Tamaño de muestra para identificar el impacto en una regresión discontinua

José A. Valderrama Torres[1*]

1 Estudios Económicos de la Oficina de Normalización Previsional (ONP) (e-mail: jvalderrama@onp.gob.pe). El documento se preparó mientras el autor laboraba en la Dirección de Calidad del Gasto del Ministerio de Economía y Finanzas (MEF) como insumo para el diseño de la evaluación de impacto del programa social "Pensión 65". Cualquier error u omisión es de exclusiva responsabilidad del autor y no compromete a la institución a la que el autor representa.

* Ingeniero Economista de la Facultad de Ingeniería Económica y Ciencias Sociales de la UNI, Magíster de Economía de la Universidad de Chile. Actualmente se desempeña como Jefe de Estudios Económicos de la Oficina de Normalización Previsional (ONP) y profesor en la Universidad de Lima. Ha sido profesor en la PUCP, UPC, USMP y CIES-INEI dictando cursos como microeconomía, microeconometría y evaluación de impacto. Ha realizado diversas investigaciones y consultorías, siendo la mayoría relacionados al análisis de datos de encuestas complejas. Sus ultimas investigaciones han estado orientadas a la determinación del tamaño de muestra requerido para una regresión discontinua teniendo en cuenta un diseño bietápico de levantamiento de datos y el efecto que tiene el incremento de la oferta laboral policial sobre la seguridad ciudadana.

Resumen

Se presenta analíticamente la determinación del tamaño de muestra necesario para identificar el impacto en una Regresión Discontinua. En particular, se considera el escenario de violación del supuesto de independencia de los errores provocados cuando los datos se encuentran distribuidos en conglomerados.

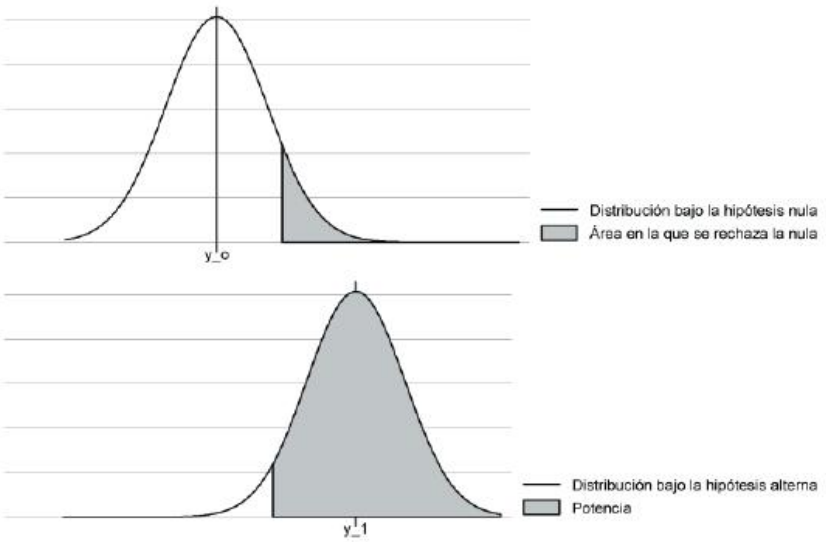
1. Introducción

El enfoque planteado por el análisis de una Regresión Discontinua (RD) permite estimar impactos de programa en situaciones en donde los candidatos son seleccionados para ser sometidos a un tratamiento, siempre y cuando el valor numérico de un puntaje, como la pobreza por ejemplo, sea inferior a un umbral o punto de corte predeterminado. En la medida que las observaciones que se encuentren en la vecindad del umbral puedan ser consideradas aleatorias, entonces la comparación entre los beneficiarios con los no beneficiarios podría ser una estimación del impacto del programa, pues la única diferencia entre ambos grupos es que un grupo recibe el tratamiento y el otro no.

En ese sentido, un escenario ideal en la etapa del diseño de una evaluación, es definir el tamaño apropiado de la muestra de tal forma que ésta permita medir el impacto si es que éste existe. Esto último vale la pena resaltar, pues no detectar impacto, no necesariamente significa que éste no exista, pues la muestra podría ser insuficiente como para detectarlo. Una definición relacionada a esto es el Efecto Mínimo Detectable (EMD). Intuitivamente, es el impacto más pequeño que un ejercicio estadístico es capaz de detectar (Bloom, 1995).

La utilidad del EMD se puede ilustrar en el siguiente ejercicio, basado en una prueba unilateral de medias, donde la hipótesis nula es $\bar{y}_0$ y la alterna es $\bar{y}_1$ (Ver figura N° 1). Así, dado un valor crítico, un nivel de significancia α y un nivel de potencia (1-β) se tiene que:

Figura 1: Potencia



$$\bar{y}_0 + z_{1-\alpha} SE(\bar{y}_0) = \bar{y}_1 + z_\beta SE(\bar{y}_1) \qquad (1)$$

$$\bar{y}_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} = \bar{y}_1 + z_\beta \frac{\sigma}{\sqrt{n}} \qquad (2)$$

$$n = \left[\frac{z_{1-\alpha} - z_\beta}{\bar{y}_1 - \bar{y}_0} \sigma \right]^2 \qquad (3)$$

Es decir, se necesita una muestra de tamaño n para alcanzar los niveles de α y β cuando se elige a $\bar{y}_1$ como valor alternativo de $\bar{y}_0$.

Del mismo modo, en una regresión en la que un parámetro identifica el impacto y en el que rechazar la hipótesis significa que existe impacto, el error estándar del mismo junto con el EMD propuesto son insumos para la determinación del tamaño de muestra, pues mientras más dispersión tenga el estimador del impacto y más exigente sea el EMD, la muestra necesaria para medir el impacto, si es que existe, deberá ser más grande.

En lo que sigue se determinará el tamaño de muestra requerida para la identificación del impacto empleando una regresión discontinua en su versión paramétrica[2]. Para conseguir

2 El ejercicio se complejiza cuando se asume un enfoque no paramétrico.

esto se estimará la varianza de interés en un contexto de dependencia de los datos generado por que el levantamiento de datos no sigue un diseño aleatorio simple, sino más bien uno en el cual los datos son elegidos al interior de conglomerados.

2. Tamaño de muestra en una regresión discontinua

2.1 Especificación del modelo

Para describir la idea básica del diseño, se debe tener en cuenta que antes del tratamiento la relación entre el puntaje a partir del cual se determina si una persona es elegible o no, como por ejemplo el nivel de pobreza (score) y la variable de resultado (y) es dado por la siguiente regresión lineal:

$$y_i = \hat{\gamma}_0 + \hat{\gamma}_2 S_i + \varepsilon_i$$

Después del tratamiento, si las unidades tratadas son afectadas por un efecto constante y_1 sobre la variable de resultado, entonces la regresión puede ser especificada como:

$$y_i = \hat{\gamma}_0 + \hat{\gamma}_1 T_i + \hat{\gamma}_2 S_i + \varepsilon_i$$

Donde T_i es una dicotómica que toma el valor de 1 si la unidad de análisis es asignado al tratamiento, o 0 si éste es asignado al grupo de control. Debido a que se asume que el efecto es constante, la pendiente de la regresión no cambia aunque si lo hace el intercepto para los tratados que pasa a ser $\hat{\gamma}_0+\hat{\gamma}_1$, es decir, la diferencia entre tratados y controles es $\hat{\gamma}_1$, parámetro que es interpretado como el efecto del programa.

Esta especificación puede ser generalizada para considerar la eventual dependencia de los errores cuando la muestra es recogida empleando conglomerados geográficos como primera unidad de muestreo y los hogares como segunda unidad (diseño bietápico). Dicho procedimiento es usual en el diseño de encuestas complejas como lo son las encuestas a hogares, que por su magnitud requiere una muestra representativa a nivel de país la misma que sería prohibitiva si esta se hiciera bajo la lógica de un muestreo aleatorio simple (MAS) y que es más accesible si sigue la lógica de muestreo en etapas. La desventaja de esta última es que la precisión de las estimaciones empeora, por lo que intuitivamente el tamaño de muestra exigido para conseguir los mismos resultados que se obtendrían bajo un MAS es superior[3].

3 Ciertamente las encuestas de hogares consideran otros elementos en el diseño como la estratificación, que provoca un efecto contrario a la "clusterización", es decir, bajo ciertas condiciones, puede mejorar la precisión de las estimaciones (Deaton (1997)). En ese sentido, esta propuesta puede ser considerada como un escenario límite, en el que la estratificación no representa mejoras en las estimaciones.

Así, para contemplar la dependencia de los datos al interior de los conglomerados, se puede añadir al error el término μc, es decir, una perturbación que es homogénea al interior del grupo "c" y distinta entre grupos, con lo cual el error compuesto v_{ic}, incluso asumiendo independencia de ε_{ic}, se encuentra correlacionado, violando así uno de los supuestos de la estimación bajo mínimos cuadrados ordinarios.

Específicamente, si ε_{ic} es un término de error no correlacionado con media cero y con varianza igual a σ_{ε}^2, mientras que se asume que es un error independiente e idénticamente distribuido entre grupos con una media de 0 y una varianza de σ_{μ}^2. Entonces, el error compuesto v_{ic} cumple con:

$$Cov(v_{ic}, v_{js}) = \sigma_{\mu}^2 + \sigma_{\varepsilon}^2;\ \forall\ i = j\ y\ c = s$$
$$Cov(v_{ic}, v_{js}) = \sigma_{\mu}^2;\ \forall\ i \neq j\ y\ c = s$$
$$Cov(v_{ic}, v_{js}) = 0;\ \forall\ i \neq j\ y\ c \neq s$$

Adicionalmente, un supuesto razonable y simplificador es que los covariados al interior de cada conglomerado sean los mismos. Este supuesto es usualmente empleado en la literatura para estimar la varianza de los parámetros en un contexto de agrupamiento de datos (Ver por ejemplo Deaton, 1997).

Finalmente, la especificación de interés quedaría definida por la siguiente expresión:

$$y_{ic} = \hat{\gamma}_0 + \hat{\gamma}_1 T_c + \hat{\gamma}_2 S_c + \mu_c + \varepsilon_{ic}$$
$$y_{ic} = \hat{\gamma}_0 + \hat{\gamma}_1 T_c + \hat{\gamma}_2 S_c + v_{ic}$$

Donde, dado el comportamiento del término de error compuesto v_{ic} la estimación eficiente exige la aplicación del método de Mínimos Cuadrados Generalizados (MCG) para lo cual, se define $Var(v)=E(vv')$ $= \sigma_v^2 \Omega$; donde $\Omega^{-1}=P'P$.

$$Py = PXB + Pv \qquad (4)$$

Donde:

$$Var(PV) = E(Pvv'P') = PE(vv')P'$$
$$Var(PV) = P\sigma_v^2 \Omega P' = \sigma_v^2 I_{nc}$$

El nuevo término de error Pv no presenta problema de correlación o dependencia. De (4) se puede deducir la nueva matriz de varianza:

$$Var(\beta_{MCG}) = \sigma_v^2 (X'\Omega^{-1}X)^{-1} \qquad (5)$$

Donde Ω^{-1} se puede obtener de:

$$Var(v) = \sigma_v^2 \Omega = E[(z_\mu \mu + \varepsilon)(z_\mu + \varepsilon)']$$
$$= z_\mu E(\mu\mu') z_\mu' + E(\varepsilon\varepsilon')$$
$$= z_\mu \sigma_\mu^2 I_c z_\mu' + \sigma_\varepsilon^2 I_{nc}$$
$$= \sigma_\mu^2 z_\mu z_\mu' + \sigma_\varepsilon^2 I_n \otimes I_c$$
$$= \sigma_\mu^2 I_c \otimes J_m + \sigma_\varepsilon^2 I_m \otimes I_c$$

Sabiendo que $J_m = m\bar{J}_m$ y $I_m = E_m + \bar{J}_m$

$$Var(v) = \sigma_\mu^2 I_c \otimes m\bar{J}_m + \sigma_\varepsilon^2 I_c \otimes (E_m + \bar{J}_m)$$

$$= \sigma_\varepsilon^2 I_c \otimes E_m + \sigma_1^2 I_c \otimes \bar{J}_m$$

Finalmente;

$$Var(v) = \sigma_v^2 \Omega$$
$$= \sigma_\varepsilon^2 Q + \sigma_1^2 S$$

Q y S y son matrices simétricas e idempotentes, por lo tanto:

$$\Omega = \frac{\sigma_\varepsilon^2}{\sigma_v^2} Q + \frac{\sigma_1^2}{\sigma_v^2} S$$

Es decir:

$$\Omega^{-1} = \left(\frac{\sigma_v}{\sigma_\varepsilon}\right)^2 Q + \left(\frac{\sigma_v}{\sigma_1}\right)^2 S \qquad (6)$$

Reemplazando en (5)

$$Var(\beta_{MCG}) = \sigma_v^2 (X'\Omega^{-1}X)^{-1}$$
$$= \left[X'\left(\frac{S}{\sigma_1^2} + \frac{Q}{\sigma_\varepsilon^2}\right)X\right]^{-1}$$
$$= \left[\frac{X'SX}{\sigma_1^2} + \frac{X'QX}{\sigma_\varepsilon^2}\right]^{-1}$$
$$= \left[\frac{X'X}{\sigma_1^2}\right]^{-1}$$

$$Var(\beta_{MCG}) = [1 + \rho(m-1)]\sigma_v^2 (X'X)^{-1} \qquad (7)$$

Donde m es el número de observaciones al interior del conglomerado y P es el coeficiente de correlación intraclase que se define como:

$$\rho = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_\varepsilon^2} \qquad (8)$$

Es decir, la varianza del parámetro de interés es $(1+P(m-1))$ veces más grande respecto a lo que se obtendría en un MCO convencional[4]. Dado que la varianza del parámetro de interés $\bar{y}_1$ bajo MCO es un insumo en la determinación de la varianza de la estimación bajo MCG, en la siguiente subsección se presenta convenientemente una versión no matricial de la estimación de la varianza bajo MCO.

2.2 Varianza en un modelo MCO

Bajo MCO la especificación vendría dado por las siguientes ecuaciones:

$$y_i = \hat{\gamma}_0 + \hat{\gamma}_1 T_i + \hat{\gamma}_2 S_i + \varepsilon_i \qquad (9)$$

$$T_i = \hat{\alpha}_0 + \hat{\alpha}_1 S_i + \mu_i \qquad (10)$$

Sabiendo que la varianza del parámetro de interés es:

$$Var(\hat{\gamma}_1) = \frac{\hat{\sigma}^2}{SCT_{TS}(1 - R_{TS}^2)}$$

4 Como lo señalan Cameron, Gelbach y Miller (2009), cuando los covariados no son constantes al interior de cada conglomerado este factor sería aproximadamente igual a $(1+P\varepsilon Pxj(m-1))$, donde P_ε es la correlación intracluster de los residuos y Pxj es la correlación intracluster de xj.

Siendo $\hat{\sigma}^2$ el estimador de la varianza de los errores, y SCT_{TS} y R^2_{TS} corresponden a la Suma de cuadrados totales y al R^2 del modelo (10) (Ver Wooldridge (2009))

Teniendo en cuenta que la varianza muestral

$$SCT_{TS} = \frac{\sum(T_i - \bar{T})^2}{n-1}(n-1)$$

Puede ser reemplazada por su equivalente poblacional:

$$SCT_{TS} = p(1-p)(n-1)$$

donde es el porcentaje de tratados. De este modo la varianza asintóticamente puede ser presentado por:

$$Asy - Var(\hat{\gamma}_1) = \frac{\bar{\sigma}^2}{np(1-p)(1-R^2_{TS})} \quad (11)$$

$$= \frac{\sigma^2_y(1-R^2)}{np(1-p)(1-R^2_{TS})} \quad (12)$$

2.3 Tamaño de muestra

Finalmente, el error estándar del parámetro de interés queda definido como:

$$SE(\hat{\gamma}_1) = \sqrt{[1+\rho(m-1)]\frac{\sigma^2_y(1-R^2)}{np(1-p)(1-R^2_{TS})}} \quad (13)$$

Con lo cual, extrapolando lo desarrollado en (1) y (2) se tendría que:

$$0 + z_{1-\alpha}SE(\hat{\gamma}_1) = \hat{\gamma}_1 + z_{\beta}SE(\hat{\gamma}_1)$$

Es decir, el tamaño de muestra queda definido como:

$$n = [1+\rho(m-1)]\frac{(z_{1-\alpha}-z_{\beta})^2(1-R^2)}{EMD^2p(1-p)(1-R^2_{TS})}$$

Puesto que en general se espera que $m>1$, entonces, la dependencia de los errores exige un mayor tamaño de muestra, a menos que $p=o$, no obstante, de acuerdo a Bloom (2006) este suele variar entre 0.01 y 0.20. Por su parte el EMD, según Cohen (1977, 1988)[5] toma un valor de 0.2 desviaciones estándar es considerado pequeño[6].

Otro parámetro para el que se puede anticipar un valor es R^2_{TS}. Siguiendo a Schochet (2008) y bajo ciertos supuestos, este parámetro podría tomar el valor de 9/16[7]. Todos los parámetros restantes, a excepción de , se les puede asignar valores usuales, por ejemplo el nivel de significancia α es igual a 5%, la potencia suele ser considerada como aceptable cuando supera al 80%, el porcentaje de tratados de la muestra p se asume que es igual a 50%. Finalmente, el R^2, al ser una estimación de corte transversal, se espera que este sea cuando menos 20% para ser considerado un buen ajuste.

5 Citado en Bloom (2006).

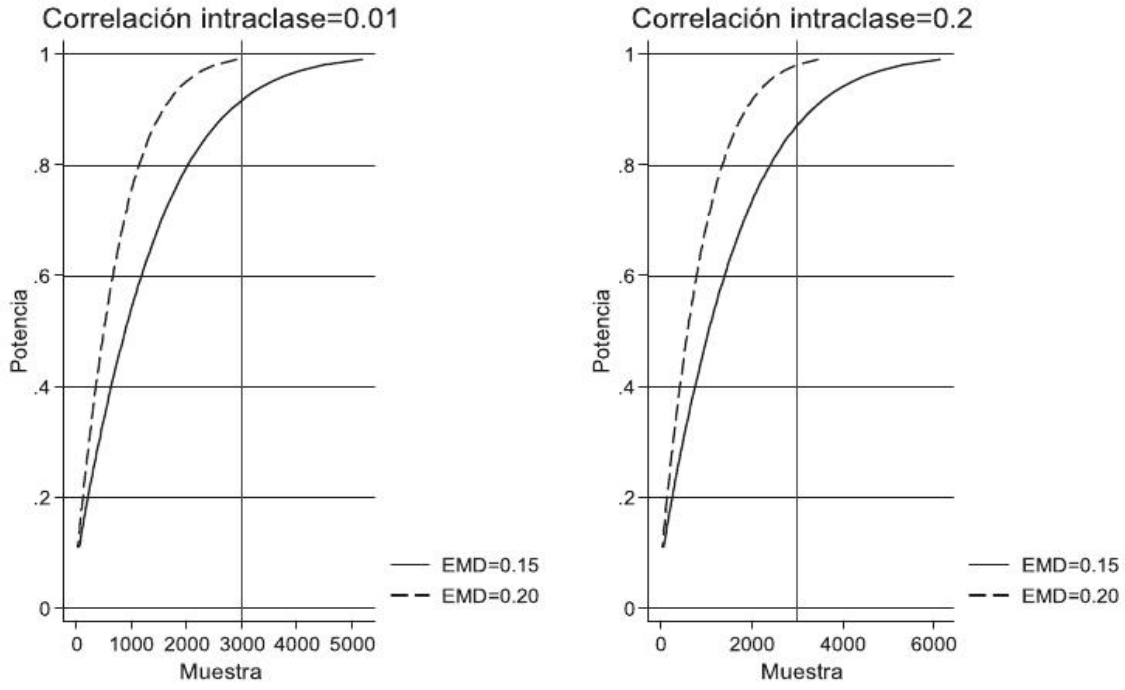
6 0.5 y 0.8 son considerados, de acuerdo al mismo autor, como moderado y grande, respectivamente.

7 Específicamente, este parámetro se alcanza cuando el score sigue una distribución uniforme y cuando el porcentaje de los tratados en la muestra es igual al porcentaje en la distribución del score que se encuentra a la derecha del punto de corte, en cuyo caso el valor es de $R\_TS\text{^}2=3p(1-p)$.

Finalmente, el valor de m es específico al diseño del muestreo. Por ejemplo, asumiendo que el levantamiento de datos es por conglomerados geográficos y que la intervención a analizar es el programa Pensión 65[8], empleando la Encuesta Nacional de Hogares 2010, se estima que en cada conglomerado geográfico hay en promedio aproximadamente 2 personas de 65 a más años calificados como pobres extremos, de acuerdo con el criterio de pobreza monetaria.

Así, fijando arbitrariamente un R^2=20%, teniendo en cuenta los casos extremos de correlación intraclase p y considerando dos valores de EMD se presenta la sensibilidad de la potencia del test ante distintos tamaños de muestra. Para un EMD=0.2 incrementar la muestra más allá de 3200 observaciones no representa mejoras significativas en la potencia estadística. Incluso considerando un EMD=0.15 la potencia supera al límite de 80%, encontrándose por arriba de 90% en el caso de un p=0.01 y cercano al mismo valor en el caso de p=0.02[9].

Figura 2: Potencia y tamaño de muestra



8 Pensión 65 es un programa que beneficia a los adultos mayores de 65 a más años que se encuentran en condición de pobreza extrema, con transferencias monetarias y acceso a servicios de salud.

9 Nótese que, de acuerdo a lo señalado en la nota 3, la muestra calculada en esta sección puede ser considerada un valor techo, pues de no haberse asumido que los covariados tienen el mismo valor al interior de cada cluster, entonces el coeficiente de correlación *pxj* sería menor a uno.

Referencias

Angrist, J., and Krueger, A 2007. Empirical Strategies in Labor Economics. Handbook of Labor Economics, ed. Orley Ashenfelter and David Card, Vol 3. North Holland, Amsterdam.

Bloom, H. 1995. Minimum Detectable Effets. A Simple Way to Report the Statistical Power of Experimental Designs}. Evaluation Review. Sage Publications, Inc, Vol 19. N° 5 October. 547-556.

Bloom, H. 2006. The Core Analytics of Randomized Experiments for Social Research. MDRC Working Papers on Research Methodology.

Cameron, A, Gellbach, J. y Miller, D 2006. "Robust Inference with Multy-way Clustering". NBER Technical Working Paper No. 327.

Deaton, A 1997. The Analysis of Household Surveys. A Microeconomic Approach to Development Policy. A Microeconomic Approach to Development Policy. World Bank. The Johns Hopkins University Press.

Imbens, G., and T. Lemieux, 2008 Regression Discontinuity Designs: A Guide to Practice. Journal of Econometrics, Vol 142(2), 615-635.

Schochet, Peter Z., 2008. Technical Methods Report: Statistical Power for Regression Discontinuity Designs in Education Evaluation. Evaluations (NCEE 2008-4026). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Wooldridge, J. 2009. Introducción a la Econometría. Un enfoque moderno. 4ta. Edición.