Problem Set 4

Econometrics Review

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on Sunday, 21 April 2019

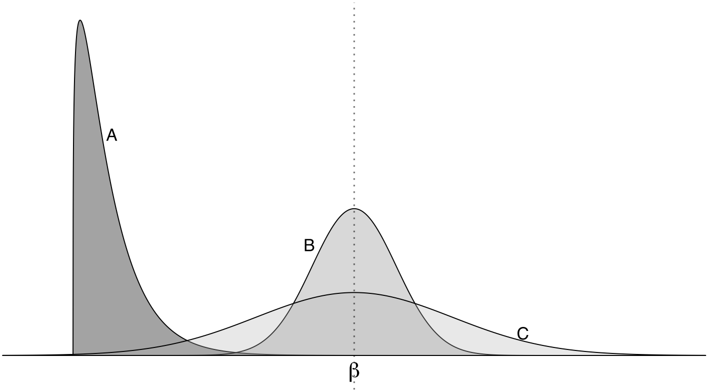
DUE Your solutions to this problem set are due *before* 11:59pm on Sunday, 21 April 2019 on Canvas. **Your answers must include two files** (**1**) your responses/answers to the question (*e.g.*, a Word document) and (**2**) the R script you used to generate any answers in R. Each student must turn in her/his own answers.

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: Bias and variance

1a. Throughout this course, we will use the OLS estimator $\hat{\beta}$ to estimate β. Explain what it means for $\hat{\beta}$ to be biased for β.

Figure 1



Note This figure shows the distributions of three estimators (A, B, and C) that each estimate the unknown parameter β. E[A]= $\beta - 3$, E[B]= β, E[C]= β

1b. Which of the estimators in Figure 1 (above) are unbiased? *Hint:* There may be more than one.

1c. Which of the estimators in Figure 1 (above) has the minimum variance?

1d. Which of the estimators in Figure 1 (above) is the best (minimum variance) unbiased estimator?

1e. Suppose we want to estimate the effect of advertising on sales. Explain what it *bias* would mean in this context.

1f. What does the term "standard error" mean?

Problem 2: Getting started

Now we will start exploring data in R.

README! The data[†] in the next two sections of 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.

2a. Open up RStudio, start a R new script (File ➡ New file ➡ R Script). You will hand in this script as part of your assignment.

2b. Load the the `pacman` package. If you haven't installed it, you will need first install it (`install.packages("pacman")`) *and then* load it (`library(pacman)`).

Now use `pacman`'s function `p_load` to load the `tidyverse` package, *i.e.*,

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

Note: If `tidyverse` is not already installed, then `p_load(tidyverse)` will automatically install it for you—this is why we're using `pacman`.

2c. Download the dataset (Canvas link). Save it in a helpful location. Remember this location.

2d. Read the data into R. What are the dimensions of the dataset (numbers of rows and columns)?

Hints: The `read_csv()` function reads CSVs into R, *e.g.*, `read_csv("file.csv")`. The `dim()` function will tell you the dimensions of a dataset, *e.g.*, `dim(some_data)`.

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.

2e. What are the names of the first five variables? *Hint:* `names(your_df)`

2f. What are the first two *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`.

[†]: 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 3: 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.

3a. 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. *E.g.*, `mean(call_df$female)` would give us the percentage of female individuals in our dataset `call_df`.

3b. 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 obsevations whose value of `race` is `"b"`.

3c. What is the difference in the groups' mean callback rate?

3d. Based upon the difference in percentages that we observe in **3b.**, can we conclude that employers consider race in hiring decisions? **Explain your answer.**

3e. 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).

3f. 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 **3c**?

Hint: Use `lm(y ~ x, data = our_df)` to regress `y` on `x` from datatset `our_df`.

3g. Conduct a *t* test for the coefficient on `i_black` in the regression above in **3f**. Write our your hypotheses (both H_0 and H_A), the test statistic, the *p*-value, the result of your test (*i.e.*, reject or fail to reject H_0), and your conclusion.

3h. 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`).

Problem 4: 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 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?

4a. Define omitted-variable bias.

4b. The central questions here is "Do employers call back individuals at different rates based upon their race (or gender)?".

If we collected data on callbacks and race, and we then regressed *Callback* on *Race*, we would likely get biased estimates due to omitted-variable bias.

Explain why this is the case and provide an example of an omitted variable in this situation.

4c. The point of experiments is to avoid omitted-variable bias. Explain how randomizing race on these (fake) résumés avoided the concerns for 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.