Statistical Investigation into the relationship between Self-perceived attractiveness and Date likelihood

Author - Aryan

Date - 15/10/2020

Introduction

The question here links of the self-perception of two concepts, attractiveness and likelihood of someone they met being interested in dating. We will refer to the former in this piece as *"Self-perceived attractiveness"* and the latter as *"Date likelihood"*. Factors such as these have been linked to the important decision of choosing a partner and is of interest to researchers across fields like psychology and behavioural economics having sparked a number of research studies over the years. In the following piece, we will consider a study conducted by researchers from Columbia, Stanford and Harvard Universities.

You can have a read of the original paper at this link. The dataset used from the study at Columbia Business school (Fisman et al. 2006) is available from Andrew Gelman's website.

Background problem and Experimental design

The researchers Fisman, Iyengar, Kamenica and Simonson were interested in the age old question of determining the factors influencing a person's choice of marriage partners. To achieve this data a few options were considered but, the researchers opted to employ a speed dating experiment due to its highly active and simple format. This allowed the researchers to control several aspects of the setup such as number of dates per individual. There were obvious drawbacks, such as the short-term dating mindset of most participants. Despite the downside, studying the reasons for an initial romantic attraction were still of interest. The study was conducted amongst the graduate students at the University of Columbia in a series of 17 successful events running from 2002 to 2004. Certain variables like number participants and gender ratio were altered for events but the 1-to-1 rotating format of the speed date. *Note only heterosexual dates were considered in this study.*

Data description

The data was collected using a series of surveys before, during, directly following and after the event. Several predictors were identified from a prior attraction research and were recorded from the both from the participant and the date's point of view. Some these included, similarity of racial backgrounds, personality matches and financial security. We are, however, interested in the attributes of *Self-perceived attractiveness* and *Date likelihood* that the raw data labels `attr3_1` and `prob` respectively. One of the data collection issues that was observed was the omission of information in the surveys. This is dealt with in the analysis conducted further in this piece.

Subsetting the data

To load the data we can use the R command:

```r
SpeedRawData <- read.csv("SpeedDatingRawData.csv")
```

The raw data has 8378 rows and 195 columns. However, as stated above, we are only interested in the two variables regarding the *self-perceived of attractiveness* and the *date likelihood*.

The following code constructs a data frame that encompasses the values of interest alone whilst removing all rows containing empty values for either variable:

```
ind <- !is.na(SpeedRawData$attr3_1) & !is.na(SpeedRawData$prob)
SpeedData <- SpeedRawData[ind, c("attr3_1", "prob")]
```

First 5 elements of the newly constructed table:

Table 1: Sample Raw Data of Variables of Interest

	Self-Perceived Attractiveness	Date Likelihood
1	6	6
2	6	5
4	6	6
5	6	6
6	6	5

We save the data frame `SpeedData` in a csv file as follows.

```
write.table(SpeedData, file = "SpeedData.csv", sep = ",", row.names = FALSE)
```

Reproducable Functions

The use of R functions ensures reproducibility and avoids cluttering the workspace with variables. Our function will take only the raw data as input. It then produces a scatterplot with the best fit line superimposed on top of it to showcase the general trend and saves the selected data in a csv file, titled `SpeedData`.

Here is the function that selects the data of interest.

```
SelectData <- function(file)
{
  SpeedRawData <- read.csv(file)
  ind <- !is.na(SpeedRawData$attr3_1) & !is.na(SpeedRawData$prob)
  SpeedData <- SpeedRawData[ind, c("attr3_1", "prob")]
  return(SpeedData)
}
```

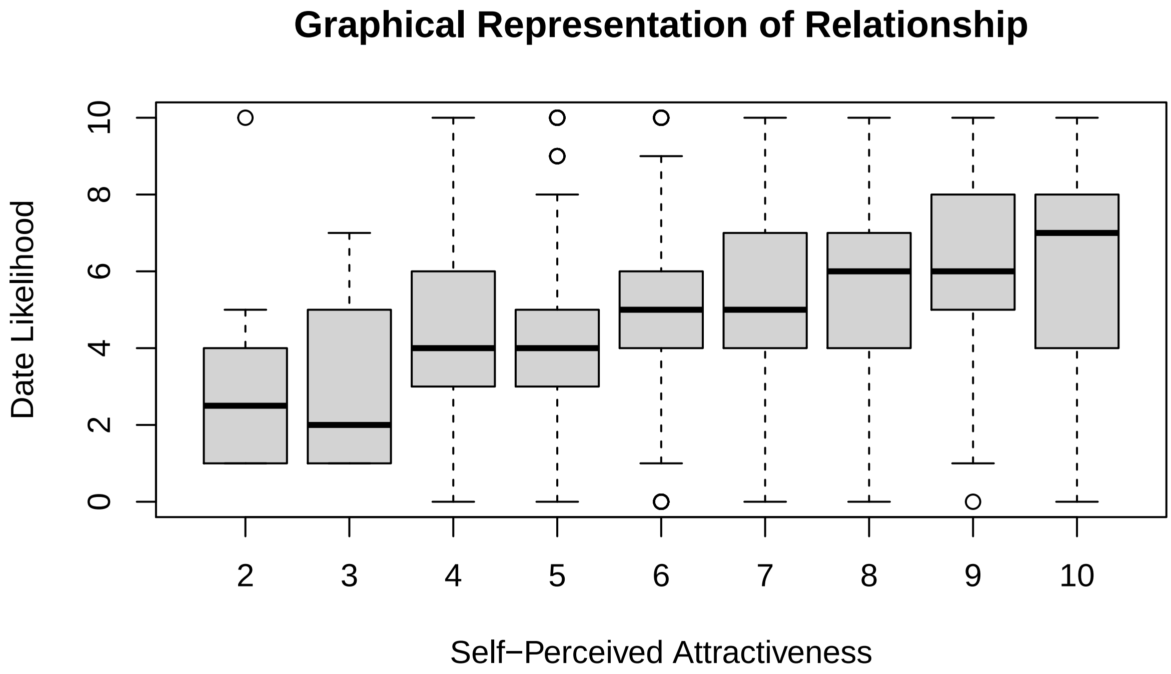
Below is the function that plots and then saves the data.

```
PlotAndSave <- function(inputfile = "input.csv", outputfile = "output.csv")
{
  data <- SelectData(file = inputfile)

  library(ggplot2)

  myggplot <- boxplot(prob~attr3_1,data=data, main = "Graphical Representation of Relationship", xlab =

#  mynamestheme <- theme(plot.title = element_text(hjust = 0.5, family = "Helvetica", face = "bold", si

  finalplot <- myggplot

  print(finalplot)

  write.table(data, file = outputfile, sep = ",", row.names = FALSE)
}
```

Resulting Plot

Utilyzing the function above.

```
PlotAndSave(inputfile = "SpeedDatingRawData.csv", outputfile = "SpeedData.csv")
```



```
## $stats
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## [1,]  1.0    1    0    0    1    0    0    1    0
## [2,]  1.0    1    3    3    4    4    4    5    4
## [3,]  2.5    2    4    4    5    5    6    6    7
## [4,]  4.0    5    6    5    6    7    7    8    8
## [5,]  5.0    7   10    8    9   10   10   10   10
##
## $n
## [1]   20  143  230  610 1068 2801 2127  713  263
##
## $conf
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
## [1,] 1.440104 1.471495 3.687454 3.872055 4.903306 4.910438 5.897223 5.822486
## [2,] 3.559896 2.528505 4.312546 4.127945 5.096694 5.089562 6.102777 6.177514
##          [,9]
## [1,] 6.610292
## [2,] 7.389708
##
## $out
##  [1] 10 10 10  9  9  9  9 10  9 10 10 10  9 10 10  9 10 10 10 10 10  0  0  0  0
## [26]  0  0  0  0  0  0  0 10 10  0 10 10  0 10 10  0
##
## $group
##  [1] 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
## [39] 5 5 8
##
## $names
## [1] "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10"
```

Concluding Analysis

A brief interpretation of the plot The plot through a general increase in median value suggests a **positive correlation** between *self-perceived attractiveness* and *date likelihood.* However, let weigh up the good with the biases that may be present here First and foremost, both these values are self-evaluated and hence could be over or under valued. However, even so this result should be expected since people who over-estimates themselves will tend to over-estimate both quantities and vice versa. It is perhaps worth restating that this was conducted within university student bodies, which restricts the ages and also makes us question the accuracy of their answers as students might tend not to take themselves or this exercise too seriously. The varied frequency of rating might also be influencing the graph. This could be controlled by studying or analyzing a fixed number of individuals from every 'self-perceived attractiveness' to see how 'date likelihood' varies and vice versa. It may also be of interest to plot the distribution of answers as some are more frequent than others, for example the rating of 7 for *self-perceived attractiveness.* On the other hand, this study does utilize a rather varied (differing backgrounds, races, subjects etc.) and engaging crowd. In conclusion, the data set shows a **positive causality** which implies higher the *self-perceived attractiveness* higher the *date likelihood* but a further study into the significance of this result is required.