What attributes influence the selection of a romantic partner?

Juanhong Cheng	W1174938
Sheng-Fu Chuang	W1189934

Table of Content

Introduction

Objective

We want to know what factors will affect we select romantic partner. We use the survey from Columbia University about speed dating. And we try to use regression form to interpret the data set.

What is the problem

We want to find out what attributes influence the selection of a romantic partner. We want to use the data set[1] to find the facts as following:

1. What are the least desirable attributes in a male partner? Does this differ for female partners?
2. How important do people think attractiveness is in potential mate selection vs. its real impact?
3. Are shared interests more important than a shared racial background?
4. In terms of getting a second date, is it better to be someone's first speed date of the night or their last?

Why this is a project related to this class

Data mining include the following technologies[2]:

1. Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
2. Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
3. Regression – attempts to find a function which models the data with the least error.
4. Summarization – providing a more compact representation of the data set, including visualization and report generation.

We want to utilize the technique of regression to find functions which models the data with the least error. And we can find out some interesting facts from these function’s coefficient.

why other approach is no good

Other approach just concern about positive attribute but not negative attributes.

why you think your approach is better

Our approach concern about both positive and negative attributes.

statement of the problem

what attributes influence the selection of a romantic partner and how important that attribute for the selection.

Area or scope of investigation

Romantic selection, speed dating, racial preference, gender difference

Theoretical bases and literature review

definition of the problem

Analyzing if particular factors related to the decision of selection and how they influence the selection

theoretical background of the problem

Clustering,Classification,Regression,Summarization

related research to solve the problem

Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment.
Racial Preferences in Dating

advantage/disadvantage of those research

Advantage: shows the fact that gender and racial is dummy variables to do the linear regression
Disadvantage: do not show enough comparison between important factors

Our solution to solve this problem

1. What are the least desirable attributes in a male partner? Does this differ for female partners?

$$Decision_{ij} = \alpha_i + \sum_{c \in C} \beta_c Rating_{ijc} + \varepsilon_{ij}$$

2. How important do people think attractiveness is in potential mate selection vs. its real impact?
   $Decision_{ij} = \alpha_i + \beta_0(Attra_{ij} - SelfExpectAttra_i)$
   $+ \beta_1(Attra_{ij} - SelfExpectAttra_i)(Attra_{ij} < SelfExpectAttra_i) + \varepsilon_{ij}$
   Observe β_1 is positive or negative and it’s value to determine its real impact
3. Are shared interests more important than a shared racial background?
   $Decision_{ij} = \alpha_i + \beta_0 SameRace_{ij} + \beta_1 SameInterest_{ij} + \varepsilon_{ij}$

4. In terms of getting a second date, is it better to be someone's first speed date of the night or their last?
   $Decision_{ij} = \alpha_i + \beta_0 FirstDate_{ij} + \beta_1 LastDate_{ij} + \varepsilon_{ij}$

Where your solution different from others

Our solution take attributes and their different level of importance into account.And the solution not only focus on the levels of various attributes influence the probability of speed dating, but also focus on the difference between the male and the female.

Why your solution is better

Our solution combines subjective attributes and objective attributes together to find which is the most important factor influence success rate of speed dating.

Hypothesis

1. Male and female have different negative attributes.
2. The expectation is related the decision.
3. Shared racial background is more important than shared interest.
4. The order of speed dating really influence, the last partner of speed dating has better feedback than the first.

Methodology

How to collect/generate input data

The data we used is based on an experiment conducted by Columbia University [1]. The researcher hold multiple Speed Dating, in which each participant will fill out a survey before and after the event. Our data is collect from these survey. To make sure each session of speed dating is consistent, the researcher control the light and music to be identical.

How to solve the problem

Algorithm design

We want to propose several regression form which can shows us the result we want.

Language used

We will use python to calculate the multiple regression

Tool used

We will like to use the package of statsmodels, matplotlib and numpy.

How to generate output

We will write a program to read data from csv file and then use statsmodels to generate the output coefficient of each attribute.

How to test against hypothesis

We can observe the the coefficient difference to determine the result. Our goal is to find the special facts from speed dating data set.

How to proof correctness

We may try to reproduce the result of paper [1] to make sure we know and understand how to write the program. And then we will have more confidence of the result from new regression form.

Implementation

code (refer programming requirements)

Python library: statsmodels, numpy, pandas

design document and flowchart

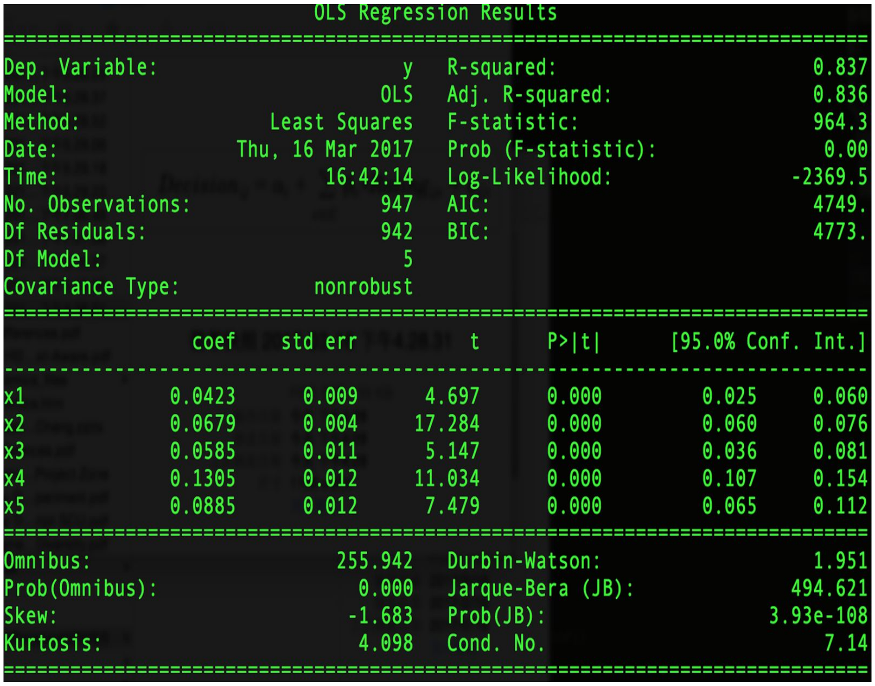
Data acquisition → Data transforming → Data cleansing → Choosing statistic model → regression

Data analysis and discussion

Output generation

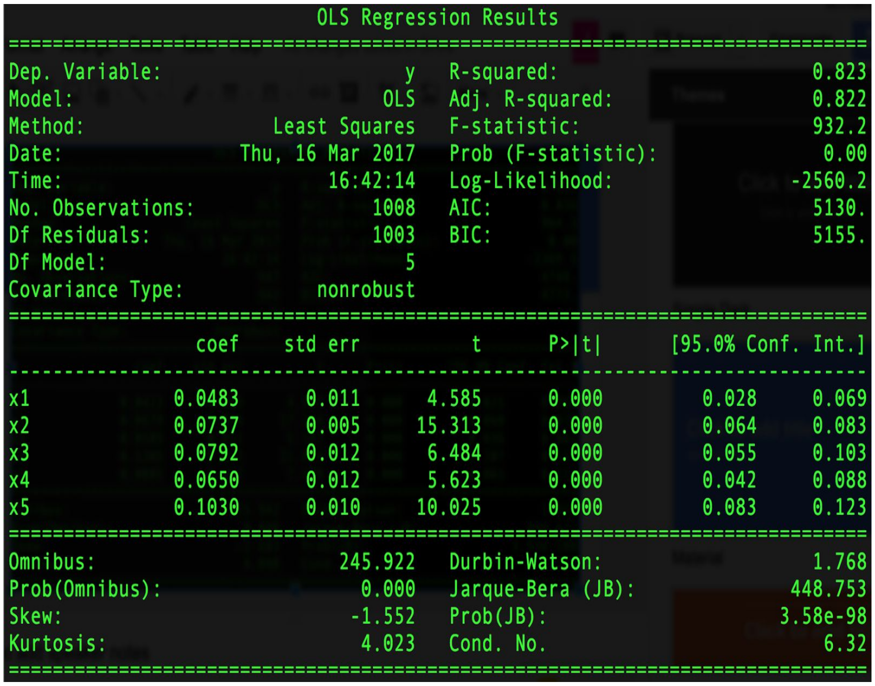
The attributes includes attractive, sincere, Intelligent, fun and shared interests/hobbies. Initially we ask people to fill out the survey and the score of each attribute should be added up to 100

Hypothesis 1: Male and female have different negative attributes.
Our result of this hypothesis is true.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.837
Model:                            OLS   Adj. R-squared:                  0.836
Method:                 Least Squares   F-statistic:                     964.3
Date:                Thu, 16 Mar 2017   Prob (F-statistic):               0.00
Time:                        16:42:14   Log-Likelihood:                -2369.5
No. Observations:                 947   AIC:                             4749.
Df Residuals:                     942   BIC:                             4773.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.0423      0.009      4.697      0.000         0.025     0.060
x2             0.0679      0.004     17.284      0.000         0.060     0.076
x3             0.0585      0.011      5.147      0.000         0.036     0.081
x4             0.1305      0.012     11.034      0.000         0.107     0.154
x5             0.0885      0.012      7.479      0.000         0.065     0.112
==============================================================================
Omnibus:                      255.942   Durbin-Watson:                   1.951
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              494.621
Skew:                          -1.683   Prob(JB):                    3.93e-108
Kurtosis:                       4.098   Cond. No.                         7.14
==============================================================================
```

male linear regression



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.823
Model:                            OLS   Adj. R-squared:                  0.822
Method:                 Least Squares   F-statistic:                     932.2
Date:                Thu, 16 Mar 2017   Prob (F-statistic):               0.00
Time:                        16:42:14   Log-Likelihood:                -2560.2
No. Observations:                1008   AIC:                             5130.
Df Residuals:                    1003   BIC:                             5155.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.0483      0.011      4.585      0.000         0.028     0.069
x2             0.0737      0.005     15.313      0.000         0.064     0.083
x3             0.0792      0.012      6.484      0.000         0.055     0.103
x4             0.0650      0.012      5.623      0.000         0.042     0.088
x5             0.1030      0.010     10.025      0.000         0.083     0.123
==============================================================================
Omnibus:                      245.922   Durbin-Watson:                   1.768
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              448.753
Skew:                          -1.552   Prob(JB):                     3.58e-98
Kurtosis:                       4.023   Cond. No.                         6.32
==============================================================================
```

female linear regression

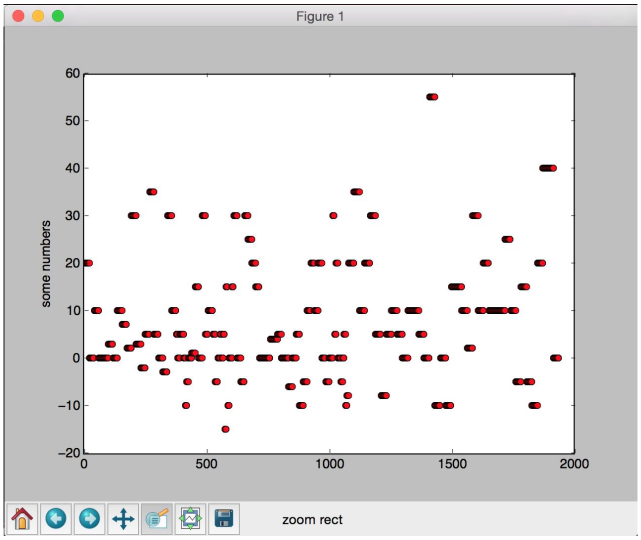
Result :

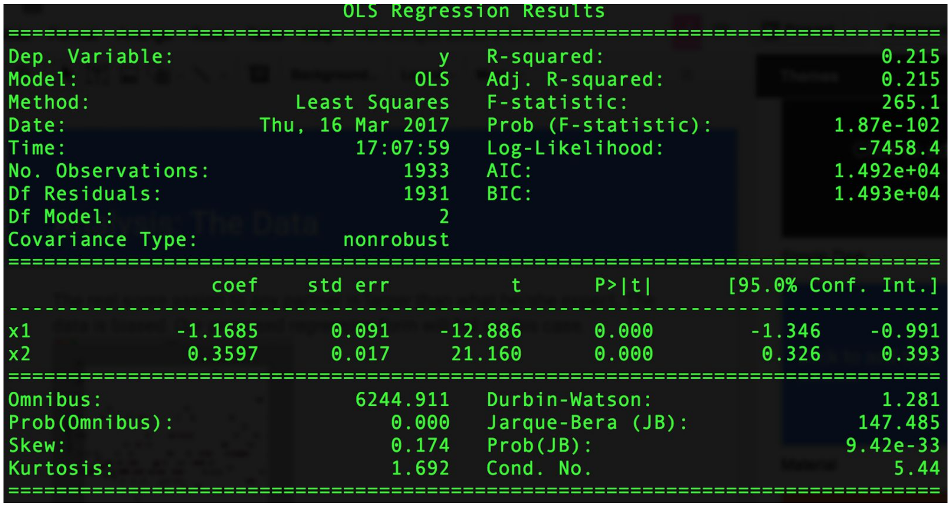
	Male	Female
Least desirable	sincere	share interest/hobbit
	share interest/hobbit	intelligent
	attractive	attractive
	intelligent	sincere
Most desirable	fun	fun

Hypothesis 2: The expectation is related the decision.

What we found out is that most of the real score assigned to any partner is larger than what he/she expect. The data is biased. Our proposed regression form will fail on this case.

The below is the graph of real score minus the expectation. As we can see, most of it is larger than zero.





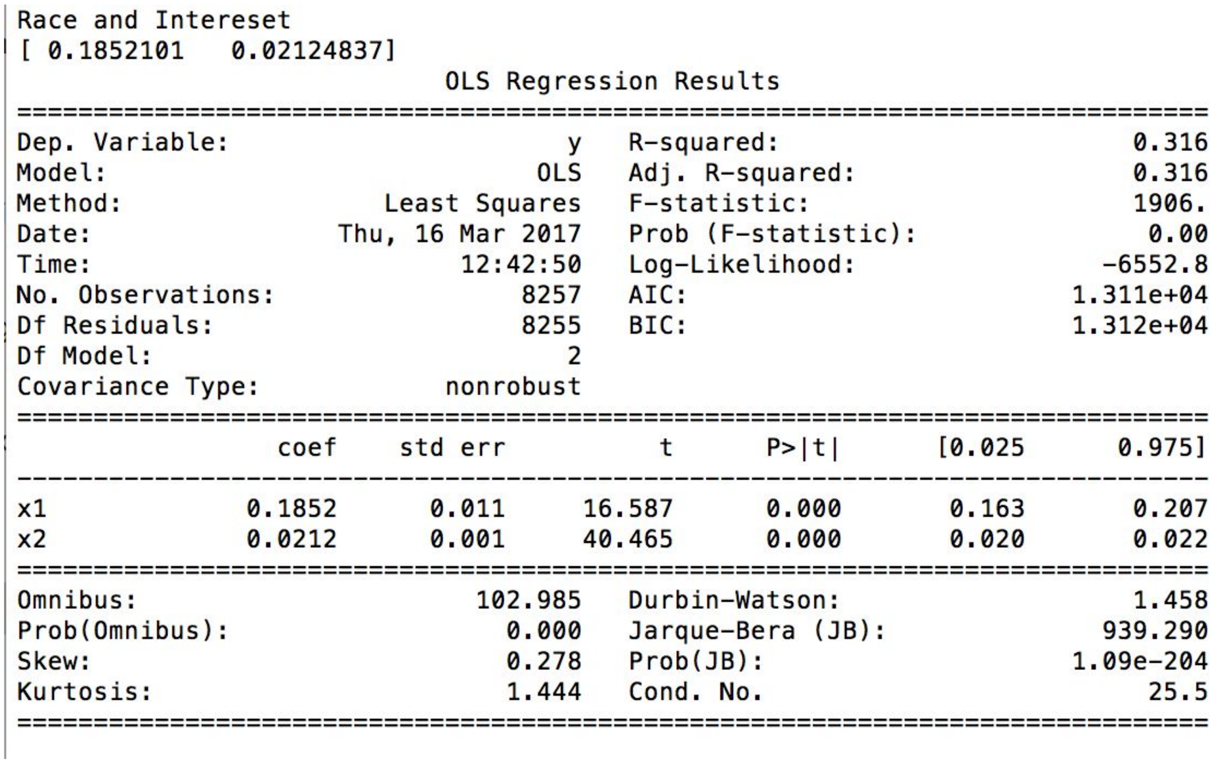
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.215
Model:                            OLS   Adj. R-squared:                  0.215
Method:                 Least Squares   F-statistic:                     265.1
Date:                Thu, 16 Mar 2017   Prob (F-statistic):          1.87e-102
Time:                        17:07:59   Log-Likelihood:                -7458.4
No. Observations:                1933   AIC:                         1.492e+04
Df Residuals:                    1931   BIC:                         1.493e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1            -1.1685      0.091    -12.886      0.000        -1.346    -0.991
x2             0.3597      0.017     21.160      0.000         0.326     0.393
==============================================================================
Omnibus:                     6244.911   Durbin-Watson:                   1.281
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              147.485
Skew:                           0.174   Prob(JB):                     9.42e-33
Kurtosis:                       1.692   Cond. No.                         5.44
==============================================================================
```

The linear regression result

shared race & shared interest

β_0 = 0.1852101

β_1 = 0.02124837

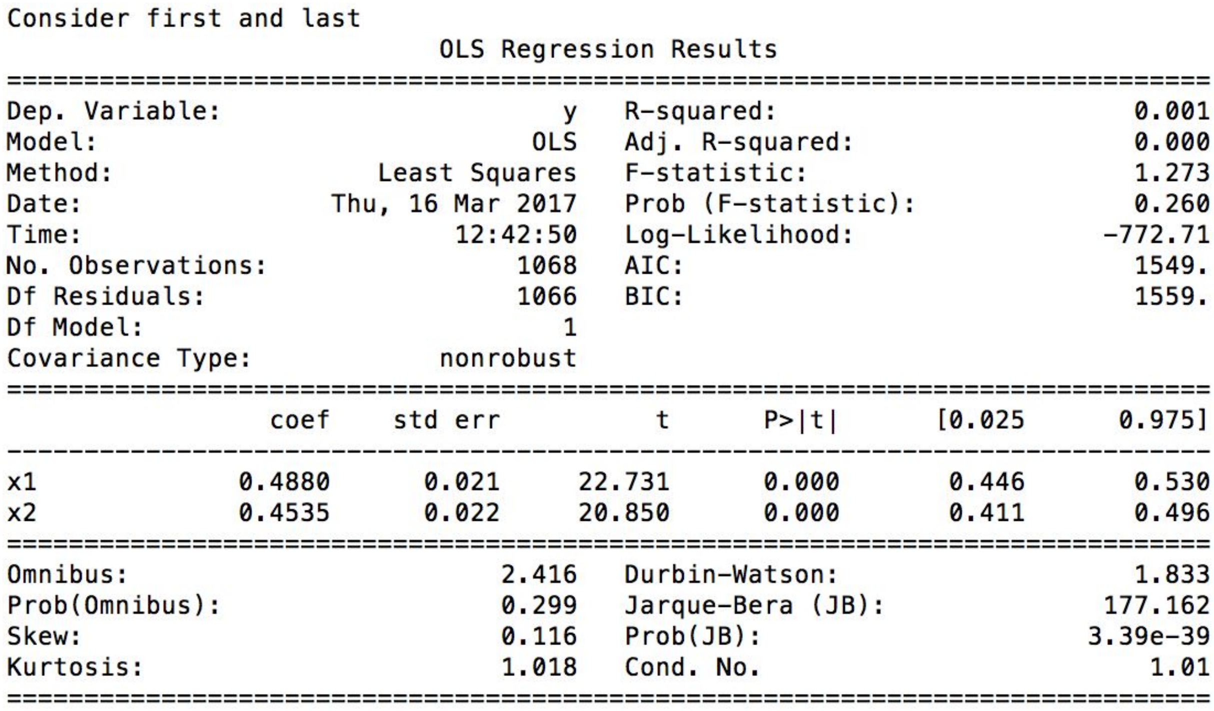


```
Race and Intereset
[ 0.1852101   0.02124837]
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.316
Model:                            OLS   Adj. R-squared:                  0.316
Method:                 Least Squares   F-statistic:                     1906.
Date:                Thu, 16 Mar 2017   Prob (F-statistic):               0.00
Time:                        12:42:50   Log-Likelihood:                -6552.8
No. Observations:                8257   AIC:                         1.311e+04
Df Residuals:                    8255   BIC:                         1.312e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.1852      0.011     16.587      0.000       0.163       0.207
x2             0.0212      0.001     40.465      0.000       0.020       0.022
==============================================================================
Omnibus:                      102.985   Durbin-Watson:                   1.458
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              939.290
Skew:                           0.278   Prob(JB):                    1.09e-204
Kurtosis:                       1.444   Cond. No.                         25.5
==============================================================================
```

First Date and last Date

$\beta_0 = 0.4880$

$\beta_1 = 0.4535$



```
Consider first and last
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     1.273
Date:                Thu, 16 Mar 2017   Prob (F-statistic):              0.260
Time:                        12:42:50   Log-Likelihood:                -772.71
No. Observations:                1068   AIC:                             1549.
Df Residuals:                    1066   BIC:                             1559.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.4880      0.021     22.731      0.000       0.446       0.530
x2             0.4535      0.022     20.850      0.000       0.411       0.496
==============================================================================
Omnibus:                        2.416   Durbin-Watson:                   1.833
Prob(Omnibus):                  0.299   Jarque-Bera (JB):              177.162
Skew:                           0.116   Prob(JB):                     3.39e-39
Kurtosis:                       1.018   Cond. No.                         1.01
==============================================================================
```

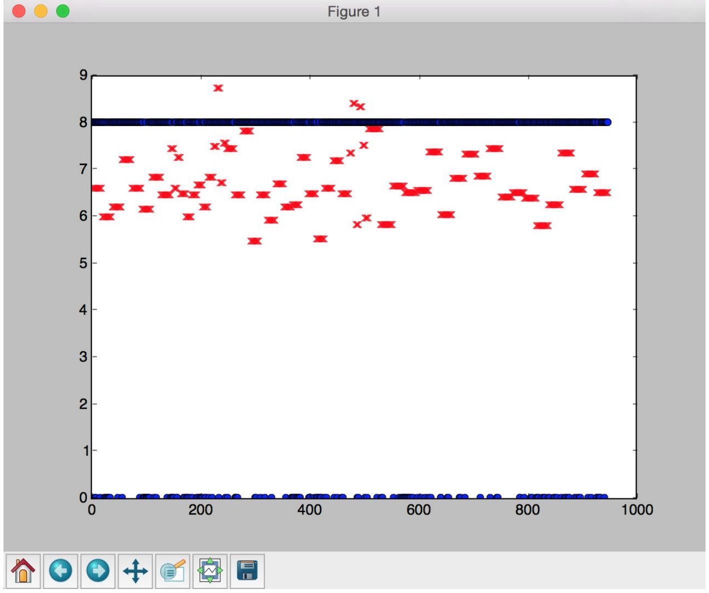
Output analysis

Hypothesis 1: Male and female have different negative attributes.
Our result of this hypothesis is true.

Result :

	Male	Female
Least desirable	sincere	share interest/hobbit
	share interest/hobbit	intelligent
	attractive	attractive
	intelligent	sincere
Most desirable	fun	fun

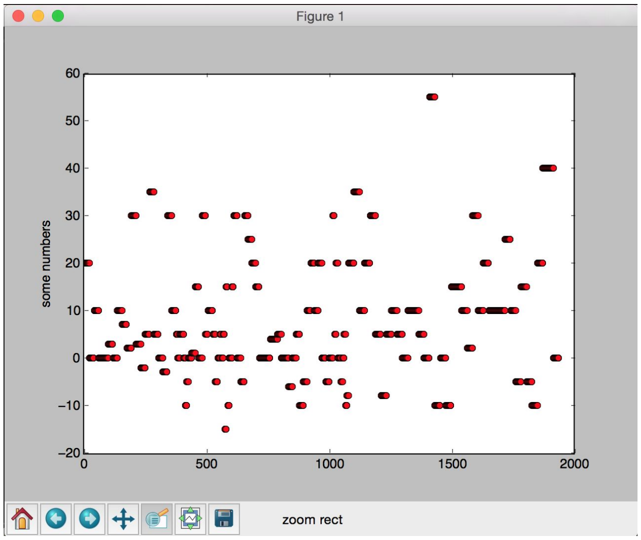
We use linear model try to fit 2 point value of decision. Although we have 82% R-square, it looks like we do not fit the decision. The red dot is our prediction and blue dot is the real value.



Hypothesis 2: The expectation is related the decision.

What we found out is that most of the real score assigned to any partner is larger than what he/she expect. The data is biased. Our proposed regression form will fail on this case.

The below is the graph of real score minus the expectation. As we can see, most of it is larger than zero.



shared race & shared interest

R - square is 0.316 and the difference between two coefficient is really big
The β_0 and β_1 worth taking into account when consider the decision of speed dating.
It shows that people prefer partner with the same race than the same interest

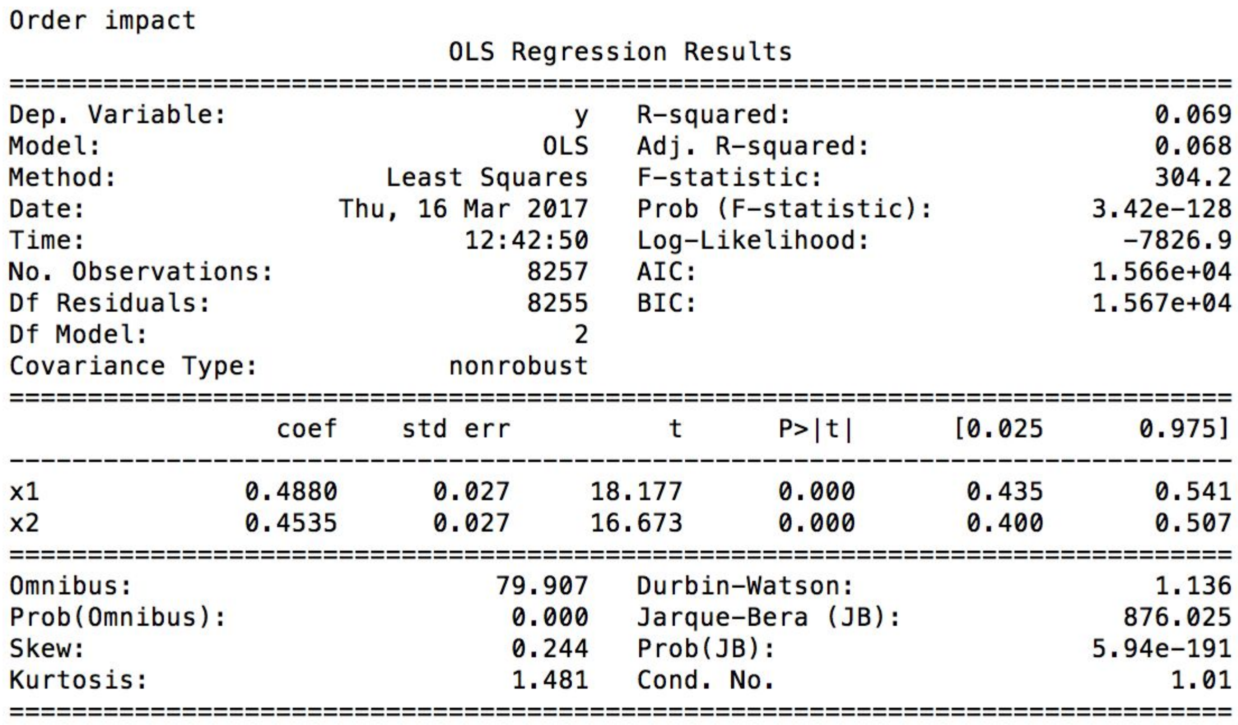
First Date and last Date

R - square close to 0 thus the 2 variable might not related to the decision. In order to analysis if the order really impact decision. The further analysis is about related order.

$$Decision_{ij} = \alpha_i + \beta_0 RelatedtFirstDistance_{ij} + \beta_1 RelatedLastDistance_{ij} + \varepsilon_{ij}$$

RelatedFirstDistance = (Round - Order)/(Round - 1)

RelatedLastDiatance = (Round - 1) /(Round - 1)



```
Order impact
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.069
Model:                            OLS   Adj. R-squared:                  0.068
Method:                 Least Squares   F-statistic:                     304.2
Date:                Thu, 16 Mar 2017   Prob (F-statistic):          3.42e-128
Time:                        12:42:50   Log-Likelihood:                -7826.9
No. Observations:                8257   AIC:                         1.566e+04
Df Residuals:                    8255   BIC:                         1.567e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.4880      0.027     18.177      0.000       0.435       0.541
x2             0.4535      0.027     16.673      0.000       0.400       0.507
==============================================================================
Omnibus:                       79.907   Durbin-Watson:                   1.136
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              876.025
Skew:                           0.244   Prob(JB):                    5.94e-191
Kurtosis:                       1.481   Cond. No.                         1.01
==============================================================================
```

The coefficient is close and R-square is really low. Thus, The order will not influence the decision

Compare output against hypothesis

shared race & shared interest

Hypothesis about comparison of shared rance and shared interest is true. People from dataset prefer partners with the same race more than the same interest

Order

Unlike hypothesis about order. The output shows order does not matters.

Abnormal case explanation (the most important task)

Use race and gender as dummy variables, Finding that Asian female do not prefer racial background like other group.

Statistic regression

Discussion

About order

The order might work as dummy variables. After find a reliable model of the decision of speed dating. It might worthy that combine order as dummy variable to do the regression.

Conclusions and recommendations

Summary and conclusions

People prefer partners with the same race.
The order of dating in round does not matter.

Recommendations for future studies

Consider more factors like survey time and position as dummy variables

bibliography

[1] RAYMOND FISMAN, SHEENA S. IYENGAR, EMIR KAMENICA, and ITAMAR SIMONSON. "GENDER DIFFERENCES IN MATE SELECTION: EVIDENCE FROM A SPEED DATING EXPERIMENT" *The Quarterly Journal of Economics, May 2006*

[2] https://en.wikipedia.org/wiki/Data_mining

[3] https://www.kaggle.com/annavictoria/speed-dating-experiment

DataSet in Folder

appendices

program flowchart

Data acquisition → Data transforming → Data cleansing → Choosing statistic model → regression

program source code with documentation

See in Folder

input/output listing

Inputting : survey about data mining. The attributes we use are gender, order, round , samerace, dec_o, wave, attr1_1, sinc1_1, intel1_1, fun1_1, amb1_1, shar1_1

Outputting: coefficient of each attributes

other related material

See in Folder

Reference

[1] RAYMOND FISMAN, SHEENA S. IYENGAR, EMIR KAMENICA, and ITAMAR SIMONSON. "GENDER DIFFERENCES IN MATE SELECTION: EVIDENCE FROM A SPEED DATING EXPERIMENT" *The Quarterly Journal of Economics, May 2006*

[2] https://en.wikipedia.org/wiki/Data_mining

[3] https://www.kaggle.com/annavictoria/speed-dating-experiment