homework 4

Yi (Chris) Chen

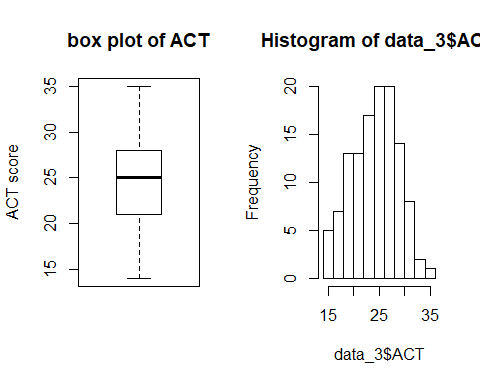
October 12, 2017

### homework four

#### problem 3.3

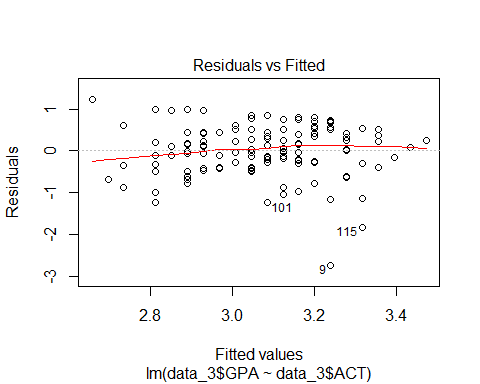
##### a. prepare a box plot for ACT score

setwd("C:/Users/cheny/Desktop/study/linear regression model/homework/homework four")  
data\_3 <- read.table('3.3.txt',header = FALSE, col.names = c('GPA','ACT','intelligence test score','high school class rank'))  
par(mfrow=c(1,2))  
boxplot(data\_3$ACT,main ='box plot of ACT',ylab='ACT score')  
hist(data\_3$ACT)



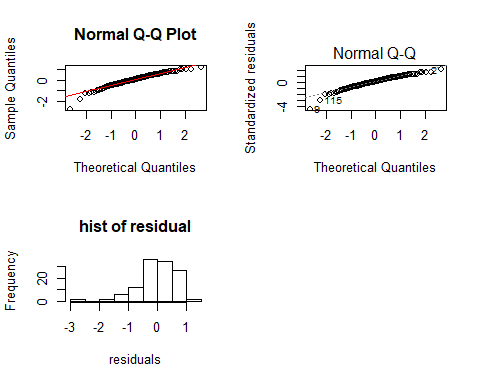
##### c. plot the residual against the fitted values

reg\_3 <- lm(data\_3$GPA ~ data\_3$ACT)  
residual\_3 <- reg\_3$residuals  
fitted\_value\_3 <- reg\_3$fitted.values  
plot(reg\_3,1)



##### d. plot the normal probability plot of residuals.

par(mfrow=c(2,2))  
qqnorm(reg\_3$residuals)  
qqline(reg\_3$residuals,col="red")  
plot(reg\_3,2)  
hist(reg\_3$residuals,main = 'hist of residual',xlab = 'residuals')



# calculate the coefficient of correlation between ordered residuals and their expected values under normality  
  
ordered\_residual <- reg\_3$residuals[order(reg\_3$residuals)]  
# according to the table B6 (n = 120, ?? = 0.05) z equals to 0.987  
MSE <- sum(reg\_3$residuals^2)/(length(reg\_3$residuals)-2)  
a <- vector()  
for(i in 1:length(residual\_3)){  
 a\_new <- ( i - 0.375)/(120 + 0.25)  
 a <- c(a,a\_new)  
}  
expected\_value <- sqrt(MSE) \* qnorm(a)  
correlation <- cor(ordered\_residual,expected\_value)  
correlation

## [1] 0.9737275

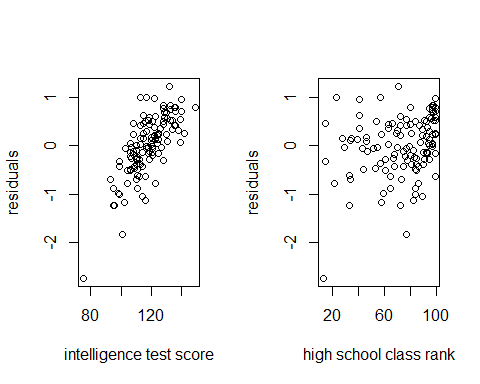
##### e. Brown-Forsythe Test

# order the ACT  
low\_26 <- which(data\_3$ACT < 26)  
residual\_low <- residual\_3[low\_26]  
  
up\_26 <- which(data\_3$ACT >= 26)  
residual\_up <- residual\_3[up\_26]  
  
median\_low <- median(residual\_low)  
median\_up <- median(residual\_up)  
  
d\_low <- vector()  
for( i in 1:length(residual\_low)){  
 d\_low\_new <- abs(residual\_low[i]-median\_low)   
 d\_low <- c(d\_low,d\_low\_new)  
}  
  
d\_up <- vector()  
for( i in 1:length(residual\_up)){  
 d\_up\_new <- abs(residual\_up[i]-median\_up)   
 d\_up <- c(d\_up,d\_up\_new)  
}  
  
mean\_d\_low <- mean(d\_low)  
mean\_d\_up <- mean(d\_up)  
  
diff\_low <- vector()  
for(i in 1:length(d\_low)){  
 diff\_low\_new <- (d\_low[i] - mean\_d\_low)^2  
 diff\_low <- c(diff\_low,diff\_low\_new)  
}  
  
diff\_up <- vector()  
for(i in 1:length(d\_up)){  
 diff\_up\_new <- (d\_up[i] - mean\_d\_up)^2  
 diff\_up <- c(diff\_up,diff\_up\_new)  
}  
  
s <- sqrt((sum(diff\_up)+sum(diff\_low))/118)  
# result of t\_BF  
t\_BF <- (mean\_d\_up - mean\_d\_low)/(0.417275 \* sqrt(1/length(residual\_up)+1/length(residual\_low)))  
#calculate t\_critical  
t\_critical <- qt(0.995,118)  
  
# test  
abs(t\_BF) <= t\_critical

## [1] TRUE

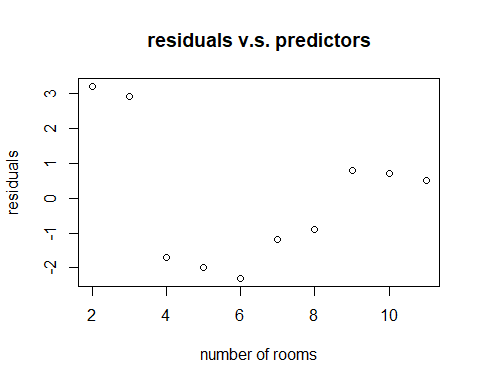
##### f. test for potential variables

par(mfrow=c(1,2))  
plot(reg\_3$residuals ~ data\_3$intelligence.test.score, ylab = 'residuals', xlab='intelligence test score')  
plot(reg\_3$residuals ~ data\_3$high.school.class.rank, ylab = 'residuals', xlab='high school class rank')

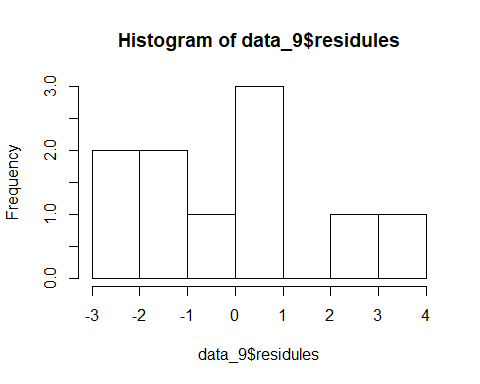


#### problem 3.9

data\_9 <- read.table('3.9.txt',header = FALSE, col.names = c('numbrs of rooms','residules'))  
plot(data\_9$residules ~ data\_9$numbrs.of.rooms,xlab='number of rooms',ylab='residuals' ,main='residuals v.s. predictors')



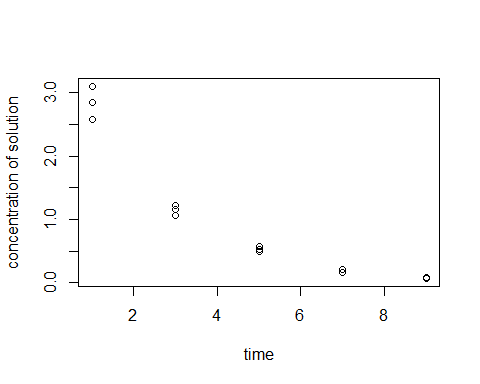
hist(data\_9$residules)



#### problem 3.16

##### a. plot the scatter

data\_16 <- read.table('3.16.txt',header = FALSE, col.names = c('concentration of solution','time'))  
plot(data\_16$concentration.of.solution ~ data\_16$time, ylab='concentration of solution',xlab='time')



##### b. Box-Cox Transformations

result <- data.frame()  
result$lambda <- vector()  
result$SSE <- vector()  
  
k2 <- prod(data\_16$concentration.of.solution)^(1/length(data\_16$concentration.of.solution))  
  
# test 0  
wi <- k2\*log(data\_16$concentration.of.solution)  
reg\_16\_0 <- lm(wi~data\_16$time)  
sse\_0 <- sum(reg\_16\_0$residuals^2)  
result[1,] <- c(0,sse\_0)  
  
# test other values  
test <- c(-0.2,-0.1,0.1,0.2)  
for(j in 1:length(test)){  
 i <- test[j]  
 k1 <- 1/(i\*k2^(i-1))  
 wi <- k1\*(data\_16$concentration.of.solution^i - 1)  
 reg <- lm(wi~data\_16$time)  
 sse <- sum(reg$residuals^2)  
 result[j+1,] <- c(i,sse)  
}  
print(result)

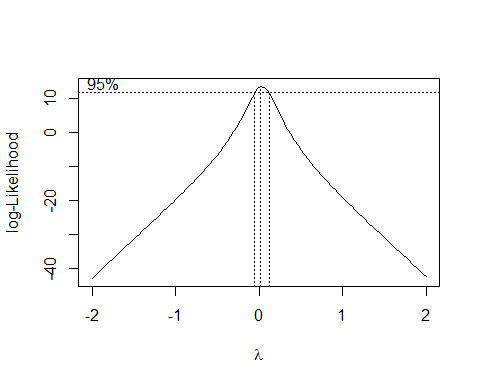
## lambda SSE  
## 1 0.0 0.03897303  
## 2 -0.2 0.12353047  
## 3 -0.1 0.06505067  
## 4 0.1 0.04396062  
## 5 0.2 0.08131793

# this problem can also be solved in this way  
library(MASS)

## Warning: package 'MASS' was built under R version 3.4.2

tr <- boxcox(data\_16$concentration.of.solution~data\_16$time,Lambda=seq(-1,+1,20))

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'Lambda' will be disregarded



# obviously, lambda should be 0

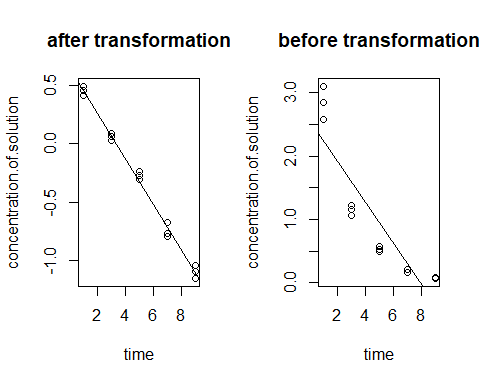
##### c. estimate the y'

y\_dot <- log10(data\_16$concentration.of.solution)  
reg\_16\_c <- lm(y\_dot~data\_16$time)  
summary(reg\_16\_c)

##   
## Call:  
## lm(formula = y\_dot ~ data\_16$time)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.082958 -0.044421 0.006813 0.033512 0.085550   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.654880 0.026181 25.01 2.22e-12 \*\*\*  
## data\_16$time -0.195400 0.004557 -42.88 2.19e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04992 on 13 degrees of freedom  
## Multiple R-squared: 0.993, Adjusted R-squared: 0.9924   
## F-statistic: 1838 on 1 and 13 DF, p-value: 2.188e-15

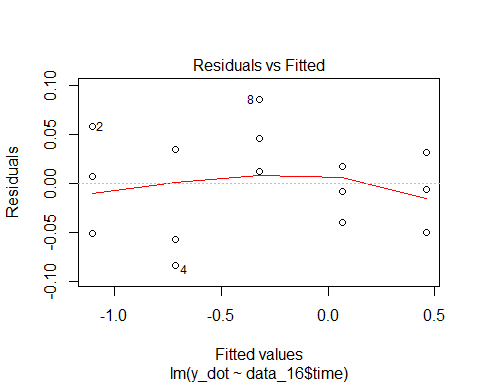
##### d. print the picture to compare the results

reg\_16\_orginal <- lm(data\_16$concentration.of.solution~data\_16$time)  
par(mfrow=c(1,2))  
plot(y\_dot~data\_16$time,main='after transformation',xlab='time',ylab='concentration.of.solution')  
abline(reg\_16\_c)  
plot(data\_16$concentration.of.solution~data\_16$time,main='before transformation',xlab='time',ylab='concentration.of.solution')  
abline(reg\_16\_orginal)

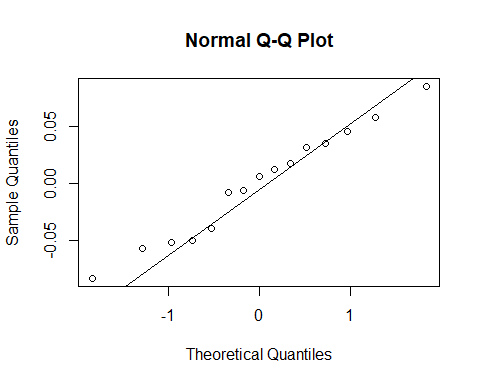


##### e. analysis the new residuals

# residual against fitted value  
plot(reg\_16\_c,1)



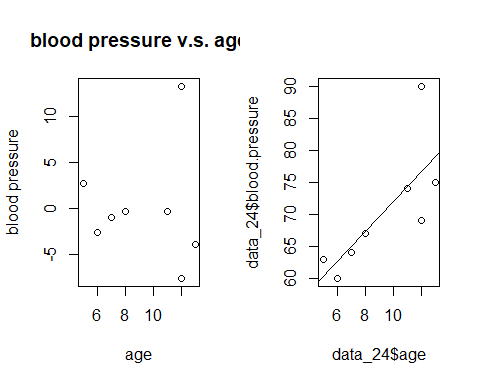
# test normality  
qqnorm(reg\_16\_c$residuals)  
qqline(reg\_16\_c$residuals)



#### problem 24

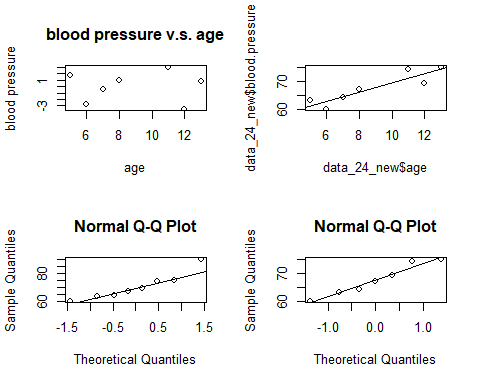
##### a. estimate the orginal data

data\_24 <- read.table('3.24.txt',header = FALSE, col.names = c('blood pressure','age'))  
reg\_24 <- lm(data\_24$blood.pressure ~ data\_24$age)  
par(mfrow=c(1,2))  
plot(reg\_24$residuals ~ data\_24$age, main='blood pressure v.s. age',xlab='age',ylab='blood pressure')  
plot(data\_24$blood.pressure ~ data\_24$age)  
abline(reg\_24)



##### b. estimate after omitting one possible outlier

data\_24\_new <- data\_24[-7,]  
reg\_24\_new <- lm(data\_24\_new$blood.pressure ~ data\_24\_new$age)  
par(mfrow=c(2,2))  
plot(reg\_24\_new$residuals ~ data\_24\_new$age, main='blood pressure v.s. age',xlab='age',ylab='blood pressure')  
plot(data\_24\_new$blood.pressure ~ data\_24\_new$age)  
abline(reg\_24\_new)  
qqnorm(data\_24$blood.pressure)  
qqline(data\_24$blood.pressure)  
qqnorm(data\_24\_new$blood.pressure)  
qqline(data\_24\_new$blood.pressure)

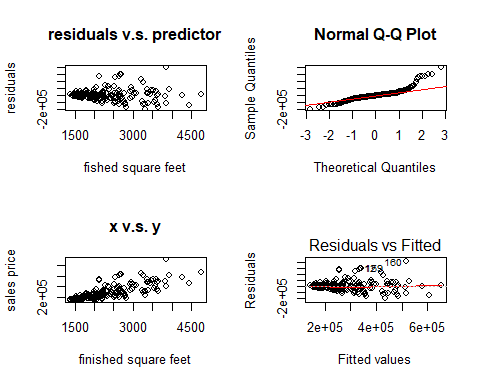


#### Problem 3.31

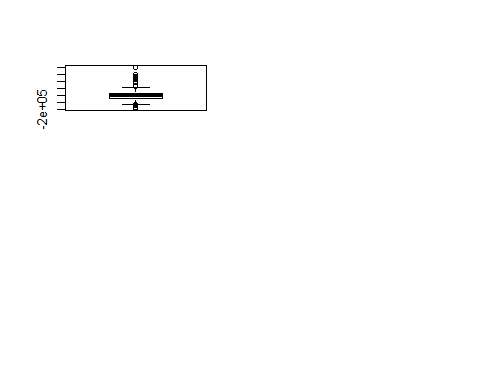
data\_31 <- read.table('3.31.txt',header = FALSE,col.names = c('indentical number','sales prive','finished square feet','nuber of bedrooms','number of bathrooms','air conditioning','garage size','pool','year built','quality','style','lot size','adjacent of high way'))  
set.seed(1)  
# obtain a random sample  
sample\_31 <- sample(500,200)  
data\_31 <- data\_31[sample\_31,]  
reg\_31 <- lm(data\_31$sales.prive~data\_31$finished.square.feet)  
summary(reg\_31)

##   
## Call:  
## lm(formula = data\_31$sales.prive ~ data\_31$finished.square.feet)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -192464 -37040 -7041 21842 401946   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -65694.5 20291.9 -3.237 0.00141 \*\*   
## data\_31$finished.square.feet 151.3 8.5 17.806 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 81630 on 198 degrees of freedom  
## Multiple R-squared: 0.6156, Adjusted R-squared: 0.6136   
## F-statistic: 317 on 1 and 198 DF, p-value: < 2.2e-16

# plot  
par(mfrow=c(2,2))  
plot(reg\_31$residuals~data\_31$finished.square.feet,main='residuals v.s. predictor', xlab='fished square feet', ylab='residuals')  
qqnorm(reg\_31$residuals)  
qqline(reg\_31$residuals,col="red")  
plot(data\_31$sales.prive~data\_31$finished.square.feet,main='x v.s. y',xlab='finished square feet',ylab='sales price')  
plot(reg\_31,1)

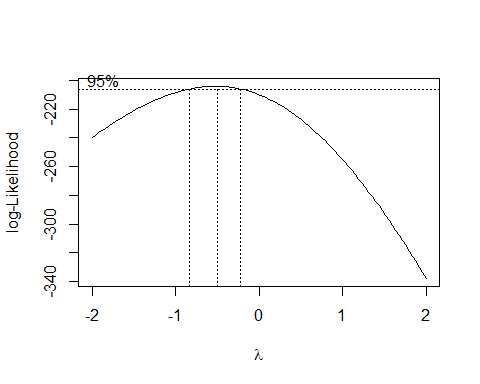


# outlier test  
boxplot(reg\_31$residuals)

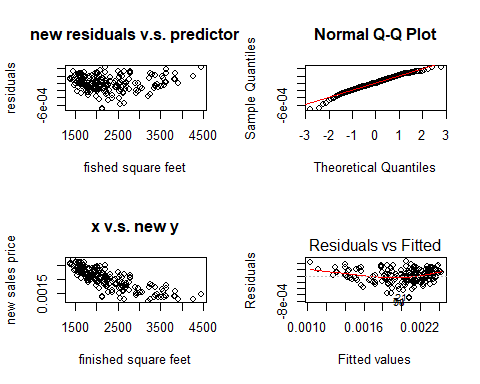


# remove the residual outliers  
outlier <- vector()  
a <- as.numeric(reg\_31$residuals)  
for(i in 1:6){  
 new\_large <- as.numeric(which.max(a))  
 outlier <- c(outlier,new\_large)  
 a <- a[-new\_large]  
  
}  
a <- as.numeric(reg\_31$residuals)  
for(i in 1:2){  
 new\_large <- as.numeric(which.min(a))  
 outlier <- c(outlier,new\_large)  
 a <- a[-new\_large]  
  
}   
data\_31 <- data\_31[-outlier,]  
  
# box cox transformation  
library(MASS)  
tr\_31 <- boxcox(data\_31$sales.prive~data\_31$finished.square.feet,Lambda=seq(-2,+2,100))

## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'Lambda' will be disregarded



# new regression  
y\_new\_31 <- data\_31$sales.prive^(-1/2)  
new\_regress\_31 <- lm(y\_new\_31 ~ data\_31$finished.square.feet )  
par(mfrow=c(2,2))  
plot(new\_regress\_31$residuals~data\_31$finished.square.feet,main='new residuals v.s. predictor', xlab='fished square feet', ylab='residuals')  
qqnorm(new\_regress\_31$residuals)  
qqline(new\_regress\_31$residuals,col="red")  
plot(y\_new\_31~data\_31$finished.square.feet,main='x v.s. new y',xlab='finished square feet',ylab='new sales price')  
plot(new\_regress\_31,1)



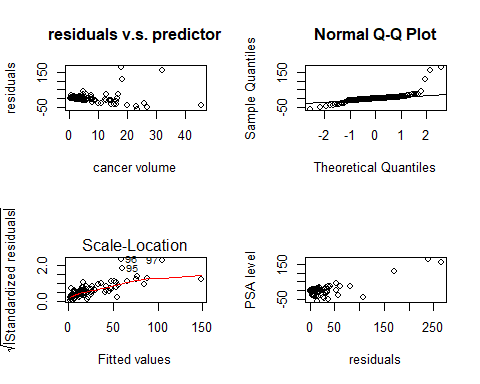
# take into the new x  
y1 <- new\_regress\_31$coefficients[1] + 1100 \* new\_regress\_31$coefficients[2]  
y1\_true <- (1/y1)^2  
y2 <- new\_regress\_31$coefficients[1] + 4900 \* new\_regress\_31$coefficients[2]  
y2\_true <- (1/y2)^2

#### Problem 3.32

data\_32 <- read.table('3.32.txt',header = FALSE, col.names = c('identifican number','PSA level','cancer volume','weight','age','benign prostatic hyperplasia','seminal vesicle invasion','capsular penetration','gleason score'))  
reg\_32 <- lm(data\_32$PSA.level~data\_32$cancer.volume)  
summary(reg\_32)

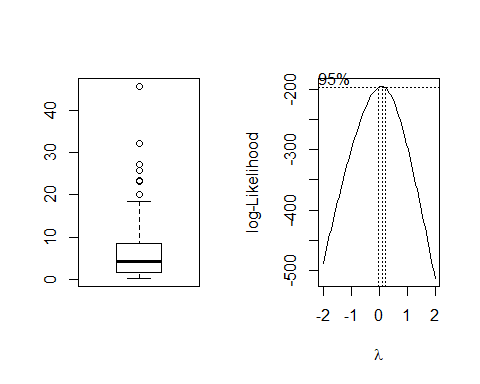
##   
## Call:  
## lm(formula = data\_32$PSA.level ~ data\_32$cancer.volume)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.619 -9.023 -1.586 3.151 181.183   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.1249 4.3596 0.258 0.797   
## data\_32$cancer.volume 3.2299 0.4148 7.786 8.47e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.03 on 95 degrees of freedom  
## Multiple R-squared: 0.3896, Adjusted R-squared: 0.3831   
## F-statistic: 60.63 on 1 and 95 DF, p-value: 8.468e-12

par(mfrow=c(2,2))  
plot(reg\_32$residuals~data\_32$cancer.volume,main='residuals v.s. predictor', xlab='cancer volume', ylab='residuals')  
qqnorm(reg\_32$residuals)  
qqline(reg\_32$residuals)  
plot(reg\_32,3)  
plot(reg\_32$residuals~data\_32$PSA.level,xlab = 'residuals',ylab = 'PSA level')



#remove the outlier of predictor  
par(mfrow=c(1,2))  
boxplot(data\_32$cancer.volume)  
data\_32 <- data\_32[order(data\_32$cancer.volume)[1:93],]  
  
tr\_32 <- boxcox(data\_32$PSA.level~data\_32$cancer.volume,Lambda=seq(-2,+2,100))

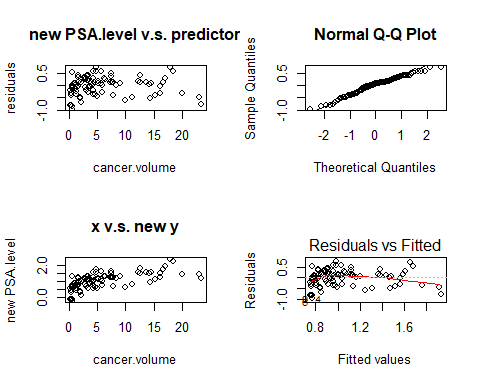
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :  
## extra argument 'Lambda' will be disregarded



y\_new\_32 <- log10(data\_32$PSA.level)  
new\_reg\_32 <- lm(y\_new\_32~data\_32$cancer.volume)  
summary(new\_reg\_32)

##   
## Call:  
## lm(formula = y\_new\_32 ~ data\_32$cancer.volume)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.95284 -0.29823 0.07392 0.26147 0.75181   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.737802 0.056161 13.137 < 2e-16 \*\*\*  
## data\_32$cancer.volume 0.051110 0.006878 7.431 5.67e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3748 on 91 degrees of freedom  
## Multiple R-squared: 0.3777, Adjusted R-squared: 0.3708   
## F-statistic: 55.22 on 1 and 91 DF, p-value: 5.672e-11

par(mfrow=c(2,2))  
plot(new\_reg\_32$residuals~data\_32$cancer.volume,main='new PSA.level v.s. predictor', xlab='cancer.volume', ylab='residuals')  
qqnorm(new\_reg\_32$residuals)  
qqline(new\_reg\_32$residuals)  
plot(y\_new\_32~data\_32$cancer.volume,main='x v.s. new y',xlab='cancer.volume',ylab='new PSA.level')  
plot(new\_reg\_32,1)



y <- new\_reg\_32$coefficients[1] + new\_reg\_32$coefficients[2] \* 20  
y\_true <- 10^y