Speed Dating Analysis

Lauren Hanlon

Introduction

My public GitHub repo including all code, images and data is here

For this analysis I will refer to a "speed dating" experiment conducted in 2006. You can find the data here.

More background on the experiment:

> "In the experiment, a few hundred students in several waves were randomly assigned to 10 short dates. At the end of each 4-minute date each participant filled out a scorecard that asked them to decide whether or not they would like to see the other person again (1=yes or 0=no). The participants were also asked to record subjective numerical ratings for the other person (attractive, sincere, intelligent, fun, ambitions, has shared interests/hobbies). A pre-interview and a follow-up survey also captured data about participants' backgrounds, preferences, and some other characteristics."

In particular, I am answering the following questions:

1. Discuss and quantify the importance, in general, of each person a's 6 ratings of person b (attractiveness, sincerity, intelligence, and so forth) in determining if person a is interested in seeing person b again.
2. Some people are more likely to want to see other people again. Discuss the implications for the general conclusions you reached in question 1 and see if you can account and correct for this kind of underlying variation in the person making the evaluation.
3. With some people, it is more likely that others will want to see them again. Is there evidence in the data for this kind of variation in the person being rated? If so, how can you account and correct for it and what are the implications for your conclusions from 1 & 2?

Data

The speed dating data set is 4435 x 195 in dimensions. There are 4435 rows, each row being a unique speed date and some of the key columns I will be working with in this analysis include:

- `dec`: Decision (1=Yes 0=No) to see the person again
- `attr`: Attractiveness rating
- `sinc`: Sincerity rating
- `intel`: Intelligence rating
- `fun`: Fun rating
- `amb`: Ambition rating
- `shar`: Shared interests/hobbies rating

Question 1

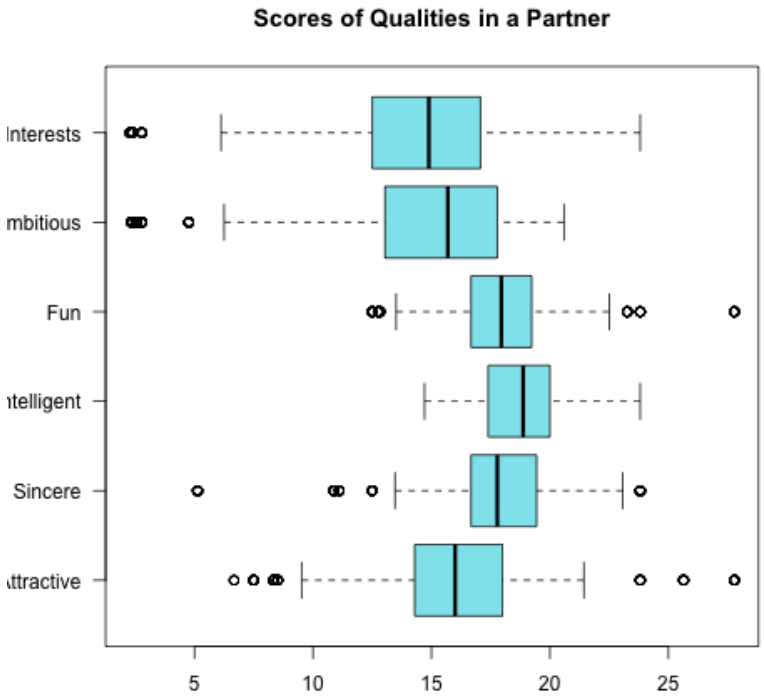
Discuss and quantify the importance, in general, of each person a's 6 ratings of person b (attractiveness, sincerity, intelligence, and so forth) in determining if person a is interested in seeing person b again.

Ranking of qualities before speed dating session

Before speed dating started in this study, each dater was requested to answer questions regarding the qualities they look for in a potential date. My thought was to look at their rankings of qualities to see if it lines up with how they eventually made their decision on whether to see a person again.

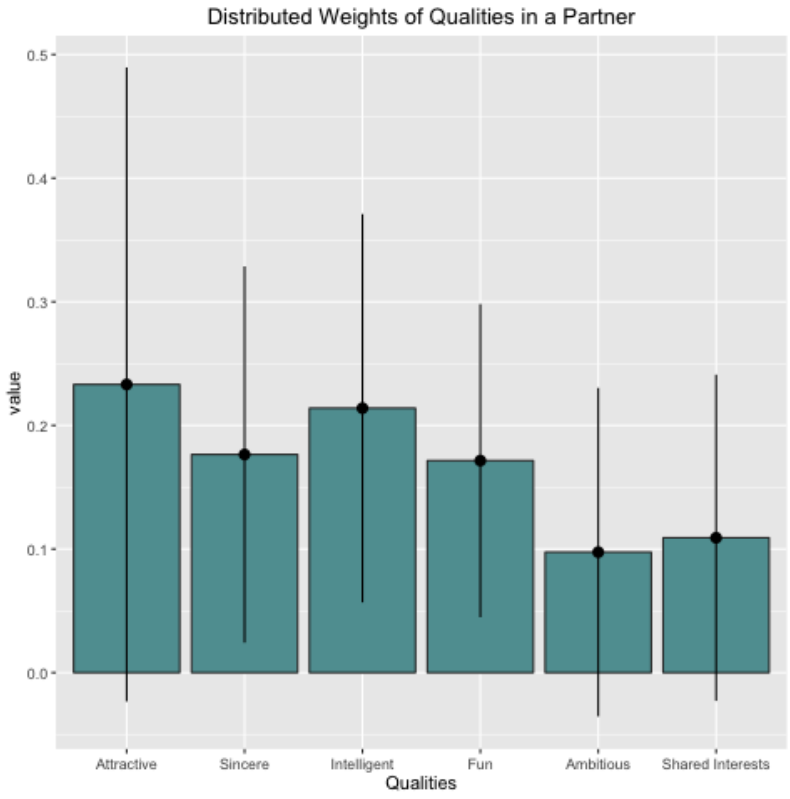
The trouble with this question was that approximately 1/3 of the daters were asked to rank qualities on a 1-10 scale, and the other 2/3 of daters were asked to distrubute 100 points amongst the qualities. I thought it was important to keep both results, but since the study uses a 1-10 scale in the scorecard while the speed dating is occuring, I will use the results from the prior method to compare results of what a dater thinks they want, versus how they actually make decisions.

Ranking qualities on a 1-10 scale



We can conlude that when ranking qualities on a 1-10 scale, the ranking is as follows: intelligence, fun, sincerity, attractiveness, shared interests followed by ambition.

Distrubuting points amongst qualities



We can conlude that when distributing 100 points amonst qualities, the ranking is as follows: attractiveness, intelligence, sincerity, fun, shared interests followed by ambition.

Ranking of qualities while speed dating

Now I turn to look at how these qualities influence a dater's decision on whether or not to continue seeing another person while speed dating.

Coefficient estimates of the least squares model

For this analysis I looked at an indvidual's rating of their speed-dater's attractiveness, sincerity, intelligence, fun, ambition and shared interests versus whether an individual wanted to see him or her again, our predictor variables will be attractiveness, sincerity, intelligence, fun, ambition and shared interests, and our Y varialbe will be their decision whether they wanted to see that person again.

$\texttt{dec} = \beta_0 + \beta_1$ x $\texttt{attr} + \beta_2$ x $\texttt{sinc} + \beta_3$ x $\texttt{intel} + \beta_4$ x $\texttt{fun} + \beta_5$ x $\texttt{amb} + \beta_6$ x $\texttt{shar}$

This table shows the least squares coefficient estimates of the multiple linear regression of a dater's decision to match with the person on attractiveness, sincerity, intelligence, fun, ambition and shared interests. How we should interpret this is that the coefficients for each of these represent the average effect of increasing that particular predictor, while holding all other predictors constant; e.g. the coefficient for `sinc` is the average effect of a dater's judgement of a person's sincerity while holding all other variables fixed.

What we note here is that the coefficient for `intel` and `amb` is negative, indicating they actually has a negative effect on a dater's decision to match when compared against the other predictors.

Correlation matrix

Table 1: Coefficient estimates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.517	0.045	-11.495	0
attr	0.117	0.005	21.410	0
sinc	0.023	0.007	3.410	0.001
intel	-0.017	0.008	-2.070	0.038
fun	0.020	0.006	3.131	0.002
amb	-0.027	0.006	-4.273	0.00002
shar	0.045	0.005	8.394	0

I constructed a correlation matrix to interpret the correlation between all variables to see how each of them interact with one another

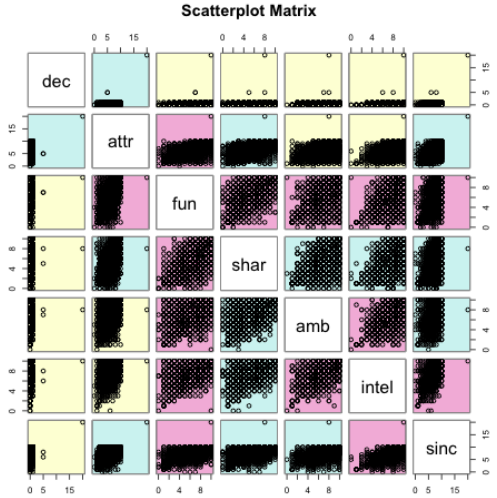
Table 2: Correlation matrix

	dec	attr	sinc	intel	fun	amb	shar
dec	1	0.479	0.258	0.191	0.360	0.158	0.351
attr	0.479	1	0.418	0.374	0.585	0.343	0.472
sinc	0.258	0.418	1	0.649	0.514	0.464	0.401
intel	0.191	0.374	0.649	1	0.485	0.605	0.411
fun	0.360	0.585	0.514	0.485	1	0.499	0.618
amb	0.158	0.343	0.464	0.605	0.499	1	0.443
shar	0.351	0.472	0.401	0.411	0.618	0.443	1

In this correlation matrix table we can clearly tell that `attr` has the strongest correlation with `dec` (0.479) while `intel` (0.191) and `amb` (0.158) are weaker.

High correlations between a dater's rating of a person include: `sinc` and `intel` (0.649), `fun` and `shar` (0.618) and `intel` and `amb` (0.605)

These relations are represented visually in the scatterplot matrix graph below. The pink scatterplots represent those with the highest correlations, whereas the relationship between `newspaper` and `sales` is represented as blue, indicating a very weak correlation.



Results: Question 1

These results made me laugh a bit. In the pre-assessment, before the speed dating began, daters rated intelligence as either their #1 or #2 indicator of what they look for in another person, yet when deciding whether or not to see a person again, intelligence actually had a *negative* correlation. I assume this can be the cause of one or two things: 1) Individuals like to say that they value intelligence to curb their own ego, or to show others that they care more about what's on the inside than the outside or 2) We saw a 0.374 correlation between attractiveness and intelligence, and maybe in the long run, attractiveness beats intelligence.

The results for question 1 are quite clear in that attractiveness plays highest role in determining whether a dater is likely to want to see a person again. The rankings of qualities are as follows: attractiveness, shared interests, sincerity, fun, intelligence and ambition. Personally, I find these ratings somewhat dismal and a little surprised that intelligence and ambition actually have a negative correlation.

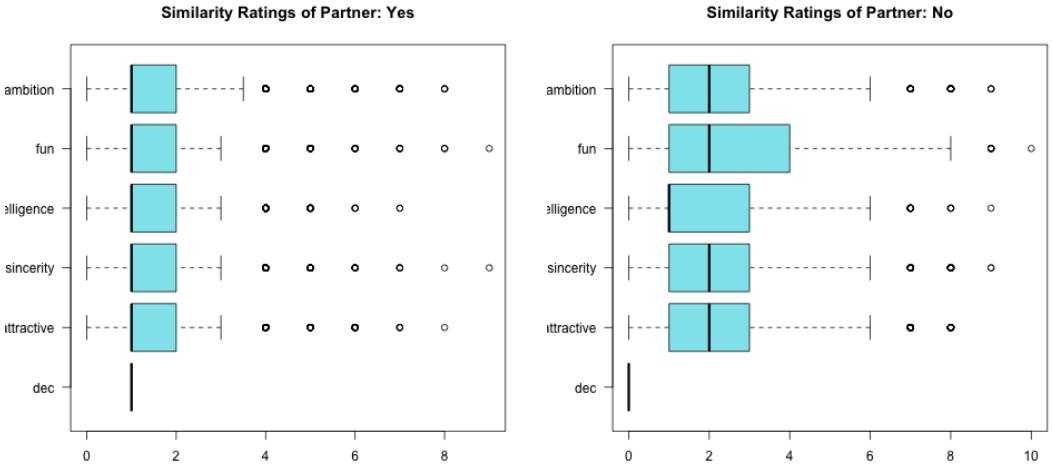
Question 2

Some people are more likely to want to see other people again. Discuss the implications for the general conclusions you reached in question 1 and see if you can account and correct for this kind of underlying variation in the person making the evaluation.

Hypothesis: People want to continue seeing people they think are similar to themselves.

This has been studied time and time again and while people might say "opposites attract", the fact of the matter is that people are interested in people like themselves. A great study that was done on eHarmony's data studies this question, and if you'd like to learn more FiveThirtyEight wrote an article about it.

To conduct this analysis, I looked at the dater's self-ratings of themselves against the ratings of the person they met. I subsetted the data based on whether the dater wanted to see that person again, then looked at the absolute values between their self-rating and the rating of the person they met. Lower values correspond to higher similarity, while higher values correspond to lower similarity.



The graph on the left, which is the graph for those who indicated they wanted to see them again, are all between 1-2, whereas the graph on the right, which is the graph for those who indicated they did not want to see them again, have much more variation, ranging from 1.5-2.5. From these graphs, are able to say that if a person felt that their qualities were in line with those of the other individual, then they were more likely to indicate that they wanted to see them again.

Question 3

With some people, it is more likely that others will want to see them again. Is there evidence in the data for this kind of variation in the person being rated? If so, how can you account and correct for it and what are the implications for your conclusions in 1 & 2.

In order to answer this question, I decided to use a Bayesian approach. My ultimate question (which is a bit more refined than this one) is: Given the other person decides to see me again, what is the percentage that I also want to see them again. I thought this was an interesting question because it speaks to A) a person's level of confidence (if they believe the other person liked them, are they more likely to like them?) and B) the variability in mutual feelings about others while dating.

I essentially took a dater's rating of the individual they met and looked at it against the individual's rating of the dater. To explain a bit further, I created duplicates of the dating table, then created a unique ID based on the `iid` and `pid` of that particular date. I then merged the tables on this key, and deleted duplicates.

Essentially I had a dataframe that looked like this:

```
##    match_id iid pid iid_dec pid_dec
## 19     1011  10  11       0       0
## 20     1012  10  12       0       0
## 21     1013  10  13       1       1
## 22     1014  10  14       0       1
## 23     1015  10  15       0       1
## 24     1016  10  16       0       0
```

- `match_id`: unique match of a dater with another individual
- `iid`: unique dater ID
- `pid`: unique partner ID
- `iid_dec`: decision of dater whether to see partner again
- `pid_dec`: decision of partner whether to see dater again

By deduping the table, we are left with unique `(iid, pid)` tuples, meaning that if `(iid(5), pid(8))` then `(pid(8), iid(5))` does not exist.

My next step was to determine conditional probabilities. To interpret the table below, `[Y_me, Y_them]` is the percentage where the dater and partner mutually liked each other. `[N_me, Y_them]` is the percentage where the dater did not like the partner, but the partner liked the dater. `[Y_me, N_them]` is the percentage where the dater liked the partner, but the partner did not like the dater. `[N_me, N_them]` is the percentage where neither the dater or partner liked each other.

```
##             Y_me      N_me    Total1
## Y_them 0.1699587 0.1548002 0.3247588
## N_them 0.3601286 0.3151125 0.6752412
## Total2 0.5300873 0.4699127 1.0000000
```

From this, we can see that in some cases, it is more likely that others will want to see them again, as evidenced by `[Y_me, Y_them]` + `[N_me, Y_them]` = 0.325. However, the case of the dater wanting to see the partnger again is a bit higher as `[Y_me, Y_them]` + `[Y_me, N_them]` = 0.530.

To answer the question: Given the partner likes the dater, what is the probability the dater likes the partner? we employ Bayes' Theorem. This is really just a simple question of conditional probability.

Pr(A|X) = Pr(X|A)Pr(A) / Pr(X|A)Pr(A) + Pr(X|not A)Pr(not A)

When we compute this with Pr(Dater Yes | Partner Yes) we obtain:

```
pAX
```

```
## [1] 0.1849908
```

Thus, given the partner chose to see them again, there is an 18.5% chance that the dater will choose to see them again as well. Interpreting this result, we might be able to apply this to the real world, outside of speed dating, in which case this predicts a slim chance for mutual liking.

BONUS

In looking at this data, my initial thought that came to mind might be gender bias (i.e. women vs. men might make decisions based on different factors). To explore this a bit more, I separated the data based on gender and looked at the correlations between qualities and a dater's decision to see an individual again or not.

Correlation matrix

I first looked at the correlation matricies for each gender, then to compare them I layered the correlation matrix for women and subtracted the correlation matrix for men.

Table 3: Correlation matrix (women - men)

	dec	attr	sinc	intel	fun	amb	shar
dec	0	-0.040	-0.074	0.008	0.039	-0.012	0.070
attr	-0.040	0	-0.045	-0.038	0.018	-0.105	0.028
sinc	-0.074	-0.045	0	-0.003	-0.040	-0.005	-0.033
intel	0.008	-0.038	-0.003	0	-0.099	0.039	-0.044
fun	0.039	0.018	-0.040	-0.099	0	-0.076	0.035
amb	-0.012	-0.105	-0.005	0.039	-0.076	0	-0.077
shar	0.070	0.028	-0.033	-0.044	0.035	-0.077	0

From this correlation matrix, I specifically want to look at the first row: the correlation between a dater's decision and each of the qualities. A positive score means that there is a stronger correlation between a dater's decision and that quality in **women** and a negative score means that there is a stronger correlation between a dater's decision and that quality in **men**.

From the table we can assume that women value intelligence, fun and sharing interests moreso than men, and men value attractiveness (not surprised), sincerity (semi shocking?), and ambition (yay!) moreso than women.

Coefficient estimates of the least squares model

From these tables, we can see the differences in qualities and their effect on whether a dater wants to see that person again. Breaking down these results by individual attributes:

`attr`: male > female

`sinc`: male > female

`intel`: male < female

`fun`: male = female

Table 4: Coefficient estimates for women

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.395	0.048	-8.211	0
attr	0.073	0.006	12.161	0
sinc	-0.004	0.007	-0.565	0.572
intel	0.007	0.009	0.760	0.448
fun	0.022	0.007	3.231	0.001
amb	-0.016	0.007	-2.273	0.023
shar	0.045	0.006	7.748	0

Table 5: Coefficient estimates for men

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.695	0.076	-9.109	0
attr	0.153	0.009	16.586	0
sinc	0.053	0.012	4.393	0.00001
intel	-0.037	0.014	-2.645	0.008
fun	0.023	0.011	2.013	0.044
amb	-0.039	0.011	-3.634	0.0003
shar	0.045	0.009	5.125	0.00000

`amb`: male $<$ female

`shar`: male $=$ female

From these results, we can conclude with some degree of certainty (since our standard errors are all <0.02 for men and <0.01 for women) that males value attractiveness, sincerity (semi shocking?), slightly more than females

Conslusion

Tada! This completes my analysis of the speed dating data. In conclusion, people are going to base their decisions primarily based off of attractiveness, similarity to themselves and on the probability that the person they're dating likes them!