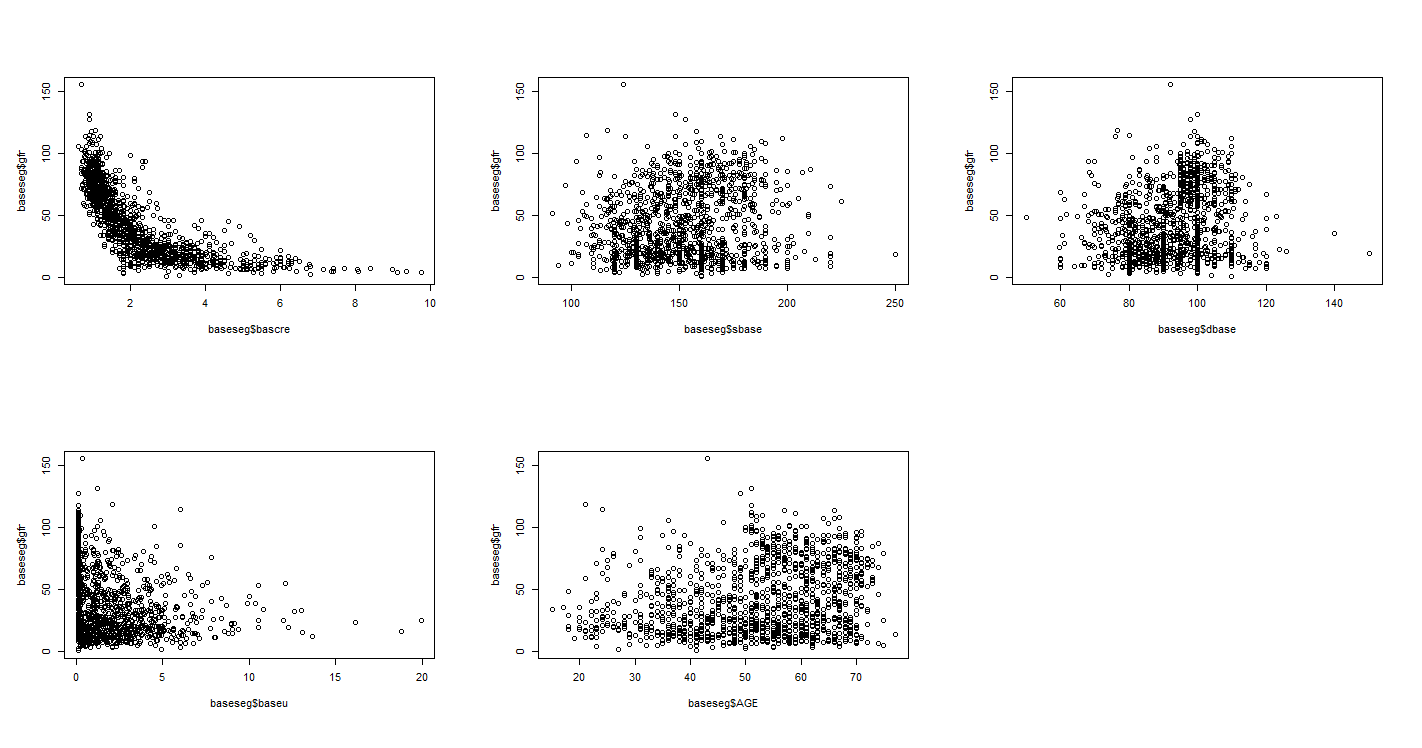
Homework2

Yue Peng

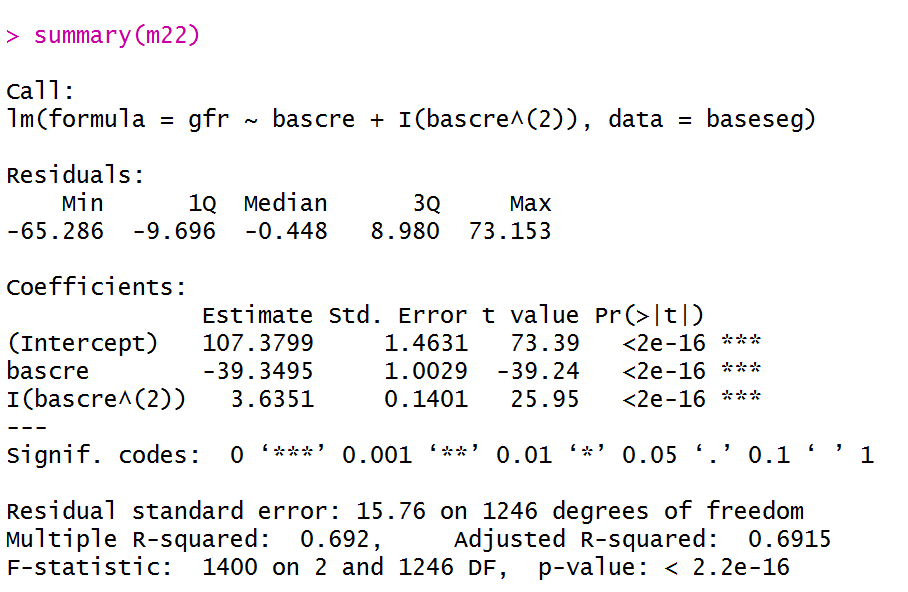
Part A:

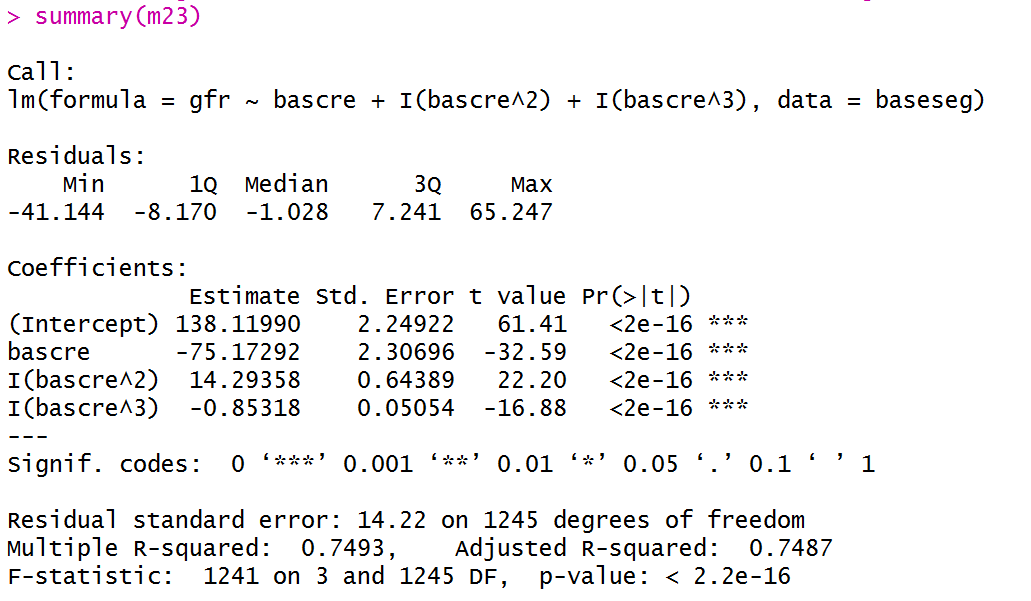
First, we eliminate the NAs in response variable gfr. Then we start to do linear regressions on those variables: "bascre", "sbase", "dbase", "baseu", "AGE", "SEX", "black". By obtaining the p-value of each slope, we extract the varibles whose p-value is less than 0.05. Finally we get five significant preditors: "bascre", "sbase", "dbase", "baseu", "AGE",

We plot the scatterplot for each potential predictor.

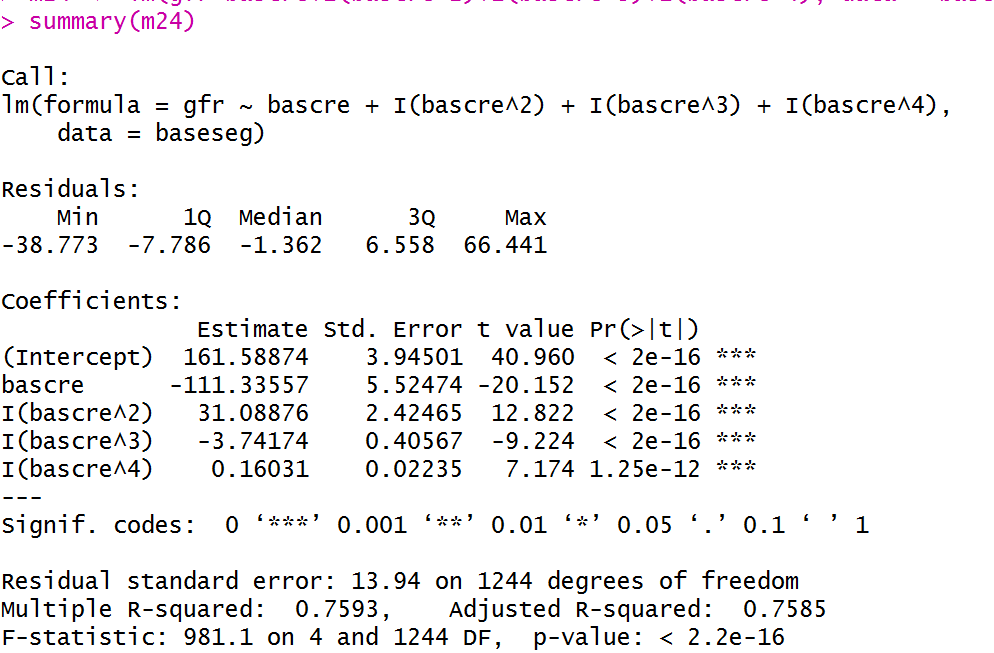


We found the “bascre” predictor had obvious nonlinearity. Thus, we add up the polynomial term step by step. The AIC and adjusted R squared tell us that it is better for us to keep the

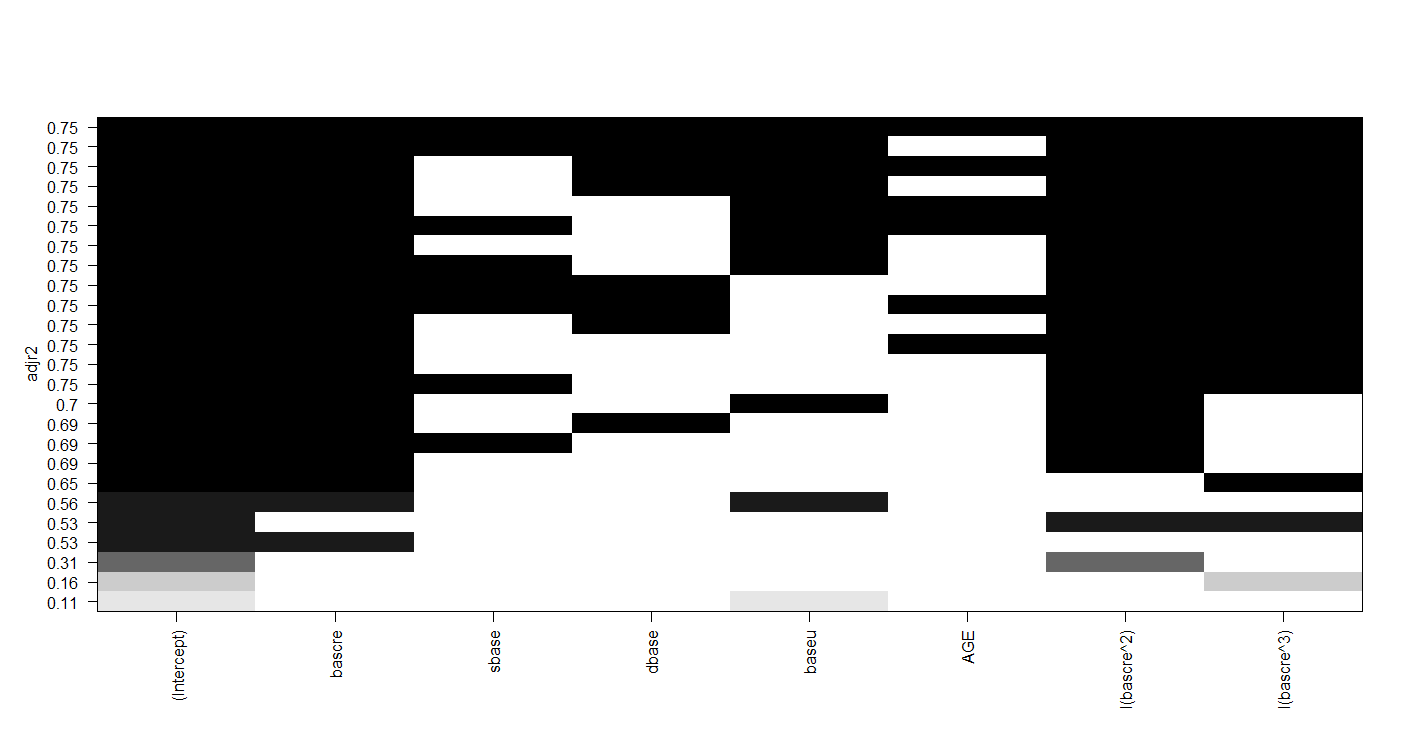




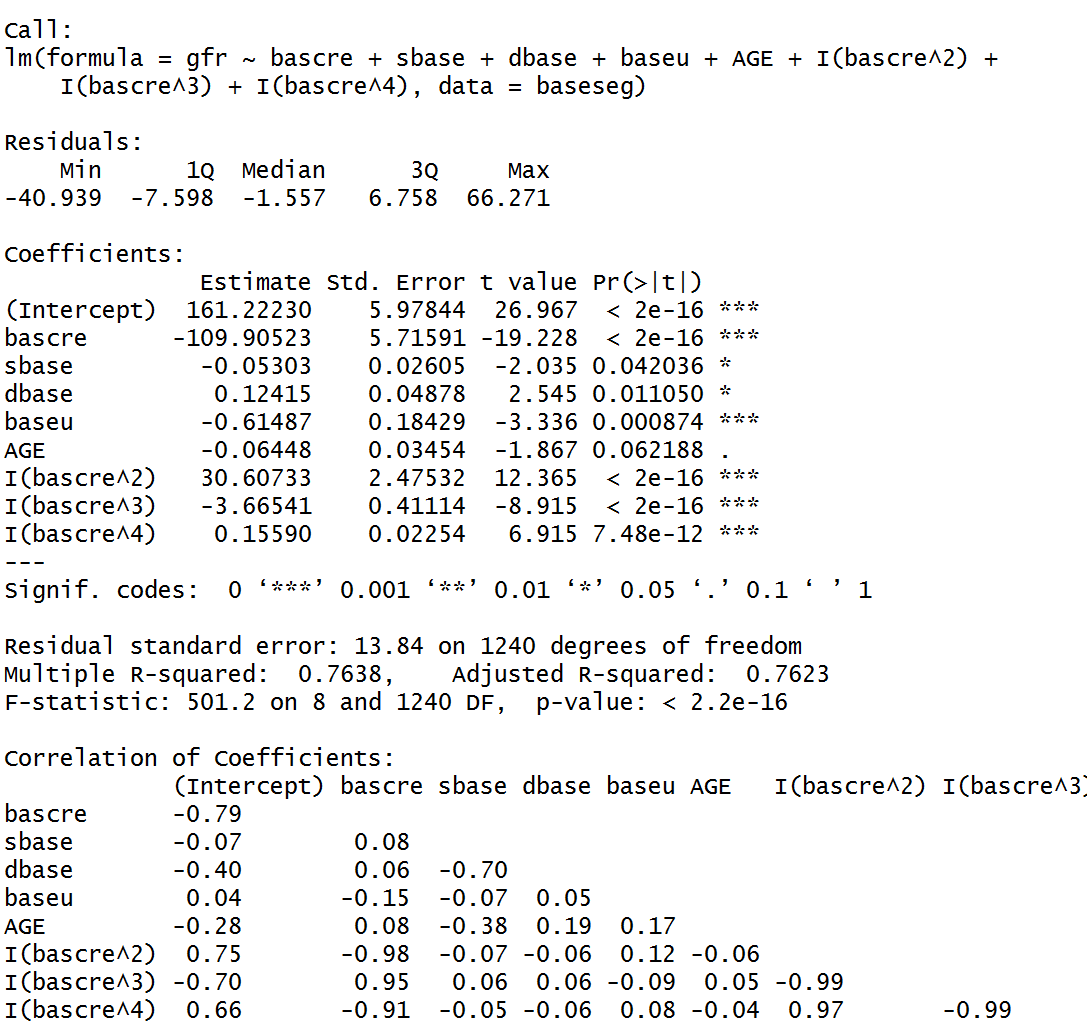
The AIC for these two models are 10438.02 and 10182.57

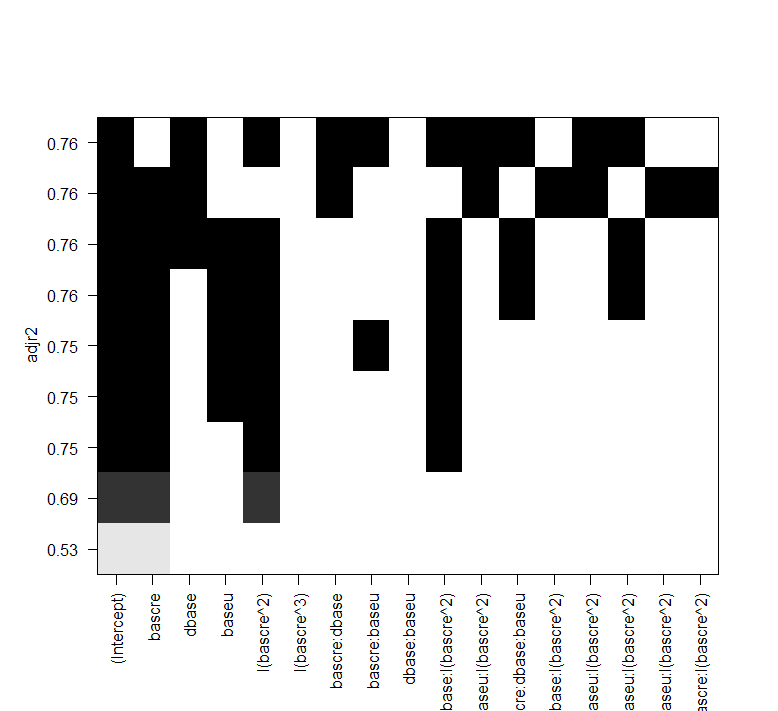


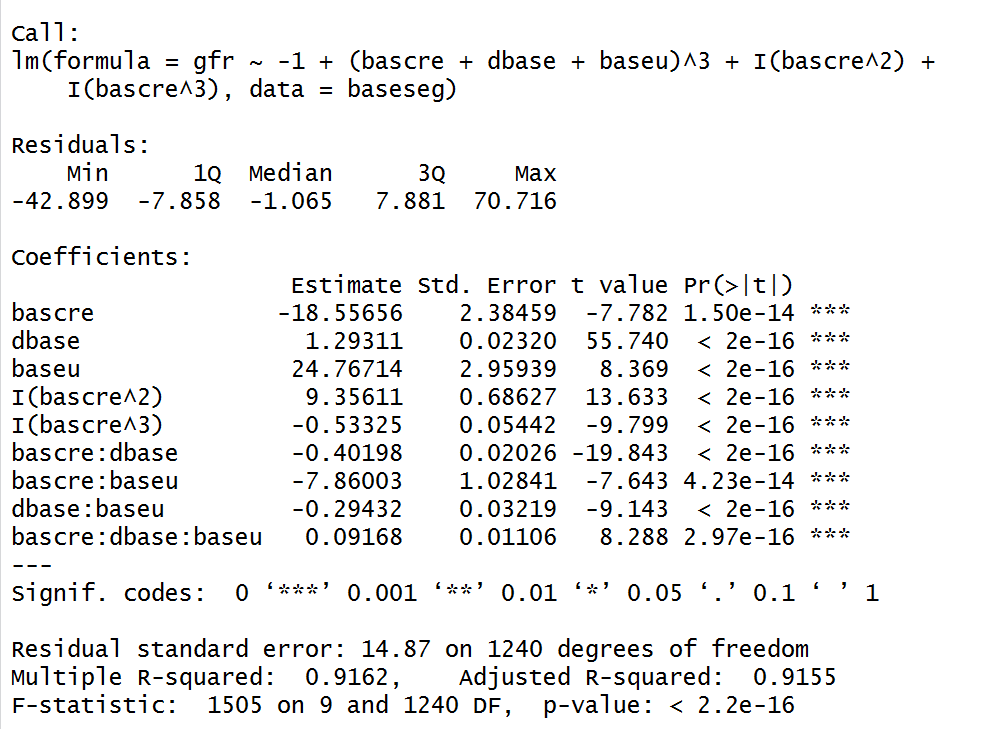
This model had an AIC=10133.93 and did not improve significantly on adjusted R squared. For our convenience, we just introduced third order term in our model. Also, we keep the “bascre” to respect the hierarchy.



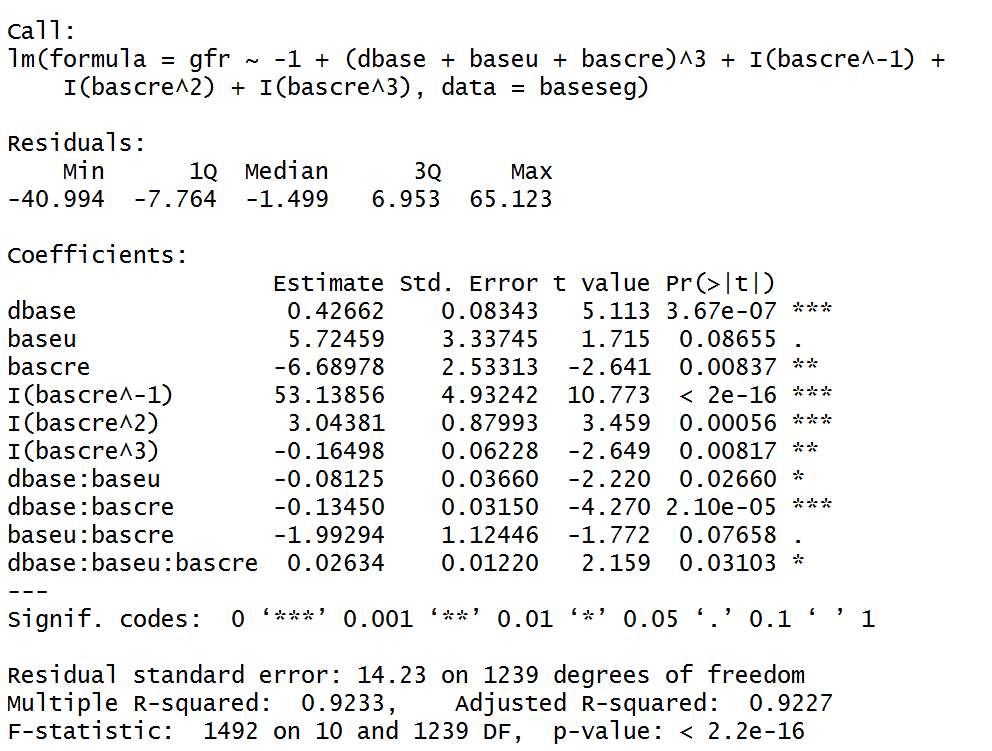
Then we implemented full subset regression with these predictors. The graph showed us all predictors could be taken into consideration.



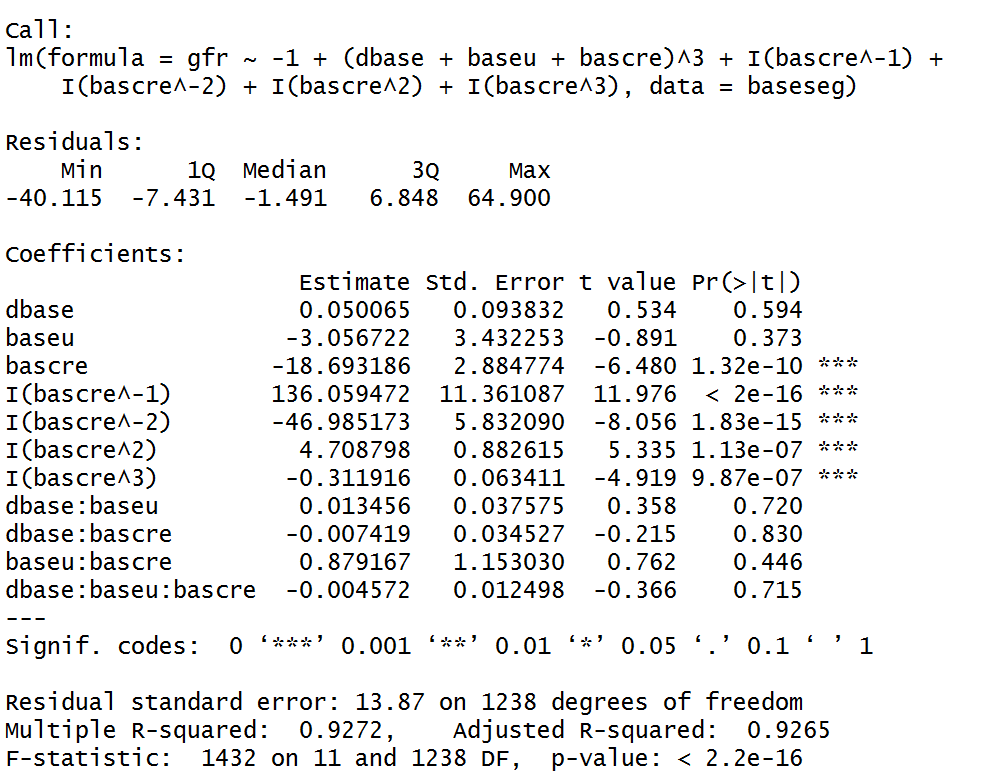
Also, we found out “sbase” and “dbase” has negative pairwise correlation and the coefficient and p-value of “sbase” are smaller than those of “dbase”. And “AGE” was not quite significant. We decided to remove “sbase” and “AGE” in our future analysis.



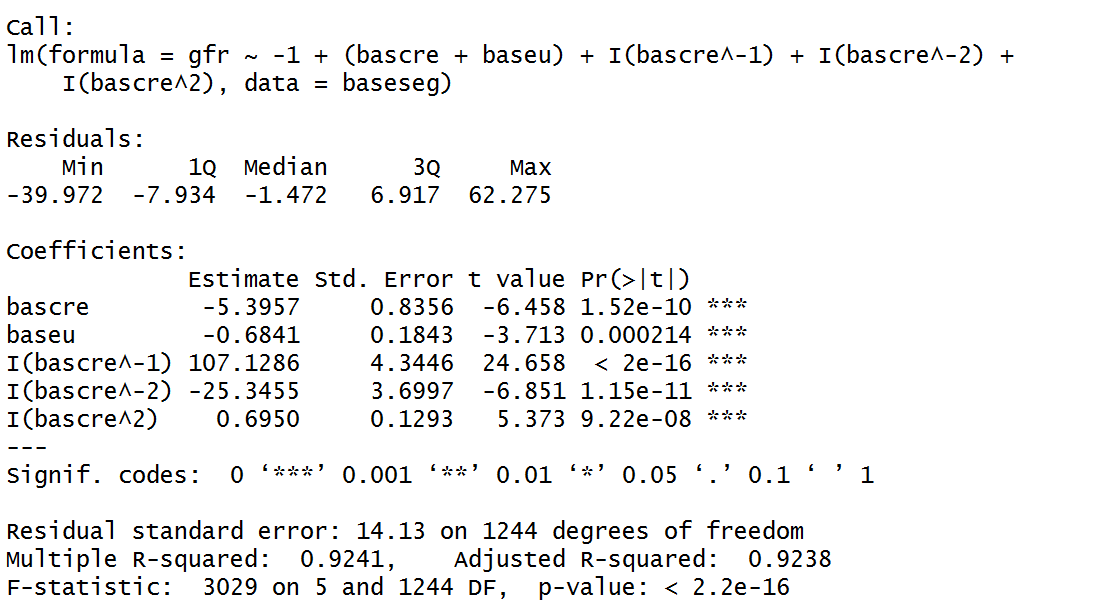
The scatter plot of “bascre” looks like the nike function. Thus, we introduce the reciprocal term.



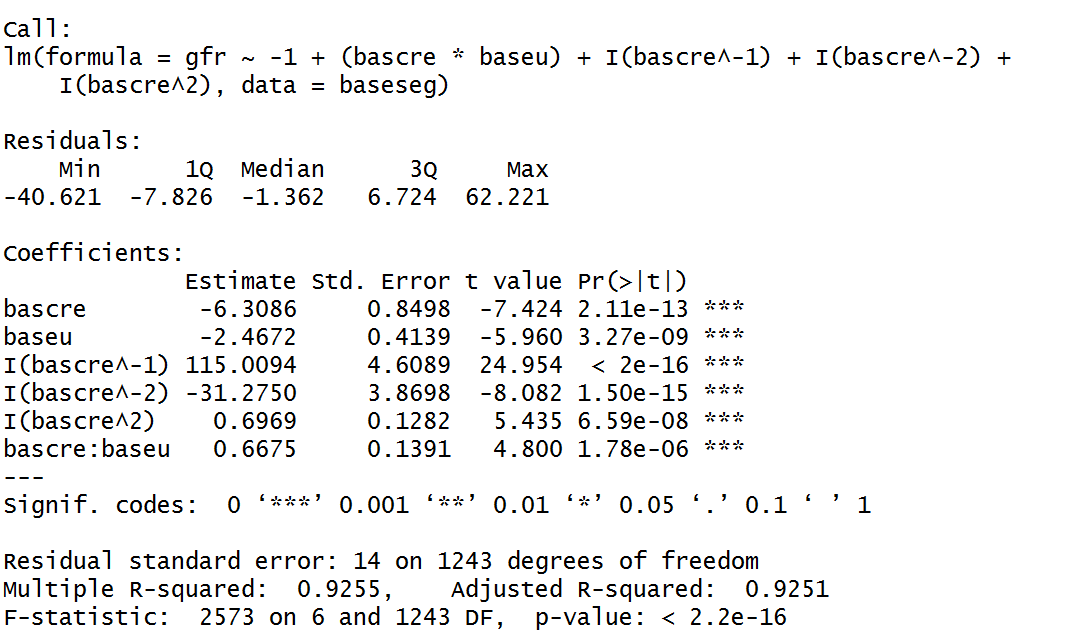
The adjusted R squared and AIC=10188.87 looks great. Then we add one more to see what will be going on in the model.



We find out all the interaction terms are not significant now. But the AIC of this model is 10127.05. And the coefficients of “bascre^-1”, “bascre”, “baseu”, “bascre^2” and “”bascre^-2” are both quite large.

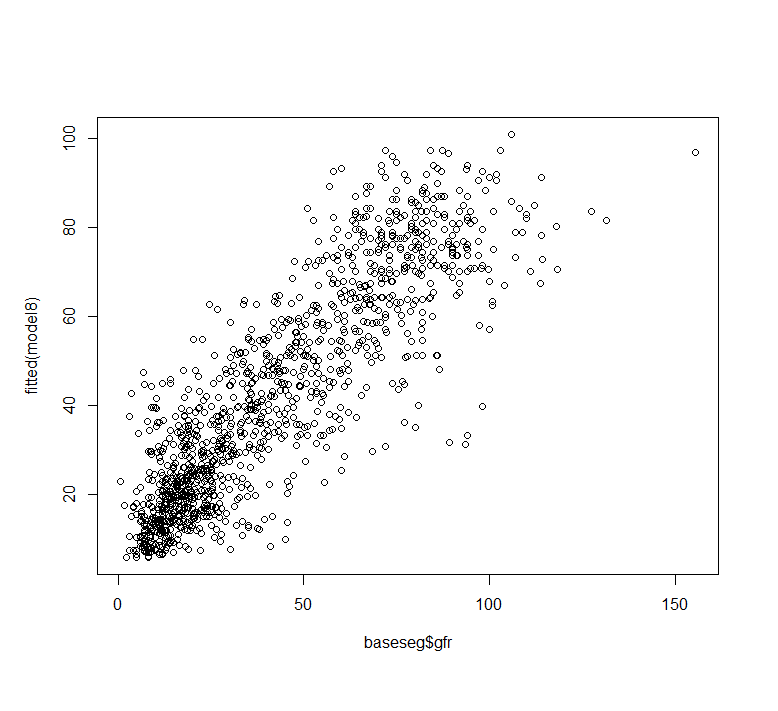


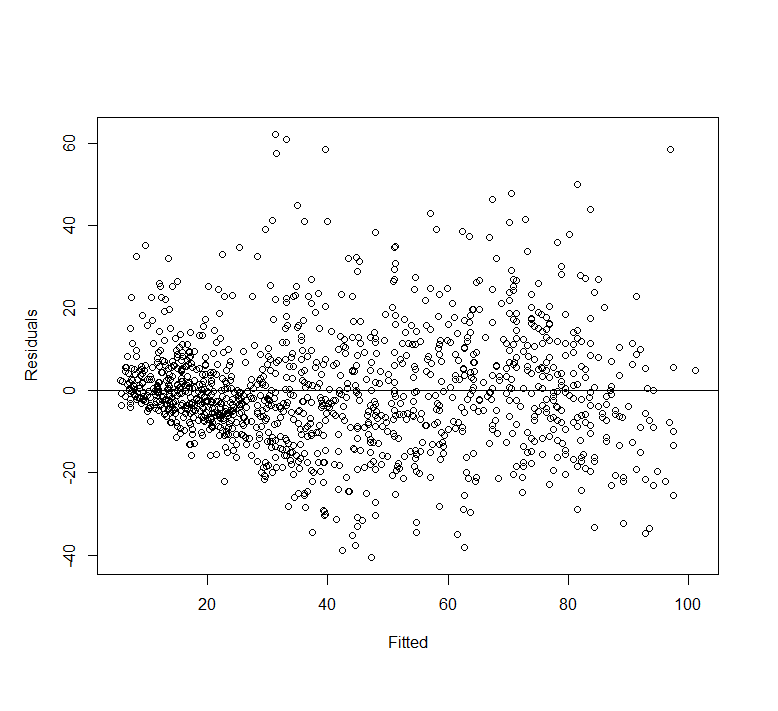
This model had an AIC with 10166.36

What’s more, we decided to add an interaction term. 

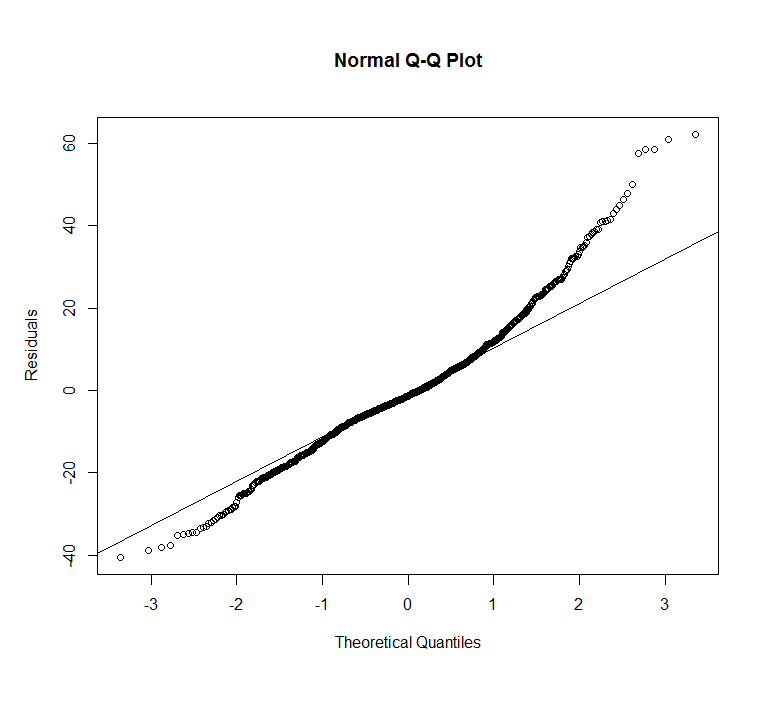
This model had an AIC with 10145.28.

In the end, our model is

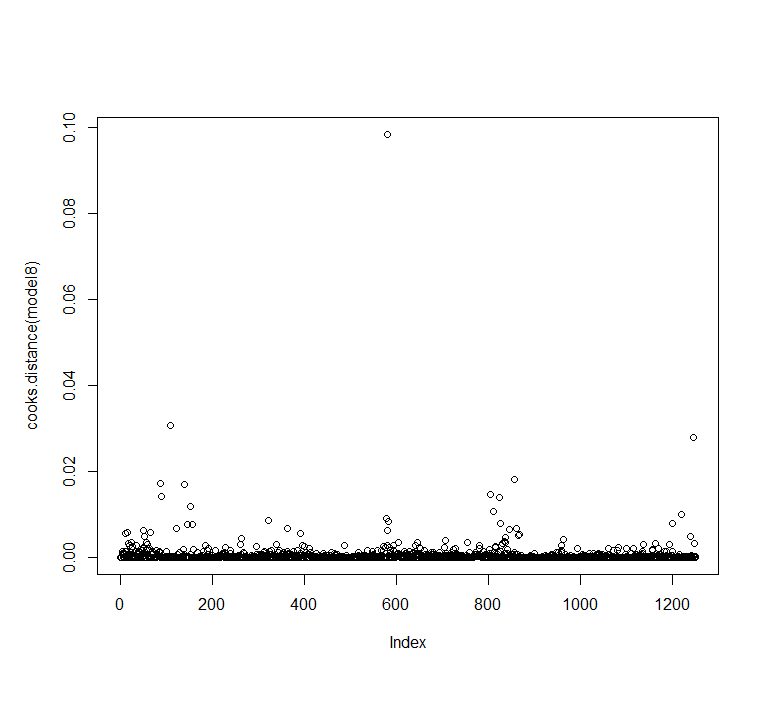




The first graph shows that the predicted values of gfr is following the line y=x, which means the model fits the data quite well. The second one shows that points lie symmetrically above and below the 0 horizontal line. And the variance is not constant.



The fitted values approximately follows the normal distribution since the curve mostly follow the line within (-2, 2).



This graph shows that there is no influential point.

Part B

1.

c. The p-value should be above 0.05

e. The proportion of times the p-value is less than 0.05 is 0.05 which matchs with my intuition. The slope could not be significant because they are from different distribution. Type I error is simulated in my simulation.

f. The proportion of times the p-value is less than 0.05 is 1. y is simulated given x so that the slope should be smaller than 0.05. 1 - Type II error is simulated here.

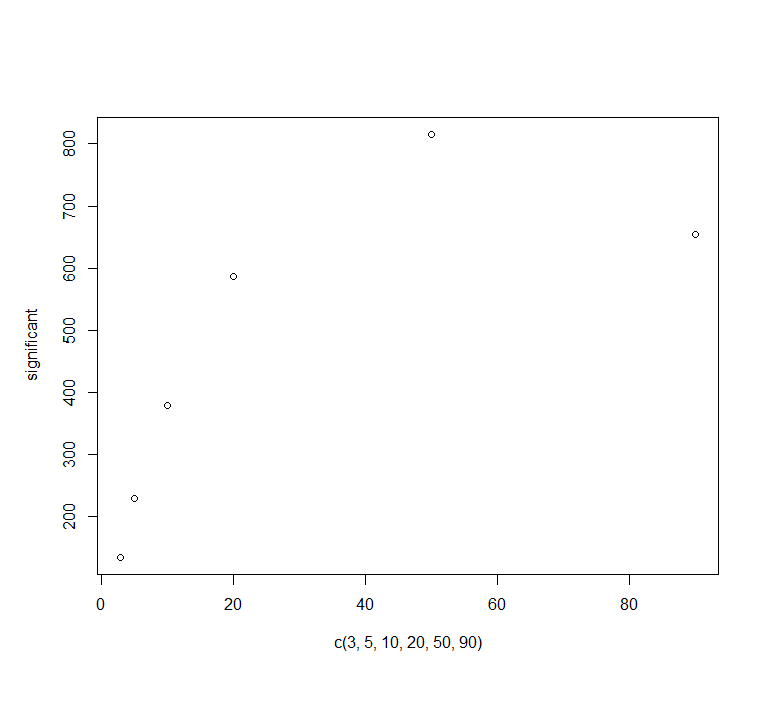
2.

c. There are 51, 51, 55 p-values for each variable are significant at the 0.05 level. Each variable has approximately 0.05 probability to be significant.

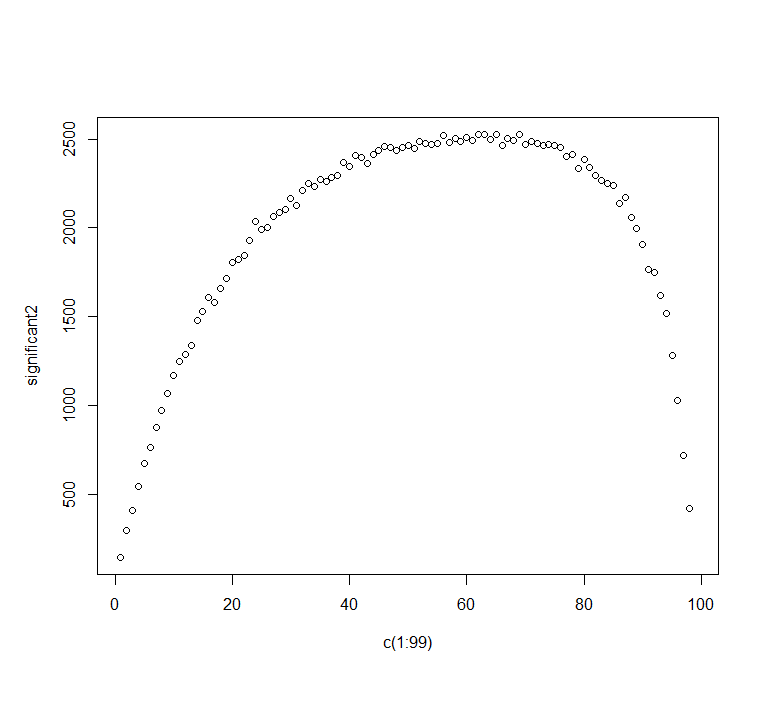
d. There are 146 minimum p-values for each regression are significant at the 0.05 level. Strictly speaking, these three variables are uncorrelated based on the assumption. Each variable have 0.05 probability to be significant according to question c. The problem now is there is at least one p-value for each regression will be less than 0.05, . This is what we simulate with R.

e. We find the number of times finding at least one significant variable for each P will go down at about 50 predictors. Before this, the curve perform well with

There are few points to show us the exact trend of this kind of simulation. And the curve should be approximately closed to 1.



f. It shows obvious pattern in the graph below with 3000 simulation and predictors from 1 to 99. The trend sudden decreases into 0. It may be due to the overfit or some problem in the simulation design.



Appendix

# Homework 2

setwd("D:/Google Drive/Fall2017/PHP2550/HW2")

baseseg <- read.csv("baseseg.csv")

potential\_predictors <- c("bascre", "sbase", "dbase", "baseu", "AGE", "SEX", "black")

#eliminate NAs in gfr

baseseg <- baseseg[-which(is.na(baseseg$gfr)),]

# extract the significant predictors which have p-values < 0.05

potentials <- c()

for (i in potential\_predictors){

if (summary(lm(paste0("baseseg$gfr~baseseg$",i)))[[4]][8] < 0.05){

potentials <- c(potentials, i)

}

}

potentials

par(mfrow=c(2,3))

plot(baseseg$bascre, baseseg$gfr)

plot(baseseg$sbase, baseseg$gfr)

plot(baseseg$dbase, baseseg$gfr)

plot(baseseg$baseu, baseseg$gfr)

plot(baseseg$AGE, baseseg$gfr)

# fit all the potentials

model1 <- lm(gfr~bascre+sbase+dbase+baseu+AGE+I(bascre^2)+I(bascre^3)+I(bascre^4),data = baseseg)

summary(model1, cor=T) # sbase and dbase has negative pairwise correlation

library(leaps)

leaps <- regsubsets(gfr~bascre+sbase+dbase+baseu+AGE+I(bascre^2)+I(bascre^3)+I(bascre^4),data = baseseg,nbest = 4)

plot(leaps,scale = "adjr2")

# drop sbase

leaps2 <- regsubsets(gfr~bascre\*sbase\*dbase\*baseu\*AGE\*I(bascre^2)+I(bascre^3),data = baseseg,nbest = 1)

plot(leaps2,scale = "adjr2")

leaps2$xnames[c(1,2,3,4,6,7,9,12)]

leaps3 <- regsubsets(gfr~bascre\*dbase\*baseu\*I(bascre^2)+I(bascre^3),data = baseseg,nbest = 1)

plot(leaps3,scale = "adjr2")

# drop AGE

model22 <- lm(gfr~bascre\*dbase\*baseu-1,data = baseseg)

summary(model22)

plot(fitted(model22), residuals(model22),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

model2 <- lm(gfr~-1+(bascre+dbase+baseu)^3+I(bascre^2)+I(bascre^3),data = baseseg)

summary(model2) # The coefficients themselves do not change residual SE does not change

# constant variance

plot(fitted(model2), residuals(model2),xlab = "Fitted", ylab = "Residuals")

abline(h=0) # Nonlinear which indicates some change in the model is necessary

full\_model <- lm(gfr~bascre\*dbase\*baseu\*I(bascre^2)\*I(bascre^3)\*I(bascre^4),data = baseseg)

reduced\_model <- step(full\_model,direction = "backward")

m1 <- lm(gfr~bascre,data = baseseg) # nonlinearity

plot(fitted(m1), residuals(m1),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

m2 <- lm(gfr~dbase,data = baseseg)

plot(fitted(m2), residuals(m2),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

m3 <- lm(gfr~baseu,data = baseseg)

plot(fitted(m3), residuals(m3),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

# Look like Nike function, add reciprocal term (respect the hierarchy)

m22 <- lm(gfr~bascre+I(bascre^(2)), data = baseseg)

plot(fitted(m22), residuals(m22),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

summary(m22)

m23 <- lm(gfr~bascre+I(bascre^2)+I(bascre^3), data = baseseg)

summary(m23)

plot(fitted(m23), residuals(m23),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

AIC(m22)

AIC(m23)

AIC(m24)

m24 <- lm(gfr~bascre+I(bascre^2)+I(bascre^3)+I(bascre^4), data = baseseg)

summary(m24)

plot(fitted(m24), residuals(m24),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

m25 <- lm(gfr~bascre+I(bascre^2)+I(bascre^3)+I(bascre^4)+I(bascre^5), data = baseseg)

summary(m25)

#add polynomial term of order 4

model3 <- lm(gfr~-1+(dbase+baseu+bascre)^3+I(bascre^-1)+I(bascre^2)+I(bascre^3),data = baseseg)

summary(model3)

plot(fitted(model3), residuals(model3),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

model4 <- lm(gfr~-1+(dbase+baseu+bascre)^3+I(bascre^-1)+I(bascre^-2)+I(bascre^2)+I(bascre^3),data = baseseg)

summary(model4)

plot(fitted(model4), residuals(model4),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

model5 <- lm(gfr~-1+(dbase+baseu+bascre)+I(bascre^-1)+I(bascre^-2)+I(bascre^2)+I(bascre^3),data = baseseg)

summary(model5)

plot(fitted(model5), residuals(model5),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

model6 <- lm(gfr~-1+(bascre+baseu)+I(bascre^-1)+I(bascre^-2)+I(bascre^2),data = baseseg)

summary(model6)

plot(fitted(model6), residuals(model6),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

plot(model6)

model8 <- lm(gfr~-1+(bascre\*baseu)+I(bascre^-1)+I(bascre^-2)+I(bascre^2),data = baseseg)

summary(model8)

plot(fitted(model8), residuals(model8),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

######

model7 <- lm(gfr~bascre+dbase+I(bascre^2)+I(bascre^3)+bascre:dbase+dbase:I(bascre^2)+bascre:I(bascre^3)+

dbase:I(bascre^3)+I(bascre^2):I(bascre^3)+bascre:dbase:I(bascre^3)+bascre:I(bascre^2):I(bascre^3)+

dbase:I(bascre^2):I(bascre^3)+bascre:dbase:I(bascre^2):I(bascre^3)+I(bascre^-1),data = baseseg)

summary(model7)

plot(fitted(model7), residuals(model7),xlab = "Fitted", ylab = "Residuals")

abline(h=0)

# Normality

qqnorm(residuals(model5), ylab = "Residuals")

qqline(residuals(model5))

qqnorm(residuals(model8), ylab = "Residuals")

qqline(residuals(model8))

shapiro.test(residuals(model8))

#outlier

library(car)

outlierTest(model8)

plot(cooks.distance(model8))

# PART B

# 1

y <- rnorm(100, mean = 10, sd=2)

x <- rnorm(100, mean = 3, sd=1)

model\_B <- lm(y~x)

summary(model\_B)[[4]][2,4]

p\_values <- c()

for (i in c(1:1000)){

y <- rnorm(100, mean = 10, sd=2)

x <- rnorm(100, mean = 3, sd=1)

model\_B1 <- lm(y~x)

p\_values <- c(p\_values,summary(model\_B1)[[4]][2,4])

}

sum(p\_values<0.05)/1000

#

p\_values2 <- c()

for (i in 1:1000){

y <- rep(0, 100)

x <- rnorm(100,mean = 3,sd=1)

for (j in 1:100){

y[j] <- rnorm(1, x[j]+10,1)

}

model\_B2 <- lm(y~x)

p\_values2 <- c(p\_values2, summary(model\_B2)[[4]][2,4])

}

sum(p\_values2<0.05)/1000

#2

library(MASS)

y <- rnorm(100,10,4)

X <- mvrnorm(n=100,c(1,2,3),diag(3))

set.seed(4)

p\_values3 <- c()

for (i in 1:1000){

y <- rnorm(100,10,4)

X <- mvrnorm(n=100,c(1,2,3),diag(3))

model\_B3 <- lm(y~X)

p\_values3 <- rbind(p\_values3, summary(model\_B3)$coefficients[-1,4])

}

sum(p\_values3[,1] <0.05)

sum(p\_values3[,2] <0.05)

sum(p\_values3[,3] <0.05)

sum(apply(p\_values3,1,min)<0.05)

significant <- c()

for (n in c(3,5,10,20,50,90)){

p\_values4 <- c()

for (i in 1:1000){

y <- rnorm(100,10,4)

X <- mvrnorm(n=100,c(1:n),diag(n))

model\_B4 <- lm(y~X)

p\_values4 <- rbind(p\_values4, summary(model\_B4)$coefficients[-1,4])

}

significant <- c(significant, sum(apply(p\_values4,1,min)<0.05))

}

plot(x=c(3,5,10,20,50,90), y=significant)

significant2 <- c()

for (n in c(1:200)){

p\_values5 <- c()

for (i in 1:2000){

y <- rnorm(100,10,4)

X <- mvrnorm(n=100,c(1:n),diag(n))

model\_B5 <- lm(y~X)

p\_values5 <- rbind(p\_values5, summary(model\_B5)$coefficients[-1,4])

}

significant2 <- c(significant2, sum(apply(p\_values5,1,min)<0.05))

}

plot(x=c(1:99), y=significant2)