##### Question 1

bea.2006=read.csv("bea-2006.csv")
library(ggplot2)
is.data.frame(bea.2006)
is.matrix(bea.2006)
dim(bea.2006)
head(bea.2006)

##### Question 2

str(bea.2006)
summary(bea.2006[1:7])
databea=na.omit(bea.2006[,1:7])
dim(databea)
View(databea)

##### Question 3

library(ggplot2)
library(GGally)
ggpairs(databea[,2:7])

##### Question 4

model = lm(pcgmp~pop, data= databea)
summary(model)
plot(model)
databea$pop
names(databea)
attach(databea)
Z=cbind(rep(1,133),pop)
beta=solve(t(Z)%\*%Z)%\*%t(Z)%\*%pcgmp
beta

##### Question 5

par(mfrow=c(2,2))
plot(model)

par(mfrow = c(1,2))
with(databea, plot(pop, pcgmp, xlab = "Population", ylab = "Per Capita GMP",
 cex = 0.5, col="blue"))
abline(model, lwd = 2, col = "red")
with(databea, plot(pop, pcgmp, xlab = "Population", ylab = "Per Capita GMP",
 cex = 0.5, log = "x", col="blue", main="log-scale"))
abline(model2, lwd = 2, col = "red")

model2 = lm(pcgmp~I(log(pop)), data= databea)
summary(model2)
plot(log(pop),pcgmp)
ypred=as.numeric(model2$fitted)
lines(log(pop),ypred,col='red')
par(mfrow=c(2,2))
plot(model2)

MSE = mean(model$residuals^2)

predict(model, newdata = data.frame(pop = databea$pop[262] + 1e5))

model3 = lm(pcgmp~., data= databea[,2:7])
summary(model3)
par(mfrow=c(2,2))
plot(model3)
View(databea)

#######################
##### Simulations #####

n=200
i=1:n
sigma=5
Z1=i;Z2=i^2;Z3=i^3
epsilon=sigma\*rnorm(n)
Y=5-0.03\*Z1+0.002\*Z2+epsilon;
par(mfrow=c(1,1))
plot(i,Y,col='blue')

Yreg=lm(Y~Z1+Z2)
summary(Yreg)
lines(Yreg$fitted.values)
par(mfrow=c(2,2))
plot(lm(Y~Z1+Z2))
names(Yreg)
Yreg$coeff
summary(Yreg)$fstat
summary(Yreg)$sigma
vcov(lm(Y~Z1+Z2))

n=100
m=5000
theta0=rep(0,m)
theta1=rep(0,m)
theta2=rep(0,m)
sigmaest=rep(0,m)
ficherstat=rep(0,m)
for (j in c(1:m))
{
 epsilon=runif(n,-30,30) # rnorm(n,0,5)# rt(n,6) # # # =rt(n,1) ##rt(n,2)#
 i=1:n
 Z1=i;Z2=i^2;Z3=i^3
 Y=5-0.03\*Z1+0.002\*Z2+epsilon;
 Y.lm=summary(lm(Y~Z1+Z2))
 theta0[j]=Y.lm$coeff[1]
 theta1[j]=Y.lm$coeff[2]
 theta2[j]=Y.lm$coeff[3]
 sigmaest[j]=Y.lm$sigma
 ficherstat[j]=Y.lm$fstat
}
hist(theta0,nclass=20)
plot(density(theta0))
mean(theta0)
hist(theta1,nclass=20)
hist(theta2,nclass=20)
hist(sigmaest,nclass=20)
hist(ficherstat,nclass=20)

shapiro.test(theta0)
shapiro.test(theta1)
shapiro.test(theta2)
shapiro.test(sigmaest)

n=100
for (j in c(1:m))
{
 epsilon=rt(n,1) #runif(n,-30,30)
 i=1:n
 Z1=i;Z2=i^2;Z3=i^3
 Y=5+epsilon;
 Y.lm=summary(lm(Y~Z1+Z2))
 ficherstat[j]=Y.lm$fstat
}
hist(ficherstat)
ks.test(ficherstat,"pf",df1=2,df2=(n-3))
ks.test(ficherstat\*2,"pchisq",df=2)

m=1000
theta0=rep(0,m)
s0=c()
theta1=rep(0,m)
s1=c()
theta2=rep(0,m)
s2=c()
sigmaest=rep(0,m)
ficherstat=rep(0,m)
for (n in c(100,200,500, 1000,2000,5000,10000))
{i=c(1:n)
sigma=5
Z1=i;Z2=i^2;Z3=i^3
for (j in c(1:m))
{
 epsilon=runif(n,-30,30)
 Y=5-0.03\*Z1+0.002\*Z2+epsilon;
 Y.lm=summary(lm(Y~Z1+Z2))
 theta0[j]=Y.lm$coeff[1]
 theta1[j]=Y.lm$coeff[2]
 theta2[j]=Y.lm$coeff[3]
 sigmaest[j]=Y.lm$sigma
 ficherstat[j]=Y.lm$fstat
}
s0=c(s0,sd(theta0))# s0=c(s0,(mean((theta0-5)^2))^0.5)
s1=c(s1,sd(theta1))
s2=c(s2,sd(theta2))
}
nn=c(100,200,500, 1000,2000,5000,10000)
plot(log(nn),log(s0),ylim=c(-20,5),"l",col="red")
lines(log(nn),log(s1),"l",col="blue")
lines(log(nn),log(s2),"l",col="green")
ls0=log(s0); ls1=log(s1); ls2=log(s2)
lnn=log(nn)
(vite0=lm(ls0~lnn)$coef[2]) ### Vitesse en n^(-0.5) ####
(vite1=lm(ls1~lnn)$coef[2]) ### Vitesse en n^(-1.5) ####
(vite2=lm(ls2~lnn)$coef[2]) ### Vitesse en n^(-2.5) ####

########################################
########## Premier exercice ############

m=1000
theta0=rep(0,m)
s0=c()
theta1=rep(0,m)
s1=c()
theta2=rep(0,m)
s2=c()
theta3=rep(0,m)
s3=c()
sigmaest=rep(0,m)
ficherstat=rep(0,m)
for (n in c(100,200,500, 1000,2000,5000,10000))
{i=c(1:n)
sigma=5
Z1=i;Z2=i^2;Z3=i^3
for (j in c(1:m))
{
 epsilon=rnorm(n,0,sigma)
 Y=5-0.03\*Z1+0.002\*Z2+epsilon;
 Y.lm=summary(lm(Y~Z1+Z2+Z3))
 theta0[j]=Y.lm$coeff[1]
 theta1[j]=Y.lm$coeff[2]
 theta2[j]=Y.lm$coeff[3]
 theta3[j]=Y.lm$coeff[4]
 sigmaest[j]=Y.lm$sigma
 ficherstat[j]=Y.lm$fstat
}
s0=c(s0,sd(theta0))
s1=c(s1,sd(theta1))
s2=c(s2,sd(theta2))
s3=c(s3,sd(theta3))
}
nn=c(100,200,500, 1000,2000,5000,10000)
plot(log(nn),log(s0),ylim=c(-40,5),"l",col="red")
lines(log(nn),log(s1),"l",col="blue")
lines(log(nn),log(s2),"l",col="green")
lines(log(nn),log(s3),"l",col="cyan")
ls0=log(s0); ls1=log(s1); ls2=log(s2); ls3=log(s3)
lnn=log(nn)
(vite0=lm(ls0~lnn)$coef[2]) ### Vitesse en n^(-0.5) ####
(vite1=lm(ls1~lnn)$coef[2]) ### Vitesse en n^(-1.5) ####
(vite2=lm(ls2~lnn)$coef[2]) ### Vitesse en n^(-2.5) ####
(vite3=lm(ls3~lnn)$coef[2]) ### Vitesse en n^(-3.5) ####

m=1000
theta0=rep(0,m)
s0=c()
theta1=rep(0,m)
s1=c()
sigmaest=rep(0,m)
ficherstat=rep(0,m)
for (n in c(100,200,500, 1000,2000,5000,10000))
{i=c(1:n)
sigma=5
Z1=i;Z2=i^2;Z3=i^3
for (j in c(1:m))
{
 epsilon=rnorm(n,0,sigma)
 Y=5-0.03\*Z1+0.002\*Z2+epsilon;
 Y.lm=summary(lm(Y~Z1))
 theta0[j]=Y.lm$coeff[1]
 theta1[j]=Y.lm$coeff[2]
 sigmaest[j]=Y.lm$sigma
 ficherstat[j]=Y.lm$fstat
}
s0=c(s0,(mean((theta0-5)^2))^0.5)
s1=c(s1,(mean((theta1+0.03)^2))^0.5)
}
nn=c(100,200,500, 1000,2000,5000,10000)
plot(log(nn),log(s0),ylim=c(-5,15),"l",col="red")
lines(log(nn),log(s1),"l",col="blue")
ls0=log(s0); ls1=log(s1);
lnn=log(nn)
(vite0=lm(ls0~lnn)$coef[2]) ### Divergence ####
(vite1=lm(ls1~lnn)$coef[2]) ### Divergence ####

########################################
########## Second exercice ############

m=1000
theta0=rep(0,m)
s0=c()
theta1=rep(0,m)
s1=c()
theta2=rep(0,m)
s2=c()
sigmaest=rep(0,m)
ficherstat=rep(0,m)
for (n in c(100,200,500, 1000,2000,5000,10000))
{i=c(1:n)
sigma=5
Z1=i;Z2=1/i;Z3=i^3
for (j in c(1:m))
{
 epsilon=rnorm(n,0,sigma)
 Y=5-0.03\*Z1+0.002\*Z2+epsilon;
 Y.lm=summary(lm(Y~Z1+Z2))
 theta0[j]=Y.lm$coeff[1]
 theta1[j]=Y.lm$coeff[2]
 theta2[j]=Y.lm$coeff[3]
 sigmaest[j]=Y.lm$sigma
 ficherstat[j]=Y.lm$fstat
}
s0=c(s0,(mean((theta0-5)^2))^0.5)
s1=c(s1,(mean((theta1+0.03)^2))^0.5)
s2=c(s2,(mean((theta2-0.002)^2))^0.5)
}
nn=c(100,200,500, 1000,2000,5000,10000)
plot(log(nn),log(s0),ylim=c(-20,5),"l",col="red")
lines(log(nn),log(s1),"l",col="blue")
lines(log(nn),log(s2),"l",col="green")
ls0=log(s0); ls1=log(s1); ls2=log(s2);
lnn=log(nn)
(vite0=lm(ls0~lnn)$coef[2]) ### Vitesse en n^(-0.5) ####
(vite1=lm(ls1~lnn)$coef[2]) ### Vitesse en n^(-1.5) ####
(vite2=lm(ls2~lnn)$coef[2]) ### Non convergence ####

##########################################
########## Troisi?me exercice ############

n=1000
Z1=rexp(n,1)
Z2=rbinom(n,10,1/3)
epsilon=rnorm(n,0,2)
Y=5+2\*Z1-3\*Z2+epsilon
reg=lm(Y~Z1+Z2)
summary(reg)
reg$coef
plot(reg)

m=1000
theta0=rep(0,m)
s0=c()
theta1=rep(0,m)
s1=c()
theta2=rep(0,m)
s2=c()
sigmaest=rep(0,m)
ficherstat=rep(0,m)
for (n in c(100,200,500, 1000,2000,5000,10000))
{i=c(1:n)
sigma=2
Z1=rexp(n,1)
Z2=rbinom(n,10,1/3)
for (j in c(1:m))
{
 epsilon=rnorm(n,0,sigma)
 Y=5-0.03\*Z1+0.002\*Z2+epsilon;
 Y.lm=summary(lm(Y~Z1+Z2))
 theta0[j]=Y.lm$coeff[1]
 theta1[j]=Y.lm$coeff[2]
 theta2[j]=Y.lm$coeff[3]
 sigmaest[j]=Y.lm$sigma
 ficherstat[j]=Y.lm$fstat
}
s0=c(s0,(mean((theta0-5)^2))^0.5)
s1=c(s1,(mean((theta1+0.03)^2))^0.5)
s2=c(s2,(mean((theta2-0.002)^2))^0.5)
}
nn=c(100,200,500, 1000,2000,5000,10000)
plot(log(nn),log(s0),ylim=c(-20,5),"l",col="red")
lines(log(nn),log(s1),"l",col="blue")
lines(log(nn),log(s2),"l",col="green")
ls0=log(s0); ls1=log(s1); ls2=log(s2);
lnn=log(nn)
(vite0=lm(ls0~lnn)$coef[2]) ### Vitesse en n^(-0.5) ####
(vite1=lm(ls1~lnn)$coef[2]) ### Vitesse en n^(-1.5) ####
(vite2=lm(ls2~lnn)$coef[2]) ### Non convergence ####