Tarea 3

Aleatorización y diferencia en diferencias

CIDE | Evaluación de programas |

Los equipos para esta tarea son:

EQUIPO 1:
González Martínez Nashellit
González Morales Romina
Romero Sánchez Christopher Daniel
Daza Vazquez Daniel Alfonso
García Viera Vanesa Janeth
EQUIPO 2:
Farías Ríos Antonio
Barrón Reyes Manuel Xicotencatl
Ortiz Peralta Diana Elizabeth
Zenteno Morales Elda Luisa
López Guerra Verónica
EQUIPO 3:
Martínez Amador Adriana
Hernández García Felipe De Jesús
Vargas Pineda Luis Roberto
Monroy Jiménez Jessica Daniela
Cruz Siachoque Jhon Jairo
EQUIPO 4:
Hernández Bernardino Juan Carlos
González Salgado Fernando
Tobón Flores Gerardo Antonio
Calderón Perez José Manuel

Cada equipo deberá idear un nombre para su equipo. No deberá contener ningún nombre de los integrantes del equipo.

Entrega de la tarea de manera anónima en este enlace

Fecha límite	28 de abril 2023, 11 pm

Recuerden no incluir ninguno de los nombres de los integrantes **en el archivo que suban**

Es **indispensable** entregar tus respuestas a cada pregunta con la siguiente estructura:

1. Respuesta escrita

2. Código (usa un tipo de letra adecuado para programación, como Courier New, Menlo, etc.)
3. *Output* de Stata o R (captura de pantalla)

Por ejemplo:

Pregunta 1. ¿Cuál es la media de *X*?

Pregunta 1. ¿Cuál es la media de *X*? La media de *X* es 4.3.

```
> mean(X)
[1] 10
```

• *Se descontarán 10 puntos porcentuales de la calificación final a los equipos que entreguen después de la fecha límite (1 hora de tolerancia).*

Aunque no es obligatorio, si utilizan R, te recomiendo entregar tu tarea en R Markdown, un formato que te permite generar documentos en PDF desde R en los que puedes incluir fácilmente tanto texto normal como código y los resultados de tu código, sin necesidad de copiar y pegar nada en Word. En el archivo .rmd, presiona el botón de "Knit" (tejer) para convertirlo en PDF. Si decides no usar R Markdown, puedes entregar en .doc o en un PDF generado con Word.

Archivos que debes entregar	
Si usas R Markdown	*Si usas Word*
Entrega tanto el archivo .rmd con todo tu código como el archivo .pdf generado.	Entrega el archivo .doc o .pdf y también el R script (.R) con todo tu código, o el log file de Stata.

I. Asignación aleatoria

El Gobierno del Estado de Yucatán está interesado en implementar un nuevo programa en toda la entidad federativa y, tras múltiples reuniones con especialistas, ha decidido otorgarlo de manera aleatoria. El programa consiste en ofrecer capacitaciones sobre diversas habilidades para el empleo a jóvenes que no estudian ni trabajan y que están en búsqueda de empleo.

Los operadores del programa solicitaron apoyo al CIDE para realizar la estrategia de aleatorización. Se determinó que la asignación aleatoria tendría que ser a nivel municipal por cuestiones de capacidad operativa. Sólo los jóvenes registrados como habitantes de los municipios de tratamiento

serán elegibles para las capacitaciones. Al buscar los datos disponibles en el INEGI sobre el estado, encuentras la base Asignación aleatoria.dta, que contiene los resultados principales del estado en el censo a nivel localidad.

En primer lugar, advierte que la base está a nivel localidad, pero tú necesitas transformarla a nivel municipal. Asegúrate de sumar, para las variables que lo necesiten, los datos de cada una de las localidades dentro de cada municipio para obtener los totales municipales correctos. Por ahora, solamente tendrás que agregar y conservar las siguientes variables (puedes descartar las demás):

NOMBRE	DESCRIPCIÓN
Variables de identificación	
mun	Código de identificación del municipio
nom_mun	Nombre del municipio
Observables poblacionales	
pobtot	Población total de la localidad
pobfem	Población femenina de la localidad
p_18a24	Población entre 18 y 24 años en la localidad
p_18a24_m	Población masculina entre 18 y 24 años en la localidad
p_18a24_f	Población femenina entre 18 y 24 años en la localidad
pnacoe	Población nacida en otra entidad
p18ym_pb	Población de 18 años y más con preparatoria en adelante
pea	Población económicamente activa
pocupada	Población ocupada
Observables de vivienda	
tvivhab	Total de viviendas habitadas
vph_c_serv	Viviendas particulares habitadas que disponen de luz eléctrica, agua entubada de la red pública y drenaje
vph_inter	Viviendas particulares habitadas que disponen de internet

1. Colapsa estas variables tomando *mun* como la variable usada para agrupar los municipios. Después, ordena tu base alfabéticamente según los nombres de los municipios resultantes y muestra los primeros cinco y los últimos cinco de la lista. Asegúrate con otras fuentes de tener el número correcto de municipios que existen en Yucatán. *Hints* para Stata. *Hints* para R.

2. Trabajaremos con las observables poblacionales y de vivienda que aparecen arriba para verificar que realmente queden balanceadas. Dado que todas ellas están en valores absolutos (número de personas o de viviendas, respectivamente), nos interesa transformarlas a valores relativos para realmente poder comparar las características observables de los municipios, nuestra unidad de análisis. Convierte todas las observables poblacionales desde *pobfem* hasta *pocupada* a porcentajes de la población total del municipio. Asimismo, convierte *vph_c_serv* y *vph_inter* en porcentajes del total de viviendas habitadas en el municipio. Nombra estas nuevas variables con el prefijo *porc_* (*porc_pobfem*, *porc_p_18a24*, etc.). Después, muestra la estadística descriptiva de todas las nuevas variables que comienzan con *porc_* (media, desviación estándar, mínimo y máximo son suficientes).

3. El Gobierno del Estado tiene recursos para implementar el programa en la mitad de los municipios. Asigna de manera aleatoria el programa a la mitad de los municipios, y asigna la otra mitad al grupo de control. Usa *set seed 12345*. Con el comando *tabulate*, muestra que la mitad de tus observaciones han quedado en el grupo de tratamiento y la otra mitad en tu grupo de control.

4. Verifica que las características observables entre los grupos estén balanceadas. Primero, realiza pruebas *t* para tres de las variables que creaste en la pregunta 2: *porc_pobfem*, *porc_p_18a24* y *porc_vph_inter*. Muestra tus resultados y completa la siguiente tabla a partir de tus hallazgos. ¿Qué concluyes a partir de estas tres pruebas?

Variable	Media del grupo de tratamiento	Media del grupo de control	Diferencia entre las medias	Estadística *t*	*p-value*	¿Rechazas o no rechazas?
porc_pobfem						
porc_p_18a24						
porc_vph_inter						

5. Realiza la misma prueba de balance que en la pregunta 4, pero esta vez usando regresiones bivariadas. Señala en tus resultados las diferencias entre las medias, las estadísticas *t* y los *p-values*. ¿Coinciden con los hallazgos encontrados con los de la pregunta 4? ¿Por qué?

6. Ve un paso más allá de lo revisado en clase y compara no sólo las medias, sino las distribuciones (solamente de manera visual). Grafica un diagrama de caja (*boxplot*) que compare las distribuciones para *porc_p_18a24_f* entre tratamiento y control, y otro diagrama de caja para *porc_pnacoe*. ¿Las distribuciones parecen ser balanceadas, por lo menos visualmente?

7. Por último, realiza pruebas de balance para las diez observables poblacionales y de vivienda que creaste que comienzan con *porc_*. Usa el paquete stargazer. Esto es como la pregunta 5, sólo que con más variables, y todas deben aparecer en una tabla. ¿Qué puedes concluir de los grupos de tratamiento y de control? ¿Qué puedes concluir del proceso de aleatorización?

8. El Gobierno del Estado ofreció los recursos humanos y materiales para impartir las capacitaciones a los municipios, pero se encontró con que los ayuntamientos gobernados por el PRI no estaban interesados en recibir este programa del gobernador —quien pertenece al PAN— y lo rechazaron. De hecho, de los 57 municipios gobernados por el PRI, 29 habían sido originalmente asignados al grupo de tratamiento. Al enterarse de que ya no se usarían estos recursos en esos 29 municipios, algunos alcaldes de partidos con menor presencia en el estado que habían sido originalmente asignados al grupo de control movieron sus influencias con el gobernador para recibir el programa, y finalmente 3 ayuntamientos del PRD, 2 de Movimiento Ciudadano y 2 de Morena lograron recibir el programa a pesar de que su estatus original era de control. Tras dos años de implementación, la tasa de desempleo en jóvenes que no estudian ni trabajan en los municipios originalmente asignados al grupo de tratamiento es de 1.3%, mientras que en los municipios originalmente asignados a control es de 3.4%. Calcula manualmente los estimadores que podríamos obtener del efecto del programa en este caso y menciona cómo se podrían interpretar.

9. El gobierno de Yucatán te pide que redactes un breve boletín de prensa (1 cuartilla max) con los resultados obtenidos con este experimento, la explicación de los resultados y posibles acciones a futuro sobre este programa derivados del estudio. Trabaja junto a tu encargado de periodismo para hacer el boletín, de manera concisa y con resultados bien presentados (sin falsos contrafactuales, etcétera).

II. Efecto promedio del tratamiento

En un experimento[1] sobre discriminación racial en el mercado laboral de Estados Unidos, los investigadores enviaron 5,000 currículum vítae (CV) falsos a ofertas laborales publicadas en los periódicos de Boston y Chicago. Para "manipular" qué raza percibían los empleadores que tenían los candidatos, los investigadores asignaron aleatoriamente nombres típicos de personas blancas (como Emily Walsh o Greg Baker) a la mitad de los CV, y nombres típicos de personas afroamericanas (como Lakisha Washington o Jamal Jones) a la otra mitad. Encontraron que los perfiles con nombres típicos de personas blancas fueron llamados para entrevista 50% más que los perfiles con nombres típicos de personas afroamericanas.

La siguiente tabla incluye algunos de los resultados de regresión de la prueba de balance que realizan los investigadores antes de calcular el efecto del tratamiento. Se trata de regresiones bivariadas en donde la característica observable es la variable dependiente, y la dummy de tratamiento es la independiente. En este ejemplo, "ser asignado un nombre afroamericano" es el tratamiento ($T = 1$) y "ser asignado un nombre blanco" es el control ($T = 0$). Las características observables entre signos de interrogación indican que se trata de variables dummy igual a 1 si el CV cumple con esa característica e igual a 0 si no.

	¿Licenciatura?	Años de experiencia	¿Voluntariado?	¿Militar?	¿Email?	¿Periodos sin empleo?	¿Estudia y trabaja?	¿Graduación con honores?	¿Habilidades especiales?
Nombre afroamericano	0.00657	-0.0267	0.00575	0.00945	0.00082	-0.00411	0.00287	-0.00287	-0.00287
	(0.0129)	(0.145)	(0.0141)	(0.00849)	(0.0143)	(0.0143)	(0.0142)	(0.00641)	(0.0135)
Constante	0.716	7.856	0.409	0.0924	0.479	0.45	0.558	0.0542	0.33
	(0.00911)	(0.102)	(0.0099)	(0.006)	(0.0101)	(0.0101)	(0.0101)	(0.00453)	(0.00952)

Errores estándar entre paréntesis.
Base de datos obtenida de: https://www.aeaweb.org/articles?id=10.1257/0002828042002561

[1] Marianne Bertrand y Sendhil Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination", *The American Economic Review*, 2004, 94 (4), p. 1000.

1. Elige tres de las características observables. ¿Existen diferencias significativas entre el grupo de tratamiento y el de control? Demuéstralo con una prueba t para cada variable. ¿Qué puedes concluir a partir de estas tres pruebas?

Los datos de los autores están disponibles en Equipos en el archivo Efecto promedio del tratamiento.dta, así como el artículo original, Bertrand y Mullainathan (2004).pdf. Las principales variables de la base son:

NOMBRE	**DESCRIPCIÓN**
id	Código de identificación del CV
call	Variable de resultado: 0 si el CV no recibió una llamada o correo electrónico para ser invitado a entrevista 1 si el CV recibió una llamada o correo electrónico para ser invitado a entrevista
black_sounding	Variable de tratamiento: 0 si el CV fue asignado un nombre que "suena" a persona blanca 1 si el CV fue asignado un nombre que "suena" a persona afroamericana
college en adelante	Desde la variable *college* hasta *linc*, se trata de distintas características observables de los CV.

2. Replica la prueba de balance, pero esta vez para todas las características observables incluidas en la base (desde *college* hasta *linc*). Esta vez, no hagas pruebas *t*, sino directamente las regresiones bivariadas. Muestra tus resultados en una o dos tablas usando el paquete *stargazer* en R.

3. Del total de regresiones que realizaste, ¿qué porcentaje indica que existe una diferencia estadísticamente significativa entre los grupos? ¿Qué podrías concluir de este hallazgo? ¿Qué puedes concluir sobre si la aleatorización se implementó de manera correcta y el balance entre los grupos?

4. ¿Qué porcentaje de los CV con nombre blanco recibieron una llamada para entrevista? ¿Qué porcentaje de los CV con nombre afroamericano lo hicieron? ¿Cuál es la diferencia en puntos porcentuales?

Recuerda que, cuando la aleatorización se implementa correctamente, todas las características observables y no observables quedan balanceadas entre los grupos, por lo que técnicamente no habría necesidad de incluirlas como variables de control en la regresión (ya han sido controladas *ex ante* con el diseño aleatorizado). En la práctica, sin embargo, generalmente se incluyen de todas maneras para incrementar la *precisión* de los estimadores, pero esto no debe alterar demasiado la *magnitud* del efecto promedio del tratamiento encontrado si la aleatorización se implementó de manera correcta.

Recuerda que por mayor *precisión* nos referimos a que los errores estándar de los estimadores sean más pequeños, pues esto determina si son significativos o no y qué tan acotado es su intervalo de confianza. En cambio, por *magnitud* nos referimos al valor del estimador ($\hat{\beta}_1$), el cual indica de qué tamaño es el efecto promedio del tratamiento que se está estimando.

5. Primero calcula el efecto promedio del tratamiento con una regresión bivariada simple, sin incluir ninguna variable de control. ¿Cuál es el efecto promedio del tratamiento? ¿Tu

resultado coincide con el de la pregunta 3? Verifica que tu respuesta coincida con el primer renglón de la Tabla 1 de los autores (p. 997). Interpreta este resultado en un enunciado.

6. ¿Cuál es el valor de la constante, $\hat{\beta}_0$? ¿Cómo interpretas este valor? ¿Qué es $\hat{\beta}_0 + \hat{\beta}_1$?

7. Ahora, calcula el efecto promedio del tratamiento con una regresión multivariada que incluya todas las variables de control, desde *college* hasta *linc*. ¿Incluir estas variables alteró demasiado la *magnitud* del estimador del ATE? ¿Cómo cambió la *precisión* del estimador? ¿Qué puedes concluir de incluir o no incluir las variables de control en un experimento cuya aleatorización se implementó de manera correcta?

8. El editor de una revista de divulgación se ha interesado en este experimento y a sabiendas de la reputación de los estudiantes de posgrado del CIDE en evaluación, contacta a tu grupo para que escriba una breve reseña del estudio que pueda captar el interés del público en general. Esta reseña no debe exceder de 1 de cuartilla.

III. Diferencia en diferencias

El gobierno central de un país lejano inició un programa nacional de subsidios para construcción de áreas verdes en zonas urbanas de los condados, la unidad más pequeña de su división política. Este país lleva algunos años en los que la violencia ha aumentado en diferentes regiones luego de cambios importantes en la política de seguridad. El programa tiene la (debatible) teoría de cambio de que las áreas verdes, como los parques y espacios de recreación, favorecen la convivencia y el espíritu comunitario, lo cual ayuda a que se recupere el tejido social y se reduzca la violencia.

En 2010 se lanzó una convocatoria abierta, a la cual podían postular todos aquellos condados que fueran urbanos o semiurbanos. Los condados que desearan ser candidatos debían preparar una propuesta de desarrollo urbano en la cual especificaran las distintas características de su proyecto (dónde se construiría el área verde, cuáles eran sus objetivos, qué efectos positivos se esperaba que pudieran tener en la zona y por qué, presupuesto y plan de trabajo, etcétera). Posteriormente, un comité técnico de empleados del programa evaluó las propuestas y decidió otorgar los subsidios a los mejores proyectos.

Como puedes ver, la asignación del programa no es aleatoria. En este ejercicio realizarás un diseño de diferencia en diferencias para evaluar el impacto del programa de áreas verdes.

La base Diferencia en diferencias.dta contiene datos longitudinales para 10,000 condados urbanos y semiurbanos de este país en tres años diferentes: 2008, 2010 y 2012. Las variables incluidas en la base son:

NOMBRE	DESCRIPCIÓN
id_condado	Código de identificación del condado
año	Año al que corresponden los datos: 2008, 2010 o 2012
homicidios	Homicidios por cada 100,000 habitantes del condado (tu variable de resultado, Y)
subsidio	Variable dicotómica (tu variable de tratamiento, T):

	0 si el condado ***no*** recibió el subsidio del programa de áreas verdes 1 si el condado recibió el subsidio del programa de áreas verdes

Entre 2010 y 2012, la mitad de los condados recibieron el subsidio.

1. Dado que el programa otorgó los subsidios después de que se obtuvieron los datos de 2010 que tienes en tu base, por ahora considera 2010 como el "antes" y 2012 como el "después". Calcula las siguientes medias condicionales de los homicidios y acomódalas en la siguiente tabla:

$$E(Y_{i1}|subsidio = 1, año = 2010)$$
$$E(Y_{i1}|subsidio = 1, año = 2012)$$
$$E(Y_{01}|subsidio = 0, año = 2010)$$
$$E(Y_{01}|subsidio = 0, año = 2012)$$

	Después (2012)	Antes (2010)
Con subsidio		
Sin subsidio		

2. Con base en tu tabla de la pregunta 1, calcula el estimador de diferencia en diferencias de manera manual.

3. Ahora, obtén el estimador de diferencia en diferencias mediante una regresión. ¿Cuál es el modelo de regresión con el que lo obtendrás? ¿Cuál de tus parámetros es el que captura el estimador de diferencia en diferencias? Utiliza notación para responder estas preguntas. Luego, muestra tus resultados de regresión. ¿Obtienes el mismo valor que en la pregunta 2? ¿Qué valor agrega obtenerlo con una regresión en comparación con la manera manual?

4. Para verificar la plausibilidad del supuesto de tendencias paralelas, utiliza los datos que tienes en tu base anteriores a la asignación del programa (2008). Muestra las tendencias de la variable de resultado para el grupo con subsidio y el grupo sin subsidio en una misma gráfica. Por lo menos visualmente, ¿las tendencias entre los grupos eran paralelas antes de la asignación del programa en 2010? ¿Qué implica esto para la comparabilidad entre los grupos y tu estimador de diferencia en diferencias?

5. Con los resultados obtenidos en este estudio, escribe una nota que será publicada en uno de los más influyentes diarios de este país lejano.