Name: Jingyi Yuan
UNI: jy2736
Class: STAT W4240

Homework 03

Problem 1

(a).

iii is correct.

Suppose the starting salary after graduation is y:

$$y = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA * IQ - 10GPA * Gender$$

for female ($X_3 = 1$):

$$y = 50 + 20GPA + 0.07IQ + 35 + 0.01GPA * IQ - 10GPA$$

$$y\_female = 85 + 10GPA + 0.07IQ + 0.01GPA * IQ$$

for male ($X_3 = 0$):

$$y\_male = 50 + 20GPA + 0.07IQ + 0.01GPA * IQ$$

For a fixed value of IQ and GPA, we need to compare (85 + 10GPA) and (50 + 20GPA) and since we do not know the exact GPA, we cannot decide whether males or females earn more on average. So i and ii are wrong. But provided that the GPA is high enough (GPA > 3.5), we can know that males earn more on average than females, so iii is correct

(b).

$$y\_female = 85 + 10GPA + 0.07IQ + 0.01GPA * IQ$$

$$y\_female = 85 + 10 * 4.0 + 0.07 * 110 + 0.01 * 110 * 4.0$$

$$y\_female = 137.1\ thousand\ dollars$$

(c).

False

Use hypothesis testing:

Suppose H_0: $\beta_4 = 0$, then we can perform this hypothesis test by computing the F-statistic. If H_0 is true, F-statistic should be close to 1, otherwise it should be greater.

Problem 2

(a).

There is not enough information to tell.

Since the cubic regression is more flexible, it may fit the data better than the linear regression even if the relationship between X and Y is linear. Because the number of the data is limited and we do not know the exact data, we cannot expect one RSS to be lower than the other.

(b).

There is not enough information to tell.

The testing RSS depends on the test data. Although we may expect the linear regression model to fit the data better and thus has lower testing RSS, the test data can also fit the cubic regression well (if the test data set is small or if it coincidentally can fit the data well).

(c).

We would expect the training RSS for cubic regression to be lower than the other.

Since the cubic regression is more flexible and the relationship between X and Y is not linear, the cubit regression may fit the data better than the linear regression.

(d).

There is not enough information to tell.

The RSS depends on the test data. We may expect the cubic regression model to fit the data better and has lower testing RSS if the relationship is closer to the cubic regression. Also we may expect the test data can fit the linear regression well if it is closer to linear regression. Since we do not know either the data set or the model, it is hard to tell which one has the lower testing RSS.

Problem 3

(a).

	1	2	3	4	5	6	7	8	9
1	0.0000000	0.4222582	4.8926648	3.5255829	1.6157577	4.0173130	4.99313534	4.7356405	4.8217840
2	0.4222582	0.0000000	4.4832414	3.2805294	1.2017552	3.7547634	4.60912059	4.3725948	4.4188412
3	4.8926648	4.4832414	0.0000000	3.0384989	3.2821468	2.8491413	0.86746538	1.3352554	0.2709822
4	3.5255829	3.2805294	3.0384989	0.0000000	2.5616948	0.5189320	2.51269749	1.9904556	2.8010084
5	1.6157577	1.2017552	3.2821468	2.5616948	0.0000000	2.9523317	3.44543841	3.2590603	3.2233971
6	4.0173130	3.7547634	2.8491413	0.5189320	2.9523317	0.0000000	2.21211835	1.6613896	2.5929574
7	4.9931353	4.6091206	0.8674654	2.5126975	3.4454384	2.2121183	0.00000000	0.5578163	0.6161452
8	4.7356405	4.3725948	1.3352554	1.9904556	3.2590603	1.6613896	0.55781634	0.0000000	1.0645004
9	4.8217840	4.4188412	0.2709822	2.8010084	3.2233971	2.5929574	0.61614524	1.0645004	0.0000000
10	4.9083691	4.5639044	1.7306693	1.8888870	3.4965704	1.4786157	0.89619151	0.4166180	1.4621086
11	3.2439888	3.2316818	4.9146705	1.9368522	3.2234934	2.3296779	4.44684306	3.9271605	4.6943326
12	6.7741772	6.3932334	2.1930352	3.9652802	5.2303056	3.5289886	1.78541086	2.0886802	2.1144262
13	4.2858379	3.8726377	0.6329516	2.7892472	2.6710005	2.7057683	1.21457856	1.4858866	0.6829752
14	0.3737693	0.2017178	4.6220315	3.4815845	1.3425654	3.9545111	4.77254054	4.5481211	4.5659434
15	5.9211394	5.5992978	2.5851580	2.5993709	4.5796875	2.0903844	1.73190885	1.4813731	2.3464807
16	1.2749250	0.8678299	3.6179554	2.6594984	0.3483426	3.0873414	3.74901047	3.5338754	3.5510615
17	2.8350718	2.6392390	3.5142714	0.7867795	2.1466006	1.3042865	3.12564303	2.6477672	3.3066193
18	5.0460794	4.6433491	0.3180300	2.9527892	3.4479485	2.7145885	0.59668630	1.1185697	0.2245621
19	4.3637225	4.0380816	2.0076964	1.2683278	3.0329658	0.9019763	1.31400902	0.7600340	1.7417798
20	4.9808835	4.5965444	0.8561590	2.5095579	3.4322490	2.2117048	0.01525788	0.5596599	0.6035401

The matrix distance is a 20*20 matrix whose element $distance_{ij} = distance_{ji}$ and the diagonal elements are all 0.

```
distance                num [1:20, 1:20]
```

(b).

```
> MSE_test
 [1] 0.0196656396 0.0268217849 0.0152777998 0.0014688370 0.0032978182 0.0008362937
 [7] 0.0192632625 0.0656325812 0.0964741491 0.1647239008

> MSE_train
 [1] 0.00000000 0.02985241 0.04586316 0.06450685 0.06215975 0.07557705 0.08332753 0.10100542
 [9] 0.10691977 0.13030996
```

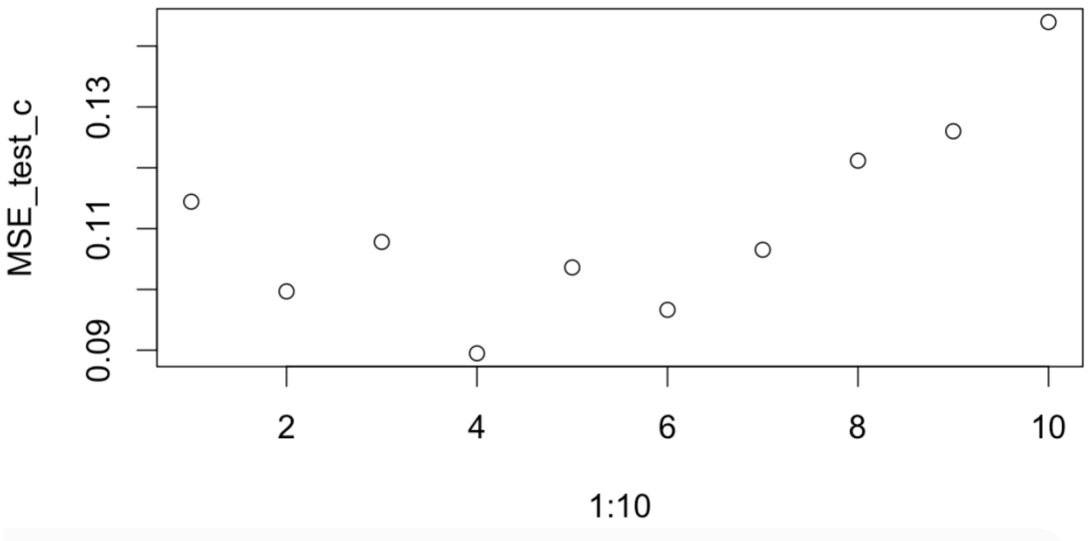
(c).

```
> MSE_test_c
 [1] 0.11442243 0.09968107 0.10780903 0.08949504 0.10361356 0.09665278 0.10652029 0.12116962
 [9] 0.12600534 0.14395572

> MSE_train_c
 [1] 0.00000000 0.02880520 0.04563393 0.05918676 0.06048102 0.07129727 0.07454968 0.08717699
 [9] 0.09846322 0.11061502
```

(d).

We should make our choice based on MSE_{test}^{k}. When we are using the training data, the MSE increases with k because of overfitting (when k = 1, the MSE is smallest, which is 0). As we can see in the picture, we choose k = 4, where MSE_{test}^{k} is smallest, which means we have smallest error.



Problem 4

(a).

Store the collection as a matrix where each row is a photo:

```
> dim(face_matrix_6a)
[1]   152 32256
```

Record the subject number and view of each row of face matrix 6a in a data frame:

```
> dim(subject)
[1]   152 32257
> class(subject)
[1] "data.frame"
```

	subjectNumber	data.1	data.2	data.3	data.4	data.5	data.6	data.7
1	1	0.31372549	0.31764706	0.31372549	0.30980392	0.31372549	0.32156863	0.30980392
2	2	0.41568627	0.40000000	0.38823529	0.39215686	0.40784314	0.39607843	0.41568627
3	3	0.36862745	0.36078431	0.35294118	0.35686275	0.37254902	0.36862745	0.38431373
4	4	0.46666667	0.46274510	0.44313725	0.44705882	0.46666667	0.46274510	0.48627451
5	5	0.33725490	0.30980392	0.30196078	0.29019608	0.28235294	0.27843137	0.26666667
6	6	0.40000000	0.40000000	0.38823529	0.36862745	0.34901961	0.34509804	0.34901961
7	7	0.35686275	0.35686275	0.34509804	0.32549020	0.30588235	0.30196078	0.30980392
8	8	0.43137255	0.43137255	0.43529412	0.41960784	0.38823529	0.38039216	0.39607843
9	9	0.31372549	0.31764706	0.31372549	0.30980392	0.30588235	0.30980392	0.30588235
10	10	0.46274510	0.47450980	0.47450980	0.45490196	0.43921569	0.43921569	0.44313725
11	11	0.43921569	0.44705882	0.44705882	0.42352941	0.41176471	0.40784314	0.41176471
12	12	0.54901961	0.56078431	0.55686275	0.53725490	0.52549020	0.53333333	0.53725490

the first 5 files in the testing set:

```
ind_test_6a          int [1:31] 5 12 18 20 21
```

which is:

```
picture_test_6a      num [1:31] 1 4 2 4 1
index_test_6a        num [1:31] 2 3 5 5 6
```

the 1st picture of the 2nd person
the 4th picture of the 3rd person
the 2nd picture of the 5th person
the 4th picture of the 5th person
the 1st picture of the 6th person

The first 5 files in the training set:

```
ind_train_6a         int [1:121] 41 57 86 136 30
```

which is:

```
picture_train_6a     num [1:121] 1 1 2 4 2
index_train_6a       num [1:121] 11 15 22 34 8
```

the same meaning as the testing set

(b).

31 subjects are identified correctly. 0 incorrectly.

```
> right
[1] 31
> wrong
[1] 0
```

(c).

Store the collection as a matrix where each row is a photo:

```
> dim(face_matrix_6c)
[1]   152 32256
```

Record the subject number and view of each row of face matrix 6a in a data frame:

```
> dim(subject_c)
[1]   152 32257
> class(subject_c)
[1] "data.frame"
```

4 subjects are identified correctly. 27 incorrectly.

```
> right_c
[1] 4
> wrong_c
[1] 27
```

One of the misidentified face (the first misidentified face) is shown below:

original **identified as**

(d).

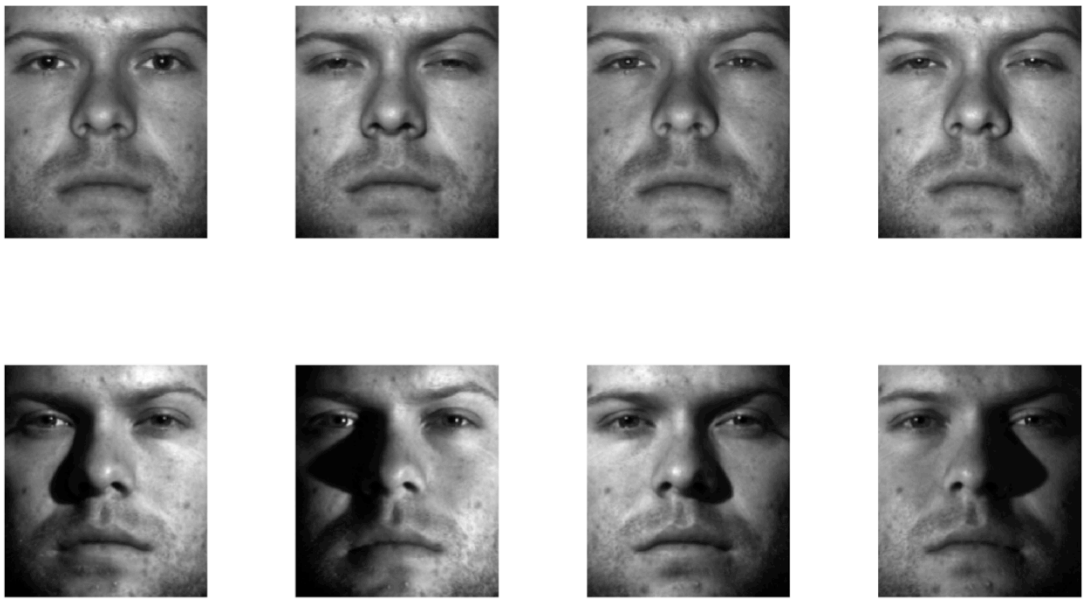
using seed from 3 to 12, we have

```
> right_d
 [1]  4  9  6  6  9  3  7 10  7  4
> wrong_d
 [1] 27 22 25 25 22 28 24 21 24 27
```

From these numbers, we know that the classification of subjects using the views P00A-035E+15, P00A-050E+00, P00A+035E+15, and P00A+050E+00 is not very accurate, no matter how testing and training sets are divided.

(e).

The testing error rates are different in (b) and (c) only because we changed the views of each person. Plotting out the 4 views we used for the 1st person in (b) and (c), we can see that in (c), the pictures have much wider lighting ranges (it has shadow in each view) and in (a), the views are closer to each other. So we can conclude that PCA works better with good lighting condition, that is, the lighting condition for all subjects are closer to each other, and is neither too light nor too dim, with little shadow and a narrow lighting range. The performance of PCA will be influenced a lot by the lighting condition.



Problem 5

(a).

when x < 0.05:

$$fraction1 = \int_0^{0.05} (x + 0.05)dx = 3.75\%$$

when x is in [0.05, 0.95]:

$$fraction2 = (0.95 - 0.05) * 10\% = 9\%$$

when x > 0.95:

$$fraction3 = \int_{0.95}^{1} (0.05 + 1 - x)dx = 3.75\%$$

so on average:

$$fraction = 3.75\% + 9\% + 3.75\% = 9.75\%$$

(b).

$$fraction = 9.75\% * 9.75\% = 9.50625\%$$

(c).

$$fraction = (9.75\%)^{100} \approx 0$$

(d).

When p is large:

$$fraction = (9.75\%)^{p} \to 0$$

that is, there are very few training observations "near" any given test observation.

(e).

for p = i (i = 1, 2, …, 100), we have:

$$length = \sqrt[i]{10\%} = 0.1^{\frac{1}{i}}$$

Problem 6

(a).

$$g(x) = \beta_0 + \beta_1 * hours + \beta_2 * GPA$$
$$g(x) = -6 + 0.05 * hours + GPA$$
$$g(x) = -6 + 0.05 * 40 + 3.5 = -0.5$$
$$p(X) = \frac{e^{g(x)}}{1 + e^{g(x)}} = 37.75\%$$

So the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an A in the class is 37.75%.

(b).

$$p(X) = \frac{e^{g(x)}}{1 + e^{g(x)}} = 0.5$$

so we have:

$$g(x) = 0$$

$$g(x) = -6 + 0.05 * hours + 3.5 = 0$$
$$hours = 50$$

So the student in part (a) would need to study 50 hours t o have a 50% chance of getting an A in the class.

Problem 7

We should prefer to use logistic regression for classification of new observations.

Since what we use is 1-nearest neighbors (i.e. K = 1), the average error rate for the training sets is 0. Also, we divided the data set into equally-sized sets. So it is obvious that the average error rate on the test data is 2*18% = 36%, which is larger than 30% when using logistic regression.