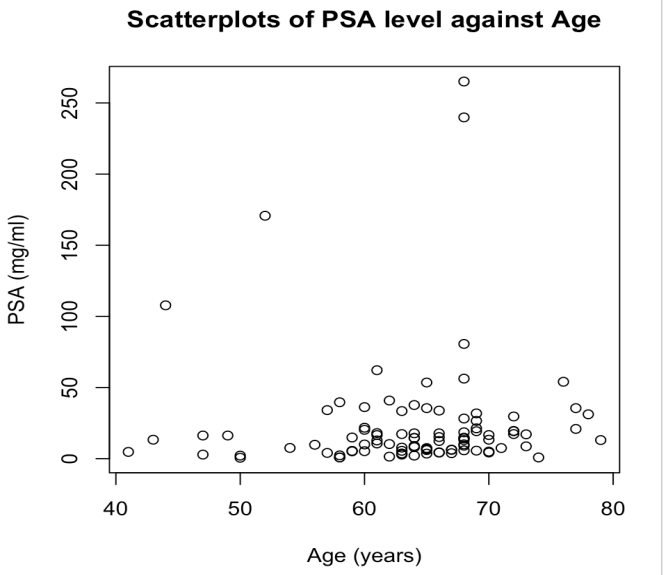
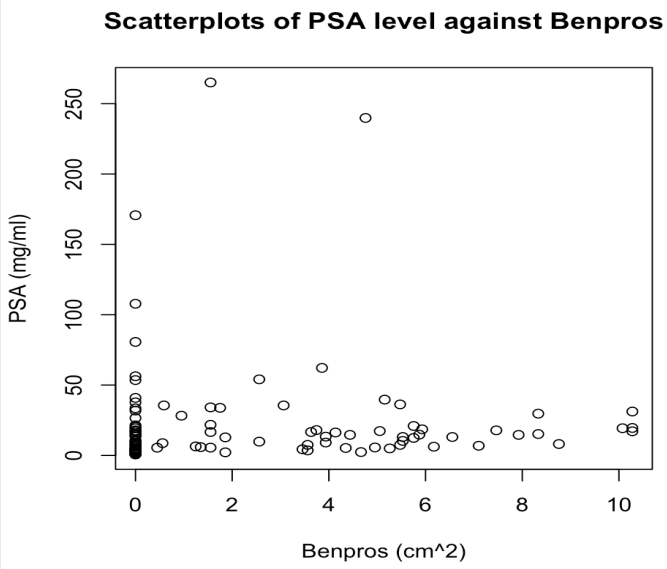
Mini project 5 You Jia(yxj161630)

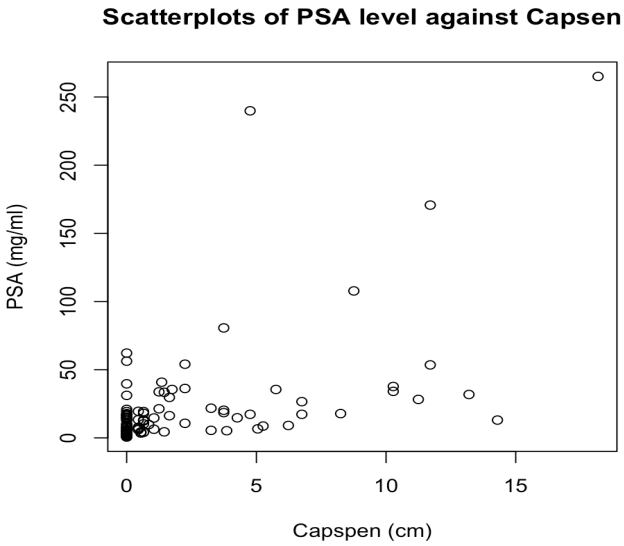
1. We make scatterplots of PSA level against each quantitative variable.

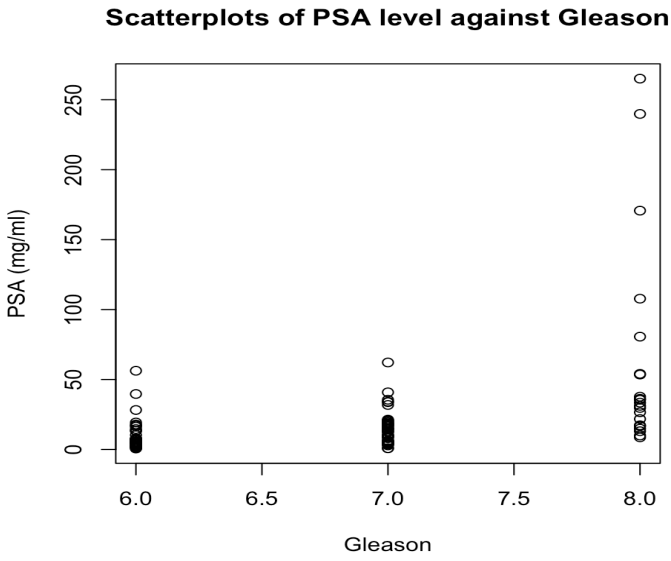












Among the five scatterplots, we can see the points in scatterplot of PSA level against weight are too dense in the corner; the points in scatterplot of Age, Benpros and Capsen don't have strong trend.
It seems like Cancervol has the strongest trend and are more likely to have a largest correlation. Hence, it is the most effectively to predict PSA level.

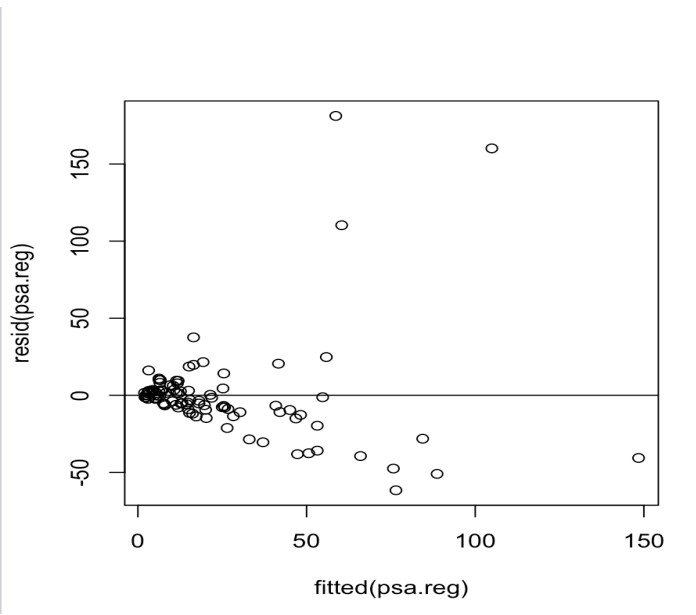
2. Then we fit the simple linear model and do the test of significance.

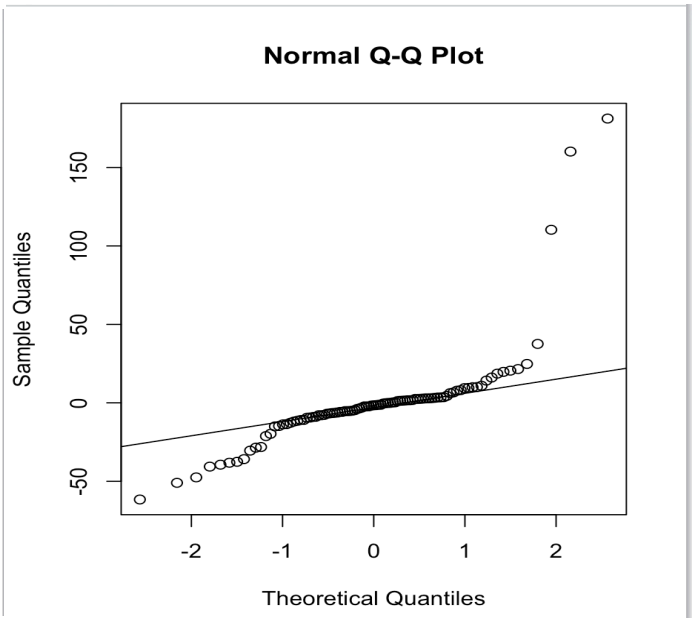
```
Call:
lm(formula = pc$psa ~ pc$cancervol)

Coefficients:
 (Intercept)  pc$cancervol
       1.125         3.230
```

```
> anova(psa.reg)
Analysis of Variance Table

Response: pc$psa
             Df Sum Sq Mean Sq F value    Pr(>F)
pc$cancervol  1  62202   62202  60.627 8.468e-12 ***
Residuals    95  97469    1026
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

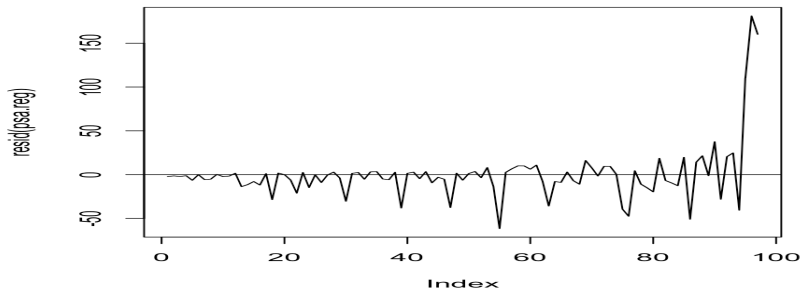
From the results, we can see that the model indeed is significant.

3. Then we make plots to see if the model is met with the three regression assumptions.
plot outcomes:
Residual plot: Normal QQ plot:



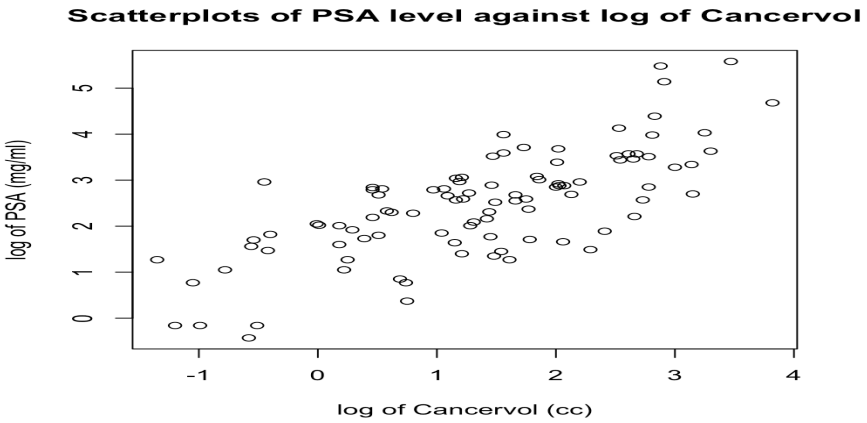


Time series plot of residual:



The model does not meet with the first two assumptions. Hence, we attempted to remedy the situation and do log transformation for the data of PSA level and Cancervol.

4. We make scatter plot after log transformation.



Compared to the scatterplot of the previous data before the log transformation, the trend is stronger.

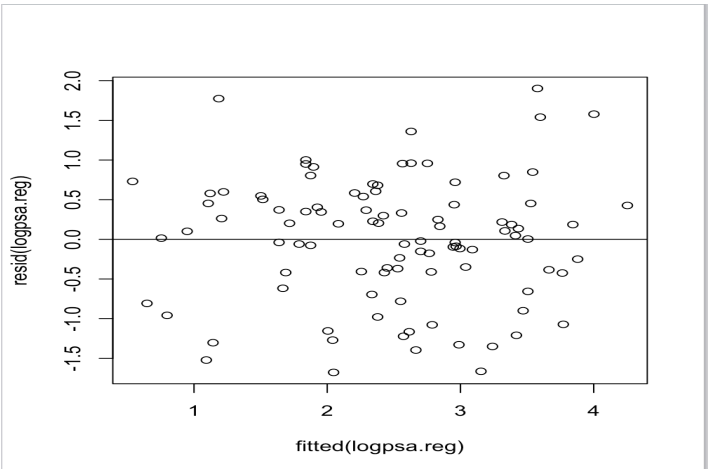
5. Then we fit the simple linear model after the log transformation and do the test of significance.

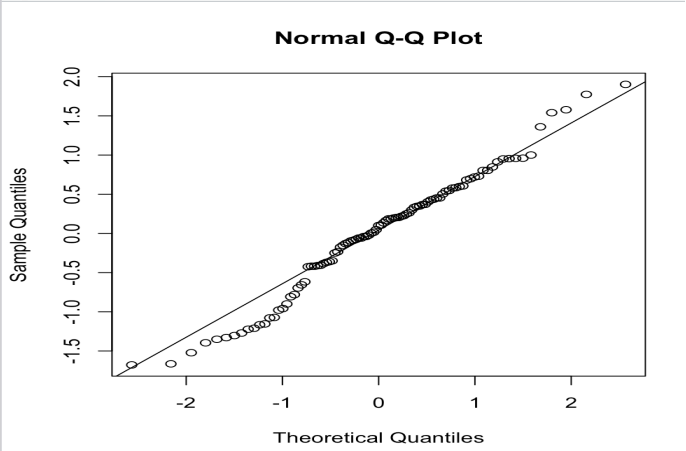
```
Call:
lm(formula = logpsa ~ logcancervol)

Coefficients:
 (Intercept)  logcancervol
      1.5092        0.7183
```

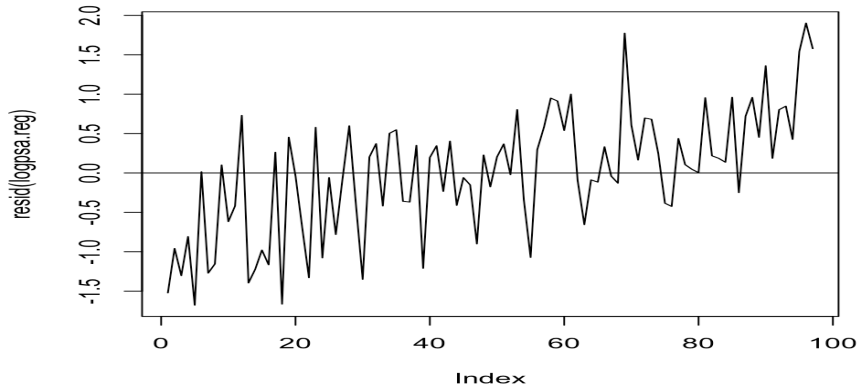
```
> anova(logpsa.reg)
Analysis of Variance Table

Response: logpsa
             Df Sum Sq Mean Sq F value    Pr(>F)
logcancervol  1 68.801  68.801  110.84 < 2.2e-16 ***
Residuals    95 58.968   0.621
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the outcome, we can see that the P-value get smaller. Hence, this model is much better.

6. Then we make plots to see if the model is met with the three assumptions.







From these three plot, we indeed successfully remedy the situation for the first two assumptions. And in the time series plot, we can see the trend, which shows the independent assumption is valid.

7. Now we use the final model to predict the PSA level for a patient whose predictors are at the sample mean of the variables.

```
> newlogcancervol<-data.frame(logcancervol=log(mean(pc$cancervol)))
> logPSAlevel<-predict(logpsa.reg,newdata=newlogcancervol)
> logPSAlevel  # 2.906784
       1
2.906784
> PSAlevel=exp(logPSAlevel)
> PSAlevel   # 18.29785
       1
18.29785
```

The PSA level for a patient whose predictor variable value is at the sample mean of the variable is 18.29785 mg/ml.

R Code:

```
getwd()
setwd("/Users/youjia/Desktop")
getwd()

pc <- read.table("prostate_cancer.csv", header = TRUE, sep = ",")
#vesinv is not quantitative predictors
plot(pc$cancervol,pc$psa,main='Scatterplots of PSA level against Cancervol',xlab='Cancervol (cc)',ylab='PSA (mg/ml)')
plot(pc$weight,pc$psa,main='Scatterplots of PSA level against Weight',xlab='Weight (gm)',ylab='PSA (mg/ml)')
plot(pc$age,pc$psa,main='Scatterplots of PSA level against Age',xlab='Age (years)',ylab='PSA (mg/ml)')
plot(pc$benpros,pc$psa,main='Scatterplots of PSA level against Benpros',xlab='Benpros (cm^2)',ylab='PSA (mg/ml)')
plot(pc$capspen,pc$psa,main='Scatterplots of PSA level against Capsen',xlab='Capspen (cm)',ylab='PSA (mg/ml)')
plot(pc$gleason,pc$psa,main='Scatterplots of PSA level against Gleason',xlab='Gleason',ylab='PSA (mg/ml)')

# According to the scatterplots, we choose cancervol as the most effectively to predict PSA level
# Get the simple linear model and do the test of significance
lm(pc$psa~pc$cancervol)
psa.reg <- lm(pc$psa~pc$cancervol)
anova(psa.reg)

# Use plots to see if the model is met with the three assumption
# residual plot
plot(fitted(psa.reg),resid(psa.reg))
abline(h=0)

# qq plot
qqnorm(resid(psa.reg))
qqline(resid(psa.reg))

#time series plot of residuals
plot(resid(psa.reg),type="l")
abline(h=0)

#This doesn't satisfied with the first two assumption
#In this case, we can do log transformation of cancervol and psa
logpsa<-log(pc$psa)
logcancervol<-log(pc$cancervol)
plot(logcancervol,logpsa,main='Scatterplots of PSA level against log of Cancervol',xlab='log of Cancervol (cc)',
     ylab='log of PSA (mg/ml)')

# Get the simple linear model after the log transformation and do the test of significance
lm(logpsa~logcancervol)
logpsa.reg<-lm(logpsa~logcancervol)
anova(logpsa.reg)

# Use plots to see if the model after the log transformation is met with the three assumption
# Residual plot
plot(fitted(logpsa.reg),resid(logpsa.reg))
abline(h=0)
```

```
# QQ plot
qqnorm(resid(logpsa.reg))
qqline(resid(logpsa.reg))

# Time series plot of residuals
plot(resid(logpsa.reg),type="l")
abline(h=0)

# Use the final model to predict the PSA level for a patient whose
# predictor variable value is at the sample mean of the variable.
newlogcancervol<-data.frame(logcancervol=log(mean(pc$cancervol)))
logPSAlevel<-predict(logpsa.reg,newdata=newlogcancervol)
logPSAlevel  # 2.906784
PSAlevel=exp(logPSAlevel)
PSAlevel   # 18.29785
```