UNIVERSITAT POMPEU FABRA

Políticas basadas en la evidencia: el papel de la evaluación y la experimentación social

José García Montalvo
Catedrático de Economía

Jornada IVALUA sobre evaluación y experimentación
26 de Septiembre de 2013

Introducción

"El diablo que hay en el mundo proviene de la ignorancia, y las buenas intenciones pueden hacer tanto daño como las malas intenciones, si falta el conocimiento"

Albert Camus, *La Plaga*

"Without data you are just one more person with an opinion"
(Anonymous)

Resumen de la presentación

- Evaluación y causalidad
- El método experimental
- Algunos ejemplos
- Conclusiones

Supuestos e inferencia estadística

- Siempre es necesario utilizar supuestos para obtener cualquier resultado. La credibilidad del resultado depende de lo apropiado de los supuestos: no hay “free lunch”
- Cuanto más tiempo y dinero se invierte en el diseño menos supuestos hace falta hacer
- Muchas bases de datos no puede soportar una inferencia causal decente

La esencia de la evaluación

- Evaluar supone medir lo que ha pasado frente a lo que habría sucedido de no haber realizado una actuación. La segunda situación es contrafactual: un individuo no puede ser tratado y no tratado a la vez. Grupo de comparación.
- Problema básico: datos faltantes.
- Evaluar supone también ser capaces de atribuir a la actuación el efecto: causalidad

Problemas básicos de los datos observacionales

- Endogeneidad
  - Omisión de variables correlacionadas con variables explicativas incluidas
  - Errores de medida
  - Simultaneidad: la relación entre instituciones y desarrollo económico
- Selección muestral

Selección muestral

- Elecciones presidenciales de 1936 en Estados Unidos
- Lista muestral: familias con teléfono y dueños de coches registrados
- Resultado de la encuesta: Landon (republicano) ganaría con el 57% de los votos frente a Roosevelt (democrata)
- La estadística no falla. Es el estadístico el que puede equivocarse.

Nuevas soluciones

- En búsqueda de una fuente exógena de variación
  - Imposible (crítica de Lucas): en el análisis económico todas las variables se fijan simultáneamente. Shocks en la función de utilidad, la productividad, etc. cambian los parámetros de las formas reducidas. Los datos son “sucios”. Es mejor la simulación utilizando la estimación de “deep parameters”. DGEM y ciclos reales
  - Construir experimentos o buscar experimentos naturales o pseudo-experimentos

Experimentos randomizados

- Es el estándar en la mayoría de las ciencias “duras”: análisis de nuevos medicamentos, procedimientos médicos, etc. FDA
- Basado en un grupo de control (al que se administra placebo) y un grupo tratado.

Experimentos randomizados

- Algunos problemas:
  - Pueden ser caros
  - Dependiendo del objeto de estudio puede ser imposible realizar un experimento
  - En ocasiones existen problemas de tipo ético

Experimentos frente a estudios observacionales

- ¿Podemos fiarnos de los resultados de estudios que no utilizan el método experimental?
  - ¿Afecta a la longevidad el hacerse un chequeo medio anual?
  - ¿Es efectiva la terapia de reemplazamiento hormonal en mujeres con menopausia?

Experimentos randomizados

- La gran ventaja es que proporciona unos resultados creíbles y difícilmente manipulables si se han seguido correctamente los protocolos
- La técnica estadística es muy simple y no manipulable: normalmente una diferencia de medias (y su desviación estándar) es suficiente.

Los criterios de la FDA

- Antes de la aprobación de un nuevo principio activo es preciso probar, por métodos experimentales:
  - que dicha molécula tiene efecto sobre el problema que pretende resolver (tamaño de un tumor, recuento vírico, etc.)
  - que el procedimiento es más efectivo que los existentes
  - que no tiene efectos secundarios perjudiciales (o que tiene efectos muy limitados)

Experimentos en ciencias sociales

- ¿Pueden hacerse experimentos randomizados en ciencias sociales? Por supuesto. Algunos ejemplos:
  - El experimento Perry (1962) y el Abcedarium
  - El efecto del tamaño de las clases sobre el aprendizaje: STAR
  - El efecto de los cheques escolares: PACES (Colombia) o “School Choice Scholarship Foundation Program” en Nueva York
  - Reducción de la pobreza: PROGRESA (México).
  - Programa “Moving to Opportunity” para familias que viven en barrios degradados de Baltimore, Boston, Chicago, Los Angeles y Nueva York.

Experimentos en ciencias sociales

- En el campo de la evaluación educativa el Instituto de Evaluación Educativa de Estados Unidos ha realizado 175 experimentos randomizados desde la aprobación de la ley Not Child Left Behind (NCLB)
- What works clearinghouse es la base de datos que contiene todos los resultados de estos experimentos (como la FDA)
- NYT, 2 de septiembre de 2013: “Guesses and Hype Give Way to Data in Study of Education”

Experimentos en ciencias sociales

- Andrew Delbanco (director of American studies at Columbia): the tension between quantitative and qualitative methodologies — or, more accurately, between methodology and non-methodology — has been explored “for many centuries” under different rubrics: “facts versus knowledge, skill versus wisdom ...information versus insight.”
- “Education in the United States ... has been moving lately toward the first term [in these pairs] and away from the second.”

Experimentos en ciencias sociales

- Por desgracia España es diferente también en esto:
  - Reformas educativas siguen polemizando sobre la asignatura de religión o educación para la ciudadanía como si esto tuviera alguna relevancia
  - Políticos siguen tozudamente haciendo caso omiso a la evidencia existente e incluso desarrollando un programa de gasto en ordenadores para todos los estudiantes cuando no existe evidencia sobre la eficacia del CAL (computer assisted learning)

Un ejemplo en detalle: STAR

- El experimento de reducción de tamaño de las clases de Tennessee (STAR: Student-Teacher Achievement Ratio)
  - Experimento de cuatro años diseñado para evaluar el efecto del aprendizaje en clases pequeñas
  - Coste: 12 millones de dólares
  - Tres tratamientos:
    - Clases tamaño normal (22-25 estudiantes)
    - Clases pequeñas (13-17 alumnos)
    - Clases tamaño normal y un profesor de apoyo.

STAR (cont.)

- Protocolo
  - Los estudiantes que entran en la guardería en los colegios participantes son asignados aleatoriamente a uno de los tres grupos
  - Los estudiantes continuarán en el mismo grupo durante los cuatro años
  - En el primer año 6.400 estudiantes participaron en 108 clases pequeñas, 101 clases normales y 99 clases con profesor de apoyo. En total (cuatro año): 11.200 estudiantes

STAR (cont.)

- La medida de resultados utilizada para comprobar el efecto de los diferentes tratamientos es la puntuación de los alumnos en el test combinado de matemáticas y lectura de Stanford (Stanford Achievement Test)
- Especificación

$$Y_i = \beta_0 + \beta_1 ClasePeque\ \tilde{n}a_i + \beta_2 Papoyo_i + u_i$$

STAR (cont.)

Reg\Curso	G	1	2	3
Clase pequeña	13.90** (2.45)	29.78** (2.83)	19.39** (2.71)	15.59** (2.40)
Normal y prof. apoyo	0.31 (2.27)	11.96** (2.65)	3.48 (2.54)	-0.29 (2.27)
Constante	918** (1.63)	1,039** (1.78)	1,157** (1.82)	1,228** (1.68)
N	5,786	6,379	6,049	5,967

STAR (cont.)

- Interpretación de los resultados:
  - Para comparar entre cursos es necesario trasladar los resultados de los tests en términos de desviaciones estándar. El efecto de las clases pequeña es reducido. Comparación con la mejora de la calidad del profesor
  - Además, el efecto se concentra en los primeros años. Continuar en una clase pequeña no produce mejoras adicionales.

La crítica de LaLonde

- Evaluación de programas de formación. LaLonde (1986) tomó los datos del National Supported Work Demonstration (NSW) y comparó los resultados del experimento randomizado con los obtenidos por varias técnicas econométricas habituales.

La crítica de LaLonde

- Heckman's two steps model (no para explicar)
- Supuesto de normalidad por todas partes

$$\text{Prob}(D = 1|Z) = \Phi(Z\gamma),$$

$$w^* = X\beta + u$$

$$E[w|X, D = 1] = X\beta + E[u|X, D = 1].$$

$$E[w|X, D = 1] = X\beta + \rho\sigma_u\lambda(Z\gamma),$$

La crítica de LaLonde

	Control	Hombres	Mujeres
Experimental	NSW	886 (476)	851 (317)
Est. 1 paso	PSD1	-1,228(896)	2,097 (491)
	CPS1	-805 (484)	1,041 (505)
Bietápicos (Heckman)	PSD1	-1,333(820)	1,129 (385)
	CPS1	-22 (584)	1,102 (323)

La crítica de LaLonde

- Los métodos no experimentales producen resultados muy diferentes dependiendo del grupo de control utilizado y de la especificación econométrica adoptada.
- Los resultados de los métodos experimentales pueden ser muy diferente de los resultados no experimentales
- Incluso cuando los estimadores no experimentales pasan los test convencionales pueden fracasar en replicar los resultados experimentales.

¿Y si no podemos hacer un experimento randomizado?

- Experimentos naturales
- Quasi experimentos y métodos de emparejamiento (selección en observables)
- Variables instrumentales (selección en no observables). Randomización es el instrumento perfecto
- RDD: “Regression discontinuity design”

¿Por qué no se evalúa?

- Pritchett (2002): los promotores de las intervenciones tienen interés en mostrar que los programas tienen un gran impacto. Si se hiciera una evaluación randomizada se revelaría el verdadero impacto, donde no existe garantía de que sea positivo ni importante.

¿Por qué no se evalúa correctamente?

- Dos teorías contrapuestas sobre la utilidad de los experimentos:
  - la teoría de los paracaídas en la prevención de un grave trauma relacionado con el reto de la gravedad: "Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials"
  - Las medicinas y la seguridad alimenticia son importantes para los políticos. La educación no se consideran tan importante. Por tanto la evaluación correcta no importa desde un punto de vista del ciclo político

La “teoría del paracaídas”

- La teoría del paracaídas aplicada a la educación o la ayuda al desarrollo supone que:
  - las intervenciones y la ayuda al desarrollo no tienen nunca efectos secundarios indeseados.
  - los procedimientos alternativos son todos igualmente eficientes o que no hay restricciones presupuestarias.

Efectos secundarios

- Maren (1997) señala que la lucha por el control de la ayuda fue una de las causas del comienzo de la conflicto de Somalia entre los señores de la guerra.
- Rajan y Subramaniam (2005) señalan que la ayuda al desarrollo reduce la competitividad
- El aumento de la proporción de ayuda sobre el PIB reduce el nivel de democracia del país.
- Easterly (2006) cita un enorme listado de efectos secundarios (White man's burden). Muchos son bastante sofisticados y difíciles de prever sin un estudio piloto de tipo experimental

¿Son todos los programas igualmente efectivos?

- No. Ejemplos en el caso de la educación y desarrollo:
  - un programa de eliminación de los parásitos intestinales en niños cuesta $3.5 dólares para conseguir un año adicional de asistencia a la escuela.
  - la provisión de uniformes y libros gratis cuesta $99 dólares por año adicional de asistencia.
  - un programa de provisión de alimentos en las escuelas cuesta $36 dólares por año adicional inducido de asistencia a la escuela.

¿Da lo mismo el método de evaluación de la ayuda?

- Glewwe, Kremer, Moulin y Zitzewitz (2004): El efecto de los “flip charts” sobre el aprendizaje de los alumnos (Kenia).
- Dos tipos de evaluación:
  - Con datos observacionales
  - Con datos experimentales
- ¿Qué es un “Flip chart”? Ayuda visual que contiene una serie de cuadros (alrededor de 12) unidos por un espiral.

¿Por qué usar “Flip charts”?

- Los libros de texto son muy escasos en las escuelas de la Kenia rural. Además están escritos en inglés (medio de instrucción en el país) y muchos estudiantes tienen un dominio limitado de dicho idioma.
- Los FC pueden promover el aprendizaje pues:
  - Todos los estudiantes recuerdan gráficos y fotografías más a menudo que haber leído palabras
  - Los estilos de aprendizaje varían mucho entre estudiantes: añadir ayudas visuales aumenta el espectro de estudiantes a los que se puede enseñar efectivamente
  - Las clases y presentaciones se benefician de tener ayudas visuales suplementarias

Flip charts: datos observacionales

- 100 colegios de un estudio sobre el efecto de la provisión de libros de texto y becas aleatoriamente en escuelas de Kenia.
- También se recogió información sobre materiales docentes en las escuelas y, entre ellos, la disponibilidad de “flip charts”
- Se dispone de información sobre el número de FC de ciencias, matemáticas, salud-negocios en cada escuela.

Flip charts: datos observacionales

- Para poder comparar con la intervención experimental que se comentará posteriormente (distribución de 4 FC) el número de FC disponibles en cada escuela se divide por 4 (bloques de FC comparables)
- Resultados: los estudiantes que utilizaron los FC aumentaron un 20% de una DE sus resultados escolares (un 10% es como pasar del percentil 50 al percentil 54). Tabla.

Flip charts: datos observacionales

- Si los resultados observacionales son correctos entonces los FC son muy efectivos: el coste por alumno de los cuatro FC (un set=$80) es solo el 10% del coste de los libros en las tres asignaturas cubiertas por los FC (un libro de texto en Kenia cuesta $3.3x3x80 estud.=$800) pero su impacto es dos veces más grande que el efecto de proveer libros de texto.
- Los FC serían, por tanto, 20 veces más efectivos en términos de coste que los libros de texto.

Flip charts: datos experimentales

- La ONG International Christelijk Steunfonds (ICS) proporcionó los FC.
- La intervención consistía en 4 FC: dos sobre ciencias (agricultura y ciencia en general), uno sobre salud, uno sobre matemáticas y un mapa del este de África para geografía.

Flip charts: datos experimentales

- El diseño experimental consistió en la selección de 178 escuelas en Busia y Teso (Kenia) por la ONG ICS
- Las escuelas elegidas tenían una media de aprendizaje de los estudiantes similar a la media del distrito en su conjunto
- 89 escuelas fueron asignadas al grupo de control y otras 89 al grupo tratado después de ser ordenadas por orden alfabético. De cada dos escuelas consecutivas una se asignaba al grupo tratado y otra al de control (sistemático)

Flip charts: datos experimentales

- Los resultados experimentales no muestran ningún efecto significativo de los “flip charts”:
  - Resultados agregados
  - Resultados por asignatura
  - Resultados por grupos de asignaturas

¿Qué funciona?

- Usos de la ayuda al desarrollo muy efectivos (en términos coste-beneficio)
  - Medicinas para acabar con los parásitos intestinales
  - Suplemento como el hierro y la vitamina A
  - Subsidios a los fertilizantes
  - Vacunación de los niños
  - Provisión de agua potable

¿Qué funciona?

- Las anteriores son intervenciones modestas, frente a las grandes promesas y el esquema utópico (hacerlo todo a la vez), que tienen una enorme incidencia sobre el bienestar de las familias pobres.

No es efectivo…

- “Flip charts”
- Dar libros (solo beneficia a los estudiantes en el segmento superior de la clase)
- Incentivos económicos para los profesores (acaban enseñando como hacer tests)

Conclusiones

- Un evaluación creíble permitiría arrebatarle a la política el protagonismo en ámbitos en los que no debería tenerlo
- Los programas pueden fallar y de eso también se aprende. Tapar los fallos evitando la evaluación de resultados o utilizando procedimiento no adecuados hace una mal servicio a la comunidad científica y a la ciudadanía en general
- La evaluación experimental puede dar resultados muy diferentes a los obtenidos por procedimientos tradicionales

Conclusiones

- La evaluación experimental (randomizada) es factible en el campo de las políticas sociales, educativas y de ayuda al desarrollo. Precisa de recursos humanos y económicos pero no más que la recolección de otros datos para otros tipos de evaluación

Conclusiones

- Agencias y ONGs deben experimentar y buscar intervenciones que funcionan verificándolo por medio de las mejores técnicas de evaluación científica disponibles y usando evaluadores externos para evitar los intereses de los gestores de los proyectos.
- Para aprender es necesario tener información.

UNIVERSITAT
POMPEU FABRA