Tarea 3 de Microeconometría Aplicada: RCTs

Miguel Lerdo de Tejada Flores

viernes 30, abril 2021

Bertrand y Mullainathan (2004) estaban interesados en determinar el grado de discriminación racial que podía darse en el mercado laboral en Estados Unidos. Para esto utilizaron un experimento aleatorizado. Su experimento consistió en preparar CVs ficticios de diversa calidad. A cada CV se le asignaría de manera aleatoria un nombre. Los nombres utilizados en su experimento fueron nombres que se utilizan mayoritariamente entre afroamericanos (e.g. Tanisha y Hakeem) y otros que se utilizan mayoritariamente entre blancos (e.g. Allison y Todd). Los CVs fueron enviados como respuesta a distintos anuncios que fueron publicados en periódicos. La idea era ver la diferencia en llamadas para entrevista (**call_back**) que podían recibir CVs con nombres afroamericanos versus CVs con nombres de blancos.

Para esta tarea utilizarás la base de datos **Names.dta** que está disponible en **Canvas**. De igual manera, en **Canvas** encontrarás la descripción de las variables de la base de datos en el archivo **nombres_des.pdf**.

1. Por qué era importante para los autores aleatorizar los nombres? Es decir, ¿por qué los investigadores no recopilaron información de postulantes verdaderos a los trabajos y codificaron si los nombres de dichas aplicaciones están más asociados a afroamericanos o blancos? ¿Qué sesgo (positivo o negativo) crees que hubiera resultado de seguir esta estrategia?

En primer lugar para evitar una sobreestimación del efecto de la percepción de los nombres y controlar desde le diseño del experimento. Por ejemplo, los afroamericanos tienen mayor incidencia criminal en promedio por lo que una menor cantidad de llamadas podría ser el resultado de que los empleadores prefieren a candidatos con pocos o nulos problemas con la ley, o también las mujeres afroamericanas tienen mayor cantidad de hijos y los empleadores tienden a rechazar a candidatas mientras más probable sientan que les tendrán que dar ausencia por maternidad, por lo que se confundiría el efecto de la discriminación por género con el de la discriminación por raza.

Además, podría haber un sesgo de selección donde en su mayoría se postulan los afroamericanos que de verdad creen que serán llamados de vuelta lo que reduciría el efecto de la discriminación en los datos.

2. Utiliza la base de datos para dar evidencia que la asignación de nombres parece haber sido aleatoria. Deberás incluir la(s) tabla(s) relevante(s) que te haya(n) permitido llegar a esta conclusión.

variables1	Media_control1	Media_trat1	p_value1
bankreal	0.085	0.085	1.000
busservice	0.268	0.268	1.000
chicago	0.555	0.555	1.000
college	0.716	0.723	0.610
compreq	0.437	0.437	0.977
computerskills	0.809	0.832	0.030
comreq	0.125	0.125	1.000
educreq	0.107	0.107	1.000
email	0.479	0.480	0.954
empholes	0.450	0.446	0.773
eoe	0.291	0.291	1.000
expreq	0.435	0.435	1.000
female	0.764	0.775	0.377
high	0.502	0.502	1.000
honors	0.054	0.051	0.654
manager	0.152	0.152	0.968
manuf	0.083	0.083	1.000
military	0.092	0.102	0.266
missind	0.165	0.165	1.000
offsupport	0.119	0.119	1.000
ofjobs	3.664	3.658	0.860
orgreq	0.073	0.073	1.000
othservice	0.155	0.155	1.000
req	0.787	0.787	1.000
retailsales	0.168	0.168	1.000
salesrep	0.151	0.151	1.000
secretary	0.333	0.333	0.976
specialskills	0.330	0.327	0.831
supervisor	0.077	0.077	1.000
trade	0.214	0.214	1.000
transcom	0.030	0.030	1.000
volunteer	0.409	0.414	0.684
workinschool	0.558	0.561	0.840
yearsexp	7.856	7.830	0.854

Note:
En la especificación black vs Detailed information on resume
El estadístico F correspodiente es: 0.67

Hay evidencia de que la asignación sí fue aleatoria ya que la difer-

encia de medias entre afroamericanos y blancos por casi cada variable es significativamente igual a cero. Para esta análisis se excluyeron las variables `firstname` y `expminreq`, la segunda ya que da información ya contenida en otras variables. Finalmente, el único valor p menor de la correspondiente *prueba t* que 0.05 es el de la varialbe `computerskills`, por lo que al parecer la asignación no es balanceada para esa variable. Además, no hay significacia conjunta entre las variables del cv y la variable `black`, ya que dicha especifiación tiene un estadístico F muy pequeño.

3. La variable **black** es una dummy creada por los investigadores para señalar si el nombre es usual de afroamericanos. Asumiendo que la distribución de nombres fue aleatoria, da evidencia de si existe discriminación racial en el **call_back** utilizando: (i) un estimador de Neyman, (ii) una estimación de OLS con errores heterocedásticos y (iii) una estimación de OLS con errores heterocedásticos y agregando controles (ustedes deberán decidir cuáles).

- Indica la prueba de hipótesis que estarás contrastando en cada estimación.

  Para el caso de estadístico de **Neyman**, contrastaré la prueba:

$$H_0: \ \tau^{Neyman} = 0 \ v.s. \ H_a: \ \tau^{Neyman} < 0$$

  Mientras que para **OLS**, donde tengo la especificación

$$call\_back_i = \beta_0 + \beta_1 black_i$$

contrastaré

$$H_0: \ \beta_1 = 0 \ v.s. \beta_1 < 0$$

Finalmente, para **OLS+controles** controlaré únicamente por `computerskills`. Como se puede ver en la tabla de balance, es la única cuya diferencia de medias $computerskills_i^{black=1} - computerskills_i^{black=0}$ es significativa. La anterior intuición fue comprobada con un processo de *Double LASSO* para escoger los controles adecuados. Como se detalla aquí y siguiendo a Urminsky, O. et. al. (2016)[1], corro una regresión LASSO que selecciona las variables que mejor explican a `call_back`, luego un LASSO para encontrar las variables que mejor explican a `black` y de la intersección de ambas listas solo sobrevive `computerskills`.

Entonces la especificación es

$$call\_back_i = \beta_0 + \beta_1 black_i + \beta_2 computerskills_i$$

con la correspondiente prueba de hipótesis

$$H_0: \ \beta_1 = 0 \ v.s. \beta_1 < 0$$

El LASSO para *call_back* selecciona 26 variables mientras que para *black* selecciona únicamente una variable.

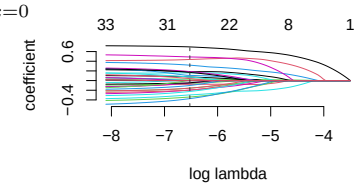

Figure 1: LASSO para callback

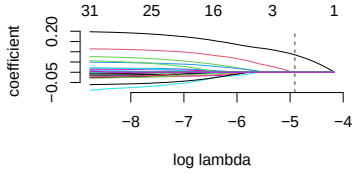

Figure 2: LASSO para black

[1] Urminsky, O., Hansen, C., & Chernozhukov, V. (2016). Using double-lasso regression for principled variable selection. Available at SSRN 2733374.

- Reporta los resultados de tus 3 estimaciones con una tabla con el formato usual que hemos empleado en el semestre.

Model:	Neyman (1)	OLS (2)	OLS + control (3)
Variables			
black	-0.0320***	-0.0320***	-0.0316***
	(0.0078)	(0.0078)	(0.0078)
computerskills			-0.0192*
			(0.0108)
(Intercept)		0.0965***	0.1120***
		(0.0060)	(0.0110)
Fit statistics			
Observations	4,870	4,870	4,870
R^2		0.00347	0.00419
F-test		16.931	10.251

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

- Asegúrate que los resultados reportados en cada columna sean comparables. Es decir, deberán estar reportados en las mismas unidades para poder hacer una comparación a lo largo de las columnas.
- Elige una de las columnas para llevar a cabo una interpretación del coeficiente relevante que estas estimando. Da evidencia como parte de esta interpretación de la importancia del efecto. Es decir, ¿consideras que es un efecto pequeño o grande?

El coeficiente de la columna 2, del método de *OLS* nos dice que C.P. que un cv tenga nombre típico afroamericano está asociado con un decremento de 0.0320329 en la probabilidad de que reciba una llamada respecto a un cv sin nombre típico afroamericano. En términos de desviaciones estándar, representa -11.773211%.

4. Planteamos ahora una prueba de hipótesis que sugiere que a nivel individual no hay un efecto de la discriminación. Es decir, un individuo **i** recibiría el mismo valor de la dummy **call_back** independientemente si es afroamericano o blanco:

$$H_0 : \ CB_{i,blanco} = CB_{i,afroam}$$

donde $CB_{i,x}$ es una dummy igual a uno si el individuo **i** de raza **x** recibió una llamada para entrevista. Utiliza un **Fischer Exact Test** para evaluar esta hipótesis. Emplea la media como estadístico para evaluar esta hipótesis. ¿Qué representa la media de las dummies? Reporta el **valor-p** y la conclusión a la que llegas.

El promedio de las dummies representa la probabilidad de que la dummy tome el valor de 1. Entonces representa la probabilidad de que a ese cv lo hayan llamado de vuelta. . Al simular la asignación del tratamiento (i.e. simular una nueva asignación de la variable `black`) y calcular la diferencia de medias entre tratamiento y control $\bar{\tau}_j^T - \bar{\tau}_j^C$ 900 veces, recordarno el valor observado del estadístico de Neyman es -0.0320329 y calculando el p-value

$$p - value \equiv \frac{1}{900} \sum_{j=1}^{900} 1\{|\bar{\tau}_j^T - \bar{\tau}_j^C| \geq |\tau^{Neyman}|\}$$

me da un valor de 0.

Igual al correr una prueba de diferencia de medias con la función **twoSamplePermutationTestLocation**, el pvalue también es 0. Entones podemos concluir que el tratamiento sí tiene un efecto i.e. que tener un nombre típico afroamericano sí impacta la probabilidad e que te llamen para un trabajo respecto a tener un nombre típico de blanco.

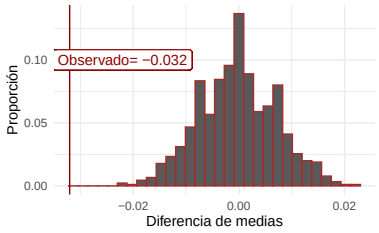

Figure 3: Simulación de asignaciones de black

5. Imagina que estratificas por: (i) sexo del aplicante (hombre o mujer), (ii) ciudad donde se postula al trabajo (Chicago o Boston) e (iii) industria de la empresa que publico el puesto (ver el pdf que indica las industrias disponibles) [Ejemplo: un posible estrato sería hombres aplicantes a trabajos en Chicago en la industria manufacturera]. Empleando todas las combinaciones posibles de las variables (i)-(iii), utiliza el método de Neyman para calcular el efecto de discrminación en cada estrato (elige el formato que quieras para reportar este resultado, tabla o gráfica). Utilizando los efectos por estrato, calcula el efecto promedio de tratamiento. Compara este estimador promedio y la varianza con el resultado que obtuviste en la pregunta (3).

El estimador, que es un promedio ponderado de las última columna de la tabla, toma el valor -0.0328092 que es muy similar al de las preguntas anteriores, pero tiene una varianza estimada de 0.0000603 que es ligeramente menor. Recordemos que en la pregunta 3 obtuvimos que τ^{Neyman} =-0.0320329 y $var(\tau^{Neyman})$ =‘0.0000606.

city	sex	industry	mean_w	mean_b	n_w	n_b	tau
boston	man	unknown	0.11	0.05	81	95	-0.06
boston	man	other services	0.00	0.00	10	8	0.00
boston	man	bussiness and personal services	0.09	0.06	81	78	-0.02
boston	man	commerce	0.06	0.04	145	135	-0.01
boston	man	finance,insurance or real state	0.25	0.00	8	8	-0.25
boston	man	transport/communication	0.16	0.06	25	16	-0.10
boston	man	manufacturing	0.02	0.03	52	39	0.01
chicago	man	unknown	0.50	0.14	6	7	-0.36
chicago	man	other services	0.00	0.00	17	14	0.00
chicago	man	bussiness and personal services	0.17	0.11	30	36	-0.06
chicago	man	commerce	0.13	0.06	87	78	-0.06
chicago	man	finance,insurance or real state	0.00	0.07	9	14	0.07
chicago	man	transport/communication	0.12	0.25	8	8	0.12
chicago	man	manufacturing	0.00	0.08	16	13	0.08
boston	woman	unknown	0.10	0.11	128	114	0.01
boston	woman	other services	0.19	0.18	117	119	-0.01
boston	woman	bussiness and personal services	0.16	0.09	199	202	-0.07
boston	woman	commerce	0.10	0.06	125	135	-0.04
boston	woman	finance,insurance or real state	0.14	0.05	44	44	-0.09
boston	woman	transport/communication	0.09	0.12	23	32	0.04
boston	woman	manufacturing	0.18	0.00	45	58	-0.18
chicago	woman	unknown	0.05	0.04	187	186	-0.02
chicago	woman	other services	0.09	0.06	233	236	-0.03
chicago	woman	bussiness and personal services	0.07	0.04	342	336	-0.03
chicago	woman	commerce	0.08	0.05	164	173	-0.03
chicago	woman	finance,insurance or real state	0.09	0.04	146	141	-0.05
chicago	woman	transport/communication	0.11	0.22	18	18	0.11
chicago	woman	manufacturing	0.06	0.07	89	92	0.01

6. Replica la primera sección de la **Tabla 7** del paper. En lugar de realizar la estimación con un Probit, realízala con un Modelo de Probabilidad Lineal (MPL) utilizando como controles las variables indicadas en la nota de la **Tabla 7**[2]. Solo para el renglón de "Total Number of Requirements" da una interpretación lo más específica posible de la columna "marginal effects."

[2] En tu estimación utiliza errores heterocedásticos.

Table 1: Effect of job requirement on racial difference in call- backs

Job requirement	sample mean (standard deviation)	Marginal effect on call-backs for African-American names
Any?	0.79 (0.41)	0.03 (0.02)
Experience?	0.44 (0.5)	0.01 (0.02)
Computer skills?	0.44 (0.5)	0.01 (0.02)
Communication skills?	0.12 (0.33)	0 (0.02)
Organization skills?	0.07 (0.26)	0.03 (0.02)
Education?	0.11 (0.31)	−0.02 (0.02)
Total number of skills	1.18 (0.93)	0.01 (0.01)

Para el renglón de `total number of requirements`, el coeficiente del modelo de probabilidad lineal corrí la especificación

$$call\_back_i = \beta_0 + \beta_1 black_i + \beta_2 Tot\_Reqs_i + \beta_3 black_i * Tot\_reqs_i$$

y el coeficiente de interés es β_3. Nos indica C.P. que un aumento de 1 en la cantidad de requerimientos incrementa en 0.01 la probabilidad de que llamen al cv para cvs con nombres de afroamericanos respecto a cvs con nombres de blancos.

7. Quisieras saber si la discriminación racial disminuye conforme aumenta la experiencia laboral de los aplicantes. Elige el método y formato que prefieras para reportar tus resultados. Muestra claramente qué parámetro o combinación de parámetros contestan tu pregunta.

El coeficiente de interés en la especificación

$$call\_back_i = \beta_0 + \beta_1 black_i + \beta_2 yearsexp_i + \beta_3 computerskills_i + \beta_4 black_i * yearsexp_i$$

es β_4 que no es estadísticamente distinto de cero. Si sí lo fuera, podríamos decir que el efecto de un año extra de experiencia laboral tiene un impacto diferenciado y *significativo* entre afroamericanos y blancos. En realidad si es que hay un impacgto diferenciado, no termina por ser significativo por lo que los años de experiencia laboral no terminan por disminuir ni acrecentar la discriminación laboral contra afroamericanos. Cabe notar que controlo por `computerskills` como en incisos anteriores.

Table 2: Mayor experiencia laboral no disminuye la discriminación racial

Dependent Variable:	call_back
Model:	(1)
Variables	
(Intercept)	0.08***
	(0.01)
black	-0.03**
	(0.01)
yearsexp	0.003***
	(0.001)
computerskills	-0.02
	(0.01)
black × yearsexp	-0.0003
	(0.002)
Fit statistics	
Observations	4,870
R^2	0.00772
F-test	9.4677

Normal standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

8. Por último, imagina que el gobierno esta interesado en replicar este estudio en México para ver posible discriminación en contra de indígenas. Te pide que lo asesores para definir el número de CVs ficticios (aplicaciones) que necesita realizar. Realiza cálculos de poder para indicar:

- Cuántos CVs ficticios necesitaría aleatorizar si es que: (i) tu anticipas que los efectos (varianza y efecto real) sean iguales a los obtenidos por Bertrand y Mullainathan, (ii) quieres un poder estadístico de 85%, (iii) asumes una significancia de 1%, y (iv) vas a dividir 50-50 tratamiento y control?

  La fórmula utilizada es

$$n_{min} = \frac{(\Phi^{-1}(\psi) + \Phi^{-1}(1 - \frac{\alpha}{2}))^2}{\frac{\tau^2}{\sigma^2}\gamma(1-\gamma)}$$

donde α es la significancia deseada, γ la proporción del grupo de tratamiento, τ el efecto del tratamiento, σ la varianza de la variable objetivo y ψ es el poder estadístico. Al evaluar obtenemos n_{min}=3766.

- En R o Stata, produce una gráfica que ilustre el tradeoff entre poder estadístico y proporción de tratamiento y control (similar a lo que hicimos con **Optimal Design**) fijando los valores que obtuviste

en el inciso anterior (número de observaciones, efectos reales y significancia).