

Metodologies d'avaluació dels resultats de programmes

Sebastià Sarasa Urdiola

PID_00197902



Els textos i imatges publicats en aquesta obra estan subjectes –llevat que s'indiqui el contrari– a una llicència de Reconeixement-NoComercial-SenseObraDerivada (BY-NC-ND) v.3.0 Espanya de Creative Commons. Podeu copiar-los, distribuir-los i transmetre'ls públicament sempre que en citeu l'autor i la font (FUOC. Fundació per a la Universitat Oberta de Catalunya), no en feu un ús comercial i no en feu obra derivada. La llicència completa es pot consultar a <http://creativecommons.org/licenses/by-nc-nd/3.0/es/legalcode.ca>

Índex

Introducció	5
1. Mètode experimental	7
2. Aparellament a partir de l'índex de similitud (<i>propensity score matching</i>)	10
3. Mètode de la doble diferència	13
3.1. Alternatives complementàries al mètode de la doble diferència	16
3.1.1. Aplicació conjunta de l'aparellament per índex de similitud i de la doble diferència	16
3.1.2. El mètode de la triple diferència	17
3.1.3. Regressions multivariables i doble diferència	17
4. Disseny de regressió en discontinuïtat	20
5. Mètode de la variable instrumental	23
5.1. Cas d'endogenitat deguda a una relació causal recíproca	25
5.2. Cas d'endogenitat deguda a l'omissió d'una variable, de la qual la participació en el programa és un mecanisme mediador	25
5.3. Construcció artificial <i>ex ante</i> d'una variable instrumental	26
6. Lectures complementàries	28

Introducció

La teoria de l'avaluació d'impactes ofereix diferents metodologies per a tractar el problema del contrafàctic; cada mètode parteix d'uns supòsits que l'avaluador ha d'assumir en referència al possible efecte selecció en els participants del programa i, en conseqüència, cada mètode ofereix una alternativa diferent per a resoldre el biaix de selecció. Veurem les metodologies més rellevants: el mètode experimental i altres de quasiexperimentals. Heu de tenir en consideració que tots aquests mètodes estan incorporats en els programes estadístics més habituals en el mercat, com ara SPSS, STATA o R, i que en aquest text es tracta només d'oferir una introducció a la seva lògica. Una comprensió més aprofundida requereix uns coneixements previs d'anàlisi estadística multivariable.

La diferència més rellevant entre el mètode experimental i el quasiexperimental és el procediment de selecció dels individus que s'han d'estudiar. En el mètode experimental hi ha una selecció aleatòria d'individus que són assignats, també aleatòriament, a un **grup experimental** que rebrà tractament i a un **grup de control** que no en rebrà. Un cop el programa ha estat desenvolupat, l'avaluador compara els resultats del grup experimental amb els del grup de control.

En els mètodes quasiexperimentals, l'avaluador no té cap oportunitat de seleccionar qui participa i qui no participa en el programa, i ha de limitar la seva tasca a observar els individus que participen en el programa i els que no hi participen, que en aquest cas no anomenarem *grup de control*, sinó **grup de comparació**. Mitjançant l'observació del grup que participa i el grup de comparació, l'avaluador especifica models estadístics que estimen l'efecte de la variable *participació* en el programa sobre el resultat hipotètic que se n'espera.

1. Mètode experimental

El mètode experimental evita l'efecte selecció perquè, fruit de l'assignació aleatòria als grups experimental i de control, els individus tenen les mateixes característiques abans d'iniciar el programa (Hinkleman i Kempthorne, 2008). D'aquesta manera, l'assignació al grup experimental no depèn de la voluntat dels individus, ni del criteri de cap administrador del programa, ja que tots els individus estan igualment disposats a participar-hi, i és l'atzar qui decideix a quin dels dos grups aniran.

El mètode experimental el podem il·lustrar de la manera següent. Suposem que Y és la variable que volem modificar amb el programa, tant si aquesta és una variable com una conducta, una actitud o un recurs dels individus; i direm el següent:

Y_{ea} és el valor mitjà de la variable abans del programa entre els que participaran en el grup experimental.

Y_{ca} és el valor mitjà de la variable abans del programa entre els que no hi participaran i formaran part del grup de control.

Si la selecció aleatòria s'ha fet sense errors s'ha de complir el següent:

$$Y_{ea} - Y_{ca} = 0$$

ja que tots dos grups, l'experimental i el de control, han de tenir un comportament similar abans de la intervenció. No obstant això, cal advertir sobre la possibilitat que hi hagi un error aleatori en la selecció¹ i que els individus del grup experimental i de control no siguin homogenis, de manera que $Y_{ea} - Y_{ca} \neq 0$. En aquest cas, és recomanable fer de nou l'assignació aleatòria als dos grups.

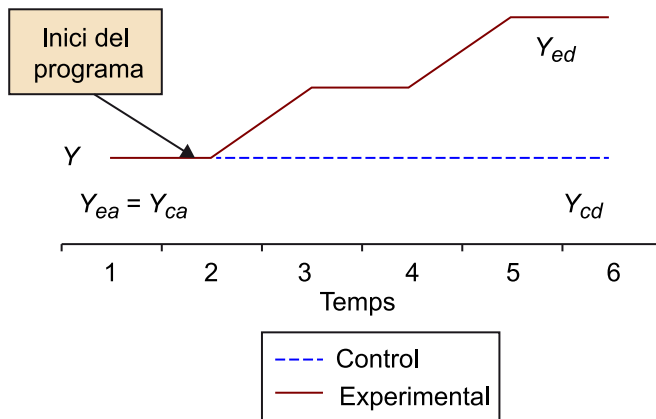
⁽¹⁾Vegeu la diferència entre *biaix* i *error aleatori*.

Si el programa té cap efecte sobre els participants, aquests valors mitjans seran diferents un cop el programa s'hagi posat en marxa, com a mínim, per al grup experimental, prenent valors Y_{ed} per al grup experimental i Y_{cd} per al grup de control.

La mesura de l'impacte del programa estarà determinada per la diferència entre els valors mitjans de Y en el grup experimental i en el grup de control.

$$Y_{ed} - Y_{cd} = \text{Impacte}$$

Figura 1. Evolució de la variable de resultats Y en els grups experimental i de control



La figura mostra el valor d'una variable Y que un programa vol modificar. El programa s'implementa a partir del moment $t = 2$ només en el grup experimental. Abans, el mesurament de la variable Y donava resultats similars per al grup experimental (e) i el de control (c) ($Y_{ea} = Y_{ca}$). A partir del moment en què el programa funciona, observem que el valor de Y per al grup experimental (Y_{ed}) és superior al valor del grup de control (Y_{cd}).

L'aplicació del mètode experimental s'ha de fer prenent algunes precaucions. Per exemple, tenint en compte els problemes derivats del **desgast de la mostra**, de la **manca de conformitat dels seleccionats** i de l'**efecte contaminació**. A més, en treure conclusions de les avaluacions fetes amb el mètode experimental hem de ser cauts pel que fa a la generalització dels resultats, atesa la major dificultat d'establir la **validesa externa** a partir d'una experiència amb un nombre reduït d'individus.

L'execució de la metodologia experimental presenta altres inconvenients que no sempre la fan aconsellable (Rossi, Lipsey i Freeman, 2004). En primer lloc, sota determinades circumstàncies el seu cost econòmic és elevat, especialment quan es vol fer una avaluació d'impacte de diferents variants o components del programa, la qual cosa exigeix la selecció aleatòria de més de dos grups; o, quan es pretén garantir la validesa externa de l'avaluació i necessita mostres molt grans representatives de tots els estrats socials. En segon lloc, la construcció del grup de control significa negar l'atenció a uns individus en situació de necessitat, quan ells volen ser atesos, i això, de vegades, xoca amb les objeccions ètiques dels responsables del programa, que consideren inadmissible negar l'atenció a una persona que la necessita i la demana. Cal advertir, però, que aquesta objecció és discutible quan no tenim cap garantia que el programa hagi de ser beneficiós. Els exemples de programes que un cop avaluats no han demostrat ser benèfics, sinó al contrari, que han causat perjudicis als beneficiaris, són molt més que anecdòtics, i de la mateixa manera que un fàrmac no es distribueix a tota la població fins que n'ha demostrat l'eficàcia i la manca d'efectes secundaris perillosos, un programa d'intervenció amb persones no hauria de ser generalitzat sense haver fet avaluacions prèvies de programes pilot.

Vegeu també

Podeu ampliar informació sobre el desgast de la mostra en l'apartat "Causes de l'efecte selecció".

Vegeu també

Tractarem dels conceptes de *manca de conformitat dels seleccionats* i de *efecte contaminació* en l'apartat "Com es dissenya l'avaluació d'un programa" del mòdul "El disseny de l'avaluació i la presentació dels resultats. Algunes notes introductòries a l'anàlisi del cost-benefici".

Vegeu també

Per aprofundir en el concepte de *validesa externa* vegeu l'apartat "Validesa de l'avaluació", del mòdul "Els reptes principals d'una avaluació d'impacte".

Alternatives a l'objecció de negar la prestació al grup de control

Hi ha, però, alternatives quan les dificultats de la metodologia experimental són no-més ètiques. La majoria tenen relació amb retards en l'assoliment de la cobertura òptima que es poden aprofitar per a construir grups de control.

Un cas és el de les llistes d'espera: quan la demanda per a participar en el programa és superior a les places ofertes, el grup que està en espera de rebre l'atenció es pot utilitzar com a grup de control.

Una altra possibilitat l'ofereixen els programes que tenen una aplicació gradual al territori: si totes les zones geogràfiques on es pensa posar en marxa són similars, les zones on encara no ha estat desenvolupat el programa es poden utilitzar com a grups de control.

En aquests casos és primordial tenir la certesa que les condicions d'accés són les mateixes per a tothom i que no hi ha diferències significatives entre els individus que reben primer l'atenció del programa i els que la reben més tard. La mínima sospita que no sigui així fa recomanable no fer ús del mètode experimental.

2. Aparellament a partir de l'índex de similitud (*propensity score matching*)

Quan no és possible seleccionar i separar aleatòriament els individus, el mètode de l'aparellament per índex de similitud ens permet construir un contrafàctic basat en la comparació d'individus que han rebut el tractament del programa amb els que no l'han rebut, però restringint la comparació a parelles d'individus que guarden la màxima similitud en les característiques de les quals disposem informació (Rosenbaum i Rubin, 1983). La idea consisteix a aparellar (*matching*) els individus que tenen la mateixa probabilitat (*propensity score*) de formar part del grup de tractament, ateses les seves característiques de les quals tenim informació, i comparar els resultats dels membres de les parelles que han participat en el programa amb els membres de les parelles que no hi han participat.

El procediment consisteix a fer una regressió logística de la variable *participa en el programa* especificant un model explicatiu en què intervenen les variables que podem observar dels individus i que considerem teòricament versemblant que puguin influir en la decisió de participar-hi. Aquest model explicatiu tindria l'especificació següent:

$$P = \beta_1 X + \varepsilon$$

En què:

P = predicció estimada de participar en el programa

X = conjunt de variables dels individus que expliquen la major o menor propensió a participar en el programa.

ε = error de mesura que recull l'efecte d'altres variables no especificades en el model.

A partir de la informació disponible sobre com covarien aquestes variables explicatives de la participació amb el fet de participar, la regressió ens donarà per cada individu i una probabilitat estimada P_i de participar-hi. Aquesta probabilitat serveix d'índex per a comparar com de similars són els individus en la propensió a participar en el programa. Amb aquest índex el programa estadístic selecciona parelles d'individus amb probabilitats similars, però amb la condició que un participi realment en el programa i l'altre no.

Bibliografia

Podeu trobar una introducció a la tècnica de regressió logística a:

A. J. Jovell (1995). *Análisis de regresión logística*. Madrid: CIS ("Cuadernos Metodológicos", 15).

Hi ha diverses tècniques d'aparellament d'individus disponibles en els programes estàndard de tractament estadístic de dades, cadascuna amb els seus avantatges i inconvenients que l'avaluador haurà de sospesar i que no veurem en aquest text in-

troductor, però sapigueu els lectors que, en general, el dilema més important per a l'avaluador és triar el marge de l'interval de similitud dins el qual considerarà que dos individus tenen la mateixa propensió a participar en el programa.

Vegem-ne un exemple. Suposem que tenim una base de dades tipus panel en què podem veure si uns individus han participat, o no, en programes voluntaris de formació professional no reglada, i tenim també els seus ingressos salarials actuals. Volem avaluar si la formació no reglada té cap impacte en els rendiments que han obtingut els individus. Com que no és possible una avaluació experimental, perquè ja no podem assignar de manera aleatòria qui va als programes, optem pel mètode de l'aparellament segons l'índex de similitud a la participació, amb el benentès, que estem convençuts que, amb la informació que tenim sobre els individus en la nostra base de dades, podem identificar **tots els factors rellevants** que, segons la teoria disponible, expliquen la propensió a participar en cursos d'aquesta mena.

Amb un model de regressió estimem la probabilitat P_i que té cada individu de fer aquests cursos de formació, a partir de l'edat, sexe, situació familiar, estudis previs i situació laboral. Llavors, l'avaluador ha de definir quin serà l'interval de similitud. És a dir, si tenim un individu A que ha fet un curs de formació professional no reglada, al qual el model ha estimat una probabilitat de fer el curs igual a 0,758, i hem de cercar l'individu més pròxim a ell, en termes de probabilitat de fer aquests cursos, però que no els ha fet, rebutjarem o acceptarem com a similar un individu B amb una probabilitat estimada de 0,789? Dependrà de com d'ample hàgim definit l'interval de similitud, si hem dit que és entre $\pm 0,050$, el valor de similitud serà entre 0,708 i 0,808, i considerarem que els dos individus són similars. Però si hem estat més restrictius i hem definit l'interval de similitud entre $\pm 0,005$, el valor de similitud seria entre 0,753 i 0,763, de manera que no considerarem l'individu B similar al A .

El dilema per a l'avaluador és que, si tria un interval molt restrictiu, corre el risc de quedar-se amb pocs casos per comparar i, quan hi ha pocs casos, és difícil obtenir diferències estadísticament significatives en els resultats del programa. En sentit contrari, si relaxa l'interval de similitud, augmentarà el nombre de casos comparables, però a costa de comparar individus que potser no són homogenis, amb la qual cosa la validesa interna de l'avaluació seria qüestionable.

D'aquesta manera es construeix una mostra d'individus amb propensions similars a participar-hi, i sobre aquesta es compara el valor que té la variable **resultats del programa** entre els que hi participen i el grup de comparació. La mesura de l'impacte del programa serà la diferència entre la mitjana dels resultats del grup de tractament i la mitjana dels resultats del grup de comparació, condicionada al fet que els individus d'un i altre grup tinguin una probabilitat estimada de participar-hi similar.

El mètode de l'aparellament per índex de similitud no és aconsellable si el nombre d'individus que satisfan la condició de similitud en l'índex de propensió és petit, o si és raonable pensar que hi ha un factor que incideix en la probabilitat de participar en el programa i que no tenim observat entre les variables que hem mesurat.

Els individus que no ha estat possible aparellar es descarten de l'avaluació. Per això és prudent fer una anàlisi dels individus que són descartats perquè no se'n troben els equivalents en el grup de participants, perquè potser ens donarà alguna pista sobre possibles factors no observats que afecten la participació en el programa.

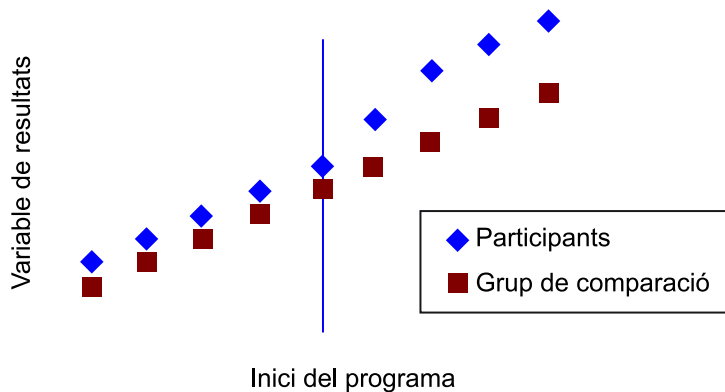
El mètode de l'aparellament per índex de similitud assumeix que l'efecte selecció, si es dona, ha de ser degut a alguna de les variables que hem mesurat i no considera la possibilitat que hi hagi altres factors, no mesurats, que puguin produir un biaix de selecció. D'altra banda, perquè pugui donar resultats estadísticament significatius, requereix que hi hagi un nombre acceptable d'individus participants i no participants que tinguin propensions similars. Per això és convenient aplegar un nombre gran d'individus no participants perquè, un cop descartats els que no són similars als participants, quedi una mostra prou gran per a oferir estimacions significatives.

3. Mètode de la doble diferència

Hem vist que el mètode de l'aparellament per l'índex de similitud assumeix que l'efecte selecció és degut només a factors que figuren entre les variables observades en el nostre model, la qual cosa és una restricció molt forta que limita la utilització d'aquest mètode. Quan aquesta assumptió és poc raonable, el mètode de la doble diferència pot ser una alternativa. Aquest mètode assumeix que hi ha un efecte selecció potencial que no podem mesurar i, per corregir-ho, compara el grup de participants amb el grup de comparació abans i després d'iniciar la intervenció del programa. D'aquesta manera, amb la comparació abans de la intervenció, podem detectar si hi ha diferències en la variable dependent entre tots dos grups amb independència del programa, i si n'hi ha, veurem que ha estat alterada un cop feta la intervenció del programa.

En la figura següent estan representats els resultats hipotètics dels participants en un programa i del grup de comparació abans i després de la intervenció. Abans del programa podem observar que els que hi participarien ja tenien uns resultats millors que el grup de comparació, i podem dir que hi ha un efecte selecció, però la diferència entre els dos grups augmenta un cop iniciat el programa, amb la qual cosa podem sostenir que el programa ha tingut efectes positius sobre els participants.

Figura 2. Exemple de resultats d'avaluació de doble diferència



A diferència d'allò que hem vist en la figura 1 fruit de l'assignació aleatòria dels individus a un grup experimental i a un altre de control, aquí veiem que el grup de participants en el programa ja mostra uns resultats una mica superiors al grup de control abans de començar la implementació del programa. Un cop el programa està en marxa, veiem que la diferència entre els dos grups s'eixampla encara més.

El mètode de la doble diferència rep el nom del fet que estima l'impacte de la intervenció mitjançant la mesura de dues diferències: una diferència en el temps, i una diferència entre grups. La figura 2 ens pot il·lustrar millor aquest mètode si la traduïm a números, com es mostra en la taula 1. Suposem que la figura representa els valors de la variable *hores d'estudi a la setmana*, ja que estem avaluant un programa que pretén augmentar el rendiment escolar incentivant l'estudi fora de l'horari escolar. La primera diferència és la que hi ha entre el grup participant i el de comparació abans de començar el programa,

que és d'una hora setmanal. La segona diferència és la que tenen els grups en el moment de l'avaluació, que ara és de tres hores a la setmana. La diferència entre les dues diferències és de dues hores. Per tant, podem concloure que l'impacte del programa ha estat positiu perquè ha augmentat una mitjana de dues hores setmanals les hores d'estudi dels participants, davant les hores que haurien estudiat en cas que el programa no hagués existit.

Taula 1

Valor mitjà de la variable que s'ha de modificar amb el programa (hores d'estudi setmanals)			
	Moment t abans d'iniciar el programa	Moment $t + n$ després d'iniciat el programa	Diferència entre moments temporals
Grup participant	10	15	+5
Grup de comparació	9	12	+3
Diferència entre grups	+1	+3	Impacte +2

Aquest mètode no està limitat a mesurar només aquestes dues diferències. Si el cas ho demana, podem augmentar les mesures de diferències en el temps i entre grups de manera que podem parlar de triples diferències o més. Un exemple d'avaluació amb una mesura de tres diferències seria el cas d'avaluar un programa de prevenció de salut per a persones més grans de setanta anys en una comunitat autònoma comparant la salut dels individus abans i després del programa (dues mesures en el temps) però de tres grups d'individus: dos grups formats pels més grans de setanta anys i els més joves d'aquesta edat de la comunitat autònoma on el programa preventiu és aplicat, la qual cosa ens permet valorar si el programa ha tingut cap impacte significatiu en el diferencial de salut que hi ha entre més grans i més joves de setanta anys, i aquest mesurament el complementaríem amb un tercer un grup de més grans de setanta anys d'una altra comunitat autònoma on no existís cap programa similar, la qual cosa ens permetria valorar l'impacte del programa entre el grup de la mateixa edat.

Cal tenir en consideració, però, que el mètode de la doble diferència assumeix un seguit de supòsits que no sempre són assumibles (Angrist i Pischke, 2008):

1) **L'efecte selecció és constant en el temps**, és a dir, que la diferència inicial entre els dos grups seria la mateixa al cap del temps si el programa no hagués existit. Per tant, qualsevol variació entre els moments previ i posterior al programa és deguda exclusivament a la intervenció del programa en el grup experimental, no al fet que l'efecte selecció hagi pogut variar amb el pas del temps amb independència del programa. Aquest supòsit no sempre és raonable. De vegades, la diferència entre el grup participant i el de control pot seguir una tendència temporal cap a la convergència que podem atribuir falsament al programa. Per exemple, imaginem un programa destinat a augmentar la participació laboral de mares amb poca qualificació laboral que pren com a grup

de comparació el conjunt de dones, i conclou que el programa ha estat un èxit perquè s'ha reduït la diferència entre la participació laboral mitjana de les dones poc qualificades i la mitjana del conjunt de dones. Però els avaluadors no han tingut en consideració que aquesta diferència ja s'estava reduint els anys anteriors i que, per tant, el programa potser no ha tingut cap efecte, llevat que veiem que el programa ha accelerat la reducció de la diferència. Casos com aquest fan prudent recollir informació prèvia sobre com estava evolucionant la variable dependent entre els dos grups abans del programa.

2) La segona assumptió és que **no hi ha interacció significativa entre l'efecte selecció i el resultat del programa**. És a dir, es parteix del supòsit que el programa té els mateixos efectes sobre el grup de comparació que sobre el grup participant, la qual cosa permet utilitzar el grup de comparació com a contrafàctic. Però no sempre és raonable pensar així. De vegades els programes van adreçats a poblacions específiques que tenen característiques no equiparables a les del conjunt de la població, i la reacció d'aquests col·lectius al programa pot ser ben diferent de la que tindria un ciutadà mitjà, de manera que l'impacte obtingut amb els participants no és extensible a l'impacte que tindria el programa amb altres persones. Pensem en l'exemple dels programes que ha finançat la Comissió Europea per a renovar barris. Els governs han triat barris en dificultats com a objecte de la intervenció, amb una composició poblacional amb més dificultats d'integració laboral que la resta de barris. Comparar simplement la mitjana de les diferències en l'ocupació laboral entre barris participants i no participants és no tenir en consideració que, malgrat el programa, les dificultats d'ocupació als barris en què s'ha intervingut són superiors. És a dir, que el mateix programa aplicat en un barri amb una altra composició social tindria uns resultats diferents i, per tant, els resultats de l'avaluació no són extensibles al conjunt de la població. En aquests casos, és recomanable fer una combinació dels mètodes de la doble diferència i de l'aparellament per índex de similitud (vegeu més avall).

3) Una variant de la interacció entre l'efecte selecció i el programa seria l'assumptió del fet que, **si hi ha desgast de la mostra (*attrition*), aquest no té cap efecte en el resultat mesurat del programa**. Això no és sempre raonable i ens obliga a controlar les característiques dels que abandonen el programa i si hi ha cap covariació entre elles i els resultat del programa. Aquesta és una situació típica en programes en què la intervenció es fa amb la intenció que els participants assoleixin un llinar que els faci autònoms respecte a alguna dependència que pateixen, bé sigui un programa de deshabitació de toxicòmans, de formació a persones a l'atur fins que obtinguin una ocupació, o bé un altra situació similar en la qual la velocitat a la qual els individus assoleixen el llinar no és homogènia, de manera que uns abandonaran el programa amb èxit en un termini de temps curt, i altres es cronificaran per llargs períodes de temps. En aquesta mena de programes, l'avaluació mitjançant el mètode de la triple diferència pot ser una bona opció, com veurem en la secció següent.

4) El quart supòsit és que **no hi ha cap altre factor extern al mateix programa que hagi influït en el resultat**. Un supòsit molt restrictiu quan no tenim el control que s'acostuma a tenir en un laboratori.

3.1. Alternatives complementàries al mètode de la doble diferència

Hem vist que els quatre supòsits assumits pel mètode de la doble diferència en la versió més simple no sempre són raonables i l'avaluador ha de sospesar si té manera de relaxar-los accedint a informació complementària. Una manera prudent d'obtenir informació complementària és en el moment de fer l'**estudi base**. El mètode de la doble diferència fa una comparació entre grups d'individus en dos moments temporals. En el moment inicial, abans de posar el programa en marxa, es fa l'estudi base.

Aquest estudi base ha de servir a l'avaluador per a analitzar si hi ha diferències entre el grup que formarà part del programa i el grup de comparació, en especial pel que fa a la variable dependent que volem modificar amb el programa. Però l'estudi base també ens ha de servir per a observar tendències anteriors, i les covariacions de la variable dependent amb altres variables tant contextuais com individuals, informació aquesta que ens serà de gran valor per a controlar altres factors, externs al programa, que puguin haver alterat el comportament de la variable dependent. Si aquesta informació complementària és adequada, alguns dels problemes anteriors es poden resoldre de manera satisfactòria. A continuació veurem tres maneres de resoldre alguns dels problemes que han estat esmentats.

3.1.1. Aplicació conjunta de l'aparellament per índex de similitud i de la doble diferència

Quan hi ha una **interacció entre l'efecte selecció i l'eficàcia del programa**, un ús combinat dels mètodes de la doble diferència i de l'aparellament per índex de similitud pot ser una bona opció (Khandker i altres, 2010). De manera succinta, la combinació es pot explicar de la manera següent: el mètode de l'aparellament és aplicat per a seleccionar els individus dels grups participants i de comparació que tenen la màxima similitud; després, només amb aquesta selecció d'individus semblants, es calcula l'impacte aplicant el mètode de la doble diferència i deixa fora de l'estimació els individus diferents. Recordem, però, que el mètode de l'aparellament exigeix tenir un nombre elevat d'individus en el grup de comparació perquè, un cop descartats els individus diferents dels participants, quedi una mida prou gran per a obtenir estimacions estadísticament significatives. Si aquest no és el cas, i el nombre d'individus no és prou gran, és més recomanable comprovar la robustesa de les estimacions fetes amb el mètode de la doble diferència utilitzant la regressió multivariable, especificant en el model les variables potencialment relacionades amb la selecció dels individus en el grup participant en el programa.

Vegeu també

Definirem l'estudi base en el mòdul "Avaluació *ex ante*: la importància d'un estudi base".

Vegeu també

Vegeu el segon supòsit en l'apartat "Mètode de la doble diferència".

3.1.2. El mètode de la triple diferència

La informació sobre factors que puguin afectar la selecció dels individus en el programa i sobre factors externs al programa que en puguin afectar els resultats és accessible si es fa un bon estudi base abans de desenvolupar el programa. Però, de vegades, és impossible fer aquest estudi base. Situacions d'emergència social, com ara un daltabaix econòmic o una catàstrofe natural, obliguen a prendre mesures urgents per a cobrir les necessitats de la població afectada, i no és possible esperar a fer cap estudi base previ.

En aquesta mena de programes, la velocitat a la qual els individus es fan de nou autònoms i deixen de dependre del programa varia, i ens trobem davant el biaix potencial derivat d'un desgast selectiu (*attrition*) de la mostra amb la qual avaluem l'eficàcia del programa. No és difícil assumir que els individus que aviat marxen del programa puguin tenir algunes característiques especials que els fan diferents dels que es cronifiquen, fins i tot, potser, que els individus haurien estat igualment autònoms en el mateix interval de temps encara que el programa no hagués existit mai.

El mètode de la triple diferència pot ser una bona opció en aquests casos. Consisteix a combinar també els mètodes de l'aparellament sobre la base de l'índex de similitud i de la doble diferència, però treballant amb tres grups d'individus: el grup de participants en el programa que ja han marxat, el grup de participants que romanen en el programa en el moment de fer l'avaluació, i un grup de comparació extern al programa. La idea és seleccionar, mitjançant l'aparellament, individus del grup de comparació que siguin similars als individus participants que ja han marxat del programa, i a la vegada individus similars als que romanen en el programa. D'aquesta manera la doble diferència es calcula comparant, d'una banda, els resultats del grup d'individus que ha marxat amb els individus del grup de comparació de característiques similars i, d'altra banda, comparant els individus que romanen en el programa amb els individus similars a ells que no han participat en el programa.

3.1.3. Regressions multivariables i doble diferència

El mètode de la doble diferència, en la versió més simplificada que hem vist en la taula 1, es pot representar com un model de regressió que tindria l'especificació² següent :

$$Y = c + P + T + T*P$$

En què:

Y és el valor que pren la variable dependent abans i després del programa.

⁽²⁾Obviem per simplicitat en la fórmula el terme d'error i els coeficients de regressió.

C és el valor mitjà obtingut pel grup de comparació en el moment anterior a la intervenció del programa.

P és també una variable dicotòmica que pren valor 0 si els individus són del grup de comparació, i valor 1 si són del grup participant en el programa. Per tant, el coeficient estimat per a la regressió de Y sobre aquesta variable ens indicarà la diferència que tenia el grup participant respecte al de comparació abans d'iniciar el programa.

T és la variable *temps*, una variable dicotòmica que pren valor 0 en el moment anterior a la intervenció del programa i valor 1 en el moment de l'avaluació. Per tant, el valor del coeficient estimat per a la regressió equival a la variació de Y produïda en el grup de comparació en el moment de fer l'avaluació.

$T*P$ és una altra variable dicotòmica fruit de la interacció entre T i P , de manera que pren valor 1 quan la mesura de la variable Y es fa entre els individus del grup participant en el moment de fer l'avaluació, i 0 en qualsevol altra situació. Aquesta variable és la que mesura l'impacte del programa, ja que ens dóna la variació de Y per al grup participant amb relació a l'experimentada per al grup de comparació.

Ara bé, el resultat d'aquesta regressió, tal com està indicada, ens donaria exactament el mateix resultat que la diferència entre mitjanes de la taula 1, però per a resoldre el problema dels efectes deguts a factors externs al programa hem d'incorporar indicadors d'aquests suposats factors en l'equació del model, de manera que la fórmula anterior hauria d'incorporar un vector X de variables, i de cadascuna obtindríem els coeficients de regressió i la seva significació estadística. La fórmula quedaria de la manera següent:

$$Y = c + P + T + T*P + X + \varepsilon$$

Un exemple hipotètic ens ajudarà a entendre millor aquesta metodologia. Suposem que els responsables d'una entitat que pretén protegir les dones maltractades estan interessats a fer una prova pilot de comunicació per a validar si una campanya mediàtica en els mitjans de comunicació locals i orientada a sensibilitzar la població té com a resultat un augment de la intolerància envers el maltractament. Abans d'iniciar la campanya contracten els serveis d'uns consultors per avaluar-la, i aquests, aprofitant que el programa encara no ha començat, opten per una metodologia de la doble diferència. Decideixen que la campanya pilot es faci només en un municipi, que serà el grup de participació, i els resultats s'avaluaran utilitzant la població d'un altre municipi com a grup de comparació.

Per fer l'estudi base, fan una enquesta a una mostra representativa de la població de tots dos municipis. En el qüestionari es fan preguntes que recullen informació sobre variables que una metanàlisi prèvia havia demostrat que tenen una relació significativa amb la tolerància envers la violència masclista; variables individuals com ara el *sexe*, la *classe social*, les *credencials educatives*, la *ideologia política*, les *creences religioses*, els *antecedents familiars* i l'*ètnia*, entre d'altres. Amb aquestes dades, els avaluadors estimen una regressió en què la variable dependent és un índex de rebuig al maltractament de les dones, contrasten si hi ha diferències estadísticament significatives entre els municipis i conclouen que, un cop controlades totes les variables del model, no hi ha diferències significatives.

Al cap d'un any de campanya, els consultors fan una primera avaluació i repeteixen l'enquesta a una mostra aleatòria i representativa de la població de tots dos municipis. Sorprenentment, en una primera descripció es veu que el valor mitjà de l'índex de tolerància al maltractament no mostra diferències estadísticament significatives entre els municipis participants en la campanya i els de control, de la qual cosa es podria inferir que la campanya ha estat un fracàs.

Els avaluadors fan una anàlisi de les dades que s'han obtingut de tots dos municipis i comproven que l'estructura social del municipi on s'ha aplicat la campanya mediàtica ha canviat. Un creixement vertiginós del sector de la construcció a la comarca ha comportat un augment del nombre de treballadors no qualificats nacionals i immigrants procedents de grups socials més tolerants amb la violència masclista. Per sort, l'estudi base havia informat que les covariacions entre aquestes variables i la variable dependent eren estadísticament significatives, de manera que els avaluadors tornen a repetir la regressió amb les dades actuals obtingudes després de la campanya. El resultat mostra que, un cop controlada la nova composició social de cada municipi, la campanya mediàtica ha ocasionat un augment significatiu en el valor mitjà de l'índex de rebuig a la violència masclista al municipi on s'ha provat.

4. Dissenys de regressió en discontinuïtat

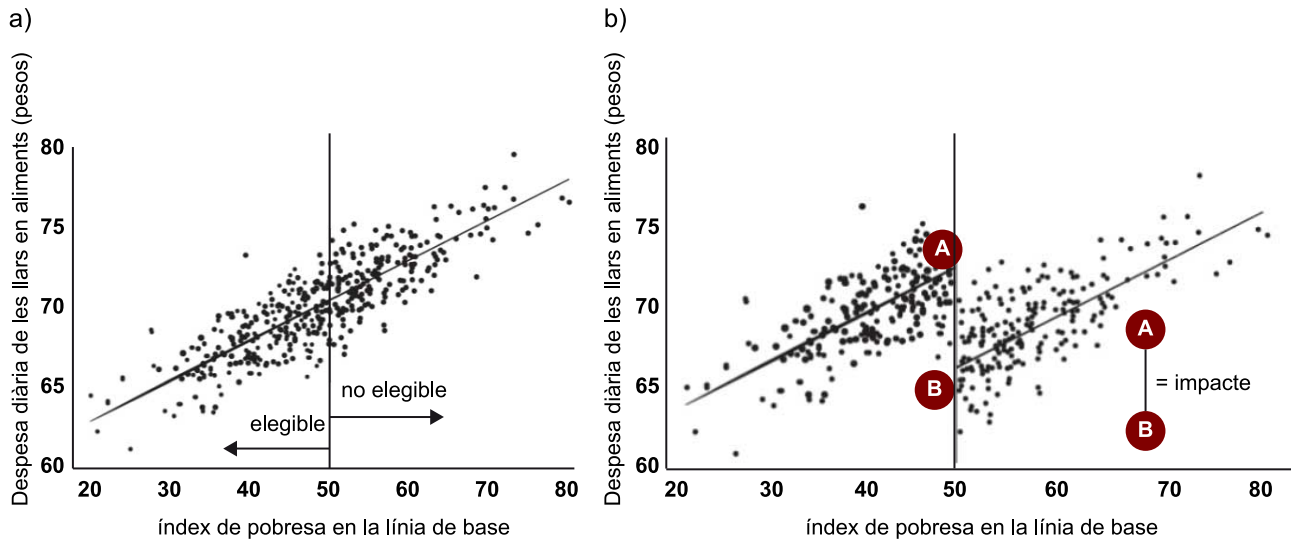
Són molts els programes en els quals l'accés al tractament està condicionat al compliment d'un llindar marcat per un valor fix. De vegades és l'edat dels individus (per exemple, l'accés a les pensions de jubilació), o l'assoliment d'un nivell de renda (prestacions socials condicionades a tenir ingressos per sota d'un llindar de pobresa) o d'un rendiment (beques d'estudis condicionades a haver obtingut una nota mitjana). En aquests casos, el mètode de la regressió en discontinuïtat pot ser una bona opció per a estimar els efectes d'un programa tot evitant el biaix de l'efecte selecció.

Per exemple, pensem en l'avaluació d'un programa de beques en què l'efecte esperat és que els beneficiaris tinguin un rendiment escolar superior i continuïtat en els estudis, i que fixa el criteri d'accés en haver superat una nota de tall concreta en els estudis anteriors en el moment de sol·licitar la beca. Una comparació simple del rendiment escolar dels becats amb el rendiment escolar dels no becats seria criticable en tant que cal suposar que els individus que ja havien superat una nota de tall determinada han demostrat tenir més motivació i capacitat que els altres, raó per la qual no ha d'estranyar que continuïn rendint més. El mètode de la regressió discontinua, però, permet obviar aquest efecte selecció perquè compara els individus que estan just per sobre i just per sota de la nota de tall. És a dir, si la nota de tall fos un 7, la comparació es podria fer entre els estudiants que haguessin obtingut una nota mitjana de 6,9 (sense dret a la beca) i els estudiants que haguessin obtingut un 7 (amb dret a beca), amb el benentès que s'assumeix que les diferències de motivació i capacitats entre els que han obtingut un 6,9 o un 7 han de ser insignificants.

Un altre cas semblant es donaria si el criteri d'accés a la beca fos el nivell de renda familiar. El llindar definit per a identificar les unitats que formaran part del programa divideix la població en dos grups a banda i banda del llindar però que podem ordenar seguint un continu creixent de la variable que serveix per a fixar el llindar, com podem veure en la figura següent, en què les famílies amb menors d'edat han estat ordenades de renda disponible més baixa a més alta i es mostra el llindar del segon decil que serveix per a identificar els nens que tindran dret a l'ajut si els seus ingressos familiars estan en el primer o en el segon decil de la distribució. L'assumpció del mètode de la discontinuïtat és que les diferències entre els individus que estan just per sota i just per sobre del llindar són insignificants i els fa quasi equivalents, de manera que podem rebutjar el risc que hi hagi cap efecte selecció i estimar els efectes del programa comparant els resultats entre els participants que estan tocant el llindar i els no participants que formen part del grup de control que també estan tocant el llindar per la banda oposada. En el cas de l'exemple que ens ocupa, si el llindar que separa el segon decil del tercer fos una renda disponible individual

mensual de 600 euros, podríem comparar els individus amb renda disponible entre 500 euros i 600 euros que reben l'assistència amb els individus amb renda disponible entre 600 euros i 700 euros i que no reben l'assistència.

Figura 3. Efecte d'un programa de transferències de rendes amb discontinuïtat en el criteri d'accés al programa: a) discontinuïtat en l'elegibilitat del programa de transferència monetària, b) despesa de les llars en relació amb la pobresa (intervenció posterior).



Font: Gertler i altres, 2011

En la figura veiem la despesa diària de les llars en alimentació segons la posició de les llars en l'índex de pobresa.

a) Ens mostra un programa de transferències de rendes en què estableix com a criteri d'accés tenir una renda per sota del llindar de 50, i b) ens mostra quina és la despesa en alimentació un cop el programa ha estat implementat.

El mètode de la regressió en discontinuïtat es basa a comparar els individus A i B que estan tot just tocant la divisòria que dona accés al programa.

La idea que hi ha al darrere d'aquest mètode és que si comparem individus que estan molt pròxims entre si en la característica que els discrimina per a entrar en el programa, estem simulant una situació semblant a la de l'experiment en què comparem un grup experimental i un altre de control. La raó és que estem comparant l'efecte d'un tractament sobre individus que són quasi idèntics als que no els reben. Però l'aplicació d'aquest mètode demana que el programa compleixi els requisits següents.

- Els individus han de ser susceptibles de ser ordenats seguint una variable contínua que fixarà el llindar d'accés al programa. Per tant, no és aplicable en els casos en què la variable discriminant és binària, o de qualsevol altre tipus, que no permeti discriminar individus molt pròxims entre si. Per exemple, si el criteri d'accés a un programa és tenir un fill, no hi ha manera de trobar situacions equivalents entre tenir-ne un o no tenir-ne cap. Les circumstàncies són molt diferents en tots dos casos.

- La regla d'accés al programa s'ha de respectar escrupolosament. Hem de tenir la certesa que els individus no poden alterar fraudulentament la seva puntuació en la variable discriminant per a tenir accés al programa. En l'exemple de les beques, el mètode no seria recomanable si hi haguessin alguns col·legis que inflessin les notes dels alumnes amb la intenció d'ajudar-los a aconseguir les beques.
- Cal que hi hagi abundància de casos a banda i banda del llindar que serveixin per a discriminar qui és objecte de tractament. Amb pocs casos serà difícil obtenir estimacions estadísticament significatives de l'impacte del programa.

La discontinuïtat entre els que són admesos en el programa i els que no pot ser una discontinuïtat precisa o imprecisa.

- La **discontinuitat precisa** es dona quan tots els individus a banda i banda del llindar discriminant passen mecànicament a la condició de participants o no participants en el programa.

Per exemple, l'edat de jubilació en determinades ocupacions, que obliga a jubilar-se tothom que assoleix l'edat que fixa la llei.

- La **discontinuitat és imprecisa** quan la fixació d'un llindar no és garantia que tothom que la compleixi opti per participar en el programa.

Per exemple, la jubilació entre els professors universitaris és possible als seixanta-cinc anys, però hi ha excepcions que en perllonguen la vida activa més anys. En aquests casos, hi ha un risc de biaix per efecte selecció, ja que la participació en el programa depèn de si es compleix l'edat mínima de jubilació, però a la vegada d'altres factors que desconexem que han empès uns individus a demanar la jubilació però no n'han empès d'altres.

En aquests casos de discontinuïtat imprecisa, la millor solució és optar per una **regressió en dues etapes amb variable instrumental** (Angrist i Pischke, 2009, pàg. 259). De manera que, en la regressió que n'estimarà l'impacte, la variable que indica si l'individu participa en el programa és substituïda per la predicció estimada sobre si hi participarà en una equació prèvia (vegeu la secció següent sobre el mètode de variables instrumentals).

La metodologia de la regressió discontinua pot tenir poca validesa externa, ja que els efectes del programa sobre uns individus pròxims a un llindar no són necessàriament extensibles al conjunt de la població. Per aquesta raó, és recomanable comprovar la robustesa de les estimacions obtingudes repetint la comparació, però ara amb individus situats més lluny del llindar que serveix per a discriminar els beneficiaris dels no beneficiaris del programa. Si els resultats del programa són els mateixos, podem estar més segurs de la validesa externa de l'avaluació que s'ha fet.

5. Mètode de la variable instrumental

Recordem que hi ha tres situacions que hem definit com a situacions d'endogenitat.

1) Una, quan hi ha **relació bidireccional** entre les variables explicativa i explicada, de manera que és difícil establir la direcció del mecanisme causal.

2) L'altra, quan el mecanisme causal és un **mecanisme mediador**, segons el qual, la participació en el programa no és més que un mecanisme mediador d'un altre factor causal que influeix en la decisió dels individus de participar en el programa, a la vegada que també influeix directament en el resultat.

3) Per acabar, hi ha endogenitat quan hi ha **variables explicatives no mesurades**, sempre que aquestes variables influeixin en el resultat del programa, i que també estan correlacionades amb el fet de participar en el programa o no. Serien variables que desconexem o que, coneixent-les, no som capaços de mesurar-les.

En els casos 2 i 3, estem davant d'una situació en la qual l'estimació de l'efecte del programa sobre la població participant pateix del biaix derivat de les variables omeses

Quan les variables no observades tenen influència en la decisió de participar i, a més, estan correlacionades amb el resultat del programa, l'estimador de l'efecte del programa està **esbiaixat** perquè recull efectes associats a la decisió de participar. El mètode de la variable instrumental tracta de corregir aquest biaix i ho fa especificant en el model explicatiu dels resultats una variable que es correlacioni amb la decisió de participar, i que per tant pugui substituir-la en el model, però que no tingui correlació amb el resultat del programa. El cas següent ens ofereix una explicació una mica més tècnica del problema del biaix a causa de variables omeses i l'alternativa de l'estimació en dues etapes utilitzant variables instrumentals.

Esquema explicatiu del mètode de variable instrumental

Model estimat amb una variable omesa que està correlacionada amb el fet de participar i directament amb la variable amb la qual mesurem el resultat del programa.

$$\text{Model estimat: } Y = \beta_1 P + \beta_2 X + \varepsilon$$

$$\text{Model real: } Y = \lambda_1 P + \lambda_2 X + \lambda_3 M + \zeta$$

En què:

P = variable dicotòmica que indica si ha participat en el programa.

X = variables exògenes que tenen incidència en el resultat Y .

M = variables omeses (que poden ser el grau d'heterogeneïtat entre els individus pel que fa a les motivacions, cost d'oportunitat o habilitats, o a l'heterogeneïtat en l'accés a la informació sobre el programa, etc.).

Si es dona que hi ha correlació entre el fet de participar en el programa i alguna variable omesa [$\text{corr}(P, M) \neq 0$] i també hi ha correlació entre la variable omesa i la part no explicada pel model estimat [$\text{corr}(M, \varepsilon) \neq 0$], llavors l'estimació de β_1 serà una estimació esbiaixada de λ_1 , perquè si no posem M en el model, β_1 incorporarà part de l'efecte de M (la variable omesa).

Model estimat amb una variable omesa de la qual la participació en el programa és un mecanisme mediador

Model estimat:

$$y = \beta_1 P + \beta_2 X + \varepsilon$$

si P és un mecanisme mediador de M ,

llavors, $P = \delta M + v$,

si a més M té influència directa en Y

Model real:

$$Y = \lambda_1(\delta M + v) + \lambda_2 X + \lambda_3 M + \zeta$$

Com en el cas anterior, en el model estimat, β_1 està esbiaixada perquè incorpora l'efecte de M .

Com ho podem corregir? Variable instrumental i estimació en dues etapes

Necessitem aïllar la variació de P que és independent de M .

És per això que hem de trobar una variable instrumental Z que compleixi el següent:

- $\text{Corr}(Z, P) \neq 0$ (la variable instrumental està correlacionada amb la de participació).
- $\text{Corr}(Z, \varepsilon) = 0$ (la variable instrumental no té cap efecte sobre el resultat Y que no sigui el que passa per la participació en el programa).

El mètode consisteix a fer una estimació en dues etapes.

Etapa 1: estimar per cada individu la predicció que participi en el programa a partir de la variable instrumental (Z) i altres factors exògens (X).

$$P = \vartheta_1 Z + \vartheta_2 X + \eta$$

Etapa 2: el valor predit \hat{P} substitueix la participació P en el model inicial $y = \beta_1 \hat{P} + \beta_2 X + \varepsilon$.

A continuació veurem tres exemples de com s'aplica aquesta tècnica. El primer és un cas d'endogenitat deguda a la bidireccionalitat en la relació entre la variable *participació* i la variable *resultat*. Els altres dos són casos de variables omeses que estan correlacionades amb el resultat del programa, però en un

cas la variable omesa condiciona la participació en el programa, i en l'altre, malgrat que la variable omesa no condiciona la participació, sí que està correlacionada amb ella.

5.1. Cas d'endogenitat deguda a una relació causal recíproca

Suposem que volem estimar si les dificultats per a pagar l'habitatge arran de la crisi econòmica iniciada el 2008 generen prou estrès per a alterar significativament la salut de les persones afectades per aquestes dificultats. Considerem que una llar té dificultats quan les despeses de lloguer o d'hipoteca superen el 50% de la renda disponible de la llar. Que es doni aquesta situació pot obeir al fet que el preu de l'habitatge sigui molt car, o que les rendes de la llar siguin molt baixes o a una combinació de tots dos. Però si estimem un model en què l'estat de salut l'expliquem directament per les dificultats per a pagar l'habitatge, tindrem biaix d'endogenitat ja que patir dificultats econòmiques i tenir una salut deteriorada són en certa manera endògenes, ja que les persones amb mala salut tenen més dificultats per a sostenir uns ingressos alts i tenen més risc de tenir problemes amb el pagament de l'habitatge. Per tant, és probable que l'estimació de l'efecte de les dificultats sigui esbiaixat.

Però podem utilitzar una variable instrumental, la data de compra de l'habitatge. Atès que els habitatges comprats entre els anys 2000 i 2007 han estat sobrevalorats i han significat una càrrega extraordinària sobre els pressupostos de les famílies, és en aquest grup de compradors on més probables seran els impagaments, però no tenim cap mena de raó per a pensar que la data de compra de l'habitatge hagi d'estar associada a l'estat de salut de les persones.

En casos com aquest, l'estimació de l'impacte d'un programa es fa amb un model de dues equacions que són estimades en dues fases. En la primera estimem la probabilitat de formar part del programa utilitzant com a predictor la variable instrumental (en aquest cas, la primera equació estimaria la probabilitat de tenir dificultats per a pagar l'habitatge incloent-hi variables rellevants com la *situació laboral* i la *renda* a més de la *data de compra de l'habitatge*). En la segona equació estimem l'impacte però substituint la variable *participació* en el programa per la predicció de participar-hi que hem estimat en la primera equació. Amb aquest procediment allò que realment estem estimant és com ha variat el risc de patir un estat de salut precari en funció de la data de compra de l'habitatge, i atribuïm el canvi a les majors dificultats per a pagar-lo.

5.2. Cas d'endogenitat deguda a l'omissió d'una variable, de la qual la participació en el programa és un mecanisme mediador

Un exemple d'endogenitat deguda al fet que la participació en el programa no és més que un mecanisme mediador d'una altra variable causal, el podem trobar quan volem avaluar l'efecte net de l'educació en el benestar econòmic dels individus. Si observem la correlació entre anys d'estudi i ingressos del

treball observem que hi ha una associació positiva, però no tota és atribuïble a l'educació. El fet que un individu decideixi continuar els estudis un cop finalitzada l'etapa de l'escolarització obligatòria depèn de factors no sempre fàcils de mesurar, com ara les motivacions, les habilitats cognitives, el nivell educatiu dels pares, entre d'altres.

Per tant, una avaluació de l'impacte de l'educació estarà esbiaixada si no corregim aquestes influències. Una manera de definir una variable instrumental que permeti controlar aquests efectes pot ser comparar els efectes de l'educació en dos moments històrics amb una oferta educativa marcadament diferent. Aquesta ha estat la tècnica utilitzada a Espanya per Arrazola i Hevia (2001) per a estimar l'efecte de l'educació en els ingressos d'homes i dones. Comparant els ingressos mitjans de la cohort que va començar l'escolarització entre els anys 1936 i 1945, amb els de la cohort que la van començar a partir de 1945, partint del fet que la primera cohort va tenir moltes més dificultats per a accedir a l'educació, ja que durant la Guerra Civil i la postguerra, el pressupost públic per a l'educació primària va caure als nivells del començament del segle XX, i que la depuració de mestres i professors republicans, com també el tancament d'escoles i instituts sota el primer franquisme, van fer que no fos fins als anys seixanta que es va recuperar el nombre d'instituts de 1939.

D'aquesta manera Arrazola i Evía, controlant les variables més rellevants que determinen els ingressos del treball, comparen els ingressos de dues cohorts que van tenir dificultats diferents per a accedir a l'educació; i assumint que les motivacions i d'altres factors eren els mateixos en les dues cohorts, atribueixen la diferència d'ingressos a l'efecte de l'educació.

5.3. Construcció artificial *ex ante* d'una variable instrumental

Trobar variables instrumentals no és una tasca fàcil quan l'avaluació la fem *ex post*, per això és millor cercar alguna solució en la fase d'avaluació *ex ante*, com la que es basa a construir la variable instrumental en funció de l'accés que han tingut els individus a una informació que controlen els responsables del programa. Aquesta estratègia consisteix a difondre informació sobre el programa abans de posar-lo en marxa convidant el públic al qual es dirigeix el programa a participar-hi, tot recollint informació rellevant i les adreces de les persones a les quals donem la informació. Després, quan fem l'avaluació d'impacte, podrem comparar els resultats de les persones que efectivament hi han participat amb els d'aquelles que no hi han participat, però ara, per a superar el problema de la endogenitat, utilitzarem com a variable instrumental haver rebut informació prèvia sobre la posada en marxa del programa. El raonament és el següent: assumim que la probabilitat de participar en el programa és més elevada entre els que han rebut informació, i assumim que haver accedit a aquesta informació subministrada per nosaltres no té cap correlació amb els resultats que els individus puguin obtenir. Per tant, estimem la probabilitat de participar-hi introduint en l'equació el fet d'haver rebut informació o no, i

en l'equació que estima els resultats obtinguts substituïm la variable **participa en el programa** per la predicció que fa la primera equació de la probabilitat de participar-hi.

Suposem el cas hipotètic següent. Una entitat filantròpica vol desenvolupar un programa educatiu amb la finalitat de promoure un ús responsable de les tecnologies de la informació i la comunicació (TIC). Abans, però, vol fer una experiència pilot i avaluar-la. L'equip encarregat de dissenyar l'avaluació valora les dificultats de fer una avaluació experimental i descarta aquesta opció, ja que, en tant que la participació en el programa serà voluntària, no pot fer una selecció aleatòria de qui va al grup experimental i qui va al grup de control. També havia sospesat l'opció de fer una assignació aleatòria dels voluntaris a dos grups que rebrien versions diferents del programa, però l'entitat promotora es nega a fer dos programes diferents.

Descartada l'opció experimental, tria el mètode quasiexperimental d'estimació en dues etapes amb construcció *ex ante* d'una variable instrumental. Per fer-ho opta per seleccionar una mostra aleatòria d'una ciutat; d'aquesta mostra selecciona, també aleatòriament, dos grups A i B. Als membres del grup A els envia informació bàsica sobre el programa, mentre que als membres del grup B els envia la informació bàsica més informació suplementària.

Uns quants mesos després d'haver acabat el programa pilot, fa una enquesta sobre usos de les TIC als membres del grup A i el B, i mesura un índex sintètic de bons usos de les TIC que oscil·la entre 0 i 100, en què 0 és un ús molt negatiu i 100, molt positiu. Amb les variables recollides en l'enquesta fa un model predictiu de la probabilitat de participar en el programa, incloent en el model si han rebut informació bàsica només o bàsica i suplementària. Amb la predicció de participar en el programa que el model assigna a cada individu, es fa la segona etapa de l'estimació, en què es fa la regressió de l'índex de bons usos de les TIC sobre la predicció de participar en el programa que ha estimat en la primera etapa.

Els resultats li mostren que del grup A, que havia rebut la informació bàsica, només un 18% va participar en el programa, i obté un valor mitjà de l'índex de 58, mentre que del grup B, un 40% va participar en el programa i obté un valor mitjà de 65, essent la diferència dels valors entre els dos grups estadísticament significativa.

En aquest cas hipotètic, la construcció de la variable instrumental *ex ante* hauria permès identificar dos grups homogenis amb l'única diferència d'haver rebut més o menys informació per a participar en el programa educatiu. Si finalment, com és el cas, es nota una diferència en els usos de les TIC de dos grups, només pot ser atribuïble al fet que en un dels grups hi ha una proporció més gran de persones que ha assistit al programa educatiu.

6. Lectures complementàries

Behrman, J. R.; Tood, P. E. (2000). "Aleatoriedad en las muestras experimentales del Programa de Educación, Salud y Alimentación (Progresas)". *Evaluación de Resultados del Programa de Educación, Salud y Alimentación* (pàg. 51-88).

Aquesta lectura mostra un bon exemple pràctic de com dur a terme una avaluació amb assignació aleatòria als grups de tractament i de control, salvant les objeccions que no hi ha suport ètic per a privar algú de l'accés a un programa que pot ser-li beneficiós.

És molt important que l'estudiant vegi:

- 1) La importància de comprovar, abans d'implementar el programa, que l'assignació aleatòria ha donat lloc a grups realment homogenis, ja que és possible que hi hagi un cert error no sistemàtic, però que cal corregir repetint l'assignació aleatòria si cal.
- 2) Els inconvenients de fer una assignació aleatòria d'àrees i no de llars, com que hi ha una mostra més petita i que el resultat avaluat no és estrictament dels que han rebut el tractament, sinó de la mitjana de tractats i no tractats d'una mateixa àrea. Amb aquest exemple l'estudiant pot entendre millor el tractament d'aquest mateix tema que fa la lectura de Vera-Hernández, en què mostra com a avantatge del mètode *matching* que es pot mesurar l'efecte del programa en els que realment han rebut tractament.
- 3) És important prestar atenció als biaixos d'aleatorització, de contaminació i de desgast (atricció).
- 4) També és important que reflexioni sobre les possibles causes i inconvenients del fet que, malgrat que la comparació entre àrees no mostra diferències significatives entre el grup de control i l'experimental, les diferències sí que ho són quan es comparen llars.
- 5) Si l'estudiant té una base estadística mínima, fóra convenient que consultés l'annex en què s'explica que si cal comparar els resultats entre els dos grups d'una variable contínua, convé fer-ho amb la prova de l'estadístic de Kolmogorov-Smirnov, i que si es tracta de variables categòriques, amb la prova de khi quadrat. Totes dues proves accessibles amb els paquets estadístics estàndard com SPSS, STATA i R.

Casado, D. (2012). "Per què no avaluem les polítiques públiques com els fàrmacs? Una aposta per l'experimentació social", *Avaluació per al Bon Govern*, núm. 3 Ivàlua.

Aquest article, introdueix l'estudiant en els avantatges de l'avaluació experimental, però bàsicament serveix perquè l'estudiant reflexioni sobre l'endarreriment que hi ha al nostre país en la implementació sistemàtica d'avaluacions i sobre com aquest escenari pot canviar en els propers anys. A més convé que presti atenció al següent:

- 1) Les crítiques que s'acostumen a fer al mètode experimental i com es poden rebatre.
- 2) Les reticències dels polítics a acceptar avaluacions rigoroses.
- 3) La diferència conceptual entre mostreig aleatori i assignació aleatòria a un grup experimental i un altre de control.
- 4) És convenient remarcar la dada que els estudis avaluatius fets als EUA de programes de justícia i prevenció de la delinqüència han mostrat una quantitat elevada de programes amb cap efecte o, fins i tot, amb efectes contraproductius per als participants, la qual cosa fa equivalents els beneficis/danys dels programes socials als dels medicaments, i per tant s'haurien de sotmetre al mateix grau de control. Si el tutor considera que els estudiants tenen prou domini de l'anglès, seria recomanable, com a estudi de cas il·lustratiu, que llegissin J. McCord (maig de 2003), "Cures that Harm. Unanticipated effects of crime prevention programs", *Annals, AAPSS* (vol. 587).

Lapuerta, I. (2013). "¿Influyen las políticas autonómicas en la utilización de la excedencia por cuidado de hijos?" *Revista Española de Investigaciones Sociológicas* (núm. 141, pàg. 29-60).

L'estudiant ha de prestar atenció als punts següents:

- 1) El treball ofereix en primer lloc un estat de la qüestió pel que fa al coneixement que hi ha sobre els efectes de les excedències. Un requisit imprescindible en tota avaluació ha de ser proveir aquest estat de la qüestió.
- 2) Aquesta avaluació és un bon exemple d'allò que es pot fer amb bases de dades longitudinals que permeten fer un seguiment dels individus durant períodes llargs de temps.
- 3) A quins problemes de biaix s'enfronta?
 - Les motivacions i costos d'oportunitat dels que prenen les baixes són diferents (biaix d'autoselecció).

- Relació causal bidireccional (endogenitat): l'oferta genera demanda, però la demanda també pressiona perquè es desenvolupi una oferta.

4) Com és que el mètode de doble diferència permet obviar aquest problema d'endogenitat? Podia haver triat un *propensity score matching*? Podia haver avaluat els programes amb el mètode experimental?

5) Per què compara homes i dones per separat?

- Castella - la Manxa només oferia la prestació als homes.
- Hi ha indicis empírics previs que mostren comportaments diferencials entre sexes.

6) Quin risc de biaix no ha pogut evitar?

- Que les dones amb més dificultats per a mantenir-se en el mercat de treball hagin abandonat la mostra de l'MCVL abans de 2006, ja que la mostra és de les treballadores actives el 2006. Però assumeix que, aquest abandó no ha de ser diferent entre comunitats autònomes, per tant, si l'assumpció és correcta, la validesa interna no es veuria afectada.
- En qualsevol cas, podria veure's afectada la validesa externa? Sí, perquè no ha pogut avaluar l'efecte dels programes sobre aquestes dones.

7) Com controla que no hi hagi pogut haver un biaix per factors externs que operin com a mecanismes moderadors? Per exemple, que alguna comunitat autònoma hagués canviat les seves condicions en els mecanismes moderadors. Per controlar-ho, inclou les taxes d'atur, l'escolarització de menors de dos anys, i les subvencions a la reducció de jornada per si han canviat en el període estudiat i han fet canviar els incentius de triar l'excedència.

Sarasa, S. (2012). "Efectos de los centros de educación preescolar en el desarrollo de los adolescentes". A: C. Gómez-Granell; P. Mari-Klose (coord.). *Familia y relaciones intergeneracionales* (pàg. 75-97). Barcelona: CIIMU.

Aquest estudi de cas ha de servir l'estudiant per a il·lustrar la metodologia de l'avaluació quan tenim indicis que hi ha endogenitat deguda al fet que la participació en el programa, en aquest cas l'escola bressol, està associada a un o més factors que a la vegada són causa del resultat del programa, en aquest cas el desenvolupament dels adolescents. L'estudiant ha de prestar atenció al següent:

1) La necessitat de fer comprovacions prèvies per a descartar, o no, la possibilitat que hi hagi endogenitat o biaix de selecció entre els participants.

2) Davant la impossibilitat de construir una variable instrumental, ja que és una avaluació feta *ex post* sense que hi hagués una previsió inicial, la necessitat de cercar una variable instrumental que estigui correlacionada amb el fet de participar en el programa, però no directament amb el resultat del programa (en aquest cas, la ideologia política de la mare).

3) Estudiar com opera el mètode d'estimació en dues fases. Primer, estimant la probabilitat que un adolescent hagi estat preescolaritzat abans de tres anys d'edat en funció de la ideologia de la mare i, segon, estimant l'efecte de l'escola bressol en el desenvolupament de l'adolescent, però substituint la variable *preescolarització* per la seva probabilitat estimada en el primer model.

4) Cridar l'atenció sobre la necessitat d'avaluar no només els efectes positius esperats del programa, sinó també els efectes negatius predits per teories crítiques amb el programa.

5) Important, fer èmfasi en la rellevància de la variable *temps* en l'avaluació de programes. En aquest cas, una avaluació tan diferida en el temps, sense haver pogut observar què ha passat des de la finalització de l'etapa preescolar, fa dubtar de la validesa interna de l'avaluació.

6) També és important cridar l'atenció sobre els efectes atribuïbles a una variable que tampoc no ha pogut ser observada, la qualitat dels centres.

Vera-Hernández, M. (2003). "Evaluar intervenciones sanitarias sin experimentos". *Gaceta Sanitaria* (vol. 17, núm. 3, pàg. 238-248).

Aquest article ha de servir perquè l'estudiant vegi l'aplicació del mètode del *propensity score matching* i com pot ser combinat amb el mètode de la doble diferència en el casos en què tenim dubtes sobre si algunes de les restriccions que assumeix aquest mètode són raonables.

L'estudiant ha de prestar atenció als temes següents:

1) Què tenen en comú el mètode experimental i el mètode de *matching*? Les variables condicionants han de tenir la mateixa distribució en el grup de control i en el de participació. Però cal que l'estudiant recordi que aquest és un supòsit molt restrictiu, ja que si bé una distribució aleatòria en el mètode experimental ha de donar una distribució homogènia entre els dos grups, i que si hi ha un error de mostreig tendirà a desaparèixer si repetim l'avaluació, no tenim la mateixa garantia amb el mètode del *matching*, ja que si hi ha condicionants no observables la distribució podria estar sistemàticament esbiaixada.

2) Quines són les limitacions que presenten altres mètodes no experimentals, com el mètode reflexiu (abans-després), semiexperimentals d'avaluació, com el de doble diferència.

3) Quins són els fonaments del mètode del *propensity score matching*?

Altres lectures recomanades

Gertler i altres (2011). *La evaluación de impacto en la práctica* (2a. part, cap. 3-9). Washington: Banco Internacional de Reconstrucción y Fomento / Banco Mundial.

Ivàlua. *Avaluació d'impacte*. Guia pràctica 5.

<http://www.ivalua.cat/>

documents/1/01_03_2010_11_33_12_Guia5_Impacte_Setembre2009_revfeb2010_massavermella.pdf