

Primer ejercicio

El Ministerio de Agricultura (MinAgricultura) implementó un experimento aleatorio controlado a gran escala con el fin de evaluar el impacto de tener buenas prácticas de cultivo y cosecha sobre la productividad de los predios agrícolas. El tratamiento asociado al experimento consistió en la entrega de dos sesiones informativas sobre las buenas prácticas de cultivo y cosecha, uso del fertilizantes, y aprovechamiento de semillas. El tratamiento incluía la entrega de un folleto que resumía la información dada en las sesiones. Dada la naturaleza del experimento, el tratamiento fue asignado de forma aleatoria para poder medir su efectividad. Todos los predios que fueron asignados a la capacitación la recibieron efectivamente. No obstante, con el fin de evitar resultados contaminados por sesgos experimentales, y, al mismo tiempo, identificar su presencia, la aleatorización del tratamiento fue hecha en dos etapas. En una primera etapa, el MinAgricultura dividió los municipios en tres grupos: Grupos 1, 2 y 3. En una segunda etapa, al interior de cada grupo, se dividieron los predios de la siguiente forma:

- **Grupo 1:** Tratamiento aleatorizado a nivel de municipio. La mitad de los municipios de este grupo fueron asignados al tratamiento y la otra mitad al grupo de control. Todos los predios de los municipios tratados participaron del programa.
- **Grupo 2:** Se realizó una asignación aleatoria estratificada por municipio: para cada uno de los municipios de este grupo, la mitad de los predios fue asignada al grupo de tratamiento y la otra mitad al grupo de control.
- **Grupo 3:** Ninguno de los predios fue asignado al tratamiento, pero fueron informados de que eran parte de un experimento.

Nota: salvo los predios del Grupo 3, ningún otro predio fue informado de que era parte de un experimento. Suponga también que no se filtró la información recibida en las capacitaciones entre predios de municipios distintos. Tampoco se filtró la información que algunos predios estaban siendo parte de un experimento.

Ustedes son contratados como consultores externos para realizar la evaluación de impacto del programa. En particular, a sus empleadores les interesa el impacto del programa sobre la productividad de los predios productores, pues se evalúa la posibilidad de extender las capacitaciones a otros predios que no hicieron parte del experimento. Para lograr este objetivo, a ustedes les entregan la base de datos “producto_predios.dta”. Esta base contiene información a nivel de predio-municipio con las siguientes variables:

- *dpto*: Código de Departamento.
- *mpio*: Código de Municipio.
- *predio*: Identificador de predio.
- *distancia_mpio*: Distancia en kilómetros desde el predio al centro del municipio al que pertenece.
- *calidad_tierra*: Índice estandarizado de calidad de la tierra.
- *productividad_LB*: Productividad del predio en línea de base (i.e. antes del experimento) medida en kilogramos cosechados por hectárea.
- *productividad*: Productividad del predio seis meses después de la finalización del experimento medida en kilogramos cosechados por hectárea. La productividad fue medida en una visita sorpresa no anunciada.
- *grupo*: Variable categórica que captura el grupo experimental al que fue asignado cada municipio.
- *T*: Dicótoma que toma el valor de uno si el predio recibió el tratamiento y cero de lo contrario.

Respondan las siguientes preguntas teniendo en cuenta el contexto y los datos disponibles.¹

¹El experimento descrito y los datos asociados a este son ficticios. Los resultados obtenidos en este taller no proveen información alguna sobre si capacitar o no a los dueños de predios productores incrementa su productividad o no.

a) Para lograr el objetivo de este estudio, ¿cuáles predios compararían? Propongan un modelo de regresión lineal que les permita recuperar el efecto de interés. Asegúrense de describir cada uno de los términos (variables y parámetros) que componen su ecuación a estimar. Mencionen y discutan los supuestos necesarios para que el estimador por MCO sea consistente e insesgado. ¿Son los supuestos plausibles en este contexto?

b) Antes de realizar la estimación, el MinAgricultura les solicita evidencia que soporte la correcta aleatorización de los predios de los grupos 1 y 2. Para esto, presenten:

I. Una tabla de balance muestral para el grupo 1.

II. Una tabla de balance muestral para el grupo 2.

Deben tener en cuenta el tipo de aleatorización que se hizo (e.g., si fue estratificada) y si deben ajustar por errores estándar tipo clúster. A partir de las tablas, argumenten: ¿hay evidencia de que la aleatorización se hizo de forma correcta? ¿Para qué grupos parece cumplirse el balance en línea base?

Pista: Para hacer pruebas de balance muestral condicionales, pueden recurrir al concepto de partialling out o pruebas de diferencias de medias condicionales.

c) Sus empleadores están preocupados por el cumplimiento del supuesto de SUTVA. Por esta razón, les solicitan presentar evidencia que les permita abogar a favor o en contra de su validez.

I. Establezcan una relación matemática (en términos de los resultados potenciales) que represente el cumplimiento del supuesto SUTVA en la muestra.

II. A partir de su respuesta en el inciso I, describan un experimento que les permitiría verificar SUTVA.

III. Describan una comparación o ecuación a estimar basada en los datos disponibles que les permita verificar el supuesto de SUTVA.

IV. Implementen y presenten en una tabla la comparación/metodología planteada en el inciso III.

A partir de los resultados encontrados, ¿parece cumplirse el supuesto de SUTVA? En caso de que no se cumpla, ¿para qué grupos parece no cumplirse?

d) Por otra parte, ustedes están preocupados por la presencia de sesgos comportamentales producto del experimento que pueden alterar los resultados de su estimación. En este inciso,

I. Establezcan una relación matemática (en términos de los resultados potenciales) que represente la presencia de efecto Hawthorne en la muestra.

II. A partir de su respuesta en el inciso I, describan el experimento ideal que les permitiría verificar si se da el efecto Hawthorne.

III. Argumenten a la luz del contexto y el experimento si es razonable que se dé o no el efecto Hawthorne. En caso de que sí, ¿para qué grupos se daría?

e) En este inciso debe presentar evidencia a favor o en contra de la presencia del efecto John Henry. Para esto,

I. Establezcan una relación matemática (en términos de los resultados potenciales) que represente la presencia de efecto John Henry en la muestra.

II. A partir de su respuesta en el inciso I, describan el experimento ideal que les permitiría verificar si se da el efecto John Henry.

III. Describan una comparación o ecuación a estimar basada en los datos disponibles que les permita verificar si se da el efecto John Henry o no.

IV. Implementen y presenten en una tabla la comparación/metodología planteada en el inciso III.

A partir de los resultados encontrados en este y el anterior inciso, ¿parece cumplirse el supuesto de no sesgos experimentales? En caso de que no se cumpla, ¿para qué grupos parece no cumplirse?

- f) Finalmente, contraste sus respuestas de los incisos b) a d) con la comparación propuesta en el inciso a). ¿Es su comparación válida? En caso de que sí, estimen el efecto de interés a partir de la comparación propuesta en el inciso a). En caso de que no, propongan una nueva comparación o ecuación a estimar que sea válida a pesar de los problemas encontrados en los incisos b) a d). Estimen el efecto de interés a partir de esta nueva comparación. Presenten en una tabla sus resultados e interprétenlos. Asegúrense de que sus resultados sean interpretados en términos de significancia estadística y económica.

Segundo ejercicio

La idea de que todos los individuos son iguales ante la ley es uno de los principios fundamentales de los sistemas judiciales modernos en países democráticos. Diferencias raciales, económicas y sociales no deberían tener injerencia al momento de juzgar a un individuo. Sin embargo, estudios sobre este tema sugieren que es común encontrar tratos diferenciados en el manejo de los procesos judiciales asociadas al nivel de ingreso de los individuos. Medir efectivamente dichas desigualdades es una tarea importante para poder intentar corregirlas.

En el pueblo de Angristville, donde recientemente han aumentado las protestas por parte de grupos que cuestionan la imparcialidad del sistema judicial, este tema está en el centro del debate público. Los grupos que protestan aseguran que las personas más pobres de la sociedad son sistemáticamente desfavorecidas en las cortes, por lo cual suelen tener mayor probabilidad de recibir penas y dichas penas son, generalmente, más severas que las impuestas contra individuos más ricos. El aumento de las tensiones sociales ha llevado a los dirigentes de Angristville a considerar una reforma al sistema judicial, pero desconocen si las protestas son realmente el reflejo de una situación real o corresponden al estado de opinión del momento. Por ello, los han contratado a ustedes, asesores del gobierno, para que estudien el tema.

En Angristville, todos los acusados son asignados aleatoriamente a uno de 8 jueces posibles acorde al día en el que se acude a una audiencia inicial. Sin embargo, cada juez está disponible para el sorteo en sólo algunos días y años. En las audiencias iniciales, se define la fianza que debe pagar el acusado a fin de permanecer en libertad durante el curso del juicio. Según la ley, las fianzas son fijadas según la naturaleza y gravedad del cargo presentado. No obstante, algunos abogados han asegurado que el carácter del juez asignado juega un papel central en el resultado de dichas audiencias. Aún así, la sentencia final, en la que se define la culpabilidad o inocencia del acusado, es decisión de un jurado conformado por personas seleccionadas aleatoriamente entre la población.

Suponga que el Gobierno les provee la base de datos “sentencias.dta”, la cual contiene información al nivel de caso sobre las sentencias dictadas, el juez encargado, y algunas características del individuo en cuestión como su raza, sexo, historial criminal y si fue o no encarcelado durante el curso del juicio (i.e. si pagó o no la fianza). Una descripción de las variables se encuentra en el pdf titulado “diccionario_de_variables.p2”:

- a) Para empezar, a partir de estadísticas descriptivas resumidas en una tabla, analicen las características demográficas de los acusados. Asegúrense de solucionar problemas evidentes en los datos y sean explícitos en cómo los resolvieron.
- b) Un mecanismo por medio del cual el nivel de ingreso puede afectar los resultados de los juicios es a través de las fianzas. ¿Creen ustedes que esta hipótesis es cierta? ¿Por qué?
- c) A raíz de lo discutido en el punto anterior, a ustedes les interesa conocer el efecto de ser encarcelado previo al juicio (i.e. no poder pagar la fianza fijada) sobre la probabilidad de ser declarado culpable. Un compañero les sugiere desentenderse pronto de este trabajo y simplemente estimar por MCO una regresión que incorpore adecuadamente sus variables de interés.
 - I) Plantee dicho modelo. Mencione qué representa cada variable incorporada. ¿Qué variables/grupos de variables de su base de datos podrían servir como controles en una posible regresión? Justifique. No es necesario que explique una a una por qué incluiría cada variable en su modelo, sino que dé una idea general de por qué sería deseable controlar por ciertas características.
 - II) Estimen el modelo propuesto y presenten sus resultados. ¿Qué sugieren los resultados sobre la existencia de sesgos en la imposición de las sentencias?

- III) ¿La anterior estimación identifica el efecto causal de interés? Sean explícitos en los supuestos que plausiblemente se están (in)cumpliendo.
- d) Propongan una estrategia de identificación alternativa que le permita atender el problema descrito en el punto anterior (*Pista: recuerden que la probabilidad de ser encarcelado depende de la severidad del juez asignado*).
- I) Expliquen con claridad cómo implementarían dicha estrategia (i.e. plantee la(s) ecuaciones a estimar), los supuestos que se deben cumplir y justifiquen su (in)plausibilidad. Por ahora, supongan que el efecto del encarcelamiento es igual para todos los individuos. ¿Qué efecto están estimando con su nueva estrategia (i.e. ¿qué representa el parámetro de interés?)
- II) ¿Cuáles serían los supuestos que necesitaría si el ser encarcelado previo al juicio tuviera efectos distintos para cada individuo? ¿Qué tipo de efecto promedio estarían estimando en este caso?
- III) Argumenten a partir de los datos si tienen evidencia que soporte el (in)cumplimiento de alguno(s) de sus supuestos. ¿Qué (problemas) ventajas puede representar esto?
- IV) Empleando la estrategia propuesta, estimen el modelo y presente sus resultados. Interprete. ¿Qué diferencias encuentran con los resultados obtenidos al efectuar la estimación por MCO? ¿Qué sugieren los resultados sobre la existencia de sesgos en la imposición de las sentencias?

Tercer ejercicio

Entre 2014 y 2018, la administración del expresidente Juan Manuel Santos implementó el programa Ser Pilo Paga (SPP). El objetivo principal de SPP fue garantizar que los mejores estudiantes del país con menos recursos económicos accedieran a Instituciones de Educación Superior (IES) de alta calidad. SPP se financió con fondos públicos. Para los beneficiarios, cubría el costo total de la matrícula de un programa de pregrado de cuatro o cinco años en cualquier universidad con acreditación de alta calidad en el país. La elegibilidad al programa se definió, entre otros criterios, con un puntaje máximo en el SISBÉN y un puntaje mínimo en el examen de salida de la educación media, Saber 11 (el puntaje mínimo para ser elegible era de 310).² Adicionalmente, para acceder a los beneficios los estudiantes tenían que ser admitidos a un programa universitario en una IES con acreditación de alta calidad. Debido a las características del programa, ustedes utilizarán la metodología de regresión discontinua para determinar el impacto de esta iniciativa estatal.

Ustedes están interesados en estudiar el efecto del programa sobre la matrícula en educación superior y la elección de universidad de los estudiantes (universidad de alta calidad o de baja calidad, universidad pública o privada). Para esto, disponen únicamente información de la primera cohorte de SPP, en la cual los estudiantes desconocían la existencia del programa hasta que recibieron los resultados de las pruebas del Saber 11. Estos datos se encuentran en el archivo “SPP_Base”, que está acompañado del diccionario “diccionario.de.variables.punto3.pdf”, el cual contiene la descripción de las variables de la base de datos.

Para cada inciso, además de contestar las preguntas respectivas, incluya el código utilizado para llegar a la solución.

- a) Importe la base de datos. Presente una tabla de estadísticas descriptivas para las variables *puntaje_saber11*, *edad*, *estrato*, *miembros_familia*, donde se reporte el número de observaciones, el promedio de la variable, su desviación estándar y el mínimo y máximo observados. Reporte toda la información con dos decimales. Esta tabla debe ser generada completamente en Stata/R para su exportación en Word o en Tex.
- b) Presente dos gráficos que le permitan analizar si hay o no discontinuidad en la probabilidad de (i) ser elegible para el programa y (ii) acceder efectivamente al programa. Para la gráfica (ii) utilice un polinomio de grado 2 diferenciado a ambos lados del punto de corte. Interprete sus resultados y discuta brevemente la relevancia de encontrar una discontinuidad en el gráfico (ii).

²La prueba Saber 11 es un examen estandarizado de salida de la educación media. Este examen lo presentan todos los estudiantes que se gradúan de la educación media independientes de si acceden o no a la educación superior.

- c) Con base en el inciso b), estime, presente en una sola tabla e interprete los efectos de la elegibilidad de SPP y de ser beneficiario de dicho programa sobre la matrícula en educación superior y la elección de universidad de los estudiantes (universidad de alta o baja calidad, universidad pública o privada). Deben estimar estas cantidades usando métodos no-paramétricos. Para ello, usen el comando `rdrobust` (quizá requieran instalarlo usando el comando “`ssc install rdrobust`” en Stata o “`install.packages(‘rdrobust’)`” en R). Además, utilicen un polinomios de grado 1 y 2, y como ponderador un kernel triangular. Incluyan en la tabla el número de observaciones disponibles y el tamaño del ancho de banda usado en cada estimación.

Una estructura sugerida para la tabla es la siguiente:

Polinomio de grado 1			Polinomio de grado 2		
Matrícula	IES alta calidad	IES privada	Matrícula	IES alta calidad	IES privada
(1)	(2)	(3)	(4)	(5)	(6)
<i>Panel A: Elegibilidad</i>					
Elegible para SPP [1=Sí]					
Observaciones					
Ancho de Banda					
Observaciones en banda					
<i>Panel B: Beneficiario del programa</i>					
Beneficiario de SPP [1=Sí]					
Observaciones					
Ancho de Banda					
Observaciones en banda					

Luego discutan

- ¿Qué tipo efecto de tratamiento se identifica en el panel A y en el Panel B? ¿Cuál es más grande? ¿Por qué?
- ¿Son los valores encontrados en la columna 1 estadísticamente significativos? ¿Son éstos económicamente significativos?

Nota: Utilicen la opción “all” del comando `rdrobust` y reporten los coeficientes cuyo sesgo ha sido corregido y cuyos errores estándar son robustos a esta corrección.

Para que las metodologías de regresión discontinua identifique el efecto causal del tratamiento se necesitan ciertos supuestos. Si bien éstos son imposibles de probar, podemos proveer evidencia empírica de su posible cumplimiento. El primero de ellos es el de continuidad local, esto es, que las variables no afectadas por el tratamiento no tienen una discontinuidad en el umbral.

- d) Para evaluar el cumplimiento del supuesto, evaluaremos si existe una discontinuidad en el umbral de otras variables que posiblemente determinan el tratamiento. Para ello, presente en una única gráfica (no tabla) los valores (con sus respectivos intervalos de confianza) resultantes de estimar de manera no-paramétrica una posible discontinuidad en las variables *mujer*, *edad*, *minoría*, *estrato*, *miembros familia* y *col.privado*.

Una gráfica sugerida viene del realizar los siguientes pasos:

- Estandarice las variables mencionadas para que los resultados sean comparables.

- ii) Utilice el comando *rdrobust* usando como variable de focalización el puntaje de la Prueba Saber y como variable dependiente cada uno de los confounders. Utilice un polinomio local de grado 1 y un kernel triangular como ponderador.
- iii) Guarde en una matriz los coeficientes y los errores estándar de cada estimación.
- iv) Muestre en una figura cada uno de los valores estimados junto con sus intervalos de confianza a un nivel de significancia del 95 %. Para ello, puede ser útil el comando *coefplot*.
- v) Interprete los resultados.

Otro posible test es evaluar que no es posible manipular la regla de asignación. A continuación vamos a implementar dos tipos de tests diferentes para evaluar la veracidad de esta afirmación.

- e) Una posible noción de no manipulación es pensar que para puntajes lo suficientemente cerca del umbral, estar a un lado o al otro es casi que aleatorio. Una manera de modelar esta intuición matemáticamente es pensar que

$$E_i = 1[\{\text{Puntaje}_i \geq 310\}] | \text{Puntaje}_i \in [310 - \lambda, 310 + \lambda] \sim \text{Bernoulli}(0.5)$$

donde E_i es una dummy que indica si un individuo es elegible o no. Esto es, bajo este modelo, caer a un lado u otro del umbral es equivalente al resultado de lanzar una moneda para aquellas observaciones que están a lo sumo a $\lambda > 0$ puntos del umbral.

Vamos a implementar un test de Monte-Carlo que nos permita poner a prueba esta idea:

- i) Fije un $\lambda > 0$.
- ii) Para cada i , genere $l = 1, \dots, 10000$ variables de asignación ficticias $D_i^{(l)} \sim \text{Bernoulli}(0.5)$
- iii) Para cada l , calcule $p^{(l)}$:

$$p^{(l)} = \frac{\#\{i : D_i^{(l)} = 1 \ \& \ \text{Puntaje}_i \in [310, 310 + \lambda]\}}{\#\{i : \text{Puntaje}_i \in [310 - \lambda, 310 + \lambda]\}}$$

- iv) Calcule p^* para la muestra original.

$$p^* = \frac{\#\{i : E_i = 1 \ \& \ \text{Puntaje}_i \in [310, 310 + \lambda]\}}{\#\{i : \text{Puntaje}_i \in [310 - \lambda, 310 + \lambda]\}}$$

- v) Utilice la distribución aproximada de los $p^{(l)}$ para estimar cuán extremo es p^* bajo este modelo. Para ello calcule:

$$\chi = \frac{\#\{l : |p^{(l)} - 0.5| > |p^* - 0.5|\}}{10000}$$

Calcule χ para $\lambda = 3$. Concluya qué significa este resultado a la luz del supuesto de no-manipulación.

- f) Otra posibilidad para descartar manipulación, es pensar que no existe una discontinuidad en la función de densidad de la variable de focalización en el umbral. Presente un figura en donde estime usando polinomios locales la función de densidad a cada lado del umbral, y reporte el estadístico T de la discontinuidad estimada así como su respectivo p-valor. Concluya qué dice esto sobre la validez del supuesto.

Nota: Una manera sencilla de hacer eso es usar el comando “*rddensity*”.

Punto doctorado

En no más de tres páginas, realice un *referee report* del artículo: [Card, D., \(2001\) Immigrant Inflows, Native Outflows, and the Local Market Impacts of Higher Immigration, Journal of Labor Economics, Vol 19, No 1 \(Enero., 2001\)](#). Sin importar que el artículo se encuentre publicado, debe ser crítico a la hora de evaluar su contenido, identificando de manera clara los puntos débiles y fuertes de la estrategia empírica.