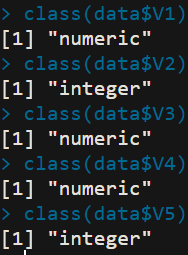
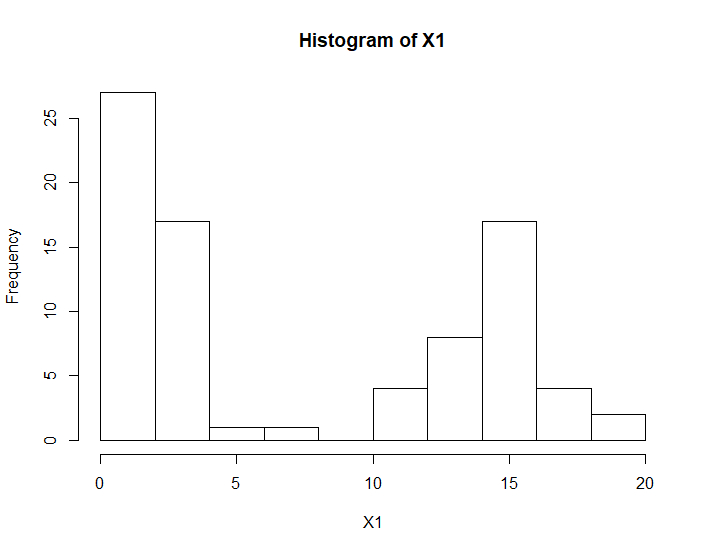
##### 6、

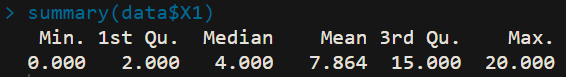
1. A

Its are all numeric or integer.

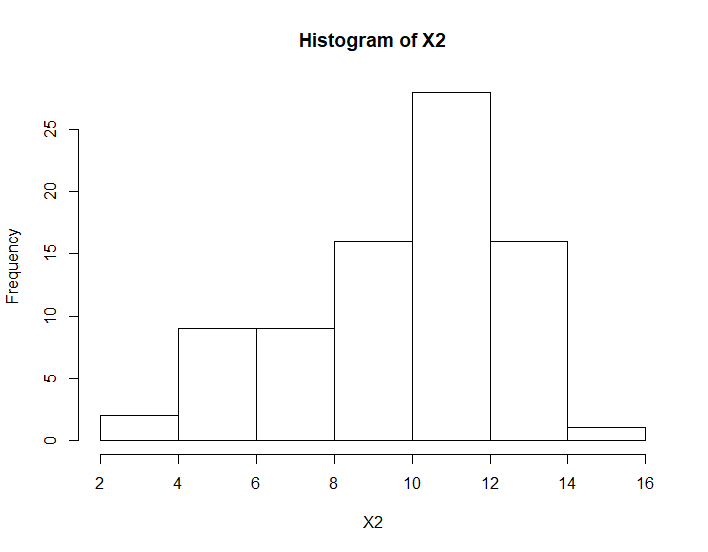


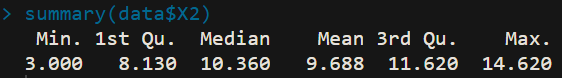
X1 variable:



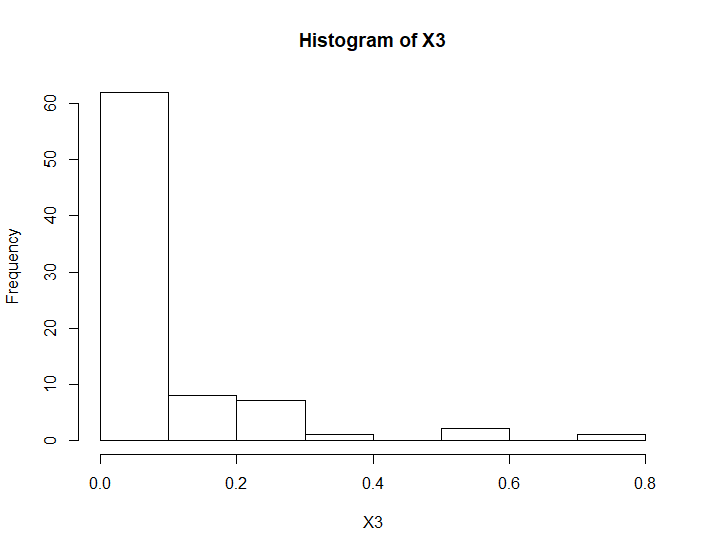


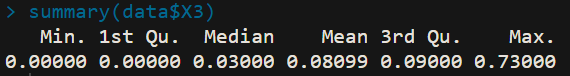
X2 variable:



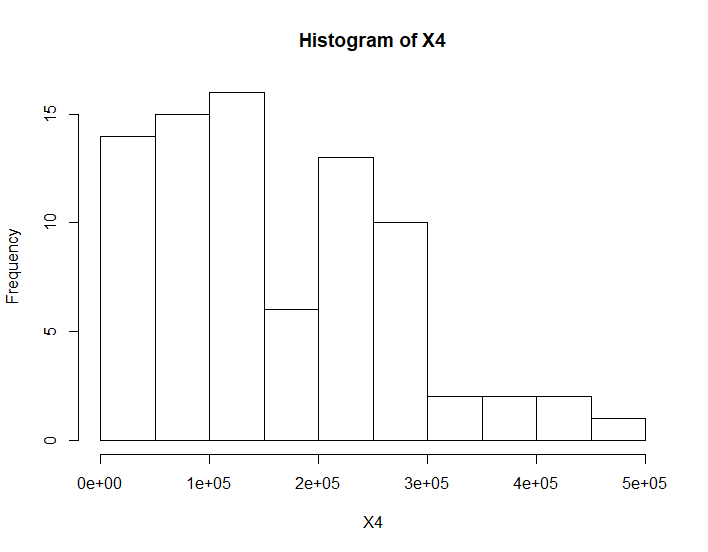


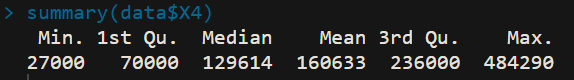
X3 variable:



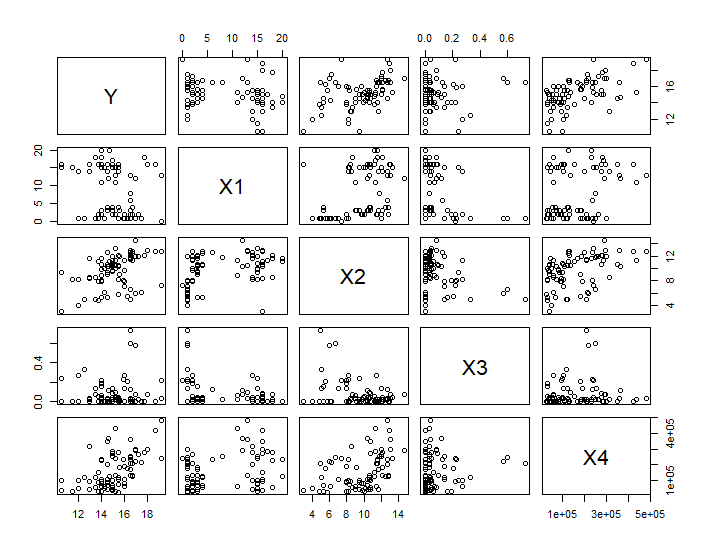


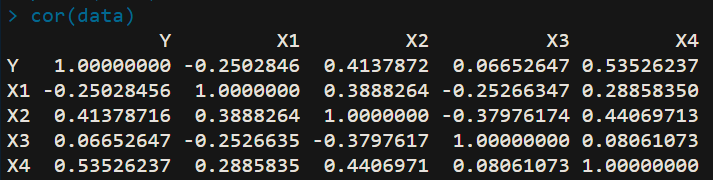
X4 variable:





1. B





I observed if the correlation coefficient is bigger, the scatter plot is more similar to a football.

1. C

Beta1 = -0.142

Beta2 = 0.282

Beta3 = 0.6193

Beta4 = 7.924\*10^-6

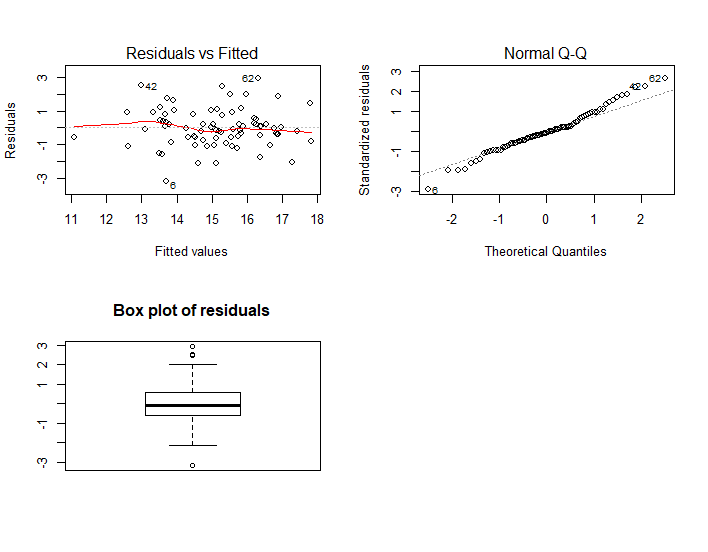


MSE = 1.293

R^2 = 0.5847

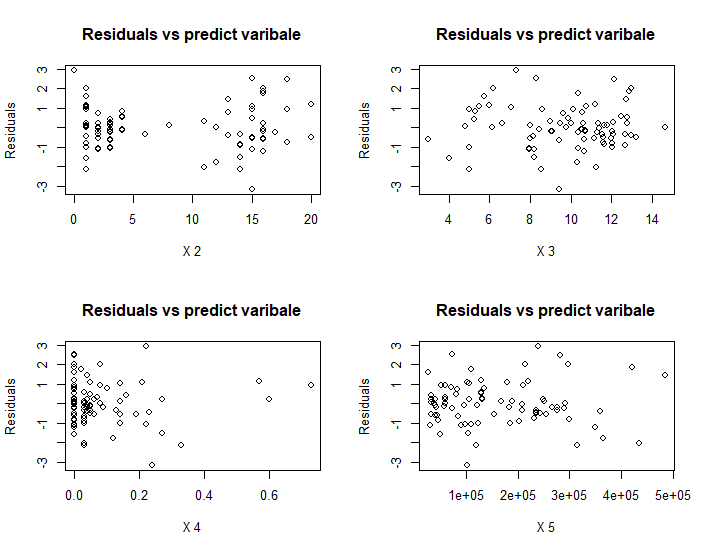
Ra^2 = 0.5629

1. D

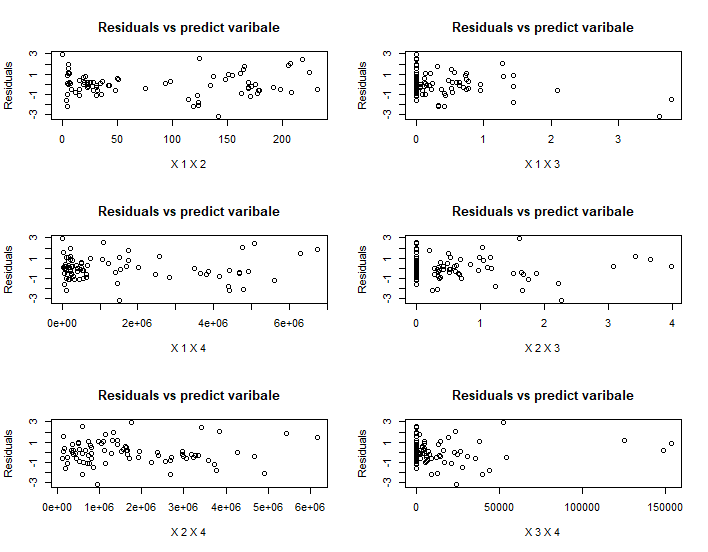


This model comply linearity, variances are the same. But it does not comply the normal distribution assumption.

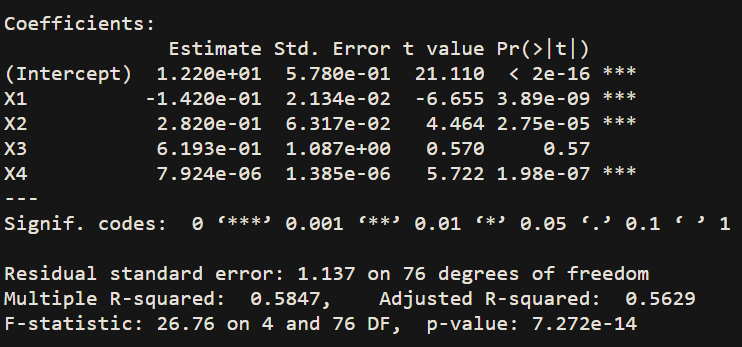
1. E



There are 6 two-way interaction.



1. F



H0: betak = 0 ; H1 = betak is not zero

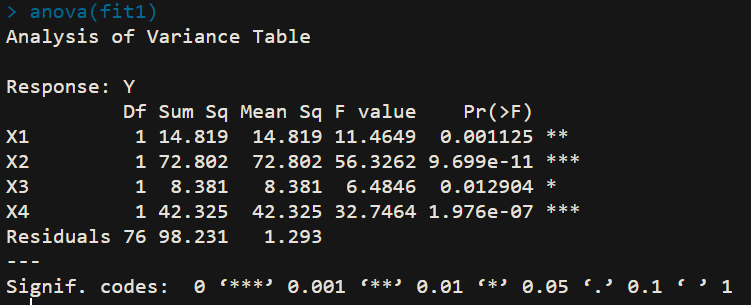


Null distribution is the T distribution

We can get from the R output that beta1 is significant under level of 0.01, beta2 is significant under level of 0.01, beta4 is significant under level of 0.01, beta3 is not significant under level of 0.01

It means that x3 has multicollineariy between x1,x2,x4.

1. G



SSTO = 236.558 d.f.ssto = 80

SSE = 98.231 d.f.sse = 76

SSR = 138.327 d.f.ssr = 4

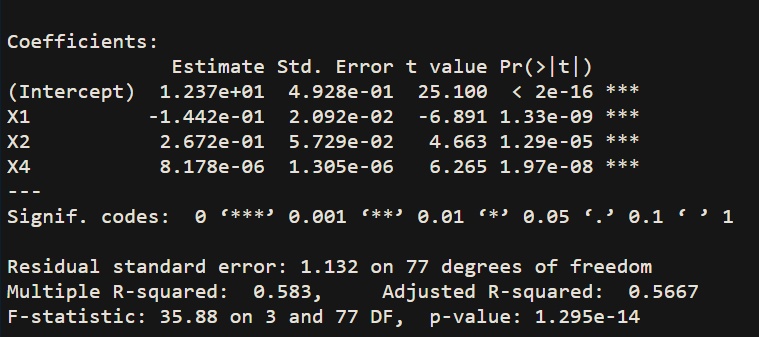
H0 : beta1 = ... = betap-1 = 0 ; H1 : not all of betak = 0

 The null distribution is F distribution

It has regression relation under the level of 0.01

1. H

Because the regression coefficient X3 is not significant under 0.01 level. So, we can drop this variable.





MSE = 1.281

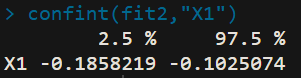
R^2 = 0.583

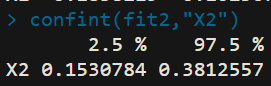
Ra^2 = 0.5667

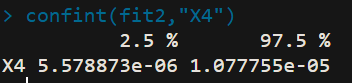
Those values are very similar with the model1.

1. I

The stander error of model1 is bigger than model2.



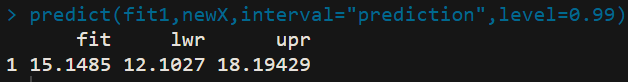




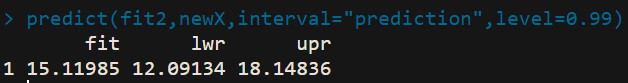
It will be wider if construct those interval under model1

1. J

Model1



Model2 :



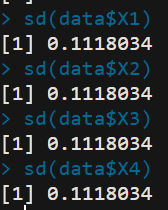
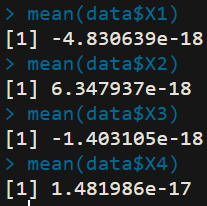
The interval of model2 is smaller than the model1.

1. K

I prefer the model 2. Because it is more concise and the interval is smaller than model1.

##### 7、

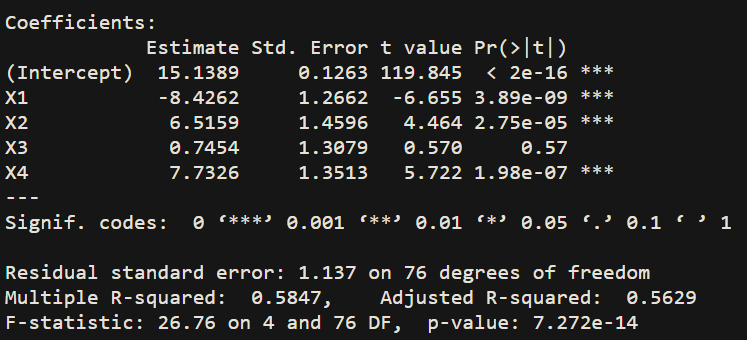
1. 1

Means are all nearly zero.

Standard error are all the same and is 1 / sqrt(80)

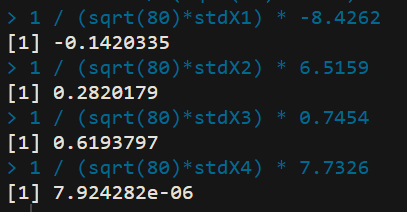
1. 2



The fit regression intercept is 15.1389

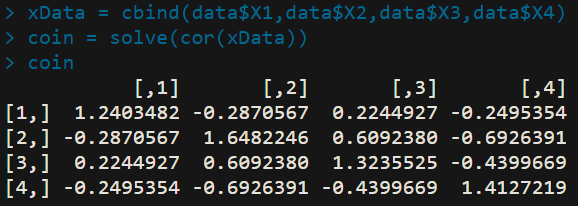


1. 3



I get the result of HW5.

1. 4



oriBetak = VIFk \* MSE / sum((dataOri$Xk - mean(dataOri$Xk))^2)

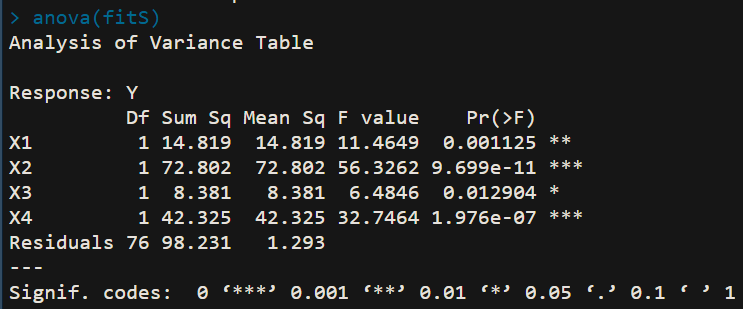
oriBeta1^2 = 1.2403482 \* 1.137^2 /sum((dataOri$X1 - mean(dataOri$X1))^2) = 4.55599\*10-4

oriBeta2^2 = 1.6482246\* 1.137^2 /sum((dataOri$X2 - mean(dataOri$X2))^2) = 3.990448\*10-3

oriBeta3^2 = 1.3235525 \* 1.137^2 /sum((dataOri$X3 - mean(dataOri$X3))^2) = 1.181569

oriBeta4^2 = 1.4127219 \* 1.137^2 /sum((dataOri$X4- mean(dataOri$X1))^4) = 1.918225\*10-12

1. 5



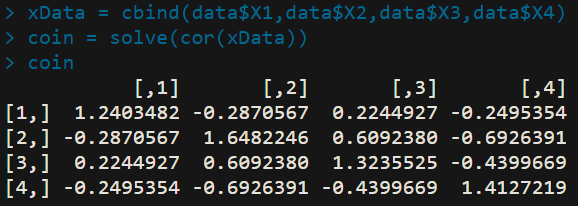
SSTO = 236.558 SSE=98.231 SSR = 138.327 it is same with original model1.

1. 6

R^2 = 0.5847 Ra^2 = 0.5629

It is also the same with original model1.

##### 8、



VIF1 = 1.2403482

VIF2 = 1.6482246

VIF3 = 1.3235525

VIF4 = 1.4127219

1. 2

estimated regression coefficients of X4 in the model with X3 and X4 is smaller than the coefficients of X4 in the model with only X4 variable.

SSR(X4) = 67.775

SSR(X4|X3) = 66.858

Those values are very similar.

Because X3 nearly has no linear association with Y but X4 have medium linear association with Y. So, the most of proportion of SSR is occupied by X4 variable.

1. 3

The X2 coefficient in the model with X2 and X4 variables is more smaller than the X2 coefficient in the model only with X2.

SSR(X2) = 40.503

SSR(X2|X4) = 9.291

Because the muticollinearity effect between X2 and X4.

##### R Code

data = read.table("d:\\property.txt",header = FALSE)

cnames = c("Y","X1","X2","X3","X4")

colnames(data) = cnames

dataOri = read.table("d:\\property.txt",header = FALSE)

cnames = c("Y","X1","X2","X3","X4")

colnames(dataOri) = cnames

hist(data$X1,main = paste("Histogram of X1"),xlab = "X1")

summary(data$X1)

hist(data$X2,main = paste("Histogram of X2"),xlab = "X2")

summary(data$X2)

hist(data$X3,main = paste("Histogram of X3"),xlab = "X3")

summary(data$X3)

hist(data$X4,main = paste("Histogram of X4"),xlab = "X4")

summary(data$X4)

pairs(data)

cor(data)

fit1 = lm(Y~X1+X2+X3+X4,data)

summary(fit1)

par(mfrow=c(2,2))

plot(fit1,which=1)

plot(fit1,which=2)

boxplot(residuals(fit1),main = "Box plot of residuals")

par(mfrow = c(2,2))

for (i in c(2:5)){

plot(x = as.vector(as.matrix(data[i])),y = residuals(fit1),

main = "Residuals vs predict varibale",xlab = paste("X",i),ylab = "Residuals")

}

par(mfrow = c(3,2))

for (i in c(2:4)){

for (j in c((i+1):5)){

plot(x = as.vector(as.matrix(data[i]\*data[j])),y = residuals(fit1),

main = "Residuals vs predict varibale",xlab = paste(paste("X",i-1),paste("X",j-1)),ylab = "Residuals")

}

}

anova(fit1)

fit2 = lm(Y~X1+X2+X4,data)

summary(fit2)

confint(fit2,"X1")

newX=data.frame(X1=4,X2=10,X3=0.1,X4=80000)

predict(fit1,newX,interval="prediction",level=0.99)

####################

meanX1 = mean(data$X1)

meanX2 = mean(data$X2)

meanX3 = mean(data$X3)

meanX4 = mean(data$X4)

stdX1 = sd(data$X1)

stdX2 = sd(data$X2)

stdX3 = sd(data$X3)

stdX4 = sd(data$X4)

data$X1 = 1 / sqrt(80)\*(data$X1-meanX1) / stdX1

data$X2 = 1 / sqrt(80)\*(data$X2-meanX2) / stdX2

data$X3 = 1 / sqrt(80)\*(data$X3-meanX3) / stdX3

data$X4 = 1 / sqrt(80)\*(data$X4-meanX4) / stdX4

fitS = lm(Y~X1+X2+X3+X4,data = data)

summary(fitS)

beta1 = 1 / (sqrt(80)\*stdX1) \* -8.4262

beta2 = 1 / (sqrt(80)\*stdX2) \* 6.5159

beta3 = 1 / (sqrt(80)\*stdX3) \* 0.7454

beta4 = 1 / (sqrt(80)\*stdX4) \* 7.7326

xData = cbind(data$X1,data$X2,data$X3,data$X4)

coin = solve(cor(xData))

summary(lm(Y~X3+X2+X4,data = data))

summary(lm(Y~X1+X2+X3+X4,data = data))