Evaluar intervenciones sanitarias sin experimentos

M. Vera-Hernández
Department of Economics, University College London. Londres. Reino Unido.

Correspondencia: Dr. Marcos Vera Hernández. Department of Economics, University College London.
Gower Street, London WC1E 6BT. Londres. Reino Unido.
Correo electrónico: uctpamv@ucl.ac.uk. http://www.homepages.ucl.ac.uk/~uctpamv.
Tel. +44-207-679-5808. Fax. +44-207-916-2775

(Evaluating health interventions without experiments)

Nota Editorial: Este artículo corresponde a un informe técnico encargado por la Junta Directiva de la Asociación de Economía de la Salud (AES) en el marco del vigente acuerdo de cooperación Gaceta Sanitaria AES, que establece un proceso de revisión editorial por expertos independientes similar al aplicado al resto de manuscritos.

Resumen

En el presente artículo se revisa la bibliografía reciente en evaluación cuantitativa de intervenciones no experimentales, poniendo especial énfasis en su aplicación a la economía y la gestión sanitarias. En particular, se han descrito las técnicas de *matching* y de doble diferencia combinada con *matching*. El parámetro elegido como objeto de la estimación es la ganancia media para los participantes en la intervención, bajo la hipótesis de heterogeneidad en las ganancias no observables que produce la intervención entre los individuos elegibles. Se ha llevado a cabo una exposición no técnica de las metodologías descritas con el espíritu de fomentar al lector una lectura más profunda de la bibliografía relevante.
Palabras clave: Estadística. Evaluación de programas. Reforma sanitaria.

Abstract

This paper summarizes recent literature on quantitative techniques for the evaluation of non experimental reforms. We closely look at the application of the methods to health economics and health management. The methods of matching and difference in differences combined with matching have been analysed in greatest detail. We have focused our attention on the estimation of the average treatment for the treated as the relevant parameter to be estimated. Along the paper, we have assumed that gains from the reform are heterogeneous in non observable variables across eligible individuals. The methods are described in a non technical manner to motivate further reading.
Key words: Statistics. Program evaluation. Health care reform.

Introducción

Este artículo tiene dos objetivos básicos: el primero es la revisión de técnicas econométricas recientes para la evaluación de políticas y reformas sociales. Recientemente, los economistas se han replanteado los supuestos, los parámetros que se deben estimar y los métodos de estimación utilizados para evaluar intervenciones sociales. Como resultado de ello han salido fortalecidas las técnicas de *matching* y de doble diferencia combinada con *matching**. Si bien ha habido aplicaciones de estas técnicas en economía laboral, de la educación y del desarrollo, todavía han sido poco utilizadas en economía de la salud. Esto enlaza con el segundo objetivo del artículo, que consiste en plantearse cómo las técnicas mencionadas pueden aplicarse en el campo de la gestión sanitaria, entendida en sentido amplio. Para abordar este objetivo hemos elegido subrayar los supuestos que subyacen en los distintos métodos, resaltar las ventajas de las nuevas técnicas frente a las más conocidas y dar ejemplos de posibles situaciones de aplicación. En estos ejemplos hemos dado más peso a su carácter pedagógico que a su coincidencia con situaciones reales. A lo largo del artículo también utilizaremos ejemplos para explicar cuándo los supuestos necesarios para la validez de las técnicas propuestas pueden ser demasiado restrictivos. Ya existen algunas revisiones de la bibliografía en cuanto a técnicas cuantitativas de evaluación, pero ninguna de ellas ha contemplado el campo de la economía de la salud[1-3].

La reciente finalización del traspaso de transferencias sanitarias a las comunidades autónomas generará diversas experiencias de gestión. Si esto ocurriera, una evaluación rigurosa de estas experiencias ayudaría a comprender mejor el funcionamiento del sistema sanitario. En este sentido conviene subrayar que ya existen programas informáticos disponibles para implementar muchas de las técnicas que aquí se mencionarán[4,5]. Ello reducirá los obstáculos de aplicarlas, ya que no se requerirá la realización de programas informáticos específicos.

Marco general

En este apartado explicaremos el tipo de intervenciones que se van a evaluar, su alcance y las características generales que tendrán los datos que examinaremos. El artículo está pensado para la evaluación de políticas, reformas o intervenciones que afecten directamente a sujetos o entidades individuales tanto del lado de la oferta (médicos, centros de salud, hospitales…) como de la demanda (pacientes). Ejemplos de posibles intervenciones que evaluar son un programa antitabaquismo impartido en un centro de salud, la introducción de un esquema de incentivos a profesionales sanitarios o un cambio en la cobertura sanitaria de un colectivo.

La intervención o reforma será entendida de una manera discreta. Aunque se pueden extender las técnicas aquí planteadas para analizar casos de intervenciones múltiples, preferimos sólo contemplar el caso donde hay una única intervención que se puede dar o no.

En el artículo asumiremos que hay una intervención o reforma que evaluar. Así, tendremos individuos que han participado en la reforma; por ejemplo, médicos sujetos a un esquema que incentive algún patrón prescriptor previamente establecido. A este grupo lo llamaremos participantes o grupo de tratamiento. También contaremos con individuos que no han participado en la reforma que se evalúa, que llamaremos no participantes, o grupo de control. Siguiendo el ejemplo, el grupo de control estaría formado por médicos que no han estado sujetos al esquema de incentivos. Para cada individuo del grupo de tratamiento y de control, contaremos con una variable de resultado, que es la variable en que se centrará la evaluación. En el caso del ejemplo, la variable resultado puede ser el porcentaje de medicamentos genéricos que prescribe cada médico. Para cada individuo también contaremos con un conjunto de variables, denominadas variables condicionantes, que afectan a la variable resultado. En ocasiones también las denominaremos variables observables. La característica fundamental de este conjunto de variables es que su valor no haya sido causado por la intervención que se evalúa o se haya alterado por ésta. En cierto sentido, son variables anteriores a la intervención. Siguiendo con el ejemplo propuesto, dentro del conjunto de variables condicionantes, se podrían incluir el porcentaje de medicamentos genéricos prescritos por cada médico en el pasado, la experiencia y la especialidad, pero no el tiempo que ha dedicado el médico a su formación durante el período que ha durado el esquema de incentivos. Esta variable podría haber sido influida por la intervención que intentamos evaluar: los médicos pueden haber aumentado el tiempo dedicado a formación para identificar en qué situaciones prescribirán especialidades farmacéuticas genéricas. Cabe resaltar que las variables condicionantes no necesariamente han de ser exógenas. Por ello, dentro del conjunto de variables condicionantes podremos introducir el valor pasado de la variable resultado (en este ejemplo, el porcentaje de medicamentos genéricos prescritos en el pasado por cada médico). Se requerirá que todos estos datos estén disponibles en un período posterior a la intervención que se va a evaluar, y en algunos estimadores que también lo estén en un período anterior a la estimación. En cada caso se harán explícitos los datos necesarios.

Nos centraremos en el estudio de técnicas estadísticas útiles para la evaluación de datos de estudios observacionales, en contraposición con los experimentales. Entendemos por estudios observacionales aquellos en que la elección entre participar o no en la intervención ha sido tomada bien por el individuo susceptible de ser expuesto, bien por alguna otra persona sin seguir un procedimiento aleatorio. También consideraremos que se trata de un estudio observacional cuando los participantes han sido escogidos mediante un conjunto de reglas que no seguían un esquema aleatorio. Cuando tiene lugar un procedimiento no aleatorio, los datos presentan características que suelen hacer más difícil y menos creíble la evaluación. Por ello nos centramos en este caso, pues es el más complejo y donde han tenido lugar los avances metodológicos a los que nos referíamos al principio del artículo. Además, las intervenciones de gestión que han tenido un esquema de asignación aleatorio son escasas, lo que no es extraño, pues para llevar a cabo un experimento social se requiere que un individuo que haya aceptado participar sea excluido de la reforma, con el objeto de llevar a cabo la evaluación (nótese que en ensayos clínicos esta exclusión se consigue por medio del placebo). Siguiendo con el ejemplo anterior, un enfoque experimental requeriría no dar el esquema de incentivos a un grupo de médicos que lo conozcan y que hayan decidido aceptarlo. Es evidente que en muchas ocasiones este planteamiento experimental es inviable desde un punto de vista práctico, e incluso puede presentar un problema ético importante. En cualquier caso, se han de mencionar algunos casos donde se ha podido llevar a cabo un enfoque experimental: el RAND Health Insurance Experiment en los EE.UU. y el programa PROGRESA en México[6,7].

Las técnicas aquí descritas servirán para evaluar el cambio que tiene la reforma sobre la variable resultado. Esta información será útil para llevar a cabo un análisis de coste-beneficio, pero generalmente no será suficiente. Por ejemplo, para evaluar cómo influye el esquema de incentivos a los médicos en la calidad de la prescripción o en los costes monetarios totales, no será suficiente con estimar cómo se ha visto afectada la proporción de especialidades farmacéuticas genéricas prescritas.

Evaluar políticas sociales. Un poco de historia

En la segunda mitad de los años noventa ha habido un avance muy considerable en la bibliografía econométrica de evaluación de políticas sociales[1-3]. El objetivo de este apartado es proporcionar una visión sobre por qué ha ocurrido este avance. Se pretende que ello ayude a comprender las técnicas de *matching* y de doble diferencia combinada con *matching* que más adelante se discutirán.

En 1986 Lalonde publica un artículo muy influyente en que se pone de manifiesto la evaluación de un programa social en los EE.UU. mediante datos experimentales[8]. Debido a la utilización de estos datos, las estimaciones deberían estar libres de la mayoría de los sesgos que se pueden dar en estudios observacionales. Por tanto, desde un punto de vista estadístico los resultados de la evaluación experimental se consideran los mejores posibles. En el mismo artículo, Lalonde realiza la evaluación del mismo programa, obteniendo el grupo de control de fuentes externas, y por tanto con datos no experimentales. Esta evaluación no experimental se lleva a cabo utilizando los estimadores más conocidos en la época: modelo de regresión, doble diferencia y modelo de selección de Heckman. Comparando los resultados de las dos evaluaciones se puede calcular la fiabilidad de los estimadores no experimentales. Si éstos son una buena alternativa, sus resultados deberían aproximarse a los experimentales, y se llegó a la conclusión de que los estimadores aplicados al conjunto de datos no experimental dieron resultados muy distintos a los experimentales. Este artículo tuvo un gran impacto en la bibliografía sobre evaluación, y llevó a muchos economistas a defender la evaluación de políticas sociales usando datos experimentales[9]. Heckman y Hotz evidenciaron que con una batería adecuada de test de especificación las estimaciones con datos observacionales podían producir resultados cercanos a los experimentales[10]. Sin embargo, el artículo de Lalonde impregnó de pesimismo a muchos científicos interesados en la evaluación de políticas sociales, lo que llevó en los años noventa a un importante avance de la bibliografía en métodos de evaluación no experimental, avance que consistió en:

– Discutir la interpretación que había que dar a los resultados de los modelos estadísticos utilizados. Es muy probable que métodos distintos den resultados diferentes, ya que simplemente estiman impactos distintos.

– Revisar los supuestos sobre los que se asentaban los modelos estadísticos. Los métodos disponibles en la época suponían que el impacto de la intervención era igual entre todos los individuos con unas características observadas idénticas. Dado que parece poco probable que este supuesto de homogeneidad se cumpla, se han considerado técnicas de estimación e interpretación de los resultados bajo la posibilidad de heterogeneidad en el impacto de la reforma, lo que significa que este impacto puede ser distinto para individuos con características observadas iguales.

– Considerar qué información de los datos estaba disponible y cuál proviene de llevar a cabo supuestos arbitrarios sobre los que el analista no tiene información. El objetivo es favorecer las técnicas de evaluación que aprovechen al máximo la información disponible en los datos y minimicen el impacto de los supuestos.

– Estudiar la evaluación bajo situaciones experimentales para poder emular, en lo posible, sus ventajas cuando se trata con datos no experimentales. Como se verá, la técnica de *matching* se centra en conseguir un grupo de control lo más parecido posible al grupo de tratamiento. Con ello, la técnica intenta conseguir lo que un experimento proporciona por sí solo.

Entender lo que medimos

Como se ha mencionado en el apartado anterior, uno de los avances que ha tenido lugar es el hecho de relajar el supuesto de homogeneidad. Este supuesto, según el cual los individuos con las mismas variables condicionantes obtienen la misma ganancia de la intervención, se ha relajado considerándose la hipótesis más realista de que la ganancia puede tomar distintos valores, incluso para individuos que tienen las mismas variables condicionantes.

Una consecuencia lógica de esta hipótesis de heterogeneidad es que los individuos que decidan participar en la reforma serán aquellos con una mayor expectativa de beneficio. Por tanto, la ganancia de los participantes será, en general, distinta de la de los no participantes. Con esto en mente, podemos definir el parámetro que normalmente es objeto de estimación: *la ganancia media de los participantes.* Este parámetro nos indica el cambio medio en la variable resultado entre los participantes; es decir, cuál es la diferencia media en la variable resultado por haber participado, en lugar de haber permanecido en el *statu quo*, en el grupo de participantes en la reforma.

Es necesario distinguir la ganancia media de los participantes del parámetro que más comúnmente se tiene en mente cuando se lleva a cabo una evaluación, pero muy pocas veces se consigue estimar: *la ganancia media de la intervención.* Este último parámetro consiste en el cambio medio de la variable resultado, que se consigue exponiendo a un individuo de la población escogido al azar a la reforma, con independencia de que sea participante o no. Por tanto, estos dos parámetros difieren en el grupo de individuos sobre los que se estima el cambio medio en la variable resultado.

En este párrafo explicaremos por qué resulta más fácil estimar la ganancia media de los participantes que la de la intervención. Para obtener una estimación de la ganancia media de los participantes es necesario estimar la media de la variable resultado de los participantes en dos escenarios distintos: al participar en la reforma y si no hubieran participado. La primera es fácil de obtener, pues es lo que observamos, los sujetos que están participando. La segunda es más complicada, pues obviamente no observamos la variable resultado para los participantes si no hubieran participado. Es decir, en el grupo de participantes observamos la variable resultado al participar en la reforma, pero no observamos cuál hubiera sido el valor de la variable resultado si no hubieran participado. Por tanto, ése será el dato que se debe estimar: el resultado medio de los participantes si no hubieran participado. Sin embargo, para estimar la ganancia media de la intervención es necesario no sólo conocer la ganancia media de los participantes, sino también *la ganancia media de los no participantes*. Con los dos datos se podría calcular una media ponderada y que proporcione la ganancia media de la población, con independencia de que sea participante o no. Siguiendo un razonamiento análogo al anterior, para estimar la ganancia media de los no participantes es necesario estimar cuánto hubiese sido la media del valor de la variable resultado para los no participantes, si hubieran participado en la intervención. Como esto no es observable, también ha de ser estimado. En resumen, para estimar la ganancia media de la intervención es necesario determinar más parámetros que los necesarios para estimar la ganancia media de los participantes. Evaluar más parámetros implica realizar más supuestos y por ello normalmente tan sólo se realiza la estimación de la ganancia media de los participantes.

La distinción entre estos dos parámetros no es una cuestión puramente técnica, sino que tiene importantes implicaciones de política económica. Si la reforma no se aplica a toda la población, sino al grupo que elija participar, tiene sentido medir la ganancia media de los participantes. Sin embargo, si la reforma se aplica con carácter general a toda la población, conocer la ganancia media de los participantes no es tan útil, ya que la ganancia media de los no participantes será menor. Por tanto, aunque la ganancia media de los participantes sea mayor que el coste marginal de la reforma, la ganancia media de los no participantes podría ser mucho menor que el coste marginal, lo que haría que la reforma no fuese beneficiosa. En este sentido, se establece una clara vinculación entre cómo se ha de llevar a cabo la evaluación y la futura implementación de la reforma.

A la luz de la discusión anterior resulta interesante considerar cuán simplificador era el supuesto de homogeneidad, que se empieza a relajar en la segunda mitad de los años noventa. Bajo la hipótesis de homogeneidad, los tres parámetros antes mencionados son iguales; es decir, que al medir la ganancia media de los participantes, también estamos midiendo la ganancia media de la intervención y la ganancia media de los no participantes. Bajo la hipótesis de homogeneidad se solucionaba la discusión anterior, a través, básicamente, de ignorar la complejidad del problema de evaluación.

Ventajas de los experimentos sociales

Los experimentos sociales presentan importantes ventajas, ya que los datos que se obtienen de ellos tienen ciertas características que facilitan la evaluación[11]. Dichas características son las responsables de que el impacto de la evaluación se estime sin sesgos relevantes. Las técnicas de *matching* que se describen posteriormente, intentan emular algunas de estas características. Por ello resulta interesante tener en mente cuáles son las ventajas de los experimentos sociales.

Siguiendo con nuestro ejemplo de incentivos a los médicos, un experimento social consistiría en, primero, determinar el conjunto de médicos elegibles, es decir, el conjunto de médicos a los que, después de la evaluación, se les piensa aplicar el esquema de incentivos. Así, se selecciona una muestra de este conjunto, y a los sujetos se les comunica la posibilidad de participar en el experimento, y se les pregunta si están de acuerdo con participar o no. Entre los que aceptan, se seleccionan un grupo de tratamiento y uno de control, que aunque ha aceptado participar se excluye del esquema de incentivos para poder llevar a cabo la evaluación según un planteamiento experimental. Esta división se efectúa de forma puramente aleatoria. Los datos para la evaluación se recogerán tanto para el grupo de tratamiento como para el de control.

Dado que el grupo de control se ha escogido de forma aleatoria dentro del grupo de elegibles que han decidido participar, se puede considerar que el grupo de tratamiento será equivalente al de control, y por tanto:

– La distribución de variables condicionantes ha de ser muy similar en los grupos de tratamiento y control. Ello se traduce en que los estadísticos descriptivos de las variables condicionantes son muy parecidos en el grupo de tratamiento y en el de control. Cuando esto ocurra la muestra estará equilibrada de acuerdo con las variables observables.

– Al igual que con las condicionantes, las variables no observables que influyan en el resultado también quedarán distribuidas de forma equivalente en el grupo de tratamiento y el de control. Cuando esto ocurra la muestra estará equilibrada de acuerdo con las variables no observables.

– Una consecuencia de lo anterior es que el rango de valores que toman las variables condicionantes en el grupo de tratamiento es muy similar al que toma en el grupo de control. Esta propiedad, que se denomina de rango común, es muy importante pues permite que la evaluación se lleve a cabo con grupos comparables de tratamiento y control.

Como se comprobará posteriormente, el método de *matching* se basa en conseguir, a través de métodos de remuestreo la primera y la tercera características previamente mencionadas. La segunda no se podrá conseguir, ya que por construcción se basa en variables que el investigador no observa, por lo que no podrá trabajar con ellas para equilibrar la muestra en ese sentido.

Algunos predecesores del método de *matching* y sus variantes

Suele ser más fácil comprender las ventajas de un método de estimación cuando se conocen las de otros métodos. Los procedimientos de diferencia de medias en sección cruzada, el de diferencia entre antes y después y el de doble diferencia se usan comúnmente en evaluaciones. En esta sección, subrayaremos las desventajas de estos métodos, lo que nos ayudará a comprender las ventajas del método de *matching* y el de doble diferencia combinado con *matching*.

En esta sección y en las restantes se asumirá que los datos a nuestra disposición provienen de un estudio observacional y no experimental. Por tanto, la separación entre grupo de tratamiento y de control no se ha llevado a cabo utilizando un procedimiento aleatorio. Nuestro objetivo, en esta sección, es analizar bajo qué supuestos los métodos que se mencionarán estiman de forma consistente el parámetro que se ha considerado clave: la ganancia media de los participantes.

Método de diferencia de medias en sección cruzada

Este método consiste en obtener la media muestral de la variable resultado después de la reforma, para el grupo de tratamiento y el de control, y obtener la diferencia de estas medias. Una versión más refinada consiste en estimar una regresión en la que se incluye una variable indicador con valor 1, si el individuo está en el grupo tratamiento, y 0, en caso contrario. El analista que utiliza este método confía en que le proporcione una estimación consistente de la ganancia media para los participantes; para que ello sea cierto, se ha de cumplir:

– Que el resultado medio para el grupo de control sea igual al que hubiera tenido el grupo de tratamiento, si este no hubiera participado en el experimento. Es común refinar el método de medias utilizando una regresión, en este caso el argumento anterior se condiciona a las variables incluidas en la regresión. Es útil poner un ejemplo de cuándo esta condición se viola. Supongamos que los médicos con mayor experiencia conocen mejor los medicamentos disponibles en el mercado y, por tanto, recetan más genéricos (esta afirmación no es necesariamente cierta, se plantea con propósito exclusivamente ilustrativo). A ellos les resultaría más fácil cumplir el esquema de incentivos, y por ello sería más probable que el grupo de tratamiento estuviese formado por médicos con más experiencia y el de control por aquellos con menos. Por tanto, la media de la proporción de medicamentos genéricos prescritos por los médicos del grupo de control no será un buen estimador de lo que los médicos del grupo de tratamiento hubieran prescrito si no hubieran participado en el esquema de incentivos. Esto se debe a que los médicos con mayor experiencia hubieran recetado más genéricos, con independencia del esquema de incentivos. El mismo problema se plantea si se hubiera utilizado una regresión, en caso de que la variable experiencia no estuviera disponible en nuestros datos. En ese caso se dice que ha habido un problema de selección por variables no observables[12].

– Que se cumpla la condición de rango común. Si los participantes son los médicos con más de 10 años de experiencia y los no participantes, los que tienen menos, entonces se está comparando lo incomparable. Es decir, el grupo de tratamiento no sería similar al grupo de control, lo que imposibilitaría realizar una evaluación, a no ser que se aceptase llevar a cabo supuestos muy fuertes que permitan extrapolar. Para realizar una evaluación con mínimos supuestos es necesario que, para cada observación en el grupo de tratamiento con una combinación de variables condicionantes determinada, exista otra observación en el grupo de control con variables condicionantes muy similares. Como veremos, el método de *matching* se asegura de que la evaluación sólo se lleve a cabo en la sección de datos donde esta condición se cumple. Sin embargo, al realizar las medias el método de diferencia de medias de sección cruzada no tiene en cuenta este factor.

– Que la muestra esté equilibrada en las variables observables. Nótese que el parámetro que se va a estimar es la ganancia media de los participantes. Para ello necesitamos una estimación del resultado de los participantes si no hubieran participado. Para obtenerla se utiliza el resultado de los no participantes. Si el grupo de no participantes es distinto en cuanto a variables condicionantes, todo indica que se necesitaría cierta ponderación de los resultados para poder llevar a cabo la estimación. Es decir, necesitamos una ponderación que permita construir un grupo de control con la misma distribución de variables condicionantes que el grupo de

participantes, pues la media que se pretende estimar es para participantes. Seguir con el ejemplo nos puede ayudar a exponerlo con más claridad. Supongamos que en el grupo de participantes hay un 90% de médicos con mucha experiencia y un 10% de médicos con poca experiencia. Supongamos, además, que en el grupo de no participantes los porcentajes se invierten, es decir, que hay un 10% de médicos con mucha experiencia y un 90% de médicos con poca experiencia. Como queremos estimar la media de lo que los participantes hubieran obtenido si no hubieran participado, hay que ponderar los resultados obtenidos por los médicos no participantes por los porcentajes de composición del grupo de participantes. Como ha quedado claro, mediante el estimador de diferencia de medias en sección cruzada esta ponderación no se realiza, por lo que sólo se obtendrán buenos resultados cuando la muestra esté equilibrada, y por tanto no sea necesario ponderar. Como se verá, con el método de *matching* sí que se lleva a cabo esta ponderación.

Método de antes y después

Este método requiere tener datos para antes y después de la reforma, pero sólo para el grupo de participantes. Se suele aplicar cuando se tienen datos de panel, aunque es posible aplicarlo cuando se dispone de un conjunto de secciones cruzadas o cortes transversales. Se obtiene hallando la diferencia entre las medias muestrales de las variables resultado antes y después de la reforma, siempre para el grupo de participantes. Por tanto, para estimar lo que los participantes hubieran obtenido después de la reforma, si la reforma no hubiera tenido lugar, se confía en los datos anteriores a la reforma, ya que en ellos ésta no existía. El supuesto clave para que este método estime la ganancia media entre los participantes es que no haya ningún otro factor que pueda afectar a la variable resultado, distinto de la reforma, entre los dos momentos de recogida de datos. Dicho de otra forma: el supuesto importante es que si no hubiese tenido lugar la reforma, la media de la variable resultado sería la misma antes y después de la reforma. Obviamente, este supuesto puede ser demasiado restrictivo. En nuestro ejemplo de incentivos a médicos, al estar los pacientes cada vez más informados sobre los medicamentos genéricos, sería plausible que la proporción de genéricos prescritos después de que tuviera lugar el esquema de incentivos fuese mayor que antes de que éste se produjera, incluso si el incentivo no hubiese tenido lugar. También quedaría invalidada la evaluación si la implementación del sistema de incentivos coincidiese con algún evento que hiciese que un mayor número de personas necesitasen medicamentos para los que no se dispone de una especialidad farmacéutica genérica.

Este estimador también está sujeto a crítica sobre la arbitrariedad en los instantes del tiempo en que se toman los datos. Por ejemplo, si se extraen cuando las personas son conscientes de que la reforma tendrá lugar, se puede estar incluyendo en la estimación un efecto de anticipación a la reforma. Sin embargo, este método presenta algunas ventajas frente al anterior, ya que, como siempre se trabaja con el mismo grupo de personas – los participantes– no existen problema de selección. Tampoco habrá problemas de rango común ni de ponderación por diferencia en la distribución de las variables condicionantes. Estos problemas surgían de tener un grupo de tratamiento y otro de control formados por personas distintas, pero en el estimador que nos ocupa sólo hay un grupo de personas. Este estimador también se puede aplicar cuando no se sigue al individuo a lo largo del tiempo, sino que se tiene un conjunto de datos de sección cruzada[12,13]. En dicho caso, se trata con un conjunto de personas que es distinto antes y después, por lo que habría que matizar las ventajas anteriormente mencionadas.

Método de doble diferencia

Se trata de una combinación de los dos métodos anteriores, y por tanto recoge ventajas e inconvenientes de ambos. Este método no es nuevo en la bibliografía de economía de la salud, y se pueden encontrar estudios que lo emplean[14,15]. En su versión más pura, son necesarios datos de participantes y no participantes en la reforma, tanto antes como después de que ésta haya tenido lugar. Su éxito se basa en que el grupo de no participantes sea lo más parecido posible al de participantes. Sin duda, esto es una tarea difícil de conseguir en un diseño no experimental, pues cuanto más parecidos sean, mayor será la probabilidad de que participen.

El estimador se obtiene aplicando la siguiente fórmula:

$$(\bar{Y}_{P,t+1} - \bar{Y}_{P,t}) - (\bar{Y}_{NP,t+1} - \bar{Y}_{NP,t})$$

donde $\bar{Y}$ se refiere a la media muestral de la variable resultado, los subíndices *P* y *NP* hacen referencia al grupo de participantes y de no participantes, respectivamente, y los subíndices *t* y *t + 1* se refieren a períodos temporales antes y tras la intervención. Entre paréntesis se presentan las diferencias de medias para cada grupo en dos momentos del tiempo. Por tanto, estas diferencias entre paréntesis eliminan el efecto de variables permanentes, incluso las no observables, que puedan ser distintas en los dos grupos. Es decir, elimina el problema de selectividad que se introducía en el método de diferencia de medias en sección cruzada. En nuestro ejemplo hipotético, elimina el posible

sesgo si los médicos participantes son aquellos con una experiencia relativamente más amplia.

Se ha discutido el efecto de la diferencia entre paréntesis, pero queda por tratar el efecto de la diferencia central, que se encarga de suprimir los sesgos introducidos por factores que hayan influido en el resultado y que coincidan en el tiempo con la implementación de la reforma, como el conocimiento que tienen los pacientes acerca de los medicamentos genéricos. Esta diferencia se encarga de eliminar la fuente principal de sesgo que criticábamos al estimador de antes y después. Esto se consigue porque tenemos un grupo de control, al que esperamos que también le afecten los factores contemporáneos a la reforma y que influyen en el resultado. Sin embargo a este grupo de no participantes no tendría que afectarle la reforma. Al comparar ambos grupos se puede identificar el efecto de la misma, aunque haya habido otros factores contemporáneos a ésta que influyan en la variable resultado.

El supuesto clave para que este método estime la ganancia media de los participantes es que la media del crecimiento entre *t* y *t + 1* de la variable resultado, para los no participantes, hubiera sido igual que para los participantes si la reforma no hubiera tenido lugar. Esto subraya la necesidad de que el grupo de no participantes sea lo más parecido posible al de participantes. En particular, es necesario que ambos grupos reaccionen de la misma forma ante acontecimientos comunes.

Como ejemplo de cuándo se violaría esta condición, supongamos que queremos evaluar el efecto de la reforma de atención primaría sobre el número de visitas de urgencia. Para ello disponemos de datos de zonas geográficas donde ha tenido lugar la reforma, y otras de donde no. Presumamos que, coincidiendo con la reforma de atención primaria, muchas personas han contraído la gripe y que las áreas no afectadas por la reforma son las de mayor renta. Podría ocurrir que, ante una gripe, personas con bajos ingresos acudan a urgencias y que otras con ingresos más elevados llamen al médico para que acuda a su domicilio, lo que sesgaría la estimación del efecto de la reforma sobre las visitas a urgencias. Esto se debe a que, si bien la renta es una variable permanente, por lo que en principio el método debería ser robusto en cuanto a la diferencia en el grupo de tratamiento y en el control, la reacción ante los acontecimientos comunes depende del valor de los factores permanentes. Es decir, no hemos contando con un grupo de control suficientemente parecido al de tratamiento para que reaccionen de forma equivalente ante eventos que ocurren al mismo tiempo que la reforma.

Conviene subrayar que, al igual que el método de diferencia de medias en sección cruzada, este estimador tampoco comprueba que se dé la situación de rango común, ni que se ponderen los resultados del grupo de control por la distribución de características del grupo de tratamiento. Al igual que el método de antes y después, también es criticable por la arbitrariedad de los momentos del tiempo en que se toman los datos. Para comprobar cómo estimar los errores estándar de las estimaciones, se sugiere que se consulte la bibliografía relevante[16].

El método de *matching*

El método de *matching* surge en la década de los setenta, pero las aplicaciones en el campo de la economía no comienzan hasta finales de los noventa[17]. Ya se han realizado aplicaciones en el marco de la evaluación de políticas activas de empleo, educación y programas de lucha contra la pobreza; sin embargo, no parece que se haya aplicado a la economía de la salud[18-20]. La reciente aparición de aplicaciones en economía no es ajena a la publicación de dos artículos en los que se relajan los supuestos para la aplicación de la técnica de *matching*, y donde se proponen nuevas variantes del método[21,22]. Estas nuevas variantes, entre las que se encuentra la combinación de *matching* con doble diferencia, hacen más plausible que se cumplan los supuestos necesarios para la aplicación de la técnica. Conviene subrayar que los métodos de *matching* no se utilizaron en el influyente artículo de Lalonde[8].

La técnica de *matching* está basada en la idea de comparar los resultados de la reforma del grupo de participantes, con los resultados obtenidos por no participantes que sean comparables a los primeros. Para su aplicación se requieren datos de después de la reforma, tanto para los participantes como para los no participantes. Esta técnica intenta replicar algunas de las características que presentan los datos que provienen de un experimento. Las ventajas fundamentales son:

– Pondera las estimaciones utilizando la distribución de variables condicionantes de la muestra tratamiento. En este sentido se dice que el método equilibra la muestra de acuerdo con las variables observables. Nótese que en un experimento la muestra está equilibrada, por lo que cuando se cuenta con datos experimentales no se requiere llevar a cabo este ejercicio.

– Impone la condición de rango común. Para estimar el impacto del programa, con este método sólo se contará con los individuos del grupo de tratamiento para los que se pueden encontrar sujetos parecidos en el grupo de control. El impacto del programa es, pues, sólo estimado para este grupo de individuos. Nótese que una de las ventajas de un experimento es que la condición de rango común siempre se cumple, lo que permite evaluar la reforma para todos los participantes elegibles, y no sólo para los que cumplan la condición de rango común, tal y como ocurre con la técnica de *matching*.

– Es una técnica flexible que impone pocos supuestos relativos a la forma funcional de las ecuaciones que determinan la variable resultado. Para ello, se dará más peso a los resultados obtenidos por los individuos del grupo de control que sean más parecidos a los del grupo de tratamiento en la estimación del impacto del programa.

De las ventajas fundamentales descritas se desprende que será necesario medir la similitud de un individuo del grupo de tratamiento y uno del grupo de control. En la bibliografía se suele usar la probabilidad de que un individuo con variables condicionantes X forme parte del grupo de tratamiento, como medida de distancia entre individuos del grupo de tratamiento y del de control[23]. Dicha probabilidad se conoce con el nombre de *propensity score*.

A continuación se describe el algoritmo para estimar la ganancia media de los participantes mediante la técnica de *matching* más sencilla: *nearest-neighbour caliper matching*[17].

– En primer lugar, utilizando un modelo de elección discreta, por ejemplo un *Probit* o un *Logit*, se estima la probabilidad de que un individuo con variables condicionantes X forme parte del grupo de tratamiento. Esta probabilidad se representará por P(X).

– En segundo lugar, para cada individuo del grupo de tratamiento, se obtiene su P(X), y se busca el individuo del grupo de control con una P(X) más cercana a la del sujeto del grupo de tratamiento. Si la diferencia en valor absoluto es menor a otro número previamente determinado, entonces se acepta la pareja, y se obtiene la diferencia de la variable resultado de los individuos de la pareja. Nótese la forma en que el método equilibra la muestra de acuerdo con las variables observables. Para los individuos del grupo de tratamiento, P(X) será generalmente elevado, ya que éste es la probabilidad de pertenecer al grupo tratamiento. Si en el grupo de control hay individuos con P(X) bajos, estos no se usarán en la estimación, y sólo se utilizarán aquellos con P(X) similares a los del grupo de tratamiento, con lo que la muestra se equilibra de acuerdo con los valores de P(X).

– En tercer lugar, se obtiene la estimación de la ganancia media de los participantes calculando la media aritmética de las diferencias para las parejas para las que se aceptó el paso anterior. Por tanto, sólo se utilizan aquellas parejas para las que se han encontrado individuos del grupo de control suficientemente parecidos. De esta manera se está imponiendo la condición de rango común. Así, la estimación del impacto no es para todo el grupo de participantes elegibles, sino para los que se ha podido encontrar una pareja suficientemente cercana. Para evaluar la importancia práctica de este problema se recomienda dibujar dos histogramas de los valores de P(X): uno para el grupo de tratamiento y otro para el grupo de control.

Existen distintas variantes del método propuesto. Una de ellas consiste en no utilizar en la estimación los valores de las variables resultado, sino en función de P(X)[19]. Con este enfoque se consigue aumentar la precisión de la estimación, ya que las variaciones que puedan ocurrir en la variable resultado se suavizan. Otra variante, conocida por el nombre de *kernel matching*, emplea todas las observaciones del grupo de control con rango común al de tratamiento, pero asignándole una ponderación decreciente en función de la lejanía que presentan respecto a la observación del grupo de tratamiento[21]. Este enfoque requiere estimar, primero, cuáles son las observaciones con un rango común[24]. En general, estas variantes contribuyen a disminuir la varianza de las estimaciones, pues utilizan más información que la del individuo del grupo de control más cercano al del grupo de tratamiento.

Todavía no se ha tratado aquí sobre bajo qué supuesto la técnica de *matching* permite estimar la ganancia media de los participantes. El supuesto básico que esta técnica necesita para realizar estimaciones consistentes es que, en media, una vez se ha tenido en cuenta el efecto de las variables condicionantes, los participantes hubieran obtenido el mismo resultado que los no participantes si la reforma no hubiera tenido lugar. Dicho de otra forma, el supuesto del *matching* es que el conjunto de variables condicionantes es suficientemente rico como para que la media condicional de la variable resultado en caso de no haber existido la reforma sea igual para los individuos que han participado y los que no lo han hecho. El supuesto fundamental es que no existe selección en variables no observables, es decir, que no hay variables no observables que influyan a la vez en la probabilidad de participación en la reforma y en el resultado. Por ello, se requiere que la selección entre participar y no participar sólo ocurra en variables observables. Por tanto, los dos supuestos que pueden justificar el uso de la técnica de *matching* es, por un lado, que el analista tenga la misma información que el sujeto en cuanto a las variables relevantes que afectan tanto al resultado como a la participación en la intervención del resultado, o por otro, que se suponga que el individuo no toma la decisión de participación en la reforma de acuerdo con las variables no observables que afectan a la variable resultado, posiblemente porque no conoce el valor de dichas variables cuando toma la decisión de participación[25].

Por tanto, de los tres supuestos que necesitábamos para que el estimador de diferencia de medias en sección cruzada proporcionara estimaciones válidas, el método de *matching* relaja los dos últimos, pero no el primero. Veremos como el método de *matching* combinado con doble diferencia soluciona, parcialmente, este problema.

Una de las cuestiones más importantes en la aplicación de la técnica de *matching* es la elección del con-

junto de variables observables, es decir, condicionantes. Como ya mencionamos en el apartado «Marco general», estas variables no necesitan ser exógenas, pero han de cumplir la condición de que no sean causadas por la reforma[1]. Esto abre la posibilidad a utilizar variables históricas que, en ocasiones, ayudan a explicar muy bien las variables resultado. Para la aplicación de la técnica, necesitamos un conjunto amplio de variables condicionantes para que el supuesto fundamental de la técnica se cumpla. Sin embargo, hay que tener en cuenta que cuantas más variables se utilicen, la probabilidad que P(X) sea distinto entre tratamiento y control es mayor, por lo que la región de soporte común disminuye. En cualquier caso, la opinión general es que para que la técnica funcione el conjunto de variables condicionantes ha de ser muy rico. De hecho, se ha comprobado, para un estudio de evaluación de políticas activas de empleo, que disminuir el número de variables condicionantes puede aumentar el sesgo de la estimación de forma importante[21].

La combinación de *matching* con doble diferencia

A no ser que el conjunto de variables condicionantes sea excepcionalmente rico, el supuesto de no selección en variables no observadas ha sido tradicionalmente considerado demasiado restrictivo. La combinación de *matching* con doble diferencia se ha propuesto para aliviar, parcialmente, este problema[20,21].

Veíamos antes que el estimador de doble diferencia eliminaba el efecto de sesgo que tenían las variables permanentes no observadas distintas en los dos grupos; decíamos que podía eliminar el efecto que los médicos participantes fueran los de mayor experiencia. Sin embargo, este estimador no tenía en cuenta la condición de rango común ni ponderaba los datos de forma adecuada de acuerdo con la distribución de las variables condicionantes en el grupo tratamiento. La ventaja de combinar el método de *matching* y el de doble diferencia es que se consigue un estimador robusto al efecto de variables permanentes no observadas y, al mismo tiempo, que equilibre la muestra de forma adecuada y que tenga en cuenta la condición de rango común. El inconveniente es que se necesitan datos para participantes y no participantes, tanto antes como después de la reforma. Si el método se aplica utilizando un conjunto de cortes transversales, y no un panel de datos, se ha de asumir que no hay cambios de composición en los grupos entre antes y después de la reforma. Este método resulta especialmente indicado cuando los datos del grupo de control se obtienen de un cuestionario distinto al de tratamiento. Por ejemplo, si los datos del grupo de tratamiento se obtienen *in situ*, pero los de control se extraen de una encuesta de salud. La diferencia en el cuestionario es una variable permanente cuyo efecto tiene en cuenta la combinación de *matching* con doble diferencia.

El supuesto necesario para que su aplicación resulte en estimaciones consistentes de la ganancia media de los participantes, es que, una vez se ha tenido en cuenta el efecto de las variables observables, la diferencia media entre el valor de la variable resultado antes y después de la intervención para el grupo de participantes hubiese sido igual que para el grupo de no participantes en caso de no haber tenido lugar la reforma. Nótese que el supuesto requerido en la técnica de *matching* era sobre la variable resultado en sí, mientras que aquí es sobre la diferencia. Al tomar esta diferencia, el estimador es robusto ante variables permanentes no observables que afecten de forma aditiva al resultado, pero no lo será ante variables transitorias ni si el grupo de control reacciona de forma distinta que el grupo de tratamiento ante eventos comunes.

El algoritmo para la estimación del método combinado de doble diferencia y *matching* es el mismo que para la técnica de *matching*, pero en lugar de utilizar la diferencia de la variable resultado entre el grupo de tratamiento y control, se utiliza la diferencia, en los grupos de tratamiento y control, de la variable resultado entre los períodos posterior y anterior a la reforma.

Evaluar los métodos de evaluación

En este apartado se intenta dar la visión de la bibliografía sobre la bondad de los métodos de evaluación descritos[21,24,26]. En este sentido, hay que afirmar que para evaluar la bondad de un método de evaluación no experimental se necesita un experimento para poder disponer de una estimación fiable del impacto de la reforma, para poder compararlo con las estimaciones que nos proporcionan los métodos de *matching* y de *matching* combinado. Esto no se ha realizado en el marco de economía de la salud, y los resultados que se mencionarán se desprenden de estudios de evaluación de políticas activas de empleo. Por tanto, no se deben extrapolar sin más al campo de economía de la salud.

Estos estudios han concluido que los resultados proporcionados por el método de *matching* se acercan a los experimentales cuando se utilizan las mismas fuentes de datos para participantes y no participantes, y cuando se tiene un conjunto de variables muy amplio para modelizar la decisión de participar[21,26]. Si el grupo de comparación no satisface estos criterios es posible que los resultados proporcionados por el método de *matching* no sean satisfactorios. Por otro lado, se ha comprobado que la combinación de *matching* con doble di-

Tabla 1. Supuestos necesarios para la validez de los métodos de evaluación no experimental

Diferencia de medias en sección cruzada
La media de la variable resultado para el grupo de control es igual a la que hubiera tenido el grupo de tratamiento si no hubiese participado en la intervención
Se cumple la condición de rango común
La muestra está equilibrada en variables observables
Diferencia entre antes y después
Si no hubiese tenido lugar la reforma, la media de la variable resultado sería la misma antes y después de la reforma
No hay efecto de anticipación a la reforma
Doble diferencia
El crecimiento de la variable resultado entre antes y después de la reforma para los no participantes es igual que para los participantes si la reforma no hubiera tenido lugar
En particular, el grupo de control reacciona a acontecimientos coincidentes con la intervención igual que el de tratamiento
Se cumple la condición de rango común
La muestra está equilibrada en variables observables
No hay efecto de anticipación a la reforma
Matching
No existen variables no observables que influyan, a su vez, en la participación en la intervención y en el resultado
La intersección de los valores del *propensity score* para los grupos de tratamiento y control es no vacía
Matching combinado con doble diferencia
No existen variables transitorias no observables que influyan a la vez en la participación en la intervención y en el resultado
La intersección de los valores del *propensity score* para los grupos de tratamiento y control es no vacía
El grupo de control reacciona a acontecimientos coincidentes con la intervención igual que el de tratamiento
No hay efecto de anticipación a la reforma

ferencia proporciona estimaciones más robustas que las del método de *matching* sin combinar; sin embargo, no elimina completamente el sesgo debido a selección por variables no observables.

En este sentido, resulta útil hacer mención al influyente artículo de Lalonde[8]. La evidencia indica que una parte importante del sesgo obtenido en su artículo era debido a diferencias en variables observables, y por tanto a su ponderación no adecuada. El sesgo introducido por variables no observables que produjeran sesgo por selección, aunque existente, era de menor cuantía[21]. Por tanto, no será extraño que este argumento sea utilizado por los artículos que utilicen la técnica de *matching* para evaluar una intervención. Sin embargo, todavía falta evidencia sobre si ello se da con carácter general o no.

Conclusiones

Este artículo ha revisado algunas técnicas cuantitativas de evaluación de estudios no experimentales, poniendo énfasis en la técnica de *matching* y de doble diferencia combinada con *matching*. Nos hemos centrado en la estimación de la ganancia media para los participantes, bajo la hipótesis de heterogeneidad no observable entre los individuos respecto a la ganancia de la reforma. Estas técnicas han cobrado protagonismo desde finales de los años noventa, y por esto hemos considerado oportuno concentrarnos en ellas. Además, no parecen haber llegado al campo de economía de la salud, aunque creemos que no tardarán en hacerlo. Uno de los objetivos del artículo ha sido proporcionar ejemplos que ayuden a la aplicación de las técnicas en economía de la salud. Hemos intentado poner énfasis en los supuestos que permiten la correcta aplicación de la técnica y en subrayar sus ventajas frente a técnicas más sencillas. Debido a la falta de espacio y a su carácter más técnico, no se ha tratado acerca de las técnicas de variables instrumentales ni de corrección de selección muestral[27].

Agradecimiento

Quiero expresar mi agradecimiento a la junta directiva de la Asociación Española de Economía de la Salud por haberme propuesto la realización de este artículo. Los comentarios detallados de Vicente Ortún y de dos evaluadores anónimos han mejorado considerablemente el artículo. Quisiera agradecer a Ingrid Vargas algunas ideas que me han servido para ejemplos, así como su ayuda con la bibliografía. También quisiera agradecer a Jaume Puig y Ricard Meneu por recomendarme algunas de las referencias bibliográficas. Cualquier error u omisión es única responsabilidad del autor. Esta investigación ha sido financiada por el programa Marie Curie de la Comunidad Europea bajo el contrato de investigación HPMF-CT-2001-01206.

*Hemos decidido traducir el término *difference in differences* por «doble diferencia». Al mismo tiempo, hemos decidido no traducir el término anglosajón *matching*.

Bibliografía

1. Heckman J, Lalonde R, Smith J, The econometrics of active labor market programs. En: Ashenfelter O, Card D, editors. Handbook of labor economics. Vol. 3. Amsterdam: North Holland, 1999; p. 1865-2097.
2. Blundell R, Costa Dias M. Evaluation methods for non-experimetal data. Fiscal Studies 2000;21:427-68.
3. Blundell R, Costas Dias M. Alternative approaches to evaluation in empirical microeconomics. Portuguese Economic Journal 2002;1:91-115.
4. Sianesi B. PSMATCH: Stata module to perform various types of propensity score *matching* [accedido el 28/10/2002]. Disponible en: http://econpapers.hhs.se/software/bocbocode/s418602.htm
5. Ichino A. Stata programs for the ATT estimation based on propensity score *matching* [accedido el 28/10/2002]. http://www.iue.it/Personal/Ichino/Welcome.html#pscore
6. Newhouse J. Free for all? Lessons from the Rand Health Insurance Experiment. Cambridge: Harvard University Press; 1993.
7. Gertler P, Boyce S. An experiment in incentive-based welfare: the impact of PROGRESA on health in Mexico, Haas School of Business, July 2001 [accedido el 28/10/2002]. http://faculty.haas.berkeley.edu/gertler/working_papers/PROGRESA%204-01.pdf
8. Lalonde R. Evaluating the econometric evaluations of training programs with experimental data. American Economic Review 1986;76:604-20.
9. Burtless G. The case for randomized field trials in economic and policy research. Journal of Economic Perspectives 1995;9:63-84.
10. Heckman J, Hotz V. Choosing among altenative nonexperimental methods for estimating the impact of social programs: the case of manpower training. J Am Stat Assoc 1989;84:862-74.
11. Hausman JA, Wise DA. Social experimentation. Chicago: University Chicago Press for National Bureau of Economic Research; 1985.
12. Heckman J, Robb R. Alternative methods for evaluating the impact of interventions. An overview. Journal of Econometrics 1985;30:239-67.
13. Heckman J, Robb R. Alternative methods for evaluating the impacts of interventions. En: Heckman H, Singer B, editors. Longitudinal analysis of labor market data. New York: Cambridge University Press for Econometric Society Monograph Series, 1985; p. 156-246.
14. Chiappori P, Durand F, Geoffard P. Moral hazard and the demand for physician services: first lessons from a French natural experiment. European Economic Review 1998;42:499-511.
15. Gray B. Do Medicaid physician fees for prenatal services affect birth outcomes? Journal of Health Economics 2001;20: 571-90.
16. Bertrand M, Duflo E, Mullainathan S. How much shoud we trust differences-in-differences estimates. Cambridge: Massachusetts Institute of Technology, Department of Economics, Working Paper Series n.º 01-34, 2001.
17. Cochrane W, Rubin D. Controlling bias in observational studies. Sankyha 1985;35:417-46.
18. Jalan J, Ravallion M. Estimating the benefit incidence of an antipoverty program by propensity score matching. Journal of Econometrics 2003;112:153-73.
19. Meghir C, Palme M. The effect of a social experiment in education. The Institute for Fiscal Studies, Working Paper n.º 01/11, 2001.
20. Blundell R, Costa Dias M, Meghir C, Van Reenen J. Evaluating the employment impact of a mandatory job search assistance program. The Institute for Fiscal Studies Working Paper n.º 01/20, 2001.
21. Heckman J, Ichimura H, Todd P. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Review of Economic Studies 1997;64:605-54.
22. Heckman J, Ichimura H, Todd P. Matching as an econometric evaluation estimator. Review of Economic Studies 1998; 65:261-94.
23. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70:41-55.
24. Smith J, Todd P. Does matching overcome Lalonde's critique of nonexperimental estimators? University of Pennsylvania Working Paper, November 2000 [accedido 28/10/2002]. Disponible en: http://athena.sas.upenn.edu/~petra/nsw.pdf
25. Heckman J, Smith J. Evaluating the welfare state. En: Strom S, editor. Frisch centenary. Cambridge: Cambridge University Press; 1997.
26. Heckman J, Ichimura H, Smith J, Todd P. Characterizing selection bias using experimental data. Econometrica 1998; 66:1017-98.
27. Heckman J. Sample selection as a specification error. Econometrica 1979;47:153-61.