Dating Experiment

Kalvin, Dan, and Jennifer

December 3, 2017

#setwd("~/Berkeley/W241/Experiment/Final\_Deliverables/final\_paper\_WIP")#Jen's personal comp  
setwd("~/Personal/Grad SChool/Courses/w241/Experiment/w241\_final\_project/Final\_Deliverables/final\_paper\_WIP") #Jen's work laptop  
#setwd("C:/MIDS/W241/final\_project/Final Paper")#Kalvin's working directory

## Experimental Design

### Research Design (using ROXO grammar)

### Randomization engineering

In order to effectively deliver the treatment variable, showing the profile to a limited audience of potential suitors was paramount. To elaborate on that point, should a potential suitor see the profile change from “MD” to “BS” that would not only ruin the data point, but also risk the injured party flagging the account to the administrators and terminating the experiment prematurely. Fortunately, the premium Tinder account enabled the research team to control who had access to the profile so only the one hundred individuals in each city selected for that week’s treatment were exposed; very effectively reducing potential spillover.

Additionally, Tinder restrict users from seeing repeat profiles to prevent communication between people who previously interacted unsuccessfully, making the trials independent. Each week a new randomly select batch of ladies and gentlemen were swiped on and put into the experiment. The potential suitors have an equal probability of receiving any given treatment (a bit of a stretch.?).

To help demonstrate the balance in treatment assignment, the distribution of suitor ages for each treatment condition are displayed in figures XX-XX below.

Figure XX (will add plots after they’re finalized– please check EDA google doc)

It is interesting to see that the age distribution has a right skew for all treatment conditions, and it is unclear that depending on Tinder to select suitors is truly random– it could be the case that Tinder selects suitors based on the age set in the test profile, and it could also be the case that the general population of Tinder users is centered around the mid-20s age range. But the consistency of this distribution across all conditions indicates that there was indeed an equal probability of receiving a given treatment condition, at least amongst the suitors selected by Tinder.

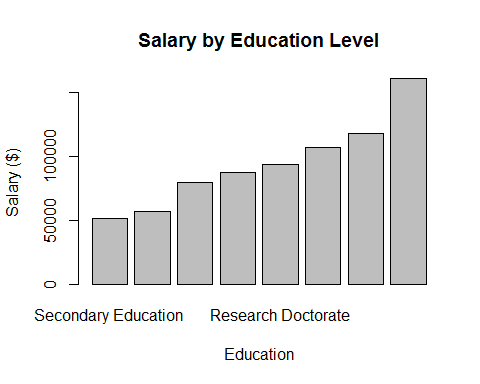
The age distribution of suitors to the Male profile (female suitors) does seem to be centered approximately 2 years younger than that of suitors to the Female profile, and the skew appears to increase slightly with increasing education level. However, statistical tests for the difference in average age by test condition yields a significant difference only between the male and female conditions (shown in figures XX-XX below), and its effect size is small– approximately half a year.

### Experimental materials (e.g. treatment materials)

Objectively measuring individual’s success in life is ambiguous and difficult, but a reasonably good proxy for research purposes could be educational attainment. As the graph below indicates, on average there is a positive association between higher education and income levels. Other benefits of education include…

-more on Reason for focus on educational attainment over job title/institution?

degree = c("Secondary Education", "Associate's Degree","Bachelor's Degree","Master's Degree", "Research Doctorate","Doctor of Law", "MBA", "Doctor of Medicine" )  
salary = c(51500,57100,79800,87700,94100,107000,118300,161200)  
  
df\_treatments = cbind(degree,salary)  
#hist(df\_treatments$salary)#Jen's original barplot of salaries  
  
#I couldn't get the above histogram command to work ("Error in df\_treatment$salary : $ operator is invalid for atomic vectors")  
#Is this what you're trying to do?  
barplot(as.numeric(df\_treatments[,"salary"]), main="Salary by Education Level", xlab="Education", ylab="Salary ($)", names.arg=df\_treatments[,"degree"])



The potential suitors were shown a profile that either read: “MD”, “PhD”, “BS” or the field to provide education information was left blank. A sample profile can be seen below.

The research team was particularly interested in the difference between the impact of having a Medical Degree versus having a Research Degree. Both degrees require about five years to obtain, however there is significant difference in earnings potential.

### Measurement of variables

As discussed previously, the treatment in this experiment is the test profile’s exposure of a given education level to a potential suitor, and the outcome is whether or not the suitor matches with the test profile. When applying an education level treatment to the test profile, the measured test subject covariates include its age, sex, and location, and the measured suitor covariates include age, sex, whether or not school information is provided, whether or not job information is provided, and whether or not instagram photos are provided (and their count, if so). These variables are described in table below.

#(kalvin) i tried to use kable and kableExtra to format the table, but i'm having trouble using kableExtra  
#(kalvin) i might try a different package later  
  
column\_names <- c("Variable", "Variable Type", "Source", "Description / Possible Values")  
variables <- c("test profile education level", "suitor profile match", "test profile sex",   
 "test profile location", "suitor age", "suitor school\*\*\*", "suitor job\*\*\*",   
 "suitor instagram\*\*\*", "number of instagram photos\*\*\*")  
variable\_type <- c("treatment", "outcome", "covariate", "covariate", "covariate",   
 "covariate", "covariate", "covariate", "covariate")  
source <- c("test profile", "Tinder application", "test profile", "test profile",   
 "suitor source code", "suitor source code", "suitor source code",   
 "suitor source code", "suitor source code")  
description <- c("4 possible values\*: No education listed (control), Bachelor's, MD, PhD",   
 "2 possible values: match or no match",   
 "2 possible values: female or male",   
 "8 possible values: Chicago, Houston, Los Angeles, New York, Philadelphia, Phoenix, San Antonio, San Diego",   
 "Due to the filters set in the test profiles, this variable can range from 24 - 34 years.",   
 "2 possible values: school information was detected or not",   
 "2 possible values: job information was detected or not",   
 "2 possible values: suitor profile included an instagram link or not",   
 "if a suitor profile contained a link to an instagram account, this variable is the number of photos in the account")  
variables\_table <- matrix(c(variables, variable\_type, source, description), nrow=length(variables))  
#variables\_table.app <- cbind(variables\_table, source)  
colnames(variables\_table) <- column\_names  
#kable(variables\_table, rownames=FALSE, padding=4)  
#kable(variables\_table, fontsize = 8, escape=FALSE, format.args=c(width=2), booktabs=TRUE)  
kable(variables\_table, fontsize = 6, caption="Table XX: Description of Variables") %>%  
 kable\_styling(bootstrap\_options = c("striped","responsive"), full\_width = FALSE)

## Currently generic markdown table using pandoc is not supported.

Table XX: Description of Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Variable Type | Source | Description / Possible Values |
| test profile education level | treatment | test profile | 4 possible values\*: No education listed (control), Bachelor’s, MD, PhD |
| suitor profile match | outcome | Tinder application | 2 possible values: match or no match |
| test profile sex | covariate | test profile | 2 possible values: female or male |
| test profile location | covariate | test profile | 8 possible values: Chicago, Houston, Los Angeles, New York, Philadelphia, Phoenix, San Antonio, San Diego |
| suitor age | covariate | suitor source code | Due to the filters set in the test profiles, this variable can range from 24 - 34 years. |
| suitor school\*\*\* | covariate | suitor source code | 2 possible values: school information was detected or not |
| suitor job\*\*\* | covariate | suitor source code | 2 possible values: job information was detected or not |
| suitor instagram\*\*\* | covariate | suitor source code | 2 possible values: suitor profile included an instagram link or not |
| number of instagram photos\*\*\* | covariate | suitor source code | if a suitor profile contained a link to an instagram account, this variable is the number of photos in the account |

#kable\_styling(bootstrap\_options = c("striped","responsive")) %>%  
 #column\_spec(4, width="10em")

\* The initial experimental plan also included the ‘Associates’ education level as a treatment condition, however, technical issues at the start of the experiment prevented proper execution of the ‘PhD’ education level treatment, so the ‘PhD’ education level treatment was repeated and the ‘Associates’ education level treatment was skipped.

\*\* The test profile age was an originally planned covariate, but due to time constraints, only one age was used for the entire experiment.

\*\*\* These variables are believed to be only partially observed– see Section for further details.

*Issues with Measurement of Outcome*

After swiping right on a suitor, the Tinder application only provides notification if that suitor likes the test profile in return, and does not provide an indication of whether or not a suitor actually saw the profile or if the suitor disliked the test profile. Our experiment therefore contains some rate of non-compliance and we are unfortunately unable to determine the compliance rate– our estimated average treatment effect is thus the effect of the intent to treat.

*Issues with Measurement of Covariates*

The structure of the relevant source code in the Tinder browser interface is such that it contains a link to a suitor’s profile image (which we use as a unique identifier for that suitor), the suitor’s name, the suitor’s age, and finally a maximum of two additional details that may include any combination of school information, job information, and instagram information.

It is known that a Tinder profile may display all three of these pieces of information, but none of the ## source codes collected contained more than two details. It is highly unlikely that amongst the ## suitors encountered in the experiment, none displayed more than two of these details, so we believe that school information, job information, and instagram information are only partially observed, and without more in-depth examination of each suitor profile at the time of experiment execution, our analysis is unable to fully determine the education and job status/level, as well as instagram information, for each suitor. Due to time constraints, thorough examination of each profile was unfortunately not an option– we were required to use an automated swiping method which acquired covariate information from source code. It is also known that a Tinder profile may additionally display the suitor’s distance and favorite spotify songs, but none of this information was present in the suitor source codes acquired during the experiment, and these two covariates were thus unobserved.

The source code indicates which suitor details are the profile image, the name, and the age, but does not indicate which details are school, job, or instagram information. Our data collection process searches the text of these details to make a best guess as to what type of detail it is (instagram information is easily identified, but not school or job information). So for some of the school and job details that were collected, it could not be determined whether the detail represented education or employment.

*Data Pre-processing* Since the non-binary variables are categorical, they were binarized using indicator variables for each category…

### Modeling choices

calculation of statistical power

## Experiment Results

In text description of your results Figures and tables that support your in text description Clean, clear, well articulated relationships between your theory, your hypotheses, the numbers that your models produce, and the figures you present