Instituto Tecnológico Autónomo de México

Econometría Aplicada

Tarea 1

Prof. Arturo Aguilar Esteva

Marco Antonio Ramos Juárez

142244

Ciudad de México 2021

Contents

Ambiente de trabajo

```
# Cargo las librerías
library(readstata13)
library(RCT)
library(sandwich)
library(EnvStats)
library(tidyr)
library(dplyr)
library(kableExtra)
library(stargazer)
library(margins)
library(ggplot2)
options(scipen = 999)

# Carga la base de datos
data <- read.dta13("names.dta")
```

1.

¿Por qué era importante para los autores aleatorizar los nombres? Es decir, ¿porqué los investigadores no recopilaron información de postulantes verdaderos a los trabajos y codificaron si los nombres de dichas aplicaciones están más asociadosa afroamericanos o blancos?

Porque *a priori* puede que los nombres de personas reales tengan alguna relación con las variables de interés, como la calidad de los CVs, las aptitudes, la educación, la experiencia, etc. De esta manera no se estaría evaluando el impacto del nombre per se si no la relación entre toda la combinación de variables y el outcome.

Al construir observaciones de manera sintética, los investigadores se están asegurando que el impacto que van a analizar es meramente el de los nombres, controlando por cualquier otro posible sesgo.

¿Qué sesgo (positivo o negativo) crees que hubiera resultado de seguir esta estrategia?

Recordemos que el efecto promeido observado se descompone de la siguiente manera:

$$D = E[Y_i^T - Y_i^C|T] + (E[Y_i^C|T] - E[Y_i^C|C])$$

$$D = RealEffect + SelectionBias$$

De haber seguido la otra estrategia, muy probablemente $E[Y_i^C|T] < E[Y_i^C|C]$, es decir el outcome seguiría siendo más favorable para los nombres blancos en el escenario hipotético donde los blancos fueran negros. El signo del sesgo es el mismo que el signo del efecto real: ambos son negativos. El sesgo estaría inflando el efecto promedio observado.

2.

Utiliza la base de datos para dar evidencia que la asignación de nombres parece habersido aleatoria.

```
# Para este inciso debemos crear algunas tablas de balance.
# Sin embargo, debemos hacer una pequeña limpieza previa.
# Una para información de cada supuesto individuo, otra para
# información del cv y otra para información de las empresas.
# Para esto necestiamos que las variables que necesitamos
# sean númericas, lo cual no es problema excepto por una
# columna, 'expminreq' a la cual le haremos una pequeña
# limpieza.

summary(factor(data$expminreq))
```

```
    0  0.5    1   10    2    3    4    5    6    7    8 some
```

2746 4 8 142 18 356 331 8 163 8 12 10 1064

```
# Como podemos observar cuenta con 2746 valores nulos (que
# aparecen con un espacio en blanco ' ') y 1064 valores con
# la categoría 'some'. Esto nos puede dar problemas a la hora
# de hacer las tablas de balance: tenemos dos alternativas
# A)ignorar la variable o B) intentar imputar los valores de
# some con el promedio de los valores numéricos e imputar con
# un 0 los valores nulos (asumiendo que su no respuesta
# indica que no se necesitaba esperiencia mínima)

# promedio
(0 * 4 + 8 * 0.5 + 1 * 142 + 10 * 18 + 2 * 356 + 3 * 331 + 4 *
    8 + 5 * 163 + 6 * 8 + 7 * 12 + 8 * 10)/(4 + 8 + 142 + 18 +
    356 + 331 + 8 + 163 + 8 + 12 + 10)
```

```
[1] 2.915094
```

```r
# El promedio de experiencia que soliticaban las vacantes que
# si respondieron ese inciso fue de 2.91 y la moda de 2
# (seguido de 3). En este sentido, creo que el promedio es un
# buen indicador de 'some' por lo que imputamos el valor.

data$expminreq[data$expminreq == " "] <- NA
data$expminreq[is.na(data$expminreq)] <- 0
data$expminreq[data$expminreq == "some"] <- 2.91
data$expminreq <- as.numeric(data$expminreq)
```

```r
# Siguiendo con el inciso, el primer paso es dividir la base
# de datos en 3 bloques con solo las variables que
# necesitamos

# A) información personal
personal_info <- data %>% select(black, female, high, chicago)
# Aquí omitimos callback pues esa es información del outcome
# y firstname, pues el treatment es black

# B) información del cv
cv_info <- data %>% select(ofjobs, yearsexp, honors, volunteer,
    military, empholes, workinschool, email, computerskills,
    specialskills, college, black)
# C) información del empleador
employer_info <- data %>% select(expminreq, eoe, manager, supervisor,
    secretary, offsupport, salesrep, retailsales, req, expreq,
    comreq, educreq, compreq, orgreq, manuf, transcom, bankreal,
    trade, busservice, othservice, missind, black)

# Creo las tablas de balance
bt_personal <- balance_table(personal_info, treatment = "black")
bt_cv <- balance_table(cv_info, treatment = "black")
bt_employer <- balance_table(employer_info, treatment = "black")

# Imprimo las tablas finales
kable(as.data.frame(bt_personal), booktabs = T, caption = "Datos personales",
    longtable = T) %>% kable_styling(position = "center", latex_options = "repeat_header")
```

Table 1. Datos personales

variables1	Media_control1	Media_trat1	p_value1
chicago	0.5552361	0.5552361	1.0000000
female	0.7638604	0.7745380	0.3766694
high	0.5022587	0.5022587	1.0000000

```
kable(as.data.frame(bt_cv), booktabs = T, caption = "Datos del CV",
    longtable = T) %>% kable_styling(position = "center", latex_options = "repeat_header")
```

Table 2. Datos del CV

variables1	Media_control1	Media_trat1	p_value1
college	0.7162218	0.7227926	0.6098859
computerskills	0.8086242	0.8324435	0.0303271
email	0.4788501	0.4796715	0.9542638
empholes	0.4501027	0.4459959	0.7732856
honors	0.0542094	0.0513347	0.6537665
military	0.0924025	0.1018480	0.2658137
ofjobs	3.6644764	3.6583162	0.8600712
specialskills	0.3301848	0.3273101	0.8309558
volunteer	0.4086242	0.4143737	0.6835925
workinschool	0.5581109	0.5609856	0.8399154
yearsexp	7.8562628	7.8295688	0.8535350

```
kable(as.data.frame(bt_employer), booktabs = T, caption = "Datos del empleador",
    longtable = T) %>% kable_styling(position = "center", latex_options = "repeat_header")
```

Table 3. Datos del empleador

variables1	Media_control1	Media_trat1	p_value1
bankreal	0.0850103	0.0850103	1.0000000
busservice	0.2677618	0.2677618	1.0000000
compreq	0.4369610	0.4373717	0.9769596
comreq	0.1248460	0.1248460	1.0000000
educreq	0.1067762	0.1067762	1.0000000
eoe	0.2911704	0.2911704	1.0000000
expminreq	2.9116008	2.9134840	0.9711680
expreq	0.4353183	0.4353183	1.0000000

Table 3. Datos del empleador *(continued)*

variables1	Media_control1	Media_trat1	p_value1
manager	0.1523614	0.1519507	0.9681839
manuf	0.0829569	0.0829569	1.0000000
missind	0.1650924	0.1650924	1.0000000
offsupport	0.1186858	0.1186858	1.0000000
orgreq	0.0726899	0.0726899	1.0000000
othservice	0.1548255	0.1548255	1.0000000
req	0.7872690	0.7872690	1.0000000
retailsales	0.1679671	0.1679671	1.0000000
salesrep	0.1511294	0.1511294	1.0000000
secretary	0.3326489	0.3330595	0.9757473
supervisor	0.0772074	0.0772074	1.0000000
trade	0.2139630	0.2139630	1.0000000
transcom	0.0303901	0.0303901	1.0000000

```
# Caber notar que la columna que limpiamos, expminreq,
# resulto balanceada si imputabamos la moda, la media, sin
# tratar los NAs y tratando los NAs,
```

Deberíass incluir la(s) tabla(s) relevante(s) que te haya(n) permitido llegar a esta conclusión.

En resumen, solamente en una variable de las 3 tablas anteriores encontramos una diferencia de medias estadísticamente significativa (*computerskills*). Sin embargo, las medias son muy parecidas (.809 para el control y .832 para el tratamiento). Es decir que esta mínimamente favoreciendo a los nombres negros. Fuera de ese detalle, todo lo demás parece indicar que la asignación de nombres fue aleatoria.

3.

Asumiendo que la distribución de nombres fue aleatoria, da evidencia de si existe discriminació racial en elcallback utilizando: (i)un estimador de Neyman, (ii) una estimación de OLS con errores heterocedásticos,(iii) una estimación de OLS agregando controles (ustedes deberán decidir cuáles) y(iv) un probit sin controles.

- **Estimador de Neyman**

Recordemos:

$$\tau = \overline{y}_i^T - \overline{y}_i^C$$

$$Var[\tau] = \frac{S_T^2}{N_T} + \frac{S_C^2}{N_C}$$

```
# para calcular el estimador de Neyman simplemente restamos
# las dos medias condicionales

# Calculo del estimador de Neyman
mean_control <- mean((data %>% filter(black == 0))$call_back)

mean_treatment <- mean((data %>% filter(black == 1))$call_back)

# Estimador de Neyman:
(neyman_t <- mean_treatment - mean_control)
```

[1] -0.03203285

```
# Calculo de la Varianza
var_control <- var((data %>% filter(black == 0))$call_back)

var_treatment <- var((data %>% filter(black == 1))$call_back)

n_control <- length((data %>% filter(black == 1))$call_back)

n_treatment <- length((data %>% filter(black == 1))$call_back)



# Varianza del estimador de Neyman
(var_neyman_t <- var_treatment/n_treatment + var_control/n_control)
```

[1] 0.00006060575

```
# Desviación estandar de Neyman
(var_neyman_t^0.5)
```

[1] 0.007784969

- **Robust OLS**

```
# para calcular el modelo con errores heterocedásticos
# primero calculo el modelo sin errores heterocedásticos y
# luego calculo los errores con el paquete sandwich
```

```
# con errores homocedásticos
(ols_con_eh <- lm(call_back ~ black, data))
```

Call: lm(formula = call_back ~ black, data = data)

Coefficients: (Intercept) black
0.09651 -0.03203

```
# con corrección heterocedástica
ols_con_eh %>% vcovHC() %>% diag() %>% sqrt()
```

(Intercept) black 0.005986531 0.007786568

- **Robust OLS con controles**

Agregamos las 3 variables con las diferencias más grandes, con miras a que tal vez estas variables puedan cambiar el coeficiente de black.

```
# con errores homocedásticos
(ols_con_controles <- lm(call_back ~ black + computerskills +
    female + military, data))
```

Call: lm(formula = call_back ~ black + computerskills + female + military, data = data)

Coefficients: (Intercept) black computerskills female military
0.10535 -0.03154 -0.02118 0.01245 -0.01321

```
# con corrección heterocedástica
ols_con_controles %>% vcovHC() %>% diag() %>% sqrt()
```

(Intercept) black computerskills female military 0.011819424 0.007785119 0.011307289 0.009288898 0.011980117

Lo que notamos esque ni aún así cambiamos el coeficiente de black, debido a la aleatorización exitosa de los investigadores.

- **Probit sin controles**

```
# con errores homocedásticos
probit <- glm(call_back ~ black, family = binomial(link = "probit"),
    data)
stargazer(probit, type = "latex", summary = FALSE, header = F)
```

Table 4

	Dependent variable:
	call_back
black	-0.217^{***}
	(0.053)
Constant	-1.302^{***}
	(0.035)
Observations	4,870
Log Likelihood	−1,354.969
Akaike Inf. Crit.	2,713.938
Note:	*p<0.1; **p<0.05; ***p<0.01

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

- **Estimador de Neyman**

Para este caso debo usar la prueba de Neyman

$H_n : \mu_c - \mu_t = 0$ $H_a : \mu_c - \mu_t \neq 0$

- **Robust OLS**

Para este caso puedo realizar una prueba de significancia individual sobre el coeficiente de black que resuelvo con el estadístico t.

$H_n : \beta_{black} = 0$ $H_a : \beta_{black} \neq 0$

- **Robust OLS con controles** Para este caso puedo realizar una prueba de significancia individual el coeficiente de black que resuelvo con el estadístico t.

$H_n : \beta_{black} = 0$ $H_a : \beta_{black} \neq 0$

- **Probit sin controles**

Para el probit, podemos hacer una prueba de hipótesis contrastando H0 que el Efecto Marginal Promedio de la variable black es cero frente a H1 de que no es cero.

$H_n : AME(black) = 0$ $H_a : AME(black) \neq 0$

```
kable(summary(margins(probit, variables = "black")))
```

factor	AME	SE	z	p	lower	upper
black	-0.0321556	0.0078716	-4.085001	0.0000441	-0.0475836	-0.0167275

Reporta los resultados de tus 4 estimaciones con una tabla con el formato usual que empleamos el semestre pasado. Asegúrate que los resultados reportados en cada columna sean comparables. Es decir, deberán estar reportados en las mismas unidades para poder hacer unacomparación a lo largo de las columnas.

```
# Para esta tarea, decidí mejor hacer manualmente mi tabla
# Para ello aprovecho que ya estime todos los modelos y
# calculé los margentes para el probit:

# Primero elaboro mis vectores individuales para hacer mi
# tabla
constant <- c("", "0.097***", " 0.105***", "")
s_dev_constant <- c("", "(0.006)", "(0.011)", "")
estimates <- c("-0.032***", "-0.032***", "-0.032***", "-0.032***")
s_dev <- c("(0.008)", "(0.008)", "(0.008)", "(0.008)")
other_1 <- c("", "", "-0.021**", "")
s_dev_other_1 <- c("", "", "(0.011)", "")
other_2 <- c("", "", "0.012", "")
s_dev_other_2 <- c("", "", "(0.010)", "")
other_3 <- c("", "", "0.013", "")
s_dev_other_3 <- c("", "", "(0.012)", "")

# Elaboro mi data frame
modelos_output <- t(data.frame(constant, s_dev_constant, estimates,
    s_dev, other_1, s_dev_other_1, other_2, s_dev_other_2, other_3,
    s_dev_other_3))

# Cambio el nombre de las filas para poder usar latex

rownames(modelos_output) <- c("Constante", "", "$\\mu^T_Y-\\mu^C_Y \\approx \\beta^{black}$",
    "", "$\\beta^{Computer}$", "", "$\\beta^{Mujer}$", "", "$\\beta^{Militar}$",
    "")

# Realizó mi tabla final:

kable(modelos_output, col.names = c("Neyman", "Robust OLS", "Controled OLS",
    "AME Probit"), longtable = T, booktabs = T, caption = "Comparación de modelos",
```

```
    escape = F) %>%
kable_styling(position = "center", latex_options = "repeat_header") %>%

add_footnote(c("Para hacer el probit comparable solamente se muestra el Efecto Marginal Promedi
    "Se redondeó a 3 digitos "), notation = "symbol")
```

Table 5. Comparación de modelos

	Neyman	Robust OLS	Controled OLS	AME Probit
Constante		0.097***	0.105***	
		(0.006)	(0.011)	
$\mu_Y^T - \mu_Y^C \approx \beta^{black}$	-0.032***	-0.032***	-0.032***	-0.032***
	(0.008)	(0.008)	(0.008)	(0.008)
$\beta^{Computer}$			-0.021**	
			(0.011)	
β^{Mujer}			0.012	
			(0.010)	
$\beta^{Militar}$			0.013	
			(0.012)	

Elige una de las columnas para llevar a cabo una interpretación del coeficienterelevante que estas estimando. Da evidencia como parte de esta interpretación la importancia del efecto. Es decir, ¿consideras que es un efecto pequeño o grande?

Tomo el coeficiente de las dos regresiones con OLS. Ceteris paribus, el solo hecho de que el nombre esté asociado a ser negro reduce la probabilidad de recibir una llamada de regreso por parte del empleador en 3.2%. Lo interesante es que el coeficiente fue virtualmente el mismo con y sin controles debido a la exitosa aleatorización de los investigadores. Asimismo, en ambos modelos resulto ser significativo a más del 99%.

Aunque a simple vista parece una diferencia diminuta, debemos considerar que la probabilidad media de recibir una llamada de regreso del grupo de control (nombres no negros) fue del 9.6% y del grupo de tratamiento del 6.4%. En este sentido, el hecho que la diferencia de medias haya sido de 3.2 quiere decir que el efecto de la discrimación es del 50% con respecto a la probabilidad de una persona afroamericana de recibir la llamada. Desde esta perspectiva si es una situación grave pues esta cifra es teniendo todo lo demás constante (habilidades computacionales, educación, sexo, etc.)

4.

Utiliza un Fischer Exact Testpara evluar la hipótesis. Reporta el valor p y la conclusión a la que llegas.

$H_n : \mu_c - \mu_t = 0.1$

$H_a : \mu_c - \mu_t \neq .1$

```
# para este inciiso usaremos el comado
# twoSamplePermutationTestLocation.  Para usar este comando
# necestiamos crear vectores con las variables de interés
vector_control <- data[data$black == 0, ]$call_back
vector_treatment <- data[data$black == 1, ]$call_back
fet <- twoSamplePermutationTestLocation(vector_control, vector_treatment,
    alternative = "two.sided", mu1.minus.mu2 = 0.01, seed = 123)

# imprimo los inputs del comando
fet$estimate
```

mean of x mean of y 0.09650924 0.06447639

```
# imprimo la hipótesis interpretada por el comando
fet$null.value
```

mu.x-mu.y 0.01

```
# imprimo el valor p
fet$p.value
```

[1] 0.0054

Con más del 99% de significancia, rechazo la hipótesis nula: no existe evidencia para sostener que la diferencia es del 1%.

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 publicó 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].

```
# para calcular el estimador de Neyman simplemente restamos
# las dos medias condicionadas a cada grupo

# calculo la tabla para el grupo de control
```

```
strat_control <- data %>% filter(black == 0) %>% select(call_back,
    female, chicago, manuf, transcom, bankreal, trade, busservice,
    othservice, missind) %>% group_by(female, chicago, manuf,
    transcom, bankreal, trade, busservice, othservice, missind) %>%
    summarise(efecto_control = mean(call_back), obs_control = n(),
        var_c = var(call_back))

# calculo la tabla para el grupo de tratamiento
strat_treatment <- data %>% filter(black == 1) %>% select(call_back,
    female, chicago, manuf, transcom, bankreal, trade, busservice,
    othservice, missind) %>% group_by(female, chicago, manuf,
    transcom, bankreal, trade, busservice, othservice, missind) %>%
    summarise(efecto_treatment = mean(call_back), obs_treatment = n(),
        var_t = var(call_back)) %>% ungroup() %>% select(efecto_treatment,
    obs_treatment, var_t)


# junto las tablas
neyman_strat <- cbind(strat_control, strat_treatment)

# le doy retoques esteticos
neyman_strat <- neyman_strat %>% mutate(neyman = efecto_treatment -
    efecto_control) %>% select(-efecto_control, -efecto_treatment) %>%
    gather(industry, match, 3:9) %>% filter(match == 1) %>% mutate(sex = ifelse(female ==
    1, "female", "male"), city = ifelse(chicago == 1, "chicago",
    "boston"), weights = (obs_treatment + obs_control)/4870) %>%
    ungroup()

neyman_strat_output <- neyman_strat %>% select(sex, city, industry,
    neyman)
```

Empleando todas las combinaciones posibles de las variables (i)-(iii),utiliza el método de Neyman para calcular el efecto de discrminacióon en cada estrato(elige el formato que quieras para reportar este resultado, tabla o gráfica).

```
kable(neyman_strat_output, caption = "Diferencias en medias mediante estratificación",
    booktabs = T, longtable = T) %>% kable_styling(position = "center",
    latex_options = "repeat_header")
```

Table 6. Diferencias en medias mediante estratificación

sex	city	industry	neyman
male	boston	manuf	0.0064103
male	chicago	manuf	0.0769231
female	boston	manuf	-0.1777778
female	chicago	manuf	0.0090376
male	boston	transcom	-0.0975000
male	chicago	transcom	0.1250000
female	boston	transcom	0.0380435
female	chicago	transcom	0.1111111
male	boston	bankreal	-0.2500000
male	chicago	bankreal	0.0714286
female	boston	bankreal	-0.0909091
female	chicago	bankreal	-0.0464879
male	boston	trade	-0.0107280
male	chicago	trade	-0.0623342
female	boston	trade	-0.0447407
female	chicago	trade	-0.0330255
male	boston	busservice	-0.0223172
male	chicago	busservice	-0.0555556
female	boston	busservice	-0.0666700
female	chicago	busservice	-0.0314327
male	boston	othservice	0.0000000
male	chicago	othservice	0.0000000
female	boston	othservice	-0.0115636
female	chicago	othservice	-0.0308067
male	boston	missind	-0.0584795
male	chicago	missind	-0.3571429
female	boston	missind	0.0124726
female	chicago	missind	-0.0158415

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).

Recordemos que calculamos el efecto promedio de la siguiente manera:

$$\overline{\tau} = \sum_{g=1}^{G} \overline{\tau}_g.(\frac{N_g}{N})$$

```
# promedio ponderado
(ate <- sum(neyman_strat$neyman * neyman_strat$weights))
```

[1] -0.03280923

Recordemos que la varianza dentro de cada estrato se calcula de la siguiente manera:

$$V_{estrato}(\tau_g) = \frac{S^2_{t,g}}{N_{t,g}} + \frac{S^2_{c,g}}{N_{c,g}}$$

y la varianza total después de estratificar:

$$V_{total}(\tau) = \sum_{g=1}^{G} V(\tau_g).(\frac{N_g}{N})^2$$

```
neyman_strat <- neyman_strat %>% mutate(variance_group = (var_t/obs_treatment) +
    (var_c/obs_control))

# varianza
(varianza_total_estratificada <- sum(neyman_strat$variance_group *
    ((neyman_strat$weights)^2)))
```

[1] 0.00006030051

```
# desviación estandar
(varianza_total_estratificada^0.5)
```

[1] 0.00776534

Como podemos observar, el valor de τ es el mismo que estimamos anteriormente y la desviación es aproximadamente la misma. Esto debido al buen balance y alteatorización de los datos.

6.

Replica la primera parte de laTabla 7 del paper.

Solo para el renglón de "Total Number of Requirements" da una interpretación lo más específica posible de la columna "marginal effects." (Ojo: Puedes considerar los errores estándar que arroja por default el software que utilices).

```r
# Para este inciso tuve que hacer un poco de rodeo pues no
# supe como calcular el efecto marginal promedio de la
# interacción con el comando margins.  El rodeo consisitó en
# desde la base de datos crear variables con la interacción
# que necesitaré, de esta manera la trato como una variable
# unica.  De esta manera , ahora margins puede calcular el
# efecto marginal promedio solamente de la interacción

# paso 1: creo la base de datos con las interacciones
table_7 <- data %>% select(call_back, black, req, expreq, eoe,
    compreq, comreq, orgreq, educreq) %>% mutate(tot_req = (expreq +
    compreq + comreq + orgreq + educreq), a = req * black, b = expreq *
    black, c = compreq * black, d = comreq * black, e = orgreq *
    black, f = educreq * black, g = tot_req * black)

# paso 2: estimo los modelos probits
A <- glm(call_back ~ black + req + a, family = binomial(link = "probit"),
    table_7)
B <- glm(call_back ~ black + expreq + b, family = binomial(link = "probit"),
    table_7)
C <- glm(call_back ~ black + compreq + c, family = binomial(link = "probit"),
    table_7)
D <- glm(call_back ~ black + comreq + d, family = binomial(link = "probit"),
    table_7)
E <- glm(call_back ~ black + orgreq + e, family = binomial(link = "probit"),
    table_7)
EFE <- glm(call_back ~ black + educreq + f, family = binomial(link = "probit"),
    table_7)
G <- glm(call_back ~ black + tot_req + g, family = binomial(link = "probit"),
    table_7)

# paso 3: extraigo el efecto marginal promedio y su
# desviación estandar

a <- summary(margins(A, variables = "a"))
b <- summary(margins(B, variables = "b"))
c <- summary(margins(C, variables = "c"))
d <- summary(margins(D, variables = "d"))
e <- summary(margins(E, variables = "e"))
```

```
f <- summary(margins(EFE, variables = "f"))
g <- summary(margins(G, variables = "g"))

# los guardo en dos vectores
marginal_avg <- c(a$AME, b$AME, c$AME, d$AME, e$AME, f$AME, g$AME)
marginal_sd <- c(a$SE, b$SE, c$SE, d$SE, e$SE, f$SE, g$SE)

# paso 4: estimo la media y desvación estandar para cada
# variable
table_7 <- table_7 %>% select(req, expreq, compreq, comreq, orgreq,
    educreq, tot_req)

# paso 5: agrupo toda la información en una sola tabla
id = c("Any requirement?", "Experience?", "Computer skills?",
    "Communication skills?", "Organization skills?", "Education?",
    "Total number of requirements")
tabla <- data.frame(id, sapply(table_7, mean), sapply(table_7,
    sd), marginal_avg, marginal_sd)

# paso 6: realizo algunos retoques estéticos

kable(tabla, caption = "Efecto de requisitos en las diferencias raciales
      en el callback",
    col.names = c("Requirement", "Sample mean", "SD", "Marginal effect",
        "SE"), booktabs = T, digits = 3, longtable = T) %>%
kable_styling(position = "center", latex_options = "repeat_header")
```

Table 7. Efecto de requisitos en las diferencias raciales en el callback

	Requirement	Sample mean	SD	Marginal effect	SE
req	Any requirement?	0.787	0.409	0.022	0.018
expreq	Experience?	0.435	0.496	0.010	0.016
compreq	Computer skills?	0.437	0.496	0.001	0.016
comreq	Communication skills?	0.125	0.331	0.000	0.024
orgreq	Organization skills?	0.073	0.260	0.025	0.035
educreq	Education?	0.107	0.309	-0.036	0.030
tot_req	Total number of requirements	1.177	0.932	0.002	0.009

7.

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

```
modelo_interaccion_naive <- lm(call_back ~ yearsexp + I(black *
    yearsexp), data)
modelo_interaccion <- lm(call_back ~ yearsexp + black + I(black *
    yearsexp), data)
modelo_interaccion_completo <- lm(call_back ~ yearsexp + black +
    I(black * yearsexp) + I(yearsexp^2) + I(black * yearsexp^2),
    data)
stargazer(modelo_interaccion_naive, modelo_interaccion, modelo_interaccion_completo,
    type = "latex", summary = FALSE, header = F)
```

Table 8

	Dependent variable:		
	call_back		
	(1)	(2)	(3)
yearsexp	0.005***	0.003***	0.011***
	(0.001)	(0.001)	(0.004)
black		−0.029**	0.002
		(0.014)	(0.025)
I(black ∗yearsexp)	−0.003***	−0.0003	−0.008
	(0.001)	(0.002)	(0.005)
I(yearsexpˆ2)			−0.0003**
			(0.0001)
I(black ∗yearsexpˆ2)			0.0003
			(0.0002)
Constant	0.055***	0.069***	0.038**
	(0.007)	(0.010)	(0.018)
Observations	4,870	4,870	4,870
R^2	0.006	0.007	0.008
Adjusted R^2	0.006	0.007	0.007
Residual Std. Error	0.271 (df = 4867)	0.271 (df = 4866)	0.271 (df = 4864)
F Statistic	15.623*** (df = 2; 4867)	11.814*** (df = 3; 4866)	7.998*** (df = 5; 4864)

Note: *p<0.1; **p<0.05; ***p<0.01

Para responder a la pregunta se propusieron tres modelos: el modelo (1) que evalúa una interacción de manera ingenua, el (2) que evalua la interacción pero controlando por black y el (3) que evalúa la interacción hasta un segundo grado, es decir si hay rendimientos. Lo que podemos concluir es que parece que en (1) si hay un efecto significativo de la discriminación en la que los beneficios de un año más experiencia son menores para los nombres negros. Sin embargo, este modelo está mal especificado pues la variable de interacción está capturando todo el efecto de black. Esto es evidente en el modelo (2) y (3) donde controlando por black, el coeficiente variable de interacción pasa a ser más pequeña y deja de tener significancia estadística.

Finalmente, en los modelos (2) y (3) notamos que ni el coeficiente de la interacción, ni el coeficiente de la interacción al cuadrado resultan ser estadísticamente significativos.

En conclusión no hay evidencia para sostener que existe un efecto de interacción entre "black" y experiencia. Por el contrario, el efecto de la discriminación parece venir preponderantemente solo por el efecto de black.

8.

¿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 significanciade 1%, y (iv) vas a dividir 50-50 tratamiento y control?

Para esta pregunta debemos analizar a detalle la derivación en el texto de Imbens. Partimos de la prueba de hipótesis sobre la diferencia en efectos:

$$H_n : E[Y_i(1) - Y_i(0)] = 0$$
$$H_a : E[Y_i(1) - Y_i(0)] \neq 0$$

lo cual podemos evaluar mediante el estadístico t:

$$T = \frac{\overline{Y}_t^{obs} - \overline{Y}_c^{obs}}{\sqrt{S_Y^2/N_t + S_Y^2/N_c}} \approx \frac{\overline{Y}_t^{obs} - \overline{Y}_c^{obs}}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}} \approx N(0,1)$$

Por lo tanto:

$$T \approx N(\frac{\tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}}, 1)$$

Rechazamos la hipotesis nula si

$$p(|T| > \Phi^{-1}(1-\alpha/2)) \approx \Phi(-\Phi^{-1}(1-\alpha/2) + \frac{\tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}})$$

(simplificando)

Queremos forzar la igualdad

$$\beta = \Phi(-\Phi^{-1}(1-\alpha/2) + \frac{\tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}})$$

de la cual derivamos la siguiente ecuación en terminos de N:

$$N = \frac{(\Phi^{-1}(\beta) + \Phi^{-1}(1-\alpha/2))^2}{(\tau^2/\sigma^2).\gamma.(1-\gamma)}$$

donde γ representa la proporción de tratados, β el poder y α el nivel de singificancia.

$$N = \frac{(\Phi^{-1}(.85) + \Phi^{-1}(.995))^2}{(-.032^2/\sigma^2).5^2}$$

```
# un paso antes es evaluar cuál varianza usaremos, en el
# texto de Imbens, asumen que las varianzas muestrales son
# las mismas, por ello, evaluamos si es pertinente mantener
# ese supuesto.

var_control
```

[1] 0.08723103

```
var_treatment
```

[1] 0.06034396

```
(var_pob <- var(data$call_back))
```

[1] 0.07402892

```
var_neyman_t
```

[1] 0.00006060575

```
# finalmente asumimos ese supuesto y computamos n
(N <- ((qnorm(0.85) + qnorm(0.995))^2)/(((neyman_t^2/var_pob)) *
    (0.5^2)))
```

[1] 3765.553

En conclusión, necesitamos al menos 3765.553 CVs ficticios si es que anticipamos que los efectos (varianza y efecto real) sean iguales a los obtenidos por Bertrand y Mullainathan, (ii) queremos un poder estadístico de 85%, (iii) asumimos una significanciade 1%, y (iv) dividimos 50-50 tratamiento y control.

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 conOptimalDesign) fijando los valores que obtuviste en el inciso anterior (número de observaciones, efectos reales y significancia).

```
# en primer lugar defino una función con la derivación de la
# formula de N del inciso anterior pero despejando para el
# poder

fun.1 <- function(x) pnorm(sqrt(N * (((neyman_t^2/var_pob)) *
    (x * (1 - x)))) - qnorm(0.995))

# en segundo lugar, grafico

ggplot(data = data.frame(x = 0), maping = aes(x = x, y = fun.1)) +
    stat_function(fun = fun.1) + xlim(0, 1) + ylim(0, 0.85) +
    labs(y = "Poder", x = "Fracción tratamiento") + ggtitle("Trade off") +
    theme(text = element_text(size = 20), panel.background = element_rect(fill = "lightblue"))
```