Problem Set 1: OLS Review

EC 421: Introduction to Econometrics

Due *before* midnight on Sunday, 27 January 2019

DUE Your solutions to this problem set are due *before* midnight on Sunday, 27 January 2019. Your files must be uploaded to Canvas—including (1) your responses/answers to the question and (2) the R script you used to generate your answers. Each student must turn in her/his own answers.

README! The data[†] in this problem set come from the paper "Are Emily and George More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination" by Bertrand and Mullainathan (published in the *American Economic Review* (AER) in 2004).[††] In their (very influential) paper, Bertrand and Mullainathan use a clever experiment to study the effects of race in labor-market decisions by sending fake résumés to job listings. To isolate the effect of race on employment decisions, Bertrand and Mullainathan randomize whether the résumé lists a typically African-American name or a typically White name.

OBJECTIVE This problem set has three purposes: (1) reinforce the econometrics topics we reviewed in class; (2) build your R toolset; (3) start building your intuition about causality within econometrics.

Problem 1: Getting started

Start here. We're going to set up R and read in the data

1a. Open up RStudio, start a R new script (File ➡ New file ➡ R Script). You will hand in this script as part of your assignment.

1b. Load the the `pacman` package. Now use its function `p_load` to load the `tidyverse` package, *i.e.*,

```
# Load the 'pacman' package
library(pacman)
# Load the packages 'tidyverse' and 'haven'
p_load(tidyverse)
```

Note: If `pacman` is not already installed on your computer, then you need to install it, *i.e.*, `install.packages("pacman")`. If `tidyverse` is not already installed, then `p_load(tidyverse)` will automatically install it for you—which is why we're using `pacman`.

1c. Download the dataset (also available on Canvas). Save it in a helpful location. Remember this location.

1d. Read the data into R. What are the dimensions of the dataset (numbers of rows and columns)?

Note: Let each row in this dataset represent a different résumé sent to a job posting. The table on the last page explains each of the variables.

Answer:

Setup

```
# The datasets's directory
dir_data ← "/Users/edwardarubin/Dropbox/UO/Teaching/EC421W19/Data/"
# Read in the data
ps1_df ← read_csv(file = paste0(dir_data, "dataPS01.csv"))
# Dimensions of the dataset with dim
dim(ps1_df)
```

```
## [1] 4870   12
```

1e. What are the names of the first three variables? *Hint:* `names(your_df)`

Answer:

Two options

```
# Using head
names(ps1_df) %>% head(3)
```

```
## [1] "i_callback" "n_jobs"     "n_expr"
```

```
# Using indexes
ps1_df[, 1:3] %>% names()
```

```
## [1] "i_callback" "n_jobs"     "n_expr"
```

1f. What are the first four *first names* in the dataset (`first_name` variable)?

Hint: `head(your_df$var_name, 10)` gives the first 10 observations of the variable `var_name` in dataset `your_df`.

Answer:

Three ways to do it:

```
# Using head
head(ps1_df$first_name, 4)
```

```
## [1] "Allison" "Kristen" "Lakisha" "Latonya"
```

```
# Using indexes
ps1_df[1:4,"first_name"]
```

```
## # A tibble: 4 x 1
##   first_name
##   <chr>
## 1 Allison
## 2 Kristen
## 3 Lakisha
## 4 Latonya
```

```
# Using head and select
ps1_df %>% head(4) %>% select(first_name)
```

```
## # A tibble: 4 x 1
##   first_name
##   <chr>
## 1 Allison
## 2 Kristen
## 3 Lakisha
## 4 Latonya
```

[†]: The data that we use in the problem set contain a subset of the variables from the original paper.

[††]: Here's a link to an article on Medium that discussed their paper.

Problem 2: Analysis

Reviewing the basic analysis tools of econometrics.

Note: When you use OLS to regress a binary indicator variable (like `i_callback`) on a set of explanatory variables, your coefficients are telling you how the explanatory variables affect the probability that the indicatory variable equals one. So if we regress `i_callback` on `n_jobs`, the coefficient on `n_jobs` tells us how the probability of a callback changes with each additional job listed on the résumé.

2a. What percentage of the résumés generated a callback (`i_callback`)?

Hint: The mean of a binary indicator variable (*i.e.*, `mean(binary_variable)`) gives the percentage of times the variable equals one.

Answer:

```
mean(ps1_df$i_callback)
```

```
## [1] 0.08049281
```

Thus, approximately 8 percent of résumés received callbacks.

2b. Calculate percentage of callbacks (*i.e.*, the mean of `i_callback`) for each racial group (`race`). Does it appear as though employers considered an applicant's race when making callbacks? Explain.

Hint: `filter(your_df, race == "b")` will select all observations (from the dataset `your_df`) where the variable `race` takes the value `"b"`. Similarly `filter(your_df, race == "b")$i_callback` will give you the values of `i_callback` for observations whose value of `race` is `"b"`.

Answer:

One method:

```
# Percentage for Black
filter(ps1_df, race == "b")$i_callback %>% mean()
```

```
## [1] 0.06447639
```

```
# Percentage for White
filter(ps1_df, race == "w")$i_callback %>% mean()
```

```
## [1] 0.09650924
```

Alternative method:

```
ps1_df %>% group_by(race) %>% summarize(mean(i_callback))
```

```
## # A tibble: 2 x 2
##   race  `mean(i_callback)`
##   <chr>              <dbl>
## 1 b                 0.0645
## 2 w                 0.0965
```

Approximately 6.4 percent of résumés with implicitly black names received callbacks, while 9.7 percent of résumés with white-sounding names received callbacks.

This difference is consistent with racial discrimination (employers considering race in hiring), but we do not know if this difference is statistically significant.

2c. What is the difference in the groups' mean callback rate?

Answer:

```
# Percentage for Black
mean_b ← filter(ps1_df, race == "b")$i_callback %>% mean()
# Percentage for White
mean_w ← filter(ps1_df, race == "w")$i_callback %>% mean()
# Difference:
mean_b - mean_w
```

```
## [1] -0.03203285
```

White-sounding names had a 3.2-percentage point higher callback rate.

2d. Based upon the difference in percentages that we observe in **2b.**, can we conclude that employers consider race in hiring decisions?

Answer: No. We have shown a difference in the groups' percentages, but we do not know if this difference is statistically meaningful (significant).

2e. Without running a regression, conduct a statistical test for the difference in the two groups' average callback rates (*i.e.*, test that the proportion of callbacks is equal for the two groups).

Hint: Back to your statistics class—difference in proportions (a Z test) or means (a t test).

Answer:

```
# Percentage for everyone
mean_all ← ps1_df$i_callback %>% mean()
# Number: Black
n_b ← filter(ps1_df, race == "b") %>% nrow()
# Number: White
n_w ← filter(ps1_df, race == "w") %>% nrow()
# The Z statistic
z_stat ← (mean_b - mean_w) / sqrt(mean_all * (1 - mean_all) * (1/n_b + 1/n_w))
# The p value
2 * pnorm(abs(z_stat), lower.tail = F)
```

```
## [1] 3.983887e-05
```

For H_0: equal callback rates *vs.* H_A: callback rates were not equal, we reject the null hypothesis at the 5-percent level with a p-value less than 0.001. We therefore conclude there is statistically significant evidence that employers responded to black- and white-sounding names at different rates.

Note: I opted for a two-sided Z test here, since we are testing unequal proportions. A t test (testing two means) would be fine, though maybe not technically correct. You could also test a one-side hypothesis if your null was that discrimination pointed in a specific direction (which it likely was).

2f. Now regress `i_callback` (whether the résumé generated a callback) on `i_black` (whether the résumé's name implied a black applicant). Report the coefficient on `i_black`. Does it match the difference that you found in **2c**?

Answer:

Simple linear regression...

```
# Regression
reg_2f ← lm(i_callback ~ i_black, data = ps1_df)
# Results
reg_2f %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   0.0965   0.00550     17.5  8.89e-67
## 2 i_black      -0.0320   0.00778     -4.11 3.94e- 5
```

The coefficient on `i_black` does indeed match the difference in callback rate across black- and white-sounding names.

2g. Conduct a t test for the coefficient on `i_black` in the regression above in **2f**. Write our your hypotheses (both H_0 and H_A), the test statistic, the result of your test (*i.e.*, reject or fail to reject H_0), and your conclusion.

Answer:

H_0: $\beta_1 = 0$ and H_A: $\beta_1 \neq 0$, where β_1 is the coefficient for the effect of race on the probability a résumé received a callback.

The point estimate for this coefficient is -0.032. Its associated t statistic is -4.11, which has a p-value less than 0.001.

We reject the null hypothesis at the 5-percent level. We conclude that there is statistically significant evidence that name's races affected callback rates for names with black or white connotations.

2h. Now regress `i_callback` (whether the résumé generated a callback) on `i_black`, `n_expr` (years of experience), and the interaction between `i_black` and `n_expr`. Interpret the estimates for the coefficients (both the meaning of the coefficients and whether they are statistically significant).

Hint: In R, `lm(y ~ x1 + x2 + x1:x2, data = your_df)` regresses `y` on `x1`, `x2`, and the interaction between `x1` and `x2` (all from the dataset `your_df`).

Answer:

```
# Regression with interaction
reg_2h ← lm(i_callback ~ i_black + n_expr + i_black:n_expr, data = ps1_df)
# Results
reg_2h %>% tidy()
```

```
## # A tibble: 4 x 5
##   term            estimate std.error statistic  p.value
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     0.0693     0.0101      6.84  8.79e-12
## 2 i_black        -0.0294     0.0144     -2.04  4.11e- 2
## 3 n_expr          0.00347    0.00108     3.20  1.36e- 3
## 4 i_black:n_expr -0.000330   0.00154    -0.214 8.30e- 1
```

The coefficient on `i_black` is quite similar to the coefficient previously found—suggesting the a black-sounding name reduced the probability of a callback by approximately 3 percentage points. This effect is still significant at the 5-percent level.

The coefficient on the number of years of experience (`n_expr`) implies that for each additional year of experience on the résumé, the probability of a callback increase by 0.3 percentage points. This effect is statistically significant at the 5-percent level.

The coefficient on the interaction between the black indicator variable and the experience variable tests whether the effect of experience on the callback rate differed between black and white résumés. The point estimate is small and is not statistically significant—meaning we cannot rule out the possibility that the interaction does not exist.

Problem 3: Thinking about causality

Now for the big picture.

This project by Bertrand and Mullainathan took a decent amount of time and effort—finding job listings, generating fake résumés, responding to the listings, *etc.* It probably would have been much quicker/cheaper/easier to just go out and get data from job applicants—whether they received callbacks and their races. So why didn't they take the easier, cheaper, and quicker route?

To answer this question, we are going to consider the model

$$\text{Callback}_i = \beta_0 + \beta_1 \text{Race}_i + u_i \tag{3.0}$$

and think about omitted-variable bias.

3a. If we go out, collect data on job applicants, and estimate the model in (3.0) using OLS, *i.e.*,

$$\text{Callback}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Race}_i + e_i \tag{3.1}$$

we should be concerned about omitted-variable bias. Explain why this is the case **and** provide at least one example of an omitted variable that could bias our estimates in (3.1).

Answer:

We should be concerned about omitted-variable bias because there likely many variables that affect whether individuals received callback **and** are correlated with race. If this is the case, then our estimates for β_1 will be biased.

Several possibilities: social connections, education, college major

3b. To avoid this potential bias, Bertrand and Mullainathan ran an experiment in which they randomized applicants' names on the résumés—thus randomly assigning the (implied) race of the job applicants. How does this randomization help Bertrand and Mullainathan avoid omitted variables bias?

In other words, why are we less concerned about omitted variable bias in the following estimated model

$$\text{Callback}_i = \hat{\beta}_0 + \hat{\beta}_1 (\text{Randomized Race})_i + w_i \tag{3.2}$$

while we were concerned about bias in (3.1)?

Answer:

Because Bertrand and Mullainathan randomize the implied race on each (fake) résumé (along with the other variables), the race variable in their study is uncorrelated with the other variables that affect callbacks. Thus, even if we omit 'important' variables (for predicting callback), they are uncorrelated with our variable of interest (race), and thus they will not cause omitted-variable bias.

Description of variables and names

Variable	Description
`i_callback`	Binary variable (0,1) for whether the resume received a callback.
`n_jobs`	Number of previous jobs listed on the application.
`n_expr`	Number of years of experience listed on the application.
`i_military`	Binary variable for whether the application included military status.
`i_computer`	Binary variable for whether the application included computer skills.
`first_name`	The first name listed on the application.
`sex`	The implied sex of the first name on the application (`'f'` or `'m'`).
`i_female`	Binary indicator for whether the implied sex was female.
`i_male`	Binary indicator for whether the implied sex was male.
`race`	The implied race of the first name on the application (`'b'` or `'w'`).
`i_black`	Binary indicator for whether the implied race was African American.
`i_white`	Binary indicator for whether the implied race was White.

In general, I've tried to stick with a naming convention. Variables that begin with `i_` denote binary indicatory variables (taking on the value of `0` or `1`). Variables that begin with `n_` are numeric variables.