Problem Set 1 *Solutions*

Econometrics Review

EC 421: Introduction to Econometrics

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

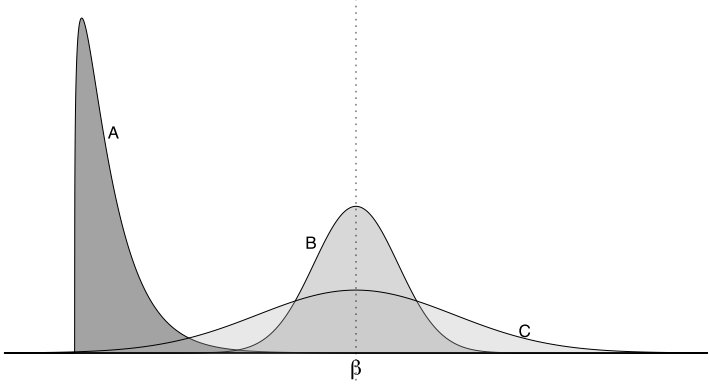
Problem 1: Bias and variance

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.

Answer If $\hat{\beta}$ is biased for β, then, on average, $\hat{\beta}$ does not return β as its estimate.

Formally, $\hat{\beta}$ is biased for β if $E\left[\hat{\beta}\right] \neq \beta$.

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.
Answer B and C

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

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

1e. Suppose we want to estimate the effect of advertising on sales. Explain what *bias* would mean in this context.
Answer Bias would mean our estimate for the effect of advertising on sales routinely misses the actual effect on sales (over-estimating or under-estimating the effect).

1f. What does the term "standard error" mean?
Answer *Standard error* gives the standard deviation of an estimator's distribution (helping us understand how noisy or precise an estimator is).

Problem 2: Getting started

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.
Answer Nothing to show.

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.*,

```
# 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`.
Answer Nothing to show.

2c. Download the dataset (Canvas link). Save it in a helpful location. Remember this location.
Answer Still nothing to show.

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)`.

Answer The dataset has 4,870 observations (rows) on 12 variables (columns).

```
# Read in the data
ps1_df <- read_csv("ps01_data.csv")
# Dimensions of the dataset:
# 1. Printed to screen (since it's a tibble)
ps1_df
```

```
#> # A tibble: 4,870 x 12
#>   i_callback n_jobs n_expr i_military i_computer first_name sex   i_female
#>        <dbl>  <dbl>  <dbl>      <dbl>      <dbl> <chr>      <chr>    <dbl>
#> 1          0      2      6          0          1 Allison    f            1
#> 2          0      3      6          1          1 Kristen    f            1
#> 3          0      1      6          0          1 Lakisha    f            1
#> # … with 4,867 more rows, and 4 more variables: i_male <dbl>, race <chr>,
#> #   i_black <dbl>, i_white <dbl>
```

```
# 2. Use dim()
dim(ps1_df)
```

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

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

Answer `i_callback`, `n_jobs`, `n_expr`, `i_military`, `i_computer`

```
names(ps1_df) %>% head(5)
```

```
#> [1] "i_callback" "n_jobs"     "n_expr"     "i_military" "i_computer"
```

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`.

Answer Allison and Kristen

```
head(ps1_df$first_name, 2)
```

```
#> [1] "Allison" "Kristen"
```

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`.

Answer Approximately 8.05 percent of résumés received callbacks.

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

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

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"`.

Answer Résumés with typically black names received a callback rate of approximately 6.4%, while white-sounding names received a callback rate of approximately 9.7%. This disparity is consistent with employers considering race when responding to résumés.

```
# 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
```

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

Answer The callback rate for résumés with black-sounding was approximately 3.2 percentage points lower than the rate for white-sounding names.

```
# 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
```

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.**

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

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).

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)))
```

```
#> [1] -4.108412
```

```
# The p value
2 * pnorm(abs(z_stat), lower.tail = F)
```

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

The *t* statistic testing the null hypothesis of no difference between the two groups callback percentages is approximately -4.11, which has a *p*-value of approximately 0.00004. Because this *p*-value is smaller than our chosen level of 0.05, we reject the null hypothesis. We conclude there is statistically significant evidence of differential callbacks for black- and white-sounding names.

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`.

Answer

```
lm(i_callback ~ i_black, data = ps1_df) %>% summary()
```

```
#>
#> Call:
#> lm(formula = i_callback ~ i_black, data = ps1_df)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -0.09651 -0.09651 -0.06448 -0.06448  0.93552
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  0.096509   0.005505  17.532  < 2e-16 ***
#> i_black     -0.032033   0.007785  -4.115 3.94e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.2716 on 4868 degrees of freedom
#> Multiple R-squared:  0.003466,    Adjusted R-squared:  0.003261
#> F-statistic: 16.93 on 1 and 4868 DF,  p-value: 3.941e-05
```

This regression finds (exactly) the same difference.

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.

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 names' races affected callback rates for names with black or white connotations.

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`).

Answer

```
lm(i_callback ~ i_black + n_expr + i_black:n_expr, data = ps1_df) %>% summary()
```

```
#>
#> Call:
#> lm(formula = i_callback ~ i_black + n_expr + i_black:n_expr,
#>     data = ps1_df)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -0.17797 -0.09011 -0.07620 -0.05874  0.95695
#>
#> Coefficients:
#>                   Estimate Std. Error t value Pr(>|t|)
#> (Intercept)     0.0692625  0.0101234   6.842 8.79e-12 ***
#> i_black        -0.0293537  0.0143684  -2.043  0.04111 *
#> n_expr          0.0034682  0.0010822   3.205  0.00136 **
#> i_black:n_expr -0.0003304  0.0015409  -0.214  0.83025
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.2712 on 4866 degrees of freedom
#> Multiple R-squared:  0.007231,    Adjusted R-squared:  0.006619
#> F-statistic: 11.81 on 3 and 4866 DF,  p-value: 1.045e-07
```

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 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.

Answer Omitted-variable bias is a specific type of *bias* in our OLS estimates that occurs when we omit a variable that (1) correlates with one of the correlated variables and (2) affects our outcome variable.

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.

Answer Using real-world collected data on callbacks and race, we would expect to have omitted-variable bias if black applicants and white applicants differ on any variable that also affects callbacks. For example, if black applicants and white applicants differ in their social networks, and their networks help them obtain callbacks, then we will misattribute the effect of social networks to race.

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.

Answer By randomizing the perceived race, the experiment breaks the correlation between race and any other variables—omitted or otherwise. By breaking this correlation, the experiment avoids 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.