Problem 1

a)

The fixed model is

$$\hat{Y} = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_1X_2 - 10X_1X_3$$

i. False. The model for male is

$$E\{\hat{Y} \mid X_3 = 0\} = 50 + 20X_1 + 0.07X_2 + 0.01X_1X_2;$$

The model for female is

$$E\{\hat{Y} \mid X_3 = 1\} = 85 + 10X_1 + 0.07X_2 + 0.01X_1X_2;$$

The deviation is

$$E\{\hat{Y} \mid X_3 = 0\} - E\{\hat{Y} \mid X_3 = 1\} = 10X_1 - 35.$$

So the male 's income are more than female when their GPA are above 3.5; Otherwise, female's income would be more than male's.

ii. False. Similar to the discussion above.

iii. True.

iv. False.

b)

$$\hat{Y}_{pred} = 85 + 10 \times 4.0 + 0.07 \times 110 + 0.01 \times 4 \times 110 = 137.1$$

c)

False. Since values of IQ data are much larger than that of GPA data, a small change on IQ value or interaction could leads to large change on response even thought its coefficient is small. To test the relative effect of every predictor on response, we can fix a model in scaled coefficient to see whether there is enough evidence on interaction which in fact increase Y about 5%.

Problem 2

$$\hat{Y}_i = X_i\hat{\beta} = X_i \frac{1}{\sum_{i'}^{n} X_{i'}^2} \sum_{i'}^{n} X_{i'}Y_{i'} = \frac{1}{\sum_{i'}^{n} X_{i'}^2} \sum_{i'}^{n} (X_i X_{i'})Y_{i'} = \sum_{i'}^{n} \left(\frac{X_i X_{i'}}{\sum_{i'}^{n} X_{i'}^2}\right)Y_{i'}$$

So $a_{i'} = \frac{\sum_{i'}^{n} X_i X_{i'}}{\sum_{i'}^{n} X_{i'}^2}$, where X_i is the ith sample.

Problem 3

a)

$X_i \sim N(0,1)$

```
> x
  [1] -0.626453811  0.183643324 -0.835628612  1.595280802  0.329507772 -0.820468384
  [7]  0.487429052  0.738324705  0.575781352 -0.305388387  1.511781168  0.389843236
 [13] -0.621240581 -2.214699887  1.124930918 -0.044933609 -0.016190263  0.943836211
 [19]  0.821221195  0.593901321  0.918977372  0.782136301  0.074564983 -1.989351696
 [25]  0.619825748 -0.056128740 -0.155795507 -1.470752384 -0.478150055  0.417941560
 [31]  1.358679552 -0.102787727  0.387671612 -0.053805041 -1.377059557 -0.414994563
 [37] -0.394289954 -0.059313397  1.100025372  0.763175748 -0.164523596 -0.253361680
 [43]  0.696963375  0.556663199 -0.688755695 -0.707495157  0.364581962  0.768532925
 [49] -0.112346212  0.881107726  0.398105880 -0.612026393  0.341119691 -1.129363096
 [55]  1.433023702  1.980399899 -0.367221476 -1.044134626  0.569719627 -0.135054604
 [61]  2.401617761 -0.039240003  0.689739362  0.028002159 -0.743273209  0.188792300
 [67] -1.804958629  1.465554862  0.153253338  2.172611670  0.475509529 -0.709946431
 [73]  0.610726353 -0.934097632 -1.253633400  0.291446236 -0.443291873  0.001105352
 [79]  0.074341324 -0.589520946 -0.568668733 -0.135178615  1.178086997 -1.523566800
 [85]  0.593946188  0.332950371  1.063099837 -0.304183924  0.370018810  0.267098791
 [91] -0.542520031  1.207867806  1.160402616  0.700213650  1.586833455  0.558486426
 [97] -1.276592208 -0.573265414 -1.224612615 -0.473400636
```

b)

$\varepsilon_i \sim N(0,0.25)$

```
> eps
  [1] -0.077545835  0.005264484 -0.113865206  0.019753597 -0.081823080  0.220910909
  [7]  0.089588435  0.113771779  0.048023170  0.210272010 -0.079467057 -0.057705591
 [13]  0.179035280 -0.081337044 -0.025922593 -0.049100991 -0.039999109 -0.034889163
 [19]  0.061773541 -0.022166310 -0.063244683  0.167879853 -0.026822426 -0.022444566
 [25] -0.012523843  0.089083288 -0.009195551 -0.004704271 -0.085207560 -0.040533784
 [31]  0.007520055 -0.073611811  0.066437024 -0.189799260  0.038319733 -0.192056228
```

```
[37] -0.037622016 -0.066034988 -0.081511848 -0.007112097 -0.239294928  0.147072914
[43] -0.208121555 -0.057941300 -0.139490013 -0.093852375  0.260895818  0.002174452
[49] -0.160787566 -0.205075692  0.056273388 -0.002319979 -0.039758547 -0.116170268
[55] -0.185932539 -0.134399037  0.125003600 -0.077658337 -0.173053356  0.233661328
[61]  0.053137547 -0.029830888  0.132310381  0.110802831 -0.077405381  0.275762808
[67] -0.031878379 -0.178061831 -0.018049950  0.025942292  0.288497300  0.013225296
[73]  0.057124851 -0.009644117 -0.041750105 -0.004340754  0.098454951  0.259405626
[79]  0.128424055  0.150988550 -0.153915428  0.122986946  0.027490600 -0.183406254
[85]  0.065127843 -0.019844326  0.183073414 -0.095760250 -0.053776469 -0.115763687
[91] -0.022137995  0.050251472 -0.091468522  0.103796646 -0.151010348 -0.130998052
[97]  0.180144713 -0.126980933  0.051496839 -0.047634506
```

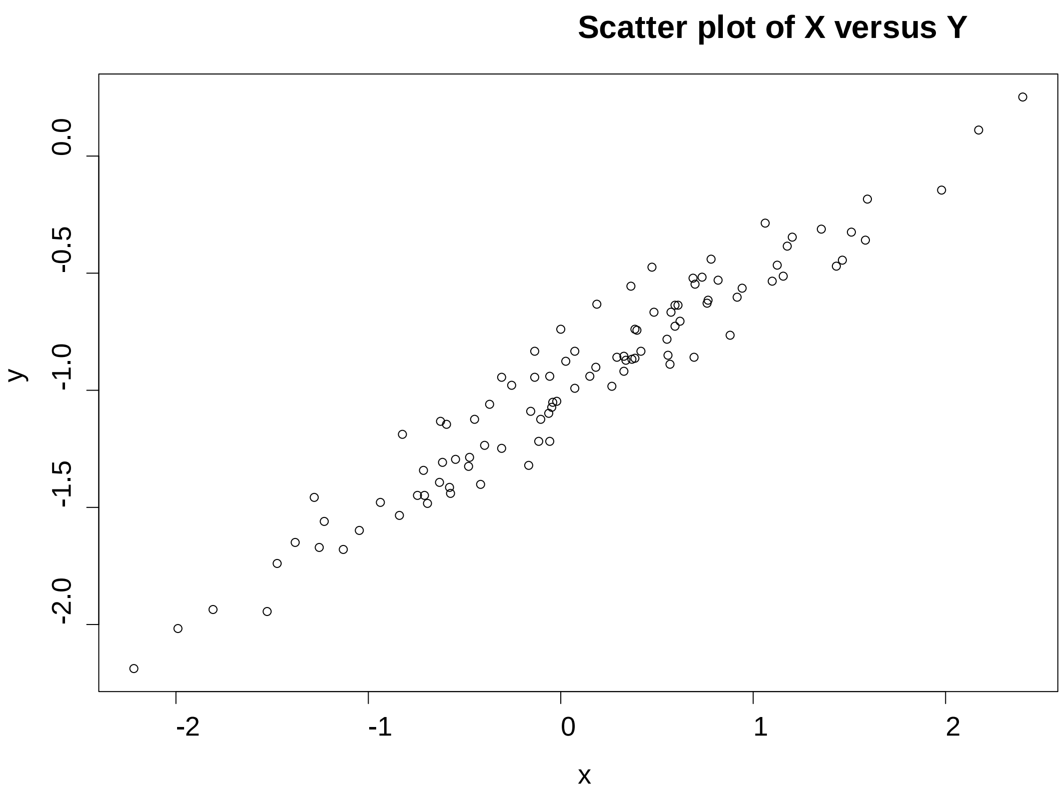
c)

$Y_i = -1 + 0.5X_i + \varepsilon_i$

$\beta_0 = -1, \beta_1 = 0.5$, the length of Y_i is 100

```
> y
  [1] -1.3907727 -0.9029139 -1.5316795 -0.1826060 -0.9170692 -1.1893233
  [7] -0.6666970 -0.5170659 -0.6640862 -0.9424222 -0.3235765 -0.8627840
 [13] -1.1315850 -2.1886870 -0.4634571 -1.0715678 -1.0480942 -0.5629711
 [19] -0.5276159 -0.7252156 -0.6037560 -0.4410520 -0.9895399 -2.0171204
 [25] -0.7026110 -0.9389811 -1.0870933 -1.7400805 -1.3242826 -0.8315630
 [31] -0.3131402 -1.1250057 -0.7397272 -1.2167018 -1.6502100 -1.3995535
 [37] -1.2347670 -1.0956917 -0.5314992 -0.6255242 -1.3215567 -0.9796079
 [43] -0.8596399 -0.7796097 -1.4838679 -1.4476000 -0.5568132 -0.6135591
 [49] -1.2169607 -0.7645218 -0.7446737 -1.3083332 -0.8691987 -1.6808518
 [55] -0.4694207 -0.1441991 -1.0586071 -1.5997257 -0.8881935 -0.8338660
 [61]  0.2539464 -1.0494509 -0.5228199 -0.8751961 -1.4490420 -0.6298410
 [67] -1.9343577 -0.4452844 -0.9414233  0.1122481 -0.4737479 -1.3417479
 [73] -0.6375120 -1.4766929 -1.6685668 -0.8586176 -1.1231910 -0.7400417
 [79] -0.8344053 -1.1437719 -1.4382498 -0.9446024 -0.3834659 -1.9451897
 [85] -0.6378991 -0.8533691 -0.2853767 -1.2478522 -0.8687671 -0.9822143
 [91] -1.2933980 -0.3458146 -0.5112672 -0.5460965 -0.3575936 -0.8517548
 [97] -1.4581514 -1.4136136 -1.5608095 -1.2843348
```

d)



The scatter plot shows the linear relationship between X_i and Y_i.

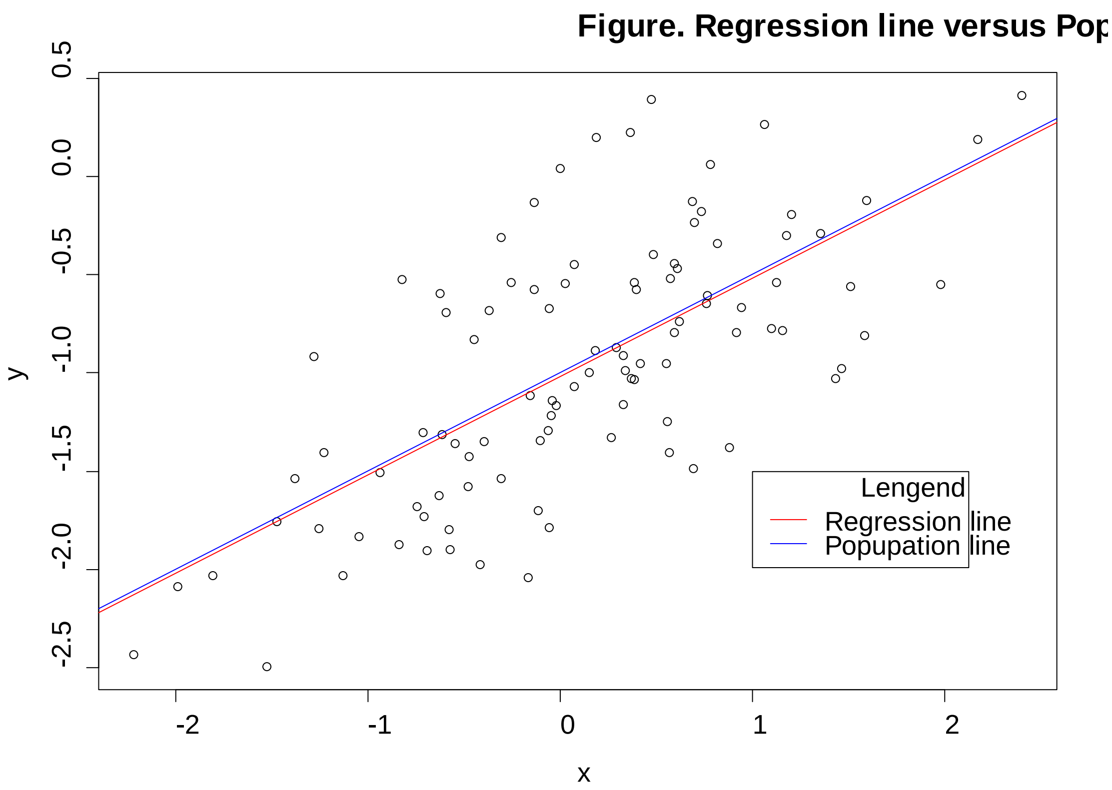
e)

```
> summary(lm.y)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.23461 -0.07672 -0.01744  0.06742  0.29327 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.00471    0.01212  -82.87   <2e-16 ***
x            0.49987    0.01347   37.12   <2e-16 ***
---
Residual standard error: 0.1203 on 98 degrees of freedom
Multiple R-squared:  0.9336, Adjusted R-squared:  0.9329 
F-statistic:  1378 on 1 and 98 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_0 = -1.00471, \hat{\beta}_1 = 0.49987$

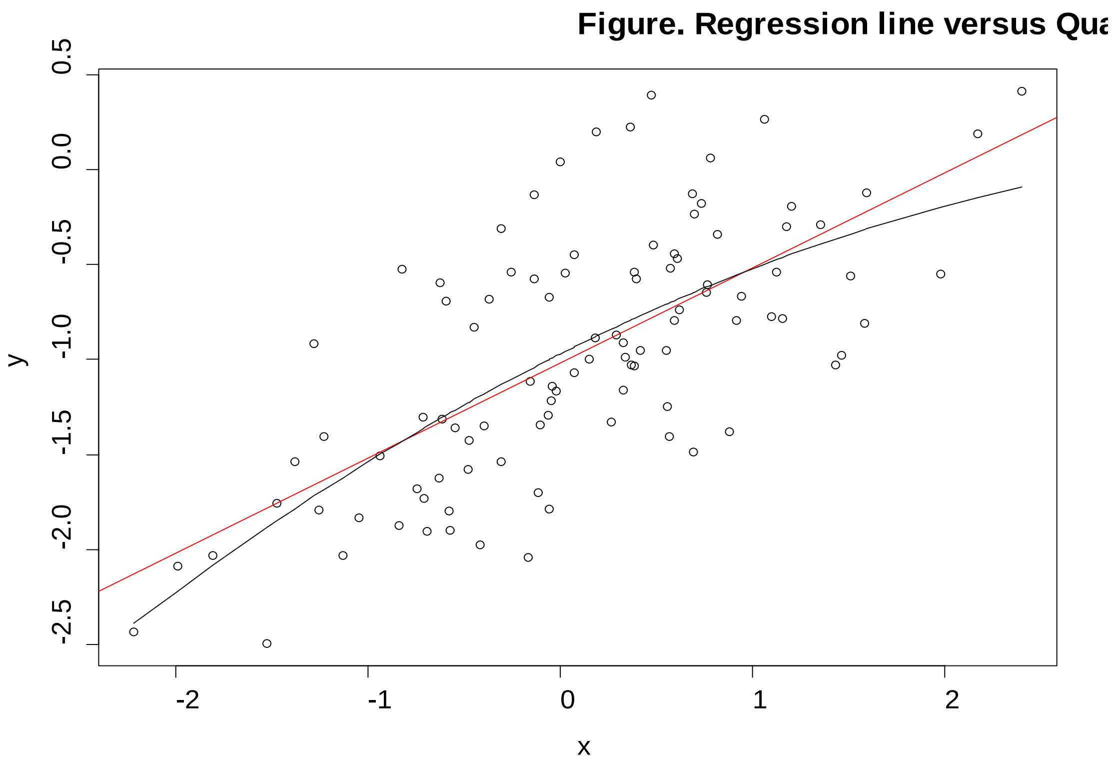
$\hat{\beta}_0$ is a little bit larger than β_0, $\hat{\beta}_1$ is a little bit smaller than β_1, which means that the regression line would be a little more gentle than the true model.

f)



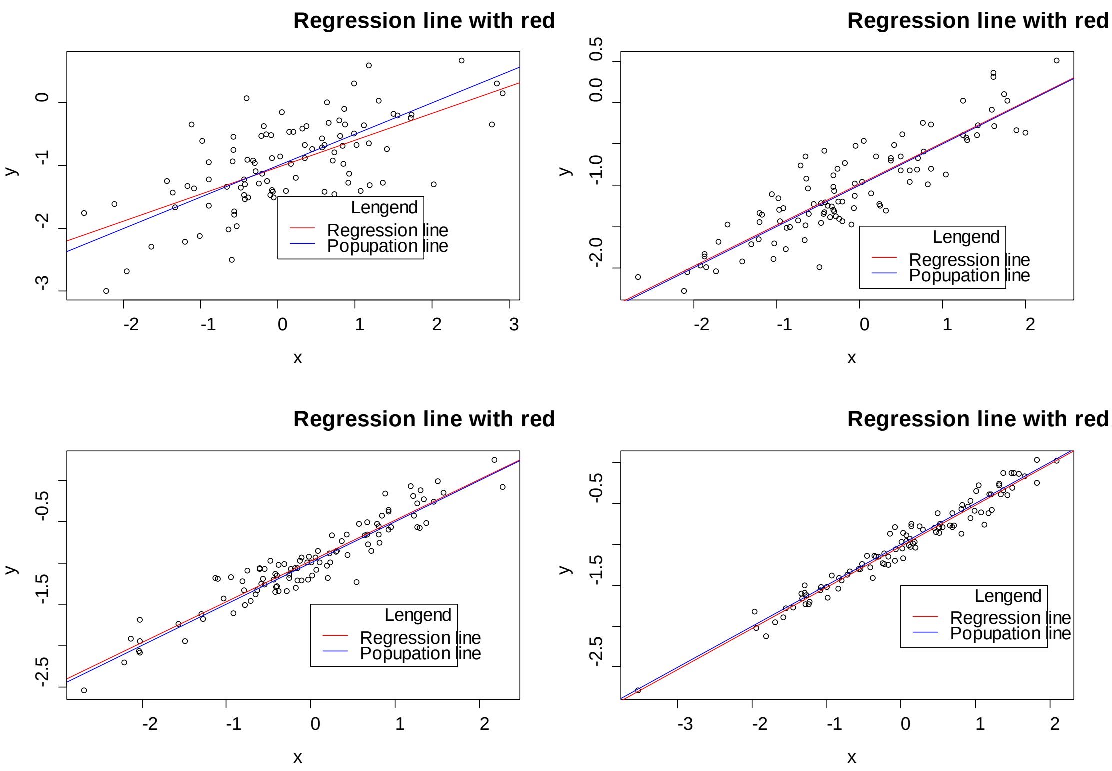
g)

```
> summary(lm.yp)
Call:
lm(formula = y ~ x + z)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98252 -0.31270 -0.06441  0.29014  1.13500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
x            0.50858    0.05399   9.420  2.4e-15 ***
z           -0.05946    0.04238  -1.403    0.164
---
Residual standard error: 0.479 on 97 degrees of freedom
Multiple R-squared:  0.4779, Adjusted R-squared:  0.4672
F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```



Since there is very slight improve in R-square value, we cannot say that the quadratic term improve the model fit.

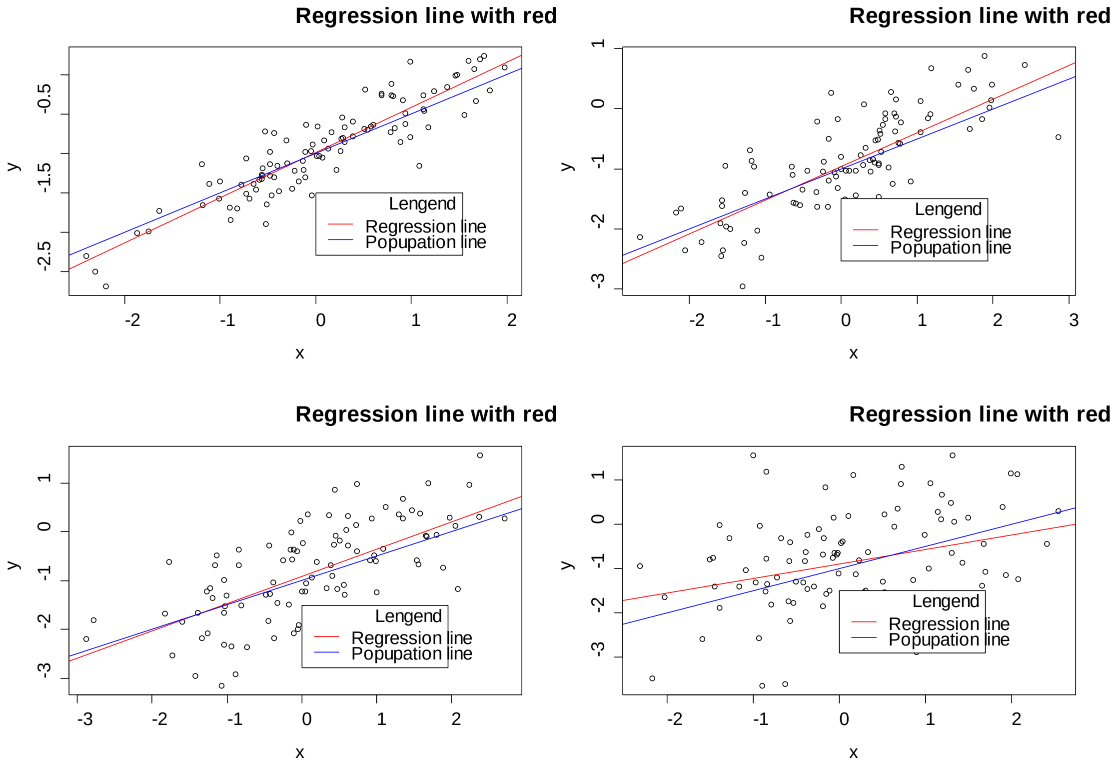
h)



Let standard deviation be 0.5,0.25,0.5/3,0.125.

If variance of error term reduces, residuals become smaller while R-square value become larger. This indicates that the population line and regression line would become more consistent as shown in plots.

i)



Let standard deviation be 0.25,0.5,0.75.1.5.

As the variance of error term increases, residuals become larger while R-square value become smaller. This indicates that the population line and regression line would detach from each other as shown in plots.

j)

Let the standard deviation of error term be 0.25, 0.5, 0.75. Let $\alpha = 0.05$

```
                2.5 %     97.5 %
(Intercept) -1.0273159 -0.9299102
x            0.4734874  0.5713821
                2.5 %     97.5 %
(Intercept) -1.0769589 -0.8454968
x            0.3710526  0.6046566
                2.5 %     97.5 %
(Intercept) -1.1196345 -0.8061289
```

```
x           0.4152323  0.6957725
```

Compare with the original data, the confidence interval of coefficients on less noisy data is narrower, and the confidence interval of coefficient on noisier data is wider. This indicates that the wide of confidence interval is positively related to the variance of error term in data.

Problem 4

a) a

The first five subjects in the training data:

```
> row.names(sample_4a.matrix[ind.ntrain[1:5],])
[1] "yaleB11" "yaleB16" "yaleB23" "yaleB35" "yaleB08"
```

The first five subjects in the testing data:

```
> row.names(sample_4a.matrix[ind.ntest[1:5],])
[1] "yaleB02" "yaleB03" "yaleB05" "yaleB05" "yaleB06"
```

b)

All the subject is correctly identified.

```
> summary(knn(sample_4a.train.scores[,1:25],sample_4a.test.scores[,1:25],cl,k=1,l=0))
yaleB01 yaleB02 yaleB03 yaleB04 yaleB05 yaleB06 yaleB07 yaleB08 yaleB09 yaleB10
      0       1       1       0       2       2       0       1       2       2
yaleB11 yaleB12 yaleB13 yaleB15 yaleB16 yaleB17 yaleB18 yaleB19 yaleB20 yaleB21
      0       0       0       2       0       0       1       1       2       1
yaleB22 yaleB23 yaleB24 yaleB25 yaleB26 yaleB27 yaleB28 yaleB29 yaleB30 yaleB31
      0       0       0       2       1       0       1       0       2       3
yaleB32 yaleB33 yaleB34 yaleB35 yaleB36 yaleB37 yaleB38 yaleB39
      1       1       0       0       0       1       0       1
```

c) c

There are just 4 subjects identified correctly.

```
> summary(knn(sample_4c.train.scores[,1:25],sample_4c.test.scores[,1:25],cl,k=1,l=0))
yaleB01 yaleB02 yaleB03 yaleB04 yaleB05 yaleB06 yaleB07 yaleB08 yaleB09 yaleB10
      0       1       1       1       0       3       2       3       0       3
yaleB11 yaleB12 yaleB13 yaleB15 yaleB16 yaleB17 yaleB18 yaleB19 yaleB20 yaleB21
      1       0       0       0       0       0       2       0       2       0
yaleB22 yaleB23 yaleB24 yaleB25 yaleB26 yaleB27 yaleB28 yaleB29 yaleB30 yaleB31
      0       3       0       0       0       0       1       0       1       0
yaleB32 yaleB33 yaleB34 yaleB35 yaleB36 yaleB37 yaleB38 yaleB39
      1       1       3       0       1       0       1       0
```

Use the result, we can find all the subjects that are misidentified and draw them out.

```
> ## 1NN classification
> # based on function dis()
> count <- 0
> for(i in 1:dim(sample_4c.test.scores)[1]){
+   distance0 <- 1000000
+   class0 <- "0"
+   index <-0
+   for(j in 1:dim(sample_4c.matrix.train.center)[1]){
+     distance <- dis(sample_4c.test.scores[i,1:25], sample_4c.train.scores[j,1:25])
+     # print(distance)
+     if(distance < distance0){
+       distance0 <- distance
+       class0 <- row.names(sample_4c.matrix.train)[j]
+       index <- j
+     }
+   }
+   if(class0==row.names(sample_4c.matrix.test)[i]){
+     print(1)
+     count <- count + 1
+   }else{
+     print(0)
+     # draw out it
+     face_misid.matrix0 <- sample_4c.matrix.test[i,]
+     dim(face_misid.matrix0) <- faces.matrix.dimension
+     face_misid.matrix1 <- sample_4c.matrix.train[index,]
+     dim(face_misid.matrix1) <- faces.matrix.dimension
+     face_misid.compare.matrix <- cbind(face_misid.matrix0, face_misid.matrix1)
+     face_misid.compare <- pixmapGrey(face_misid.compare.matrix)
+     filename = sprintf("compare of %s and %s.png", row.names(sample_4c.matrix.test)[i],
row.names(sample_4c.matrix.train)[index])
+     plot(face_misid.compare, main=filename)
+     dev.copy(device=png, file=filename, height=600, width=800)
+     dev.off()
+   }
+ }
```

The 4 correctly identified subjects are

```
"yaleB02"
"yaleB02"
"yaleB32"
"yaleB38"
```

Plots the faces remain that are misidentified

1.compare of yaleB01 and yaleB03.png

3.compare of yaleB02 and yaleB34.png

4.compare of yaleB03 and yaleB10.png

5.compare of yaleB04 and yaleB07.png

6.compare of yaleB09 and yaleB08.png

7.compare of yaleB09 and yaleB08.png

8.compare of yaleB11 and yaleB28.png

9.compare of yaleB11 and yaleB20.png

10.compare of yaleB13 and yaleB10.png

11.compare of yaleB17 and yaleB23.png

12.compare of yaleB17 and yaleB06.png

13.compare of yaleB19 and yaleB07.png

14.compare of yaleB23 and yaleB11.png

16.compare of yaleB24 and yaleB18.png

17.compare of yaleB24 and yaleB18.png

18.compare of yaleB26 and yaleB06.png

19.compare of yaleB26 and yaleB06.png

20.compare of yaleB28 and yaleB20.png

21.compare of yaleB28 and yaleB30.png

22.compare of yaleB28 and yaleB23.png

23.compare of yaleB29 and yaleB10.png

24.compare of yaleB30 and yaleB04.png

26.compare of yaleB33 and yaleB34.png

27.compare of yaleB34 and yaleB36.png

28.compare of yaleB34 and yaleB33.png

29.compare of yaleB35 and yaleB08.png

30.compare of yaleB36 and yaleB34.png

d) d

I set seed from 3 to 12 separately. Here is the output of count that the numbers of faces identified correctly.

```
[1] 4
[1] 9
[1] 6
[1] 6
[1] 9
[1] 3
[1] 7
[1] 10
[1] 8
[1] 4
```

The average of the count is 6.6, standard deviation is 2.4129.

e)

The numbers of successful identified faces tells us that the PCA works differently on the same subjects set under different azimuth(such: A+035) and elevation degree(such: E+15).

With comparison of b and c, PCA identified poorly in high azimuth and high elevation conditions because the light are not normally distributed on faces under such situations. Such that, the Faces Identification Processing is very sensitive to the direction of faces.

The reason of this result could be various. If the light direction changes, which means the distribution of light on faces becomes abnormal, the matrix of the face might became singular matrix(too bright at one side of a picture) so that the eigenvalue would not exist or not such significant. On the other hand, the noisy in scores might also effect the result of the identification.

f)

The result might be worse if we use the uncropped pictures since that the light condition would be more obviously reflected on the matrix so that the principle components could contain more noisy than cropped data sets.