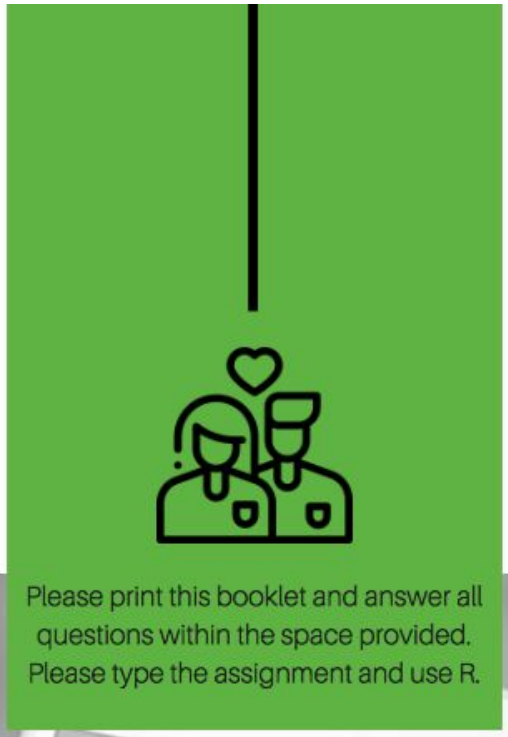
Statistical Foundations of Data Analytics

LAB 6: Classification

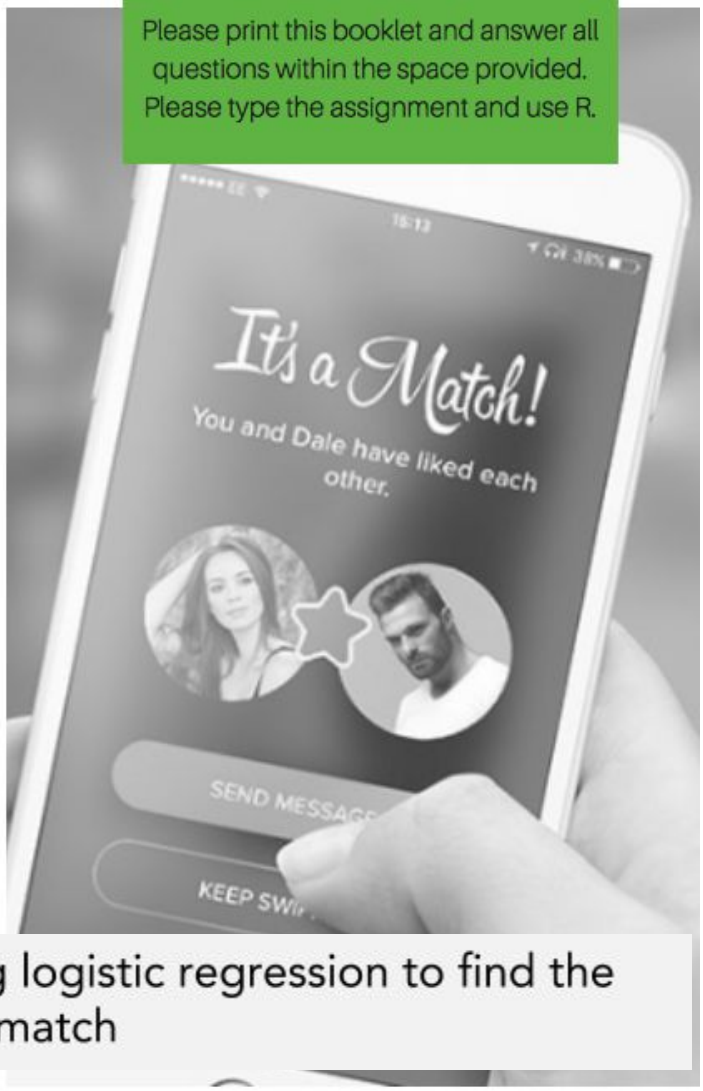


YOUR TEAM

Last name: Toronga
First name: Nicholas
Student ID: 260715831

Last name: Tian
First name: Yuan
Student ID: 260727992

Last name: Park
First name: Jamie
Student ID: 260376390



Part I: Swipe Right or Left? Using logistic regression to find the perfect match

Online dating apps have changed the dynamics of dating in the western world. For one, these apps have made it more efficient to find dates, and meet people you wouldn't have met otherwise. But these apps have also created a consumerist approach to romantic relationships, by promoting the idea that finding a partner is all about efficiency and that dating is "a numbers' game." Dating apps also promote superficiality, by bringing forward someone's looks as the first criterion when forming connections (e.g., Tinder, Bumble). With all its pros and cons, online dating apps continue to grow in popularity, and are harnessing the power of analytics to become more efficient. In fact, one of McGill's comp-sci professors is the head of analytics research for Tinder.

For today's lab, I have gathered two different datasets from users of the top online dating sites. In part I of today's lab, we use data from thousands of "swipes" from hundreds of men and women. Note, the data has been anonymized for privacy purposes, and every subject involved in this dataset gave his/her consent for analysis. In other words, these data weren't scraped, but given away by volunteers.

In this dating app, you're presented with someone's profile (i.e., a picture and a short bio). If you swipe this person's profile **right**, it means you're interested. If you swipe this person's profile **left**, it means you're not interested. If both people swipe right, then they match (you can't see anyone else's swipe, they're secret). In case you're not familiar with how these apps work, here's a link with an explanation of how tinder---the most popular online dating app---works.

In part I of this lab, we will use logistic regression to determine the probability that someone right-swipes a profile, based on the characteristics of the profile and the swiper. We will refer to the person who's swiping as the **swiper**, and to the person who appears on the screen as the **candidate**.

Response Variable

Swipe: 1= Swiped right (i.e., the swiper is interested in the candidate); 0= Swiped left (i.e., the swiper is uninterested in the candidate).

Predictors

Continuous predictors

- Swipe_number: How many swipes has the swiper performed before arriving at this swipe.
- Age_swiper: Age of the person swiping
- Age_candidate: Age of the person who appears on the screen

Categorical predictors

- Gender_swiper: The gender of the person who's swiping. (Note: This sample includes only swipes from heterosexual individuals. In other words, if the swiper is a male, then the "candidate" is a female--and vice versa.)
- Career_swiper: The swiper's career?
- Looking_for_swiper: What the swiper is looking for (friendship, short-term dating, long-term relationship, marriage)?
- Photo_candidate: In the candidate's main picture, what is she/he doing (sports, hiking, etc.)?
- Photo_with_other_people: In the candidate's main picture, does he/she appear with other people?
- Photo_with_sunglasses: In the candidate's main picture, does she/he appear with sunglasses?

1. The linear probability model (5 points)

We will explore the following relationship:

Swipe=f(Age_swiper, Age_Candidate, Swipe_number ,looking_for, Gender_swiper)

A) To begin, we will use a linear probability model. Describe (in your own words) what a linear probability model is.

Your answer:

A linear probability model is a model that predicts the likelihood of a given outcome based on different predictors.

B) Run a linear probability model using the variables above. Paste your regression results below:

Results

	Dependent variable:
	Swipe Right
age_swiper	-0.007***
	(0.002)
age_candidate	0.003**
	(0.002)
swipe_number	-0.003**
	(0.001)
looking_for_swiperlong_term	-0.136***
	(0.033)
looking_for_swipermarriage	-0.093***
	(0.017)
looking_for_swipershort_term	-0.011
	(0.017)
gender_swipermale	0.070***
	(0.012)
Constant	0.551***
	(0.062)
Observations	7,538
R^2	0.019
Adjusted R^2	0.018
Residual Std. Error	0.485 (df = 7530)
F Statistic	20.304*** (df = 7; 7530)
Note:	*p<0.1; **p<0.05; ***p<0.01

C) Interpret the value of the above regression's coefficients:

Age_swiper: -0.007: For every year the swiper ages, the probability of them swiping right decreases by 0.7%.

Age_Candidate: 0.003: For every year the candidate ages, the probability of the swiper swiping right increases by 0.3%.

Swipe_number: -0.003: For every swipe that the swiper makes, the probability that they swipe right decreases by 0.3%.

Looking_for (summarize the interpretation of all dummies): If the swiper is not looking for friendship, the probability of them swiping right decreases.

Gender_swiper: 0.07: If the swiper is male, the probability of them swiping right increases by 7%.

D) What are the two problems of linear probability models? Are any of these problems relevant in your regression? If so, provide an illustrative example of how an *lpm* would be problematic in your regression:

Your answer:

The linear probability model does not account for the fact that probabilities are bounded between 0 and 1. The linear probability model also assumes that the variables affect the predictor in a linear way.

2. Logistic Regression (15 points)

A) Suppose I wish to test the following relationship using logistic regression:

Swipe=f(Age_swiper, Swipe_number)

And that I have four observations in my dataset:

- Observation 1: Swipe=1, Age_swiper=25, Swipe_number=4
- Observation 2: Swipe=0, Age_swiper=18, Swipe_number=9
- Observation 3: Swipe=1, Age_swiper=31, Swipe_number=1
- Observation 4: Swipe=0, Age_swiper=21, Swipe_number=6

What is the value of the likelihood function $L(b_o,b_1,b_2)$ if we let b_o=-8, b_1=0.3,b_2=0.4? Estimate the likelihood manually, and show your work below:

Your work:

```
bo = -8
b1 = 0.3
b2 = 0.4
e = exp(1)
#Observation 1#
p1 = e^(bo+b1*(25)+b2*(4))/(1+e^(bo+b1*(25)+b2*(4)))
p1
#Observation 2#
p2 = e^(bo+b1*(18)+b2*(9))/(1+e^(bo+b1*(18)+b2*(9)))
p2
#Observation 3#
p3 = e^(bo+b1*(31)+b2*(1))/(1+e^(bo+b1*(31)+b2*(1)))
p3
#Observation 4#
p4 = e^(bo+b1*(21)+b2*(6))/(1+e^(bo+b1*(21)+b2*(6)))
p4
#MLE#
MLE = (p1*p3)*((1-p2)*(1-p4))
MLE
```

Your answer:

0.05661003

B) What is the value of the likelihood function $L(b_o,b_1,b_2)$ if we let b_o=-5, b_1=0.3,b_2=0.1? Estimate the likelihood manually, and show your work below:

Your work:

```
bo = -5
b1 = 0.3
b2 = 0.1
e = exp(1)
#Observation 1#
p1 = e^(bo+b1*(25)+b2*(4))/(1+e^(bo+b1*(25)+b2*(4)))
p1
#Observation 2#
p2 = e^(bo+b1*(18)+b2*(9))/(1+e^(bo+b1*(18)+b2*(9)))
p2
#Observation 3#
p3 = e^(bo+b1*(31)+b2*(1))/(1+e^(bo+b1*(31)+b2*(1)))
p3
#Observation 4#
p4 = e^(bo+b1*(21)+b2*(6))/(1+e^(bo+b1*(21)+b2*(6)))
p4
#MLE#
MLE = (p1*p3)*((1-p2)*(1-p4))
MLE
```

Your answer: 0.02609111

C) Which one of the two cases above more closely fits the data? Explain your reasoning:

Your answer:

The first case better describes the data.

Your explanation:

Neither cases fit the data perfectly, however the value of the likelihood function was higher with the first set of coefficients.

D) In your own words, explain how the maximum likelihood estimator finds the optimal coefficients of the logistic regression:

Your answer:

It plugs random numbers as the coefficients to try to find the maximum value. It keeps plugging in random numbers in the direction in which the value of the function is increasing, until it finds the 'peak' or true maximum value.

E) We, again, wish to test the following relationship

```
Swipe~f(Age_swiper, Age_candidate Swipe_number, looking_for, Gender_swiper )
```

This time, however, we wish to use a logistic regression. Run a logistic regression for the model above, and paste the output below:

Results

	Dependent variable:
	Swipe Right
age_swiper	-0.029***
	(0.007)
age_candidate	0.013**
	(0.007)
swipe_number	-0.013**
	(0.006)
looking_for_swiperlong_term	-0.605***
	(0.150)
looking_for_swipermarriage	-0.397***
	(0.072)
looking_for_swipershort_term	-0.048
	(0.070)
gender_swipermale	0.296***
	(0.049)
Constant	0.246
	(0.263)
Observations	7,538
Log Likelihood	-4,997.712
Akaike Inf. Crit.	10,011.420
Note:	*p<0.1; **p<0.05; ***p<0.01

F) How many iterations did the model take to find the coefficients that best fit the data?

Your answer: It took 4 iterations to find the coefficients that best fit the data

G) Predict the probability that, on her eighth swipe, a 25-years old woman who's looking for a long-term relationship will right-swipe a 28-years old man (please paste code and answer):

Your R code:

```
values = data.frame(age_swiper=25,age_candidate=28,
                    swipe_number=8,looking_for_swiper='long_term',
                    gender_swiper='male')
predict(logit,values,type='response')
```

Your answer:

37.31%

H) What is the R-squared value of this logistic regression?

Your answer:

I) Suppose we want to test whether people are more likely to right-swipe people of similar age. Create a variable called absolute_age_difference, which measures the age difference (in absolute value) between the swiper and the candidate.

Create a simple logistic regression that tests the following relationship:

```
Swipe=f(absolute_age_difference )
```

And answer the following question:

1) If the age difference between the swiper and the swiped person decreases from 3 years to 2 years, then the probability of a match increases (increases/decreases) by 0.547 %

2) If the age difference between the swiper and the swiped person decreases from 2 years to 1 year, then the probability of a match increases (increases/decreases) by 0.550%

3) If the age difference between the swiper and the swiped person decreases from 1 year to 0 year, then the probability of a match increases (increases/decreases) by 0.553%

J) Suppose I want to test the following relationship

Swipe=f(Age_swiper, Age_candidate, looking_for, Gender_swiper, absolute_age_difference, photo_candidate)

Run a logistic regression that includes these variable, and include your output below:[1]

[1] You could take interactions of the above variables in the model, to make it more sophisticated. For the purpose of this exercise, don't create any interactions.

Results

	Dependent variable:
	Swipe Right
age_swiper	-0.027***
	(0.007)
age_candidate	0.012*
	(0.007)
looking_for_swiperlong_term	-0.564***
	(0.151)
looking_for_swipermarriage	-0.444***
	(0.073)
looking_for_swipershort_term	-0.059
	(0.071)
gender_swipermale	0.294***
	(0.051)
absolute_age_difference	
photo_candidateclubbing	0.186
	(0.168)
photo_candidateconcert	-0.248*
	(0.131)
photo_candidateeating	-0.389***
	(0.088)
photo_candidatehiking	-0.186
	(0.133)
photo_candidatemuseum	-0.326***
	(0.124)
photo_candidatepaintig/drawing	-0.280*
	(0.146)
photo_candidateplaying_music	-0.532***
	(0.119)
photo_candidatereading	-0.659***
	(0.102)
photo_candidateshopping	-0.582***
	(0.211)
photo_candidatesports	-0.131
	(0.087)
Constant	0.317
	(0.263)
Observations	7,538
Log Likelihood	-4,960.699
Akaike Inf. Crit.	9,955.398
Note:	*p<0.1; **p<0.05; ***p<0.01

K) Using the output above, tell me what's the probability that there will be a match between (i) a 30-years old woman who's looking for a long-term relationship and has a picture where she's eating and (ii) a 28-years old man who's looking for a short_term relationship and has a picture where he's in a museum. For a match to happen, both a man and a woman need to right-swipe each other. H

int: You will need to use your knowledge about probability (from your intro stats course) to figure this one out. You may assume that neither the men or the women know each other's swipe, and that all swipes are independent. Show your work.

Your answer: 11.47%

Your explanation:

Based on the model, the probability that the woman swipes right on the man is 25.97%, and the probability that the man swipes right on the woman is 44.2%. So, to find the probability that they match (that they both swipe right on each other), we multiplied the two probabilities together.

Part II: What are they looking for? A discriminant analysis

As we saw previously, people in online dating apps are looking for different things: just meeting friends, short-term relationships, serious relationships, or even marriage. Most often, people don't reveal their preferences, and this might create ambiguity. And if you are looking for a serious relationship, you don't want to spend your time talking with someone who just wants to use the app for a short-term relationship, or someone who's just looking for friends.

So you came with an idea! You're going to create an app called "What are they looking for?" In this app, you'll use the characteristics of Tinder profiles, and their profile pictures, to predict the intensions of the person. Your goal is to get information about a given profile, and classify the person by her/his intentions: (i) looking for just friendship, (ii) looking for short-term dating, (iii) looking for a long-term relationship, or (iv) looking to marry.

This app will ultimately allow people to input a given tinder profile; then, the app will predict the intentions of the person's profile (by probability). (Note: this app was actually planned and developed, but never came to the market. At the end, we will discuss the ethical implications behind this app.)

To build this predictive model, you have gathered a dataset with thousands profiles (looking_for.csv). This dataset doesn't contain swipe data, only data about the profile itself. For each profile, you have the following variables:

Response Variable

Looking_for: (i) just_friendship, (ii) short_term_dating, (ii) long_term_dating, (iv) marriage.

- **Gender_profile:** male or female
- **Age_profile:** Age of the person who appears on the screen
- **Hours_week:** How many hours per week is the person spending on the app (on average).
- **Photo_profile:** In the candidate's main picture, what is she/he doing (sports, hiking, etc)?
- **Photo_with_other_people:** Are there any other people in the person's main profile picture?
- **Photo_with_sun_glasses:** Is the person in the profile using sunglasses?

3. Linear discriminant analysis with two predictors (15 points)

We want to understand the probability that profile *i* has Y={friendship,short_term,long_term, marriage}, given the prifle's age and the person's engagement in the app:

- Pr(Y=k | age, hours_week)

A) What are the prior probabilities of each class?

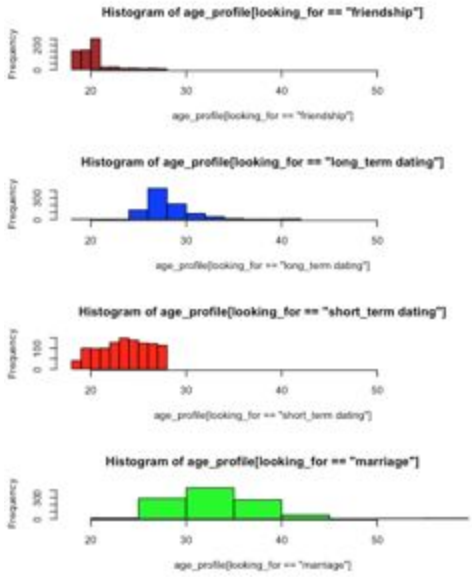
$\pi_{friendship} = 0.1746117$

$\pi_{short} = 0.2437065$

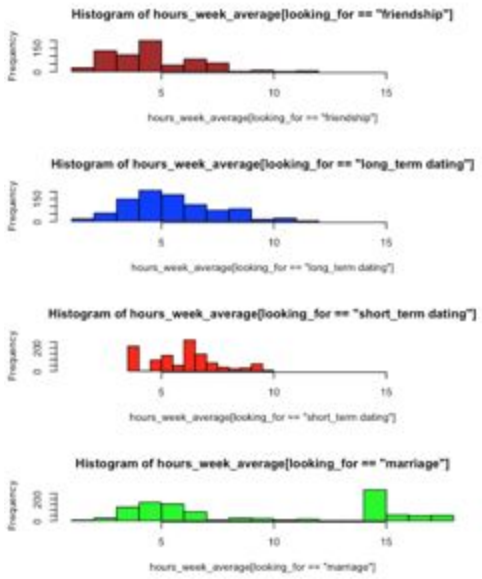
$\pi_{long} = 0.2836101$

$\pi_{marriage} = 0.2980718$

B) We want to find the probability density functions f_k(age) for k={friendship, short-term, long-term, marriage}. Plot the histograms of age level for each class. Make sure all histograms have the same range on the x axis (please ensure all axis information is legible):



C) We want to find the functions f_k(hours_week). Plot the histograms of hours_week level for each class. Make sure all histograms have the same range along the x axis:



D) Is it reasonable to assume that all $f_k()$ functions are normally distributed?

Your answer:

Yes, it is reasonable to make that assumption. From the two plots for the predictors age and hours/week, we can see that more than half of the plots exhibit a somewhat normal distribution.

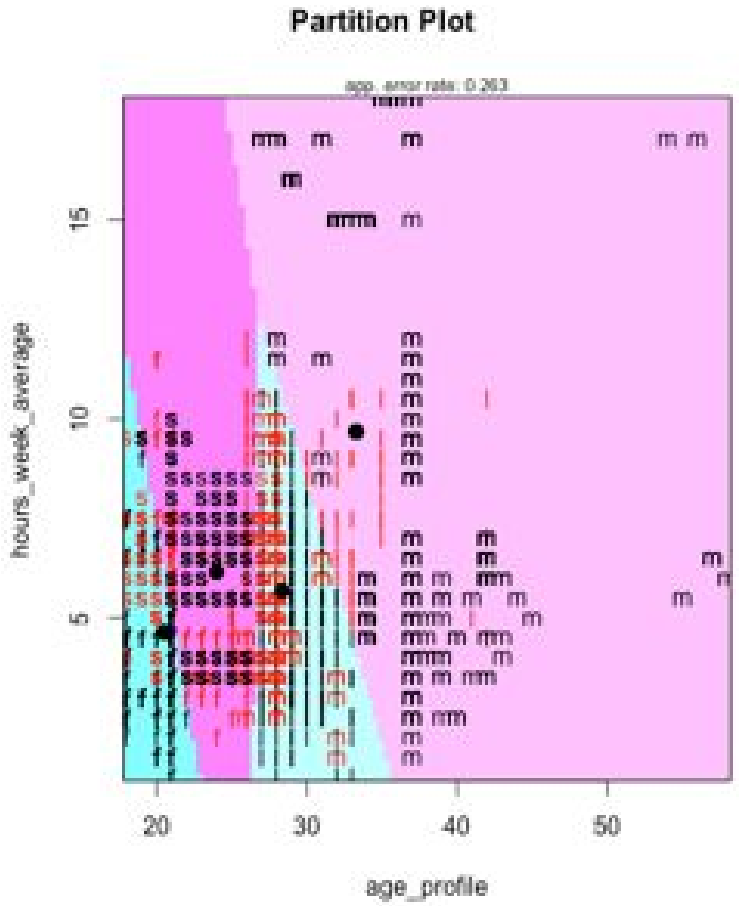
E) Run a linear discriminant analysis to find

$$Pr(Y=k| \text{ age, hours\_week})$$

Paste the lda's output below:

```
> model_lda
Call:
lda(looking_for ~ age_profile + hours_week_average)

Prior probabilities of groups:
       friendship  long_term dating          marriage short_term dating
        0.1746117         0.2437065         0.2836101         0.2980718

Group means:
                  age_profile hours_week_average
friendship           20.52301           4.642638
long_term dating     28.35495           5.682418
marriage             33.28990           9.680831
short_term dating    23.98473           6.153639

Coefficients of linear discriminants:
                         LD1        LD2
age_profile        0.3086571 -0.1083066
hours_week_average 0.1237855  0.2911136

Proportion of trace:
   LD1    LD2
0.9757 0.0243
```

F) Plot the classification regions using the *partimat* function:



G) Interpret the results above from the partimat() function. What would you tell to a manager of a dating app based on your analysis above? No jargon:

Your answer:

The app can correctly classify whether someone aged between 18 and 58 is looking for a friendship, short-term relationship, longterm relationship or marriage given the time they spend on the app. The app is performing well; it can accurately classify about 74% of any given dataset.

H) What is the error rate of the linear discriminant model? What does this mean?:

Your answer:

Your answer:

The error rate is 26.3%. It means that the model can inaccurately classify 26.3% of the observations in a dataset

l) Suppose a profile has age=26 and hours_week=5. What is the probability that this profile will belong to each class (i.e., looking for friendship, short-term dating, serious dating, marriage)?

```
$class
[1] short_term dating
Levels: friendship long_term dating marriage short_term dating

$posterior
  friendship long_term dating    marriage short_term dating
1 0.07787845        0.3962863 0.009946766         0.5158885
```

4. Quadratic discriminant analysis with two predictors (7 points)

A) What is the difference between the linear and the quadratic discriminant analysis?

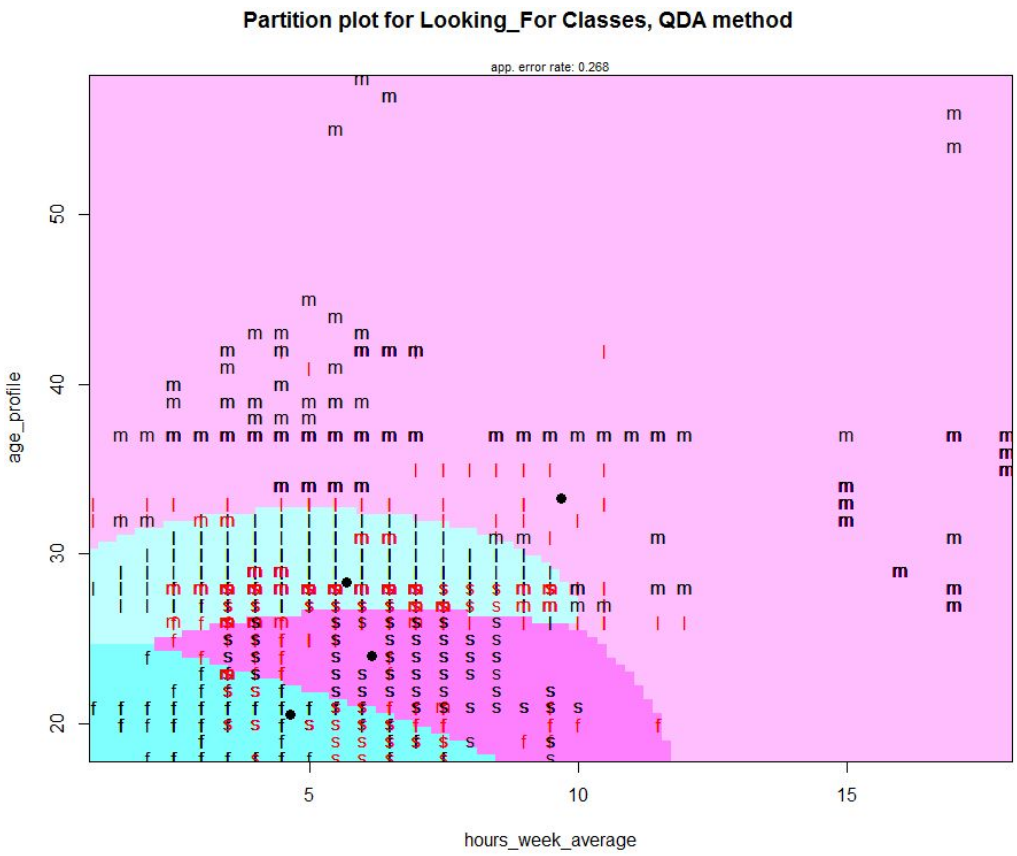
Discriminant analysis is used for predicting results when the dependent variable is categorical. In LDA, all classes in the dependent variables have the same covariance whereas in QDA each class can have its own unique covariance. Thus, the dependent variables have widely varying standard deviations, QDA allows for a more accurate representation of the distribution.

B) Run a quadratic discriminant analysis to find

$$Pr(Y=k|age, hours\_week)$$

```
> myqda = qda(looking_for~age_profile+hours_week_average, data = looking_for_data)
> myqda
Call:
qda(looking_for ~ age_profile + hours_week_average, data = looking_for_data)

Prior probabilities of groups:
       friendship  long_term dating          marriage short_term dating
        0.1746117         0.2437065         0.2836101         0.2980718

Group means:
                  age_profile hours_week_average
friendship           3.523006           8.285276
long_term dating    11.354945          10.364835
marriage            16.244570          16.033994
short_term dating    6.984726          11.307278
```

C) Plot the classification regions using the partimat() function:



D) What is the error rate of the linear discriminant model?:

0.263

E) Suppose a profile has age=26 and hours_week=5. What is the probability that this profile will belong to each class (i.e., looking for friendship, short-term dating, long-term dating, marriage)?

```
> predict(myqda, data.frame(age_profile=26, hours_week_average=5))
$`class`
[1] short_term dating
Levels: friendship long_term dating marriage short_term dating

$posterior
  friendship long_term dating   marriage short_term dating
1 0.01007554        0.3898064 0.02504612          0.575072
```

Friendship: 10.07%

Long-term dating: 38.98%

Marriage: 2.50%

Short-term dating: 57.51%

F) Which model performs better in the training data: the linear or the quadratic model? How did you reach that conclusion?

As shown in the respective partition plots, QDA has a higher error of 0.268 than the LDA model's 0.263 thus LDA performs better.

5. Discriminant analysis with all predictors (3 points)

A) Run a linear discriminant model including all continuous and categorical predictors. Paste the output below:

```
Call:
lda(looking_for ~ gender_profile + age_profile + hours_week_average +
    photo_profile + photo_with_other_people + photo_with_sun_glasses,
    data = looking_for_data)

Prior probabilities of groups:
       friendship  long_term dating          marriage short_term dating
        0.1746117         0.2437065         0.2836101         0.2980718

Group means:
                  gender_profilemale age_profile hours_week_average
friendship                 0.3481595    20.52301           4.642638
long_term dating           0.4131868    28.35495           5.682418
marriage                   0.3994334    33.28990           9.680831
short_term dating          0.5561545    23.98473           6.153639
                  photo_profileclubbing photo_profileconcert
friendship                 0.013803681           0.06441718
long_term dating           0.026373626           0.02747253
marriage                   0.009442871           0.04060434
short_term dating          0.016172507           0.04761905
                  photo_profileeating photo_profilehiking
friendship                 0.2453988          0.04754601
long_term dating           0.1934066          0.06703297
marriage                   0.2105760          0.05193579
short_term dating          0.2084456          0.03953279
                  photo_profilemuseum photo_profilepaintig/drawing
friendship                 0.05061350                   0.03834356
long_term dating           0.04725275                   0.00989011
marriage                   0.10292729                   0.04910293
short_term dating          0.03863432                   0.02964960
```

```
                 photo_profileplaying_music photo_profilereading
friendship                       0.07668712           0.10889571
long_term dating                 0.09450549           0.14835165
marriage                         0.03871577           0.09065156
short_term dating                0.06199461           0.12309075
                 photo_profileshopping photo_profilesports
friendship                  0.01380368           0.2223926
long_term dating            0.01208791           0.2219780
marriage                    0.01605288           0.2115203
short_term dating           0.01437556           0.3117700
                 photo_with_other_people photo_with_sun_glasses
friendship                    0.59815951              0.5444785
long_term dating              0.42307692              0.5219780
marriage                      0.07365439              0.4929178
short_term dating             0.51392633              0.6010782

Coefficients of linear discriminants:
                                    LD1        LD2         LD3
gender_profilemale            0.16424571  0.1166080  1.57588009
age_profile                   0.30205735 -0.1161148  0.03087552
hours_week_average            0.11741071  0.2305180  0.03057394
photo_profileclubbing         0.08527249 -0.9452692  1.09835011
photo_profileconcert         -0.12248939  0.5279140 -0.22391388
photo_profileeating          -0.03620406  0.1258011  0.11088169
photo_profilehiking          -0.31074176 -0.4090344  0.08990419
photo_profilemuseum          -0.04200963  0.7007511 -0.45385187
photo_profilepaintig/drawing -0.18216981  1.2424185 -0.73353811
photo_profileplaying_music   -0.19194430 -0.8490119 -0.09379563
photo_profilereading         -0.28370919 -0.3507348  0.50462039
photo_profileshopping         0.00744853  0.1156164  0.51839919
photo_profilesports          -0.24934779  0.2793341  0.72647716
photo_with_other_people      -0.62447450 -0.6194681  0.60114659
photo_with_sun_glasses       -0.16518058  0.2144354  0.58780927

Proportion of trace:
   LD1    LD2    LD3
0.9590 0.0297 0.0113
```

B) What is the in-sample error rate for this model?

Your answer: in-sample error is 25.36%

	Predicted Group			
Actual Group	friendship	long_term dating	marriage	short_term dating
friendship	461	9	0	182
long_term dating	0	627	72	211
marriage	3	191	837	28
short_term dating	60	190	1	862

C) Run a quadratic discriminant model by including all predictors. What is the error rate for this model in the training data?

Your answer: the error rate for this model is 26.27%

	Predicted Group			
Actual Group	friendship	long_term dating	marriage	short_term dating
friendship	502	31	1	118
long_term dating	4	682	91	133
marriage	2	193	836	28
short_term dating	101	239	40	733

D) As I told you, this type of app was actually created. In this exercise, I removed all physical traits and race about the profile, but the actual app incorporated these features to classify profiles. In other words, the app looked at a person's skin colour, facial features, job, pose, and profile bio to predict her/his intentions. This app, however, never came to fruition. There was too much criticism about the ethical implications behind it. What do you think is morally questionable about an app that uses your profile picture and bio to predict your dating intentions?

-The privacy of users is harmed. The private data is highly likely tracked by some groups of people. These data could perhaps be used to detect unusual activities.

-By imposing a "likelihood" that you will match with a specific other type of person, you force bias onto that person, when in reality preference for a significant other is subjective and completely up to the person's own preferences.

-The technology that has become central to modern dating has stripped away bits and pieces of people's humanity.

Note: Big Data an with a dark side. I would highly recommend you to read the book "Weapons of Math Destruction," which discusses the ethical issues behind machine-learning algorithms that are used to profile people. Also, if you want to see an interesting documentary on online-dating algorithms, I recommend you to watch HBO's "Swiped: The Dark Side of Online Dating."