Matriculation No. A0169388H
Name. LI ZHI

AS.3

Q1:

1. ##--------Q1----------##

```
mydata<-read.csv('Sediment.csv')
#1------scatterplot matrices---------
#install.packages('GGally')
library(GGally)
pairs(mydata)
corelationship<-cor(mydata,method='pearson')
write.xlsx(corelationship,file = 'cor.xlsx')
```

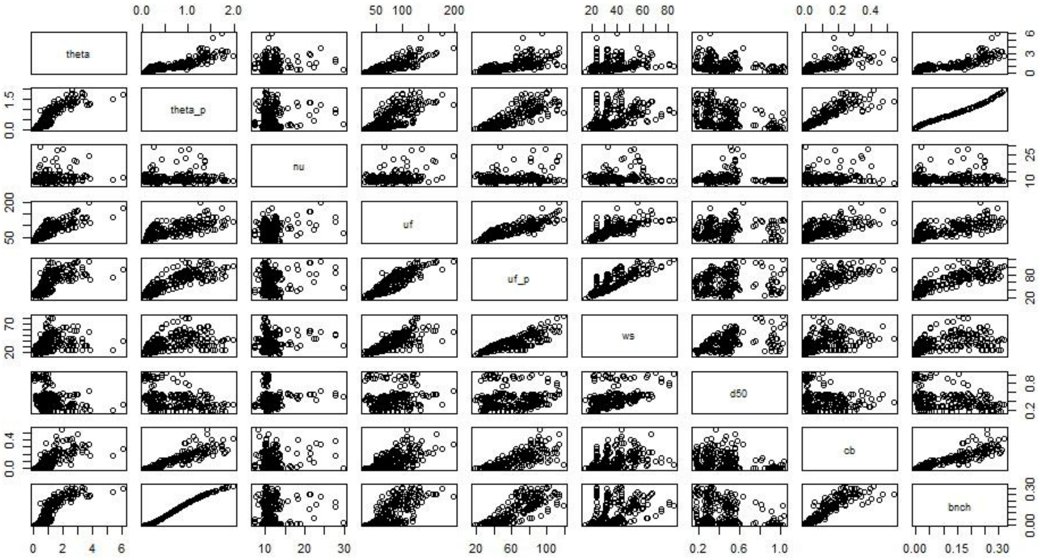

Fig.1 scatterplot matrices

Tab.1 correlation of all variables

	theta	theta_p	nu	uf	uf_p	ws	d50	cb	bnch
theta	1	0.845908258	0.152097888	0.782360562	0.660920907	0.27824871	-0.282756202	0.710913552	0.828671333
theta_p	0.845908258	1	0.075677581	0.698698532	0.835765443	0.443326319	-0.289941362	0.893676834	0.996189352
nu	0.152097888	0.075677581	1	0.266900973	0.175539725	0.133498763	0.071190141	0.151786753	0.086686635
uf	0.782360562	0.698698532	0.266900973	1	0.866942743	0.742154315	0.199731291	0.62793473	0.694165354
uf_p	0.660920907	0.835765443	0.175539725	0.866942743	1	0.84694099	0.144914123	0.784080845	0.846247562
ws	0.27824871	0.443326319	0.133498763	0.742154315	0.84694099	1	0.455686658	0.430143066	0.461769829
d50	-0.282756	-0.289941	0.071190	0.199731	0.144914	0.455686	1	-0.231695	-0.270209

	202	362	141	291	123	658		236	813
cb	0. 7109135 52	0. 8936768 34	0. 151786 753	0. 627934 73	0. 784080 845	0. 430143 066	−0. 231695 236	1	0. 8977457 69
bnch	0. 8286713 33	0. 9961893 52	0. 086686 635	0. 694165 354	0. 846247 562	0. 461769 829	−0. 270209 813	0. 8977457 69	1

As we can see, theta_p has the largest correlation with theta.

2.

```
#2---------linear model---------
library(UsingR)
plot(mydata$theta_p,mydata$theta)
modelize<-lm(mydata$theta~mydata$theta_p)
abline(lm(mydata$theta~mydata$theta_p))
lm_r<- simple.lm(mydata$theta_p,mydata$theta)
simple.lm(mydata$theta_p,mydata$theta,show.ci = T)
summary(lm_r)
```

mathematical equation of this model is y=1.54x+0.18;

the proportion of predictor and response variable is R square which is 0.7156;

3.

```
residual_mydata<-resid(lm_r)
summary(residual_mydata)
qqnorm(residual_mydata)
```

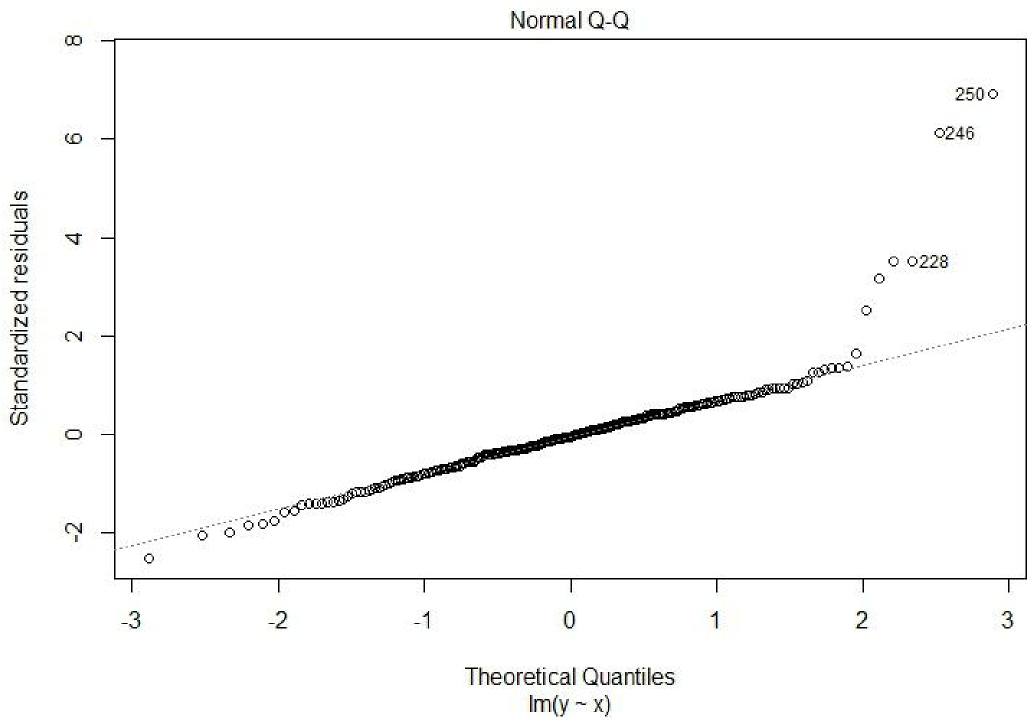

Fig.2. residual distribution

From this figure, the residuals distribution approximately approach to normal distribution and it validates our model.

4.

```
theta<-mydata$theta
theta_p<-mydata$theta_p
pre_theta<- data.frame(theta_p=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7))
prediction<-predict(lm(theta~theta_p),pre_theta)
write.xlsx(prediction,file = 'prediction.xlsx')
```

Tab.2 prediction of theta

	1	2	3	4	5	6	7
x	0. 339027387	0. 493439283	0. 647851178	0. 802263073	0. 956674969	1. 111086864	1. 265498759

Q2:

1.

```
##1-----split the data into 2 data sets-----
my_data<-read.csv('Sediment.csv')
dim(my_data)
index<- sample(1:nrow(my_data),size = round(0.3*nrow(my_data)))
test_data<- my_data[index,]
train_data<-my_data[-index,]
```

2.

```
multi_model<- lm(train_data$theta~train_data$theta_p+train_data$uf+train_data$uf_p)
plot(multi_model)
summary(multi_model)
coef(multi_model)
coef(summary(multi_model))
write.xlsx(coef(multi_model),file = 'coef.xlsx')
```

Tab.3 coefficients of regression

	(Intercept)	train_data$theta_p	train_data$uf	train_data$uf_p
x	-0. 185736	1. 972069	0. 029421	-0. 038886

3.

```
theta<-train_data$theta
theta_p<-train_data$theta_p
uf<-train_data$uf
uf_p<-train_data$uf_p
predict_data<-predict(lm(theta~theta_p+uf+uf_p),test_data)
write.xlsx(predict_data,file = 'prediction_data.xlsx')
```

Tab.4 prediction of theta in multiple regression model

order	x
102	0. 93159172
122	2. 241143277
228	0. 713006117
169	0. 373286967

187	0.789085671
230	1.073375584
161	0.63761097
217	0.788662343
72	1.630723901
33	1.089343053
248	0.55437153
18	0.410191328
47	1.065920476
66	1.521104313
126	1.389524098
32	0.338707449
68	3.160124662
45	0.379485709
50	0.796046996
202	0.871159131
129	2.043002139
116	1.213646647
151	2.865762558
105	0.63709957
256	0.425930596
106	1.104875906
212	3.451311706
23	0.279320299
81	0.183973668
82	0.02511623
55	1.500502726
215	1.046310074
188	0.577261285
98	0.331947446
54	0.714899531
14	1.27116902
232	0.541869708
132	0.857455792
90	0.143885471
210	0.932284923
164	1.151940831
70	3.148491918
197	1.200954938

138	0.456487604
206	0.620126898
34	0.802896417
140	0.48424187
238	0.889696975
3	1.762053099
172	0.841540474
239	1.032272124

```
predict_data<-as.data.frame(predict_data)
X<-seq(1:length(predict_data$predict_data))
theta_test<-as.data.frame(test_data$theta)
data_new<-cbind(theta_test,predict_data)
data_new$order<-X
data_new<-melt(data_new,id='order')
ggplot(model,aes(order,theta))+geom_line(aes(color=source,group=source))+ggtitle('comparison of predict data and observed data')
```

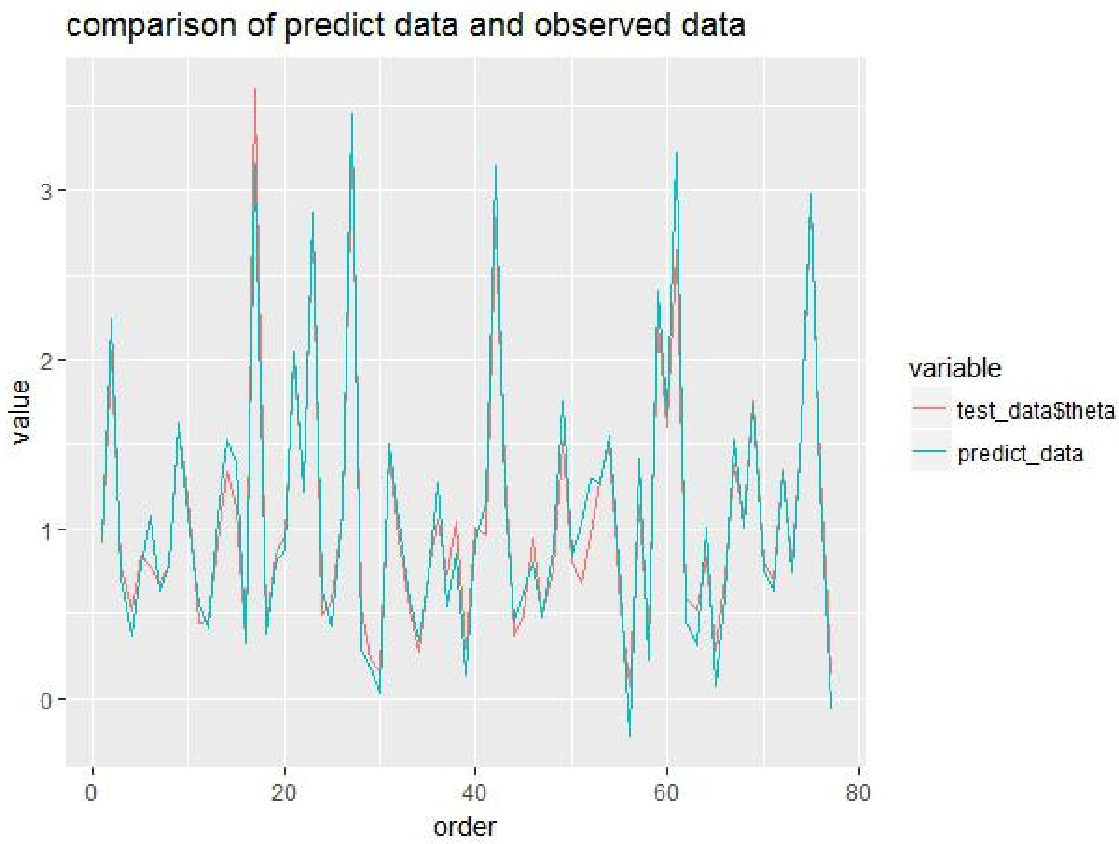

Fig.3 model validation