Lab #3 - Racial Bias in the Labor Market

Econ 224

September 4th, 2018

Introduction

Today we'll replicate a well-known paper published 2004: "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination" by Marianne Bertrand and Sendhil Mullainathan. The paper, which I'll refer to as BM for short, appears in Volume 94, Issue #4 of the *American Economic Review* (AER). Before beginning this lab, visit the website of the *American Economic Review* and search for the paper. Once you've found it, download both a pdf of the paper and the associated dataset, linked under "Additional Materials."

Exercise #1

Read the introduction and conclusion of BM, and then write a one-paragraph summary addressing the following points: 1. What question does the paper try to answer? 2. What data and methodology are used to address the question? 3. What are the key findings?

Solution to Exercise #1

Write your solutions here

Importing the Dataset

The dataset posted on the AER website is stored in a zip archive containing a single file: `lakisha_aer.dta`. You'll need to unzip this archive, and save the file `lakisha_aer.dta` to an appropriate directory on your machine. I suggest creating a directory called `econ224` with a subdirectory called `labs` and storing the data there along with your `.Rmd` file.

The extension `.dta` indicates that `lakisha_aer.dta` is a STATA datafile. STATA is a commercial statistics package that is much less powerful than R but very expensive. Because of this, its makers need to resort to other means to try to encourage people to buy their program. For example, they lock data used in STATA in a proprietary file format that is incompatible with other statistical software packages. This way, if I want to open your dataset and you are a STATA user, I'll have to buy a copy of STATA myself. Fortunately, enterprising open-source programmers have written software that can decode `.dta` files and convert them into other formats. We'll use the function `read_dta` from the `readr` package to convert `lakisha_aer.dta` into a tibble that we can manipulate with `dplyr`. We'll do this using the function `read_dta` in the package `haven`, which contains functions for converting data from SPSS, STATA, and SAS formats. When we want to read in data that is already in a non-proprietary format, we'll use the package `readr` that is included as part of `tidyverse`. See the chapter "Data Import" in *R for Data Science*.

Make sure to install `haven` before proceeding. Start by loading both `tidyverse` and `ggplot2`.

```
library(haven)
library(tidyverse)
library(ggplot2)
```

Now we can read in the data as follows by using the `read_dta` function. Note that you'll have to specify the directory where you've saved the data.I've saved my copy in `~/econ224/labs` but you may have saved yours somewhere else. If you're using Windows, it's a little trickier to specify the right file path: you may need to Google this.

```
bm <- read_dta('~/econ224/labs/lakisha_aer.dta')
```

Each row in `bm` corresponds to a single fake resume.

Exercise #2

1. Display the tibble `bm`. How many rows and columns does it have?
2. Display only the columns `sex`, `race` and `firstname` of `bm`. What information do these columns contain? How are `sex` and `race` encoded?
3. Add two new columns to `bm`: `female` should take the value `TRUE` if `sex` is female, and `black` should take value `TRUE` if `race` is black.

Solution to Exercise #2

Write your code and solutions here

```
bm
```

```
# A tibble: 4,870 x 65
   id    ad    education ofjobs yearsexp honors volunteer military
   <chr> <chr>     <dbl>  <dbl>    <dbl>  <dbl>     <dbl>    <dbl>
 1 b     1             4      2        6      0         0        0
 2 b     1             3      3        6      0         1        1
 3 b     1             4      1        6      0         0        0
 4 b     1             3      4        6      0         1        0
 5 b     1             3      3       22      0         0        0
 6 b     1             4      2        6      1         0        0
 7 b     1             4      2        5      0         1        0
 8 b     1             3      4       21      0         1        0
 9 b     1             4      3        3      0         0        0
10 b     1             4      2        6      0         1        0
# ... with 4,860 more rows, and 57 more variables: empholes <dbl>,
#   occupspecific <dbl>, occupbroad <dbl>, workinschool <dbl>,
#   email <dbl>, computerskills <dbl>, specialskills <dbl>,
#   firstname <chr>, sex <chr>, race <chr>, h <dbl>, l <dbl>, call <dbl>,
#   city <chr>, kind <chr>, adid <dbl>, fracblack <dbl>, fracwhite <dbl>,
#   lmedhhinc <dbl>, fracdropout <dbl>, fraccolp <dbl>, linc <dbl>,
#   col <dbl>, expminreq <chr>, schoolreq <chr>, eoe <dbl>,
#   parent_sales <dbl>, parent_emp <dbl>, branch_sales <dbl>,
#   branch_emp <dbl>, fed <dbl>, fracblack_empzip <dbl>,
#   fracwhite_empzip <dbl>, lmedhhinc_empzip <dbl>,
#   fracdropout_empzip <dbl>, fraccolp_empzip <dbl>, linc_empzip <dbl>,
#   manager <dbl>, supervisor <dbl>, secretary <dbl>, offsupport <dbl>,
#   salesrep <dbl>, retailsales <dbl>, req <dbl>, expreq <dbl>,
#   comreq <dbl>, educreq <dbl>, compreq <dbl>, orgreq <dbl>, manuf <dbl>,
```

```
#   transcom <dbl>, bankreal <dbl>, trade <dbl>, busservice <dbl>,
#   othservice <dbl>, missind <dbl>, ownership <chr>
```

```
bm <- bm %>%
  mutate(female = (sex == 'f'),
         black = (race == 'b'))
```

Checking for Balance

Because The variable `computerskills` takes on the value `1` if a given resume says that the applicant has computer skills. The variables `education` and `yearsexp` indicate level of education and years experience, while `ofjobs` indicates the number of previous jobs listed on the resume.

Exercise #3

1. Read parts A-D of section II in BM and answer the following:

- How did the experimenters create their bank of resumes for the experiment?
- The experimenters classified the resumes into two groups. What were they and how did they make the classification?
- How did the experimenters generate identities for their fictitious job applicants?

2. Is sex balanced across race? Use `dplyr` to answer this question. Hint: what happens if you apply the function `sum` to a vector of `TRUE` and `FALSE` values?
3. Are computer skills balanced across race? Hint: the summary statistic you'll want to use is the *proportion* of individuals in each group with computer skills. If you have a vector of ones and zeros, there is a very easy way to compute this.
4. Are `education` and `ofjobs` balanced across race?
5. Compute the mean and standard deviation of `yearsexp` by race. Comment on your findings.
6. Why do we care if `sex`, `education`, `ofjobs`, `computerskills`, and `yearsexp` are balanced across race?
7. Is `computerskills` balanced across `sex`? What about `education`? What's going on here? Is it a problem? Hint: re-read section II C of the paper.

Solution to Exercise #3

Write your code and solutions here

```
# Part 2
bm %>%
  group_by(black) %>%
  summarize(n_female = sum(female))
```

```
# A tibble: 2 x 2
  black n_female
  <lgl>    <int>
1 FALSE     1860
2 TRUE      1886
```

```r
# Part 3
bm %>%
  group_by(black) %>%
  summarize(avg_computerskills = mean(computerskills))
```

```
# A tibble: 2 x 2
  black avg_computerskills
  <lgl>              <dbl>
1 FALSE              0.809
2 TRUE               0.832
```

```r
# Part 4
bm %>%
  group_by(black) %>%
  summarize(avg_numjobs = mean(ofjobs), avg_educ = mean(education))
```

```
# A tibble: 2 x 3
  black avg_numjobs avg_educ
  <lgl>       <dbl>    <dbl>
1 FALSE        3.66     3.62
2 TRUE         3.66     3.62
```

```r
# Part 5
bm %>%
  group_by(black) %>%
  summarize(avg_exp = mean(yearsexp), sd_exp = sd(yearsexp))
```

```
# A tibble: 2 x 3
  black avg_exp sd_exp
  <lgl>   <dbl>  <dbl>
1 FALSE    7.86   5.08
2 TRUE     7.83   5.01
```

```r
# Part 7
# These aren't balanced across sex because:
# "we use nearly exclusively female names for
# administrative and clerical jobs to increase
# callback rates"
bm %>%
  group_by(female) %>%
  summarize(avg_computerskills = mean(computerskills),
            avg_educ = mean(education))
```

```
# A tibble: 2 x 3
  female avg_computerskills avg_educ
  <lgl>               <dbl>    <dbl>
1 FALSE               0.662     3.73
2 TRUE                0.868     3.58
```

Callbacks by Race and Sex

The outcome of interest in `bm` is `call` which takes on the value `1` if the corresponding resume elicts an email or telephone callback for an interview. Check your results in this section against Table 1 of the paper.

Exercise #4

1. Calculate the average callback rate for all resumes in `bm`.
2. Calculate the average callback rates separately for resumes with "white-sounding" and "black-sounding" names. What do your results suggest?
3. Repeat part 2, but calculate the average rates for each combination of race and sex. What do your results suggest?

Solution to Exercise #4

Write your code and solutions here

```
# Part 1
bm %>%
  summarize(avg_callback = mean(call))
```

```
# A tibble: 1 x 1
  avg_callback
         <dbl>
1       0.0805
```

```
# Part 2
bm %>%
  group_by(black) %>%
  summarize(avg_callback = mean(call))
```

```
# A tibble: 2 x 2
  black avg_callback
  <lgl>        <dbl>
1 FALSE       0.0965
2 TRUE        0.0645
```

```
# Part 3
bm %>%
  group_by(female, black) %>%
  summarize(avg_callback = mean(call))
```

```
# A tibble: 4 x 3
# Groups:   female [?]
  female black avg_callback
  <lgl>  <lgl>        <dbl>
1 FALSE  FALSE       0.0887
2 FALSE  TRUE        0.0583
3 TRUE   FALSE       0.0989
4 TRUE   TRUE        0.0663
```

Comparing Returns to Quality

Bertrand and Mullainathan write: "for most of the employment ads we respond to, we send four different resumes: two higher-quality and two-lower quality ones." The column `h` takes on the value `1` if a resume as classified *a priori* as "high quality" and `0` if it was classified as "low quality." The columns `col`, `military`, `email`, `volunteer` are indicators for: college degree, has an email address, has done volunteer work, and served in the military.

Exercise #5

1. Compare the average value of `col`, `military`, `email`, and `volunteer` across "high quality" and "low quality" resumes. Discuss your findings.
2. Calculate average callback rates for black versus white-sounding names *separately* for "high-quality" and "low-quality" resumes. Discuss your findings

Solution to Exercise #5

Write your code and solutions here

```
# Part 1
bm %>%
  group_by(h) %>%
  summarize(mean(col), mean(military), mean(email), mean(volunteer))
```

```
# A tibble: 2 x 5
      h `mean(col)` `mean(military)` `mean(email)` `mean(volunteer)`
  <dbl>       <dbl>            <dbl>         <dbl>              <dbl>
1     0       0.715          0.00330        0.0305             0.0272
2     1       0.724          0.190          0.924              0.792
```

```
# Part 2
bm %>%
  group_by(h, black) %>%
  summarize(mean(call))
```

```
# A tibble: 4 x 3
# Groups:   h [?]
      h black `mean(call)`
  <dbl> <lgl>        <dbl>
1     0 FALSE       0.0850
2     0 TRUE        0.0619
3     1 FALSE       0.108
4     1 TRUE        0.0670
```

Interpreting the results

Read Section IV of the paper. Also read intro and conclusion of Fryer and Levitt.

Solution to Exercise #???

Write your code and solutions here