MÉTODOS DE EVALUACIÓN

Dina Pomeranz[*]

Agosto 2011

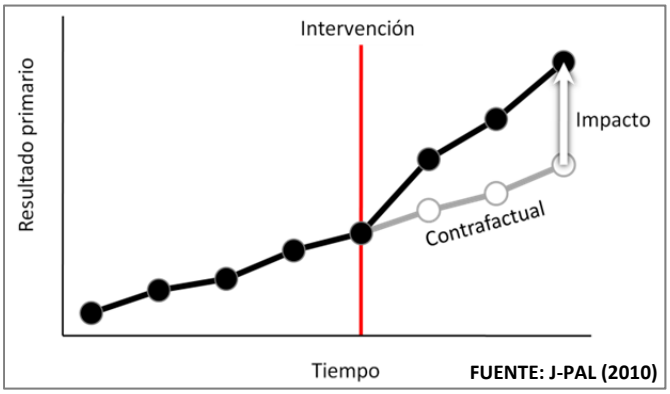
En la administración tributaria se toman decisiones diariamente que pueden afectar a la economía de todo el país. ¿Cómo se toman estas decisiones? ¿Son decisiones buenas o malas? La meta de una evaluación es informar sobre los efectos de políticas actuales y potenciales. Existen varios métodos de evaluación con diferentes niveles de fiabilidad. La calidad de la evaluación es de suma importancia para poder entregar resultados correctos. Este documento ofrece un breve resumen de los métodos más comunes con las ventajas y desventajas de cada uno y una descripción de las condiciones bajo las cuales cada método produce resultados fiables.

Antes de presentar los métodos específicos, se presentarán algunos conceptos básicos:

El objetivo de cada evaluación de impacto es demonstrar un *efecto causal*: Se quiere medir el impacto de un programa o una política en alguna variable de interés. Por ejemplo, cuál es el impacto de una notificación en las rectificaciones de impuestos. Existe una causa y un efecto. La causa es el cambio de una política o la implementación de un programa nuevo. El efecto es el resultado que se atribuye directamente a la política o al programa nuevo.

La dificultad en medir el impacto está en que solamente se puede observar lo que ocurrió, no lo que hubiera ocurrido sin el programa. Vemos si un contribuyente que recibió una notificación hizo una rectificación, pero no vemos que hubiera hecho si no hubiera recibido la notificación, es decir no sabemos si el contribuyente hubiera hecho la misma rectificación. Esta situación imaginaria, lo que hubiera pasado sin el programa, se llama *el* ***contrafactual***. Entender el contrafactual es clave para entender el impacto de un programa.

FIGURA 1 - EL CONTRAFACTUAL



FUENTE: J-PAL (2010)

Si existiera una representación correcta del contrafactual la estimación del impacto sería fácil. El impacto del programa o de la política es la diferencia entre el resultado que observamos con el programa y el resultado que hubiera ocurrido sin el programa - el contrafactual.

Como en la realidad el contrafactual no existe, ya que es lo que hubiera pasado en un escenario distinto, cada evaluación intenta - de manera explícita o implícita - construir una estimación del contrafactual para compararlo con lo que ocurrió. Normalmente, la estimación del contrafactual se representa con un grupo que se denomina *el* ***grupo de control*** **o** ***de comparación***. El grupo de control consiste de personas o empresas que no participaron en el programa, mientras que el ***grupo de tratamiento*** es el grupo que participó en el programa. Para estimar el impacto de la intervención se compara el grupo de tratamiento con el grupo de control.

[*] Harvard Business School , Rock Center 213, Soldiers Field Road, Boston, MA 02163, dpomeranz@hbs.edu.

La evaluación produce resultados fiables, si el grupo de control es igual al grupo de tratamiento en todas las características - observables y no observables - salvo una: su exposición al programa. En este caso, cualquier diferencia después de la intervención se le puede atribuir al programa, ya que en su ausencia los dos grupos serían iguales.

Cada método utilizado para la construcción del grupo de comparación impone ciertos supuestos bajo los cuales el grupo de control y el grupo de tratamiento serían comparables. Cuando los supuestos son realistas, el grupo de control es una buena representación del contrafactual. Pero cuando los supuestos no son realistas, la estimación del impacto del programa resulta ***sesgada***. Una evaluación sesgada puede resultar en malas decisiones, y genera pérdidas de esfuerzo, tiempo y fondos públicos.

Por lo tanto, es importante hacer explícitos los supuestos involucrados en cada método de evaluación y trabajar con métodos de alta calidad. En la siguiente parte del documento, se presentarán los diferentes métodos de evaluación con una descripción de sus características, cualidades y limitaciones.

1. EVALUACIÓN ALEATORIA

Las evaluaciones aleatorias (o evaluaciones experimentales) construyen un grupo de comparación de máxima calidad: la asignación aleatoria tiene como objetivo que no exista ninguna diferencia entre los individuos del grupo de tratamiento y del grupo de control, salvo el hecho que uno ha sido escogido al azar para participar en el programa y el otro no. Por lo tanto, las evaluaciones aleatorias representan el caso ideal de una evaluación de impacto. Es por esta razón que en la evaluación de nuevas medicinas y en las investigaciones de ciencias naturales, se usa casi exclusivamente este método.[1]

FIGURA 2:
EL DISEÑO DE UNA EVALUACIÓN ALEATORIA

Sin embargo, la asignación aleatoria requiere que la evaluación se prepare ***antes*** de iniciar el programa. Por esto, este método también se denomina evaluación prospectiva. En un proceso aleatorio se asigna individuos (o empresas u otras entidades) al grupo de tratamiento y aquellos que no se seleccionan forman parte del grupo de control. El proceso aleatorio puede ser algo tan simple como tirar una moneda o un sorteo. Normalmente, la asignación aleatoria se hace a través de un simple proceso en Excel o Stata. No es necesario que los dos grupos sean de igual tamaño.

Aleatorización en la práctica

Es importante que el proceso de aleatorización sea realmente aleatorio y no un proceso que simplemente "parece" arbitrario. Por ejemplo, asignar personas con apellidos con la primera letra "A-L" a tratamiento y "M-Z" a control parece aleatorio, pero no lo es. Tal asignación requiere el supuesto que personas "A-L" son iguales a personas "M-Z". Pero es posible que las familias con apellidos con la primera letra "A-L" sean distintas de las familias con apellidos que comiencen con "M-Z". Por ejemplo, la composición étnica puede variar. Para evitar esta situación se recomienda el uso de un proceso automatizado, por ejemplo usando una computadora para generar números aleatorios y asignar tratamiento en base a estos números.

Una computadora también facilita procesos de aleatorización más complicados, como la aleatorización estratificada. La estratificación se recomienda cuando el número de participantes potenciales es pequeño, y en general para asegurarse que los dos grupos sean equilibrados con respeto a las variables más importantes. En la estratificación, se divide la muestra en sub grupos con característica similares y se aleatoriza dentro de cada subgrupo. Por ejemplo, si se divide la población por género y se asigna 30% de los hombres y 30% de las mujeres al tratamiento, la asignación será perfectamente equilibrada por género. El grupo de tratamiento tendría la misma composición de género que el grupo de control.

[1] Hay que distinguir entre una evaluación aleatoria y un muestreo aleatorio: Muchos estudios hacen muestreos aleatorios para sacar información representativa de la población. El muestreo aleatorio no intenta medir impacto. La característica distintiva de una evaluación aleatoria es la *asignación* aleatoria del tratamiento.

Según la ley de los grandes números cuando hay suficientes personas en cada grupo una asignación aleatoria genera dos grupos que se parecen en todas las características observables (como educación), y no observables, (como motivación). Por lo tanto, cualquier diferencia que surja posteriormente entre el grupo de tratamiento y el de control se puede atribuir al programa y no a otros factores. Por esta razón, si se diseñan e implementan adecuadamente las evaluaciones aleatorias son el método más confiable para estimar el impacto de un programa.

¿Cómo se determina el número requerido de participantes a un estudio aleatorio? Según la ley de los grandes números mientras más individuos están en el estudio más probable es que los dos grupos serán parecidos. Esta es una de las razones por la cual el tamaño de la muestra es importante. Un mayor tamaño siempre es mejor porque reduce la probabilidad de que, por casualidad, se obtengan grupos desbalanceados. Sin embargo, un estudio de mayor tamaño puede ser más costoso y no siempre es factible. Por lo tanto se recomienda hacer cálculos de poder estadístico para determinar cuál es el tamaño necesario para tener una buena esperanza de poder medir los impactos en las principales variables de interés.

Estudios aleatorios: Pasos a seguir

1) Escoger un programa y una población de interés, y las principales variables de interés.
2) Cálculos de poder estadístico: Determinar el tamaño requerido de los grupos de tratamiento y de control, para tener una buena esperanza de medir los impactos en las variables de interés.
3) Asignación aleatoria al tratamiento. Verificar que la asignación resultó equilibrada con respeto a las principales variables de interés.
4) Piloto: implementación del programa a pequeña escala para evitar problemas inesperadas (si posible).
5) Implementación: Asegurarse no haya ninguna otra diferencia entre grupos de tratamiento y de control.

Los cálculos de poder incorporan los distintos factores que afectan el número de participantes requeridos. Entre los factores a ser considerados están la varianza de la variable de interés y el efecto mínimo que se espera detectar. Mientras más alta es la varianza de la variable dependiente más observaciones son necesarias para poder detectar un efecto estadísticamente significativo. Mientras el efecto que se quiere medir es más pequeño más grande es el número de participantes requerido. Finalmente, el diseño de la aleatorización puede afectar el tamaño del grupo que se necesita. Si se aleatoriza a nivel de grupos (***diseño conglomerado***), por ejemplo todas las empresas de un mismo contador juntas, se requiere más empresas que si la aleatorización es a nivel individual.

Después de haber determinado el número de participantes requeridos, se puede proceder al proceso de asignación aleatoria. Es importante verificar que los grupos estén balanceados con respeto a las principales variables de interés. Los artículos académicos con estudios experimentales normalmente incluyen una tabla de balance que muestra que las principales características son parecidas en los dos grupos.

Finalmente se pasa a la implementación del programa o de la política a evaluar. En muchos casos, se recomienda hacer un piloto de la intervención a pequeña escala, para testear todos los procedimientos y evitar que se presenten problemas inesperados en la implementación. Durante la implementación es importante asegurarse que se respete la asignación aleatoria y que no se cambien participantes de un grupo a otro.[2] Lo más importante en este proceso es asegurarse que no haya *ninguna* otra diferencia entre el grupo de tratamiento y el grupo de control

[2] En el caso que no se haya respetado la asignación aleatoria en la implementación, se puede ocupar la metodología del "Intent-to-Treat", y con variables instrumentales observar el efecto "Treatment-on-the-Treated". Esto puede darse por ejemplo si se quiere medir el impacto de una fiscalización pero al momento de intentar de fiscalizar resulta que algunos contribuyentes en el grupo de tratamiento son no ubicables. O si se mandan cartas a contribuyentes y una parte de las cartas no llega. Es muy importante que en la evaluación de los datos se trabaje con la asignación aleatoria original, es decir comparar los que se *asignaron al tratamiento* con los que se *asignaron al control.* Nunca es válido comparar los que *de hecho* se trataron con los que se intentó tratar pero que al final no formaron parte del programa. La razón es que estos dos grupos ya no son iguales ex ante.

salvo la aplicación del programa. Por ejemplo, se perdería la validez del estudio si se detienen las otras actividades fiscalizadoras en el grupo de control pero se siguen aplicando en el grupo de tratamiento o al revés.

Esto concluye el resumen de los estudios aleatorios. Sin embargo, muchas veces no es posible asignar políticas o programas al azar. En los siguientes apartados se describen otros métodos de evaluación que intentan construir una aproximación del contrafactual bajo ciertos supuestos. La validez de cada método dependerá de que tan parecido sea el grupo de tratamiento al grupo de control antes de la intervención.

En resumen: Evaluación aleatoria

Descripción: Método experimental que sirve para medir relaciones causales entre dos variables comparando los tratados con los no tratados cuando la participación fue determinada aleatoriamente.

Representación del contrafactual: El grupo de comparación es seleccionado de forma aleatoria antes del comienzo del programa dentro de un grupo de participantes potenciales.

Supuestos claves: La aleatorización es válida. Es decir, los dos grupos son estadísticamente idénticos (en factores observables y no observables). No se le aplica ningún otro tratamiento diferente a alguno de los grupos.

Ventajas: La estimación del impacto del programa es muy creíble cuando se diseñó e implementó correctamente.

Desventajas: Requiere la asignación aleatoria antes del programa por lo que usualmente no se pueden hacer evaluaciones retrospectivas. El tamaño de la muestra debe ser lo suficientemente grande para poder detectar un resultado significativo.

2. DIFERENCIA SIMPLE (TRATADOS V. NO TRATADOS)

El ***método de diferencia simple*** es uno de los más comunes. La metodología es simple: comparar el grupo que recibió el programa con otro grupo que no lo recibió. Sin embargo, para ser una buena representación del contrafactual el grupo de control debería representar lo que hubiera pasado con el grupo de tratamiento sin el programa. ¿Esto es un supuesto creíble? Lamentablemente, muchas veces la respuesta es no.

En muchos programas hay un proceso de selección de quién recibe el tratamiento. A veces la selección es explícita; por ejemplo, un programa de fiscalización para el cual se seleccionan sólo a los contribuyentes con un indicador de riesgo alto. La selección también puede resultar de algo no explícito o no observable; por ejemplo, si los fiscalizadores escogen a aquellos contribuyentes que presienten tienen algún comportamiento irregular. En cualquier caso, esta asignación no aleatoria introduce un ***sesgo de selección***. Es decir, el grupo no tratado y el grupo tratado dejan de ser igual antes de la implementación del programa. La diferencia que se observa entre los grupos podría ser el resultado del impacto del programa, o de la diferencia original entre los dos grupos o de una mezcla de los dos.

Por ejemplo, existe un programa de tutores gratis para niños con problemas escolares y queremos medir su impacto. Si se compara simplemente las notas de los niños que recibieron la ayuda de un tutor con los que no la recibieron es posible que se observe que los niños con tutores tienen notas más bajas que los niños sin tutores. Concluir, en base a esta observación, que los tutores hicieron daño al logro escolar de los niños muy probablemente sería

erróneo. Lo más probable es que hubo una selección inicial en la cual niños con notas más bajas tenían mayor probabilidad de recibir la ayuda de un tutor. En este caso, el sesgo de selección introduce una subestimación del impacto tan fuerte que el impacto parece negativo en lugar de positivo.

En resumen: Diferencia simple

Descripción: Mide las diferencias después del programa entre aquellos que participaron en el programa y aquellos que no participaron.

Representación del contrafactual: El grupo de comparación corresponde a los individuos que no participaron en el programa (por alguna razón), y para los cuales tenemos datos después del programa.

Supuestos claves: Los no participantes son idénticos a los participantes excepto por la intervención del programa. No hay ninguna selección en el tipo de persona que entró al programa.

Ventajas: Muchas veces ya existen datos administrativos que se pueden analizar retrospectivamente. No requiere datos de la situación anterior al programa.

Desventajas: Necesita un grupo no afectado por el programa. Si los grupos tratados y no tratados son distintos antes del programa, el método puede sub estimar o sobre estimar el impacto verdadero de la política; es decir se introduce un sesgo de selección en la estimación.

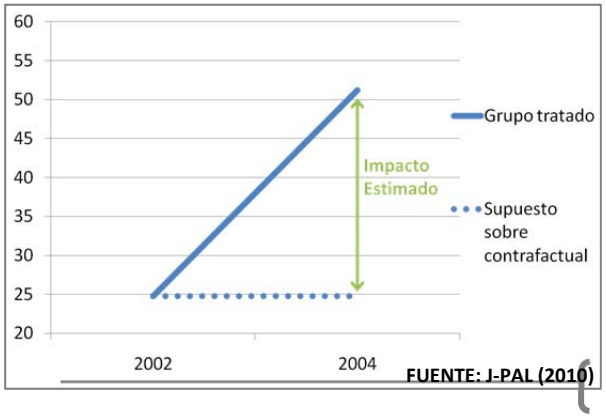
3. PRE-POST (ANTES VS. DESPUÉS)

Una ***evaluación pre-post*** es un tipo particular de evaluación de diferencia simple. En vez de usar otro grupo de personas como grupo del control se usa el mismo grupo de personas *antes del comienzo del programa.*

Por tanto, una evaluación pre-post mide el cambio en el tiempo tomando en cuenta el estado inicial del grupo. En este caso, se mide el impacto como la diferencia entre la situación anterior y la situación posterior a una intervención. El análisis pre-post es una manera muy común de evaluar programas. Muchas veces este tipo de análisis retrospectivo parece conveniente si los datos de la situación anterior al programa existen.

Una evaluación pre-post nos permite tomar en cuenta el nivel escolar original de los estudiantes. Pero, ¿el grupo de personas antes del comienzo del programa es una buena representación del contrafactual? Es decir, ¿es correcto suponer que sin el programa, durante este periodo no se hubiera dado ningún cambio en los resultados del grupo tratado?

FIGURA 3: SUPUESTO SOBRE CONTRAFACTUAL PARA PRE-POST



FUENTE: J-PAL (2010)

Miremos la situación en el ejemplo de los tutores gratuitos para estudiantes. ¿Es creíble asumir que en los 2 años del programa, los niños no hubieran mejorado sus notas sin los tutores? En realidad, es probable que los estudiantes hubieran seguido aprendiendo y mejorando sus conocimientos. Si se hace una evaluación pre-post, se atribui-

ría este aprendizaje, normal del desarrollo del niño, al programa de tutores.

Esta evolución natural del resultado a través del tiempo se llama ***tendencia secular***. Además de la tendencia secular, puede haber "choques" que cambian el resultado pero no tienen que ver con el programa. Por ejemplo, si hay una crisis económica durante el periodo de implementación de una política fiscalizadora el comportamiento tributario puede variar independientemente de esta política. En este caso no sería correcto atribuir el cambio del comportamiento tributario a la política. No se sabe si el cambio en el tiempo se debe a la crisis, a la política, o a una mezcla de las dos.

En resumen: Evaluación pre-post

Descripción: Mide el cambio en los resultados de los participantes de un programa en el tiempo. Es la diferencia entre la situación anterior y posterior a un tratamiento.

Representación del contrafactual: El grupo de comparación consiste en los mismos participantes del programa antes de su inicio.

Supuestos claves: El programa es el único factor que influyó en el cambio del resultado. Sin el programa el resultado se hubiera mantenido igual.

Ventajas: Muchas veces ya existen datos administrativos que se pueden analizar retrospectivamente. No requiere datos de personas que no participaron al programa.

Desventajas: Muchos factores cambian con el tiempo y pueden afectar el resultado, lo que va en contra del supuesto clave. En particular, la comparación pre-post no controla por el efecto de la tendencia secular o de choques, ajeno al programa, que afectan el resultado.

4. DIFERENCIAS EN DIFERENCIAS (DIFF-IN-DIFF)

Una evaluación de ***diferencias-en-diferencias*** combina los dos métodos anteriores para tomar en cuenta tanto las diferencias de nivel entre los dos grupos como las tendencias seculares.

La metodología de diferencias en diferencias usa las dos variaciones. La diferencia en el tiempo y la diferencia entre los dos grupos. Para calcular el efecto, primero se debe encontrar el cambio en el tiempo del grupo tratado (1) y el cambio del grupo no tratado (2) y luego restar estos dos resultados (3).

FIGURA 4 – CALCULANDO DIFERENCIAS-EN-DIFFERENCIAS

	Resultado antes del programa	Resultado después del programa	Diferencia
Grupo tratado	24,80	51,22	**26,42** (1)
Grupo no tratado	36,67	56,27	**19,60** (2)
Estimación diferencias-en-diferencias:			**6,82** (3)

FUENTE: J-PAL (2010)

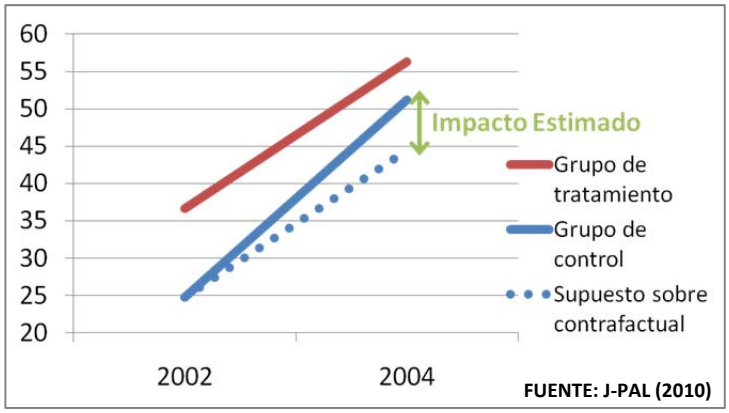
En una regresión múltiple la diferencia en diferencias se ve en el término de interacción entre el grupo tratado y el periodo post-tratamiento:

$$Y_{it} = \alpha + \beta_1 T_i + \beta_2 post_t + \beta_3 T_i * post_t + \epsilon_{it},$$

donde Y_{it} representa la variable de interés del individuo i en el periodo t, T_i es una variable dicotómica indicando si el individuo i recibió el programa o no, y $post_t$ es un dicotómica indicando el periodo posterior al programa, β_3 representa el estimador de la diferencia en diferencias.

En esencia, la estimación de diferencias en diferencias usa tanto el cambio en el tiempo del grupo no tratado como una estimación del contrafactual para el cambio en el tiempo del grupo tratado. El supuesto clave es que sin el programa la tendencia en los dos grupos hubiera sido igual. Esto es el ***supuesto de tendencias comunes o de tendencias paralelas*.** Se viola el supuesto si el grupo de tratados hubiera tenido una tendencia diferente al grupo de control en la ausencia del programa.

FIGURA 5: SUPUESTO SOBRE CONTRAFACTUAL EN DIFERENCIAS-EN-DIFERENCIAS



FUENTE: J-PAL (2010)

En el caso del programa de tutores para estudiantes el supuesto implicaría que sin la ayuda adicional los niños con tutor y sin tutor hubieran mejorado su rendimiento escolar al mismo ritmo. Pero es posible que aun sin el programa los niños lentos hubieran mejorado más que los avanzados, ya que tenían mucho que mejorar. O al revés, es posible que sin los tutores la distancia entre los niños lentos y los niños avanzados hubiera aumentado aún más. En los dos casos, no sabemos si la diferencia de la diferencia se debe a la característica de los grupos, al programa de los tutores o a una mezcla de ambas.

En Resumen: Diferencias-en-diferencias

Descripción: Compara el cambio en los resultados de los participantes con el cambio en los resultados de los que no participaron en el programa.

Representación del contrafactual: El cambio de los que no participaron en el programa sirve como representación del contrafactual del cambio de los participantes del programa.

Supuestos claves: Supuesto de tendencias comunes: Asume que sin el programa los dos grupos tendrían trayectorias idénticas a lo largo de este periodo.

Ventajas: Controla por todas las características que no cambian con en el tiempo (tanto observables como no observables) y por todos los cambios en el tiempo que afectan al grupo tratado y no tratado de igual manera.

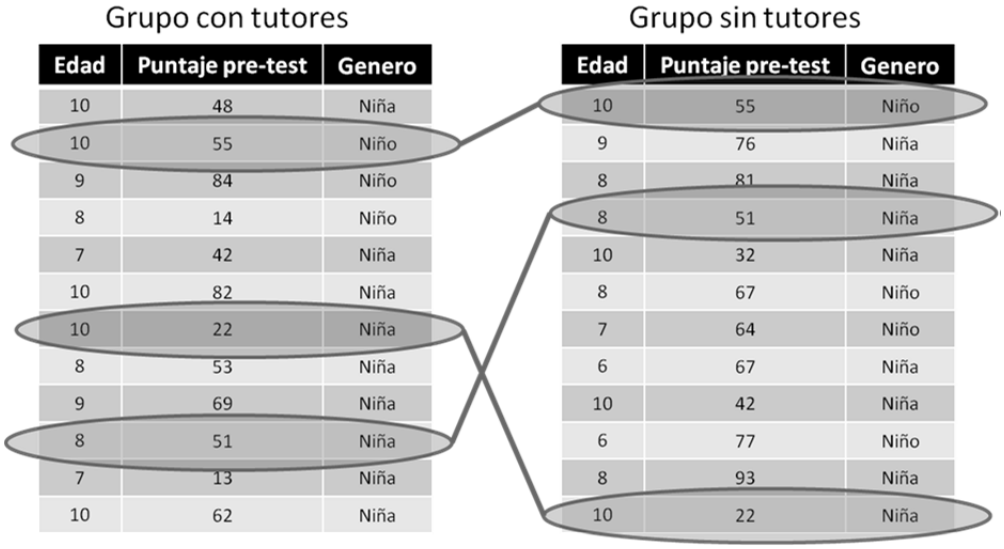
Desventajas: Si los dos grupos se hubieran desarrollado de manera diferente en la ausencia del programa existe un sesgo de selección. Se necesita un grupo no afectado por el programa y datos anteriores a la intervención.

5. MATCHING Y PSM

Con el ***matching*** volvemos a la meta original de construir una representación del contrafactual y crear un grupo igual al grupo tratado. El matching construye un grupo idéntico en características observables antes del programa. Existen varios métodos de matching. A continuación se describe el caso básico donde para cada individuo en el grupo tratado se busca un individuo con las mismas características observables en el grupo no tratado. Para estimar el impacto del programa se comparan los resultados entre el grupo de tratamiento y el grupo de control que está compuesto de individuos con características idénticas a los individuos tratados. Dado que los dos grupos tenían las mismas características observables antes del programa, se espera que la única diferencia después del programa sea la exposición al mismo.

En el caso de tutores, por ejemplo, podemos encontrar niños que no se inscribieron en el programa, pero que antes de la intervención tenían las mismas notas que un niño que recibió la ayuda de un tutor. De esta manera, creamos un grupo con todos los tratados y otro grupo con los pares de los tratados, es decir individuos no tratados que tienen las mismas características observables. La figura 6 muestra el proceso de selección de pares con tres características: edad, puntaje pre-test y género.

FIGURA 6 - EL PROCESO DE MATCHING EN EL EJEMPLO DE TUTORES



FUENTE: J-PAL (2010)

En ciertos casos, el matching puede ser mejor que diferencia en diferencias porque el proceso de encontrar pares nos asegura que los dos grupos son iguales en los factores observables que consideramos importantes. Pero, ¿es creíble asumir que el grupo tratado es igual al grupo que se le parece según las variables observables?

El problema es que el matching nunca puede controlar por las variables no observables. En el ejemplo de tutores, habrá alguna razón para qué dos niños con notas iguales reciban un tratamiento distinto. ¿Será qué la maestra sabe que uno tiene más potencial que otro? ¿Será qué uno tiene papás que lo apoyan más y le buscan un tutor? Si hay algo que no está en nuestros datos o que es difícil medir (por ejemplo, la motivación de los papás) que influye en el resultado, entonces volvemos al problema de sesgo de selección. Es probable, por ejemplo, que un niño con papás que lo apoyan hubiera mejorado más que su compañero con notas iguales aun sin el programa de tutores.

Aparte del problema de no observables, otro desafío del matching es que necesitamos encontrar individuos con las mismas características tanto en el grupo no tratado como en el grupo tratado. Este requerimiento se llama *la **con-***

***dición de apoyo común* (*common support condition*)**. En el ejemplo de los tutores, si fuera el caso que todos los estudiantes con notas muy bajas recibieron la ayuda de un tutor no sería posible hacer el matching por notas.

Finalmente, mientras más características queramos incluir en el matching más difícil es hacerlo. Con muchos datos (por ejemplo el censo de todos los estudiantes en el país) podría ser imposible encontrar un estudiante comparable que no recibió un tutor. Por otro lado, con menos datos puede ser que ciertos individuos en el grupo tratado no tengan un par exacto en el grupo no tratado.

Por estas razones se ha desarrollado el ***"Propensity Score Matching"*** (PSM*).* El PSM permite hacer un matching con muchas características. Se reduce el número de características a solo un índice que predice la probabilidad de formar parte del programa. En efecto, el índice es un promedio ponderado de las características subyacentes. El matching luego se hace entre individuos que tenían igual probabilidad de participar en el programa.

En Resumen: "Matching"

Descripción: Compara los resultados de individuos tratados con los resultados de individuos similares pero que no fueron tratados.

Representación del contrafactual:

Matching exacto: Para cada participante, se escoge al menos un no participante que es idéntico en las características seleccionadas.

Propensity score matching (PSM): Se compara participantes del programa a no participantes que según sus características observables tenían la misma probabilidad de participar en el programa.

Supuestos claves: Los no participantes en promedio son idénticos a los participantes "emparejados", excepto por la participación en el programa.

Ventajas: No requiere una aleatorización anterior al programa. Nos puede dar no solo el impacto promedio del programa, sino también la distribución del impacto del programa.

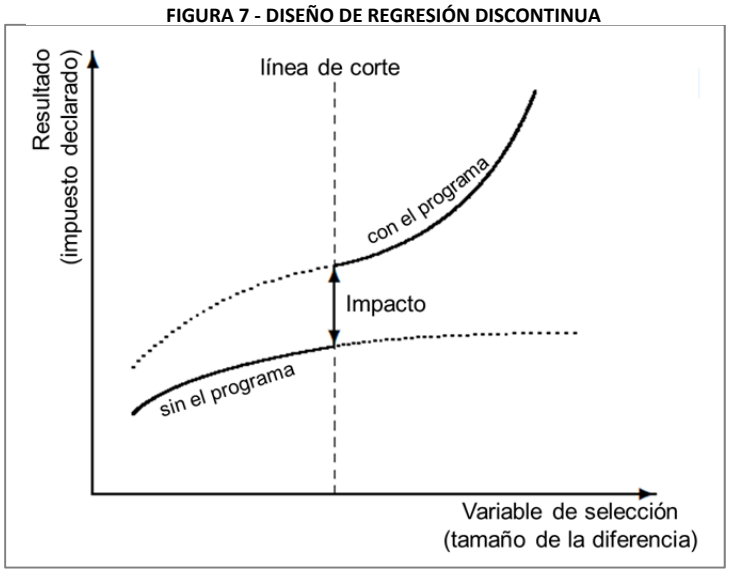
Desventajas: Pueden existir características no observables que afectan la probabilidad de participar al programa y al mismo tiempo afectan los resultados. Esto introduce un sesgo de selección. Saber si es probable que las características no observables sean importantes en este contexto requiere conocer muy bien la manera en la cual se seleccionaron los participantes del programa.

6. REGRESIÓN DISCONTINUA

Existe una metodología que permite sacar conclusiones causales tan fiables como el experimento aleatorio, que se puede aplicar en ciertos casos especiales. A veces programas o políticas tienen un umbral específico que determina quién recibe un tratamiento. Un diseño de ***regresión discontinua*** se aprovecha del hecho que los individuos o empresas muy cercanas al umbral son básicamente iguales. Bajo ciertos supuestos, se puede interpretar la diferencia entre los resultados de los individuos justo debajo del umbral (que no reciben el programa) y los resultados de los individuos justo encima del umbral (que reciben el programa) como el impacto de la intervención.

Supongamos por ejemplo que se diseña un programa que manda una carta de notificación a todas las empresas que tengan una diferencia con información de terceros superior a 100 dólares. En este caso, el tamaño de la diferencia es *la variable de selección* porque la línea de corte está definida por esta variable.

FIGURA 7 - DISEÑO DE REGRESIÓN DISCONTINUA



La figura 7 muestra el concepto de una evaluación de regresión discontinua. La línea sólida representa la relación entre el tamaño de la diferencia y el monto de impuesto declarado: mientras más grande la diferencia, más impuesto se declara. Se puede ver que en la región de la línea de corte, el umbral sobre el cual se manda la carta, hay *una discontinuidad* o "salto" en el pago de impuestos. Bajo ciertas condiciones se puede atribuir este salto al envío de la carta.

Uno de los supuestos más importante para usar el diseño de regresión discontinua es que no hubo un cambio estratégico en el comportamiento de las firmas alrededor del umbral. Si por ejemplo, las empresas justo debajo de 100 dólares de diferencia tenían buenos contadores que sabían cómo manejarse para quedar debajo del límite, existe una diferencia entre las empresas justo debajo y justa arriba del umbral. Tal diferencia entorno al umbral introduce un sesgo de selección. La manipulación alrededor del umbral se denomina una ***respuesta conductual al umbral***.

La ventaja de la regresión discontinua es que el supuesto que no hay respuesta conductual al umbral se puede testear. Si hubiera manipulación se produciría una concentración más alta de empresas justo arriba o justo debajo del umbral, lo que se puede verificar. Al igual, se puede verificar que no haya diferencias en las características claves entre las empresas justo debajo y sobre el umbral.

Finalmente, un diseño de regresión discontinua también requiere que no haya otros programas o políticas que se apliquen al mismo umbral. Por ejemplo, si las empresas con diferencias mayores a 100 dólares además recibieron una visita de un fiscalizador no se puede distinguir el impacto de la visita del impacto de la carta.

Ambos problemas, la respuesta conductual al umbral y otras políticas que se aplican al mismo umbral, se presentan con mayor frecuencia cuando el umbral es un número conocido por todos. Por lo tanto, los umbrales óptimos para el uso de esta metodología son secretos, o definidos ex-post, y se aplican en la implementación de un solo programa.[3]

En el análisis de regresión discontinua no se compara simplemente los resultados de las empresas o individuos justo debajo del umbral con los resultados de los que están justo encima. Se corre una regresión en la cual se controla por el cambio en la variable de selección de manera lineal y también con potencias de la variable de selección. Para ver los detalles refiérase a la bibliografía.

[3] Un ejemplo de un umbral escogido ex-post sería si se define un monto de ventas declaradas, debajo del cual se aplica un cierto tratamiento, después de que las declaraciones se hayan presentado. En este caso, las empresas no pueden ajustar sus ventas declaradas al umbral porque en la fecha de hacer la declaración no se conocía el punto de corte. En el caso que el umbral sea público y conocido anteriormente es muy importante testear si hay respuesta conductual al umbral antes de aplicar la regresión discontinua. Si existió manipulación alrededor del punto de corte el estimador calculado con una regresión discontinua no es válido. .

En resumen: Regresión discontinua

Descripción: Compara los resultados de individuos que están justo debajo de un umbral que los califica para el tratamiento con los resultados de individuos que están justo arriba de este umbral.

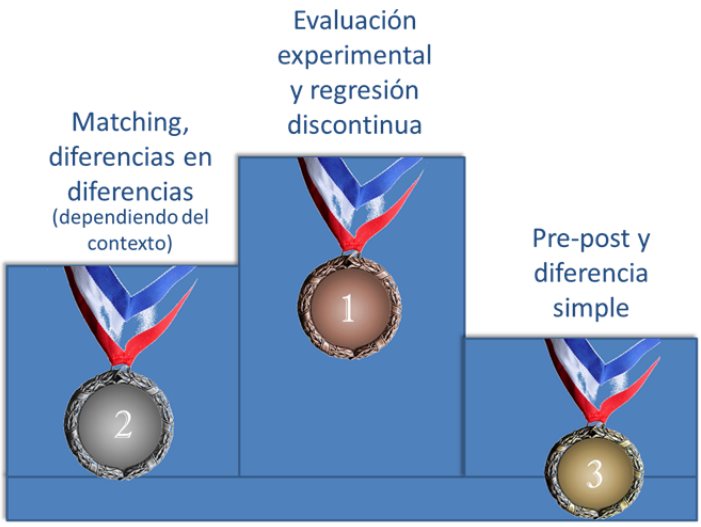
Representación del contrafactual: Los resultados de los individuos que están cerca de la línea de corte, pero que caen en el otro lado y por tanto no pueden participar en el programa, representan el contrafactual de los individuos que caen justo encima del umbral y por lo tanto reciben el tratamiento.

Supuestos claves: Los individuos justo arriba de la línea de corte son iguales a los individuos que caen justo debajo de la línea de corte. No hay ni manipulación alrededor del umbral ni otras políticas que se aplican a partir del mismo corte.

Ventajas: Produce estimaciones muy fiables del impacto. En las administraciones tributarias, existen muchas políticas que se aplican según un corte y muchas veces ya existen los datos administrativos que se requieren para el análisis. La mayoría de los supuestos se dejan testear.

Desventajas: Las conclusiones solamente se aplican a individuos o empresas alrededor del corte. No se puede saber cuál sería el impacto en aquellos que están muy lejos del umbral.

FIGURA 8: COMPARACION (MUY SIMPLIFICADA) DE LOS MÉTODOS



Bibliografía

Abdul Latif-Jameel Poverty Action Lab (J-PAL). "¿Por qué aleatorizar?" *La evaluación de programas sociales*. Universidad de los Andes. Bogota, Colombia. 14 Jul 2010.

Textos generales

Angrist, Joshua D., and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press, 2009.

Imbens, Guido, and Jeffrey Wooldridge. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*. 47.1 (2009): 5-86.

Evaluación experimental

Banerjee, Abhijit, and Esther Duflo. "The Experimental Approach to Development Economics." *Annual Reviews of Economics*. 1. (2009): 151-178.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. "Using Randomization in Development Economics Research: A Toolkit." *Handbook of Development Economics*. 4. (2007): 3895-3962.

Ludwig, Jens, Jeffrey Kling, and Sendhil Mullainathan. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives*. Forthcoming (2011).

Diferencias en diferencias

Duflo, Esther. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review*. 91. (2001): 795-813.

Abadie, Alberto. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies*. 72. (2005): 1-19.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How Much Should We Trust Differences-In-Differences Estimates?." *Quarterly Journal of Economics*. 119.1 (2004): 249-275.

Matching

Dehejia, Rajeev, and Sadek Wahba. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*. 94. (1999): 1053-1062.

Diseño de regresión discontinua

Imbens, Guido and Thomas Lemieux. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*. 142. (2008): 615-635.

Lee, David, and Thomas Lemieux. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*. 48.2 (2010): 281–355.