CAPÍTULO 28

ESTIMACIÓN NO PARAMÉTRICA DE FUNCIONES DE REGRESIÓN EN POBLACIONES OCULTAS A TRAVÉS DE *RESPONDENT-DRIVEN SAMPLING (RDS)*

Pilar Olave Rubio

1. INTRODUCCIÓN

Este capítulo pretende ser mi pequeño reconocimiento a Vicente Salas, gran investigador y mejor persona, que siempre ha repetido "los matemáticos tenéis un gran potencial y las herramientas imprescindibles en modelos económicos". Algo que hoy es una obviedad, pero no lo era a finales de los setenta, cuando nos conocimos.

Este trabajo presenta un estimador de la función de regresión de determinadas variables de interés en el contexto de la Economía Social particularizando al caso de la población de Personas sin Hogar (PSH), utilizando un muestreo alternativo a los tradicionales, ya que en las poblaciones denominadas "ocultas", como es el caso de las PSH, el muestreo probabilístico usual o cualquier otro basado en él, son inviables por la imposibilidad de un marco muestral. Así pues, se presenta el muestreo "impulsado" por los propios encuestados, *Respondent Driven Sampling (RDS),* que al igual que en otros muestreos basados en el rastreo de enlaces, hace uso de un muestreo en red en donde varios individuos de la población objetivo se inscriben en el estudio como "semillas", y los individuos posteriores se seleccionan en función de sus conexiones de red con los miembros anteriores de la muestra, que se encargan de reclutarlos a través de diversas oleadas. A pesar de su reciente uso en temas de importancia sobre Salud Pública, Economía y Política Social, las propiedades estadísticas y las estrategias inferenciales óptimas para los estimadores resultantes del *RDS* aún requieren de estudios adicionales. Aquí se presenta su aplicación a un modelo de regresión y de alguna forma, es un ejemplo de la frase del profesor Salas, que tan acertadamente señalaba "la metodología está en vuestras investigaciones",

y este método creo que tiene un gran alcance para todo tipo de poblaciones ocultas tan habituales en nuestra sociedad.

A modo de introducción, quiero señalar cómo fue mi primer contacto con Vicente Salas. Mi maestro el profesor Cano Sevilla, comisionado junto con otras personas para esbozar el primer Plan de Estudios de la Facultad de Económicas y Empresariales, fue al primero que le oí hablar del gran investigador Salas Fumás, cuando nos comentó a varios de sus entonces jóvenes becarios, que si algún economista se tendría que encargar de impartir Investigación Operativa sería él. Finalmente, hablaron con el catedrático de Estadística e Investigación Operativa Miguel Sánchez, recién llegado de Madrid y él se encargó de diseñar e impartir dicha docencia, ya que Vicente siempre ha buscado el óptimo en cada situación anteponiéndolo a sus propias satisfacciones personales.

Luego, vinieron nuestras conversaciones sobre funciones de utilidad, primas de riesgo, tipos de interés y política monetaria en España…, y las correcciones de mis primeros trabajos en revistas y reuniones de economía que como él decía tenían potencial. Años más tarde me invitaron a un seminario sobre mercados financieros, organizado por BBVA en la universidad Carlos III, y acepté participar una vez que Vicente había revisado el *preprint* de "Prima de riesgo y volatilidad en el mercado de valores español".

Así pues, intentaré con este trabajo poner de manifiesto que las interrelaciones entre áreas del conocimiento son siempre bienvenidas y no hacen sino enriquecernos en la búsqueda de nuevos instrumentos. Gracias Vicente por creer en mis posibilidades y la de mis colegas matemáticos y por supuesto, al Departamento de Dirección y Organización de Empresas por permitir mi incursión en este sentido homenaje.

2. POBLACIONES OCULTAS. MUESTREO *RDS*

2.1. Poblaciones ocultas

Se dice que una población está "oculta" cuando no existe un marco de muestreo y el reconocimiento público de la pertenencia a la población puede suponer un estigma para el individuo. Por ejemplo, la población de individuos que se inyectan drogas en una determinada ciudad es de enorme interés si se desea estudiar la incidencia del virus VIH en dicha población, pero es difícil encontrar un marco para muestrearla y además algunos individuos puede encontrar amenazante reconocer su pertenencia a la misma.

Acceder a tales poblaciones es difícil porque los métodos de muestreo probabilísticos estándar producen tasas de respuesta bajas (por ejemplo, una muestra aleatoria de los individuos de toda la ciudad incluirá un número muy escaso de personas que son usuarios activos de drogas inyectables) y las respuestas pueden no ser sinceras. Sin embargo, la identificación de estos grupos es crucial para el desarrollo de intervenciones efectivas de prevención del sida, y encontrar formas válidas y fiables de muestreo es esencial para evaluar las intervenciones de política pública.

Las "poblaciones ocultas" tienen dos características: primero, no existe un marco de muestreo, por lo que el tamaño y los límites de la población son desconocidos, y además los individuos a encuestar suelen estar geográficamente dispersos; y segundo, existen fuertes preocupaciones sobre la privacidad, porque la pertenencia a ellas implica un comportamiento estigmatizado o ilegal, lo que lleva a las personas a negarse a cooperar, o a dar respuestas poco fiables para proteger su privacidad. Los métodos tradicionales, como las Encuestas de Hogares, no pueden producir en estas poblaciones ocultas muestras fiables y son ineficientes, porque la mayoría de las poblaciones ocultas son raras.

Hay una gran cantidad de ejemplos de poblaciones ocultas cuyo análisis es de un gran interés dentro de campos tan diferentes como la Biomedicina, la Economía, la Política Pública, la Economía Social, etc. Por ello, se han desarrollado métodos de muestreo especiales para estas poblaciones ocultas, de los que forma parte fundamental el Muestreo Impulsado por los Encuestados (*RDS,* por sus iniciales en inglés: *Respondent-Driven Sampling*). Dicho método fue desarrollado como parte de una intervención de prevención del sida, el proyecto ECHO de Salud Oriental de Connecticut, dirigido a Usuarios Activos de Drogas Inyectables, consistente en entrevistas, educación sobre prevención del sida y pruebas y asesoramiento sobre el VIH (Broadhead y Heckathorn, 1994). En cuanto a poblaciones ocultas en las que se ha realizado RDS dentro de diversas áreas de la Economía y la Política Pública, podemos citar la población de inmigrantes ilegales, trabajadores no declarados, participantes de algunos movimientos sociales ilegales, personas sin hogar, determinados grupos de artistas, o desertores de un ejército o de una institución.

2.2. Muestreos basados en redes sociales

2.2.1. Primeros enfoques

Dado el carácter especial que presentan las poblaciones ocultas, se han ideado métodos de muestreo adaptados a ellas, como pueden ser el muestreo basado en informantes clave (elementos exteriores que se supone conocen el comporta-

miento de esa población oculta) y otros muestreos dirigidos, que presentan un gran número de problemas muy difíciles de resolver y que resultan ser muestreos no probabilísticos, por lo que no pueden ser de ayuda para realizar inferencias probabilísticas acuradas sobre toda la población.

Otro enfoque muy conocido es el muestreo de bola de nieve (Goodman, 1961), en el que se parte de una muestra para fijar unos elementos iniciales ("semilla") que, aunque se supone elegida al azar, en la práctica suele estar determinada por su facilidad de acceso; posteriormente, estos "individuos iniciales" proporcionan los nombres de un número fijo de otras personas que cumplen con los criterios de pertenencia a la población que se analiza. El proceso se continúa tantas etapas como se desee.

Hay una serie de problemas que afectan al muestreo de bola de nieve. En primer lugar, las inferencias finales pueden depender de las semillas elegidas, ya que los individuos adicionales encontrados mediante el rastreo en cadena nunca se encuentran al azar o incluso con sesgos que no son conocidos (el problema no es la existencia del sesgo, sino su desconocimiento; si se conociera la naturaleza y magnitud del sesgo, es posible eliminar su influencia en el resultado final). Tengamos en cuenta que los reclutamientos se producen a través de enlaces de red, por lo que los sujetos con redes personales más grandes serán sobremuestreados. Por tales motivos, las muestras de bolas de nieve generalmente se ven como "muestras de conveniencia" que carecen de cualquier soporte válido para producir muestras imparciales y consistentes. Es por ello que el muestreo del punto siguiente tiene gran interés.

2.2.2. Métodos de muestreo impulsado por los encuestados (respondent-driven sampling, RDS)

El *RDS* es un método de muestreo en red que se usa normalmente para inferir proporciones de la población de rasgos binarios en poblaciones humanas ocultas, difíciles de alcanzar. El *RDS* se ha adoptado ampliamente para estimar la prevalencia de enfermedades o conductas de riesgo dentro de poblaciones humanas de alto riesgo, incluidas las citadas en el punto anterior. A pesar de su amplio uso en temas de importancia sobre la Salud Pública, Economía y Política Social, las propiedades estadísticas y las estrategias inferenciales óptimas para los datos resultantes del *RDS* aún requieren de estudios adicionales.

En *RDS,* varios individuos de la población objetivo se inscriben en el estudio como "semillas", y las muestras posteriores se seleccionan en función de sus conexiones de red con los miembros anteriores de la muestra.

Las redes se utilizan para representar sistemas de individuos interrelacionados. En las redes sociales, las personas (o grupos de personas) están representadas por nodos, y las interrelaciones están representadas por enlaces, que representan el conocimiento mutuo; en la muestra *RDS* cada enlace indica el reclutamiento de un individuo a partir del otro. *RDS* toma su nombre del hecho de que los encuestados son responsables del reclutamiento mediante la distribución de cupones identificando a los miembros de la población que conocen, y a los que luego se pide que se inscriban en la muestra.

El reclutamiento puede ser modelado como un proceso de Markov, una forma de proceso estocástico donde la probabilidad de que el próximo recluta provenga de un grupo determinado depende del grupo del que proviene el reclutador actual y es un proceso sin memoria, lo que significa que las características de un nuevo individuo reclutado (los estados) dependen únicamente de las características de quien lo ha reclutado, pero no de los anteriores individuos que han llevado hasta él.

Además, se supone que el proceso es "regular", lo que significa que cuando el proceso se mueve de un estado a otro, puede alcanzarse cualquier estado, y existe una probabilidad cero de que cualquier estado vuelva a repetirse. En esencia, esto quiere decir que el reclutamiento no puede quedar atrapado dentro de un solo grupo, como ocurriría si una vez que la cadena de reclutamiento ingresara en ese grupo, no pudiera salir de allí.

En este tipo de procesos, para cualquier distribución de probabilidad inicial sobre los estados se obtiene, tras una consecutiva realización del mismo, una distribución final, denominada "de equilibrio". Así que a medida que el proceso de reclutamiento continúa de ola en ola, eventualmente se logrará una mezcla (de equilibrio) de reclutas, independientemente de las características del sujeto o conjunto de sujetos a partir del cual comenzó el reclutamiento.

Por lo tanto, permitir que el reclutamiento opere hasta que se alcance el equilibrio en una muestra correspondiente a un proceso regular de Markov evita el problema central para muestrear poblaciones ocultas, y se consigue que la composición de la muestra sea totalmente independiente de los sujetos iniciales. Es decir, a medida que avanza el muestreo, el efecto del punto de partida se debilita progresivamente hasta que se vuelve insignificante.

Notemos que este tipo de muestreo tiene una gran potencia para acceder a miembros de poblaciones ocultas. Como se muestra en la literatura sobre "la pequeñez del mundo", incluso en una nación tan grande como Estados Unidos, cada persona se asocia indirectamente con cualquier otra persona a través de apro-

ximadamente seis intermediarios (Killworth y Bernard, 1978). Por lo tanto, todos los habitantes del país podrían ser hipotéticamente alcanzados por la sexta ola de una muestra en cadena como se ha indicado arriba.

La implementación de un muestreo *RDS* viene caracterizado por los siguientes puntos:

1. El personal de investigación recluta a un grupo de sujetos que sirven como "semillas".

2. A las semillas se les dan varios cupones de reclutamiento y se les ofrecen incentivos financieros para reclutar a sus pares.

3. A los nuevos individuos reclutados se les ofrecen los mismos incentivos duales que a las semillas. Este mecanismo crea un sistema en expansión de referencias en cadena en el que los sujetos reclutan más sujetos, en oleadas sucesivas. Para evitar que surjan reclutadores semiprofesionales, se limita el número de cupones por encuestado.

4. El rasgo que define la pertenencia a la población debe ser objetivamente verificable, para evitar el reclutamiento fuera de la población en estudio, solo por motivos económicos.

5. Debe evitarse la inclusión de individuos con identidades múltiples, mediante un amplio registro de características físicas de los sujetos.

6. El proceso de selección es concebido como un proceso de Markov de modo que las características de un nuevo individuo reclutado (los estados) dependen únicamente de las características de quien lo ha reclutado, pero no de los anteriores individuos que han llevado hasta él.

7. Se supone que el número de oleadas de reclutamiento es suficientemente grande para que las probabilidades de los estados hayan llegado a la situación de equilibrio (usualmente, este número es bastante modesto).

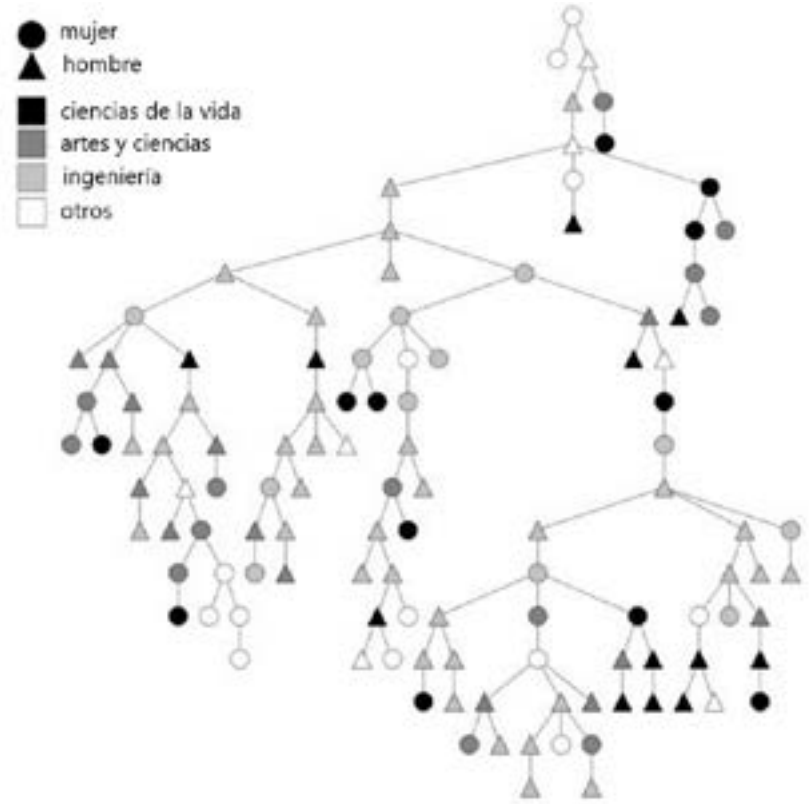
8. El muestreo finaliza cuando la comunidad objetivo está saturada, o cuando se ha alcanzado un tamaño de muestra fijado.

En la figura 1 representamos un ejemplo del grafo de una red muestral publicado por Wejnert (2009) sobre un caso práctico muy simple a partir de una sola semilla, de *Undergraduate Enrollment in Cornell University, 2008* donde las variables que se consideran son el sexo y el tipo de College en el que están inscritos los miembros de una cierta población de estudiantes de dicha Universidad.

Debido a que el proceso de muestreo *RDS* depende de la estructura de la red, la media de la variable, si ésta es binaria (presentar la enfermedad o no, estar parado o no, tener tarjeta sanitaria o no,…) en la muestra así reclutada será en general un estimador sesgado de la media (o proporción) de la variable en la población. En efecto, la probabilidad de ser elegido en la muestra un individuo depende del número de enlaces que tenga dicho individuo en su red social, y por lo tanto, no hay equiprobabilidad de pertenencia a la muestra en todos los individuos reclutados. El número de vínculos sociales de un individuo en la red se denomina grado, y la probabilidad de que un elemento sea elegido para la muestra será proporcional a su grado. Algo análogo sucede si la variable de estudio tiene más de dos categorías.

Figura 1

Acceso a los grados en la Universidad de Cornell, 2008



Fuente: Wejnert, 2009.

3. UN CASO ILUSTRATIVO: *RDS* SOBRE LA POBLACIÓN OCULTA DE PERSONAS SIN HOGAR (PSH)

Citando textualmente la Estrategia Nacional Integral para Personas sin Hogar 2015-2020, aprobada por Acuerdo de Consejo de Ministros de 6 de noviembre de 2015, "las condiciones que sufren las personas sin hogar constituyen probablemente el peor rostro de la exclusión social en nuestro país. No obstante, el "sinhogarismo" es uno de los fenómenos peor conocidos y que ha adolecido de falta de políticas integrales en su intervención". La creación de esta Estrategia "no está dirigida a la mera asistencia o a la supervivencia de las personas bajo mínimos de garantía vital, sino que aspira a que las personas sin hogar restauren su proyecto de vida y se reincorporen a una sociedad que, sin duda, para incluirles, debe cambiar".

Comenzando por la definición de persona sin hogar, en Europa existe un amplio consenso, aunque no oficial, en usar la categorización denominada *ETHOS (European Typology on Homelessness and Housing Exclusion).* Según esta clasificación, hay cuatro tipos generales para identificar las PSH:

A. Personas sin alojamiento (sin techo), constituida por personas que viven en las calles o lugares públicos, o que esporádicamente utilicen alojamientos de emergencia, como albergues nocturnos.

B. Personas sin vivienda (*houseless*), o aquellas que viven en alojamientos temporales o transitorios con apoyo. Aquí se incluyen también las mujeres que utilizan refugios por motivo de haber sufrido violencia de género, los inmigrantes que viven en alojamientos temporales y las personas dependientes de instituciones penitenciarias, sanitarias o tuteladas que carecen de vivienda a donde ir.

C. Personas en viviendas inseguras, esto es, aquellas que viven en alojamientos temporalmente sin derechos legales (como en condiciones de ocupación), personas con requerimiento de abandono de la vivienda, o amenazados por personas convivientes.

D. Personas en viviendas inadecuadas, aquellas que viven en alojamientos móviles, chabolas, cabañas, las que se alojan en viviendas sin permiso de habitabilidad o las que viven en viviendas sobreocupadas.

Las políticas sociales en Europa sobre el colectivo de PSH han experimentado un importante impulso en los últimos años, formando parte de la agenda política del Parlamento Europeo, con el objetivo de ayudar a estas personas a integrarse en su comunidad, formarse, encontrar trabajo y acceder a las prestaciones sociales.

En lo que respecta a España, la Administración General del Estado inició desde hace más de una década una línea de trabajo con las administraciones autonómicas y locales, con el objeto de tener una visión conjunta sobre las políticas dirigidas a las PSH, para lo que se creó un Grupo de Cooperación Técnica entre todas estas administraciones, con la colaboración del Instituto Nacional de Estadística (INE) para la incorporación de sus *Encuestas sobre Personas sin Hogar (EPSH),* que comenzaron a implementarse en el año 2004.

Lo primero que se advierte es la escasez de datos referidos a las PSH: son pocos los estudios existentes y poco generalizables y los datos ofrecidos por el INE sobre las personas atendidas en los centros que ofrecen alojamiento o restauración, son difícilmente extrapolables a toda la población que está sin hogar. Concretamente, en la encuesta realizada por INE en 2012, se detectaron 23.000 PSH, pero se estima que son muchos más por las limitaciones ya señaladas de este tipo de encuestas. De este modo, dentro de esta población en España hay una parte visible y mayoritaria vinculada a centros de alojamiento y restauración y otra parte más reducida y desconocida que pernocta en espacios inadecuados y se encuentra en peores condiciones de vida. El primer grupo puede analizarse en base a las encuestas del INE, y para el segundo existen algunas otras fuentes como son los diversos recuentos nocturnos organizados en distintas ciudades, en los que se intenta localizar a personas que no están pernoctando en centros asistenciales y se realiza una encuesta al conjunto de personas sin hogar de la ciudad (sirva como ejemplo el realizado por Cabrera en Zaragoza en el año 2012). Así pues, se plantea la necesidad de otras metodologías de muestreo (una de las cuales representaría el *RDS,* como se sugiere en este trabajo) para la obtención de datos que permitan analizar las características y tendencias más relevantes que perfilan la situación de todas las PSH en España.

Como se señala en la Estrategia Nacional para PSH arriba citada, "Un enfoque basado en evidencias y en la satisfacción de necesidades y resolución de problemas, exige investigación continua, innovación metodológica y organizativa, desarrollo de nuevas competencias y herramientas y la formación adecuada de los profesionales que tienen que aplicarlas".

El European Observatory on Homelessness (EOH) hizo público a finales de 2014 un documento en que analizaba las tendencias de las PSH en la Unión Europea. Se fijaba fundamentalmente en variables como la edad, ya que las personas jóvenes tenían características propias que cambiaban con el tiempo de manera diferente (en las encuestas del INE entre los años 2005 Y 2012, los jóvenes sin hogar aumentan las pernoctaciones en calle aunque disminuyen como usuarios de centros asistenciales), y asimismo los mayores de 45 años presentaban también sus características especiales (aumento mucho más elevado que el resto, tanto en base a las

encuestas del INE entre 2005 y 2012 como en los recuentos en Madrid entre 2009 y 2014). Por tal motivo, entendemos que sería de gran interés estudiar cómo influye esta variable edad sobre las demás variables socioeconómicas, sanitarias, etcétera. (sexo, ingresos, impagos, desempleos, violencia en el hogar, tiempo transcurrido sin alojamiento propio, posesión de tarjeta sanitaria, consumo de alcohol y drogas, esperanza de vida, agresiones y robos, etc.). Es decir, a partir de un muestreo *RDS,* nos interesa estimar funciones de regresión en la población oculta (PSH), cuando la variable explicativa es la edad del individuo, lo que se estudia en el punto siguiente. Por supuesto, se presenta la metodología para cualquier otra variable exógena, siendo la edad tan solo un caso particular.

4. NUEVOS ESTIMADORES PARA UNA FUNCIÓN DE REGRESIÓN A TRAVÉS DE UN MUESTREO *RDS*

4.1. Estimador de una función de regresión, generalizando el método de Volz y Heckathorn (2008) para una proporción

Supongamos que se realiza un muestreo RDS de una población oculta donde interesa analizar una variable binaria Z_i estimando la media $\mu = \frac{1}{N}\Sigma_{K=1}^{N} z_k$, siendo d_i el grado (número de enlaces) del nodo i-ésimo en la red muestral.

Volz y Heckathorn(2008) propusieron el estimador:

$$\widehat{\mu_{VH}} = \frac{\Sigma_{i=1}^{n} z_i / d_i}{\Sigma_{i=1}^{n} 1 / d_i} \qquad [1]$$

"estimador natural" utilizando el método de Horvitz–Thompson para evitar el sesgo proporcional al grado.

Este estimador lo podemos generalizar al caso de una función de regresión en la siguiente forma:

Partimos ahora de z_i (x), una variable binaria sobre covariable x, y deseamos estimar:

$$m(x) = E[\{Z|x\}] = P\{Z = 1|x] \qquad [2]$$

(por ejemplo, en un *RDS* sobre PSH, x puede ser la edad del individuo y Z la variable dicotómica de estar desempleado o no).

Un modo de proceder consiste en utilizar el estimador lineal local ponderado (WLLE) propuesto por Cristóbal y Alcalá (2000) que evita el sesgo proporcional al grado). Este estimador es $\widehat{m(x)} = \alpha_0$ definido como la solución en α de:

$$min\, \Sigma_i \left\{ z_i - \alpha - \beta \left(x_i - x \right)^2 \right\}^2 K_h \left(x_i - x \right) d_i^{-1} \qquad [3]$$

donde K_h es una función Kernel, *h* el parámetro de suavizado y d_i^{-1} la ponderación extra para evitar el sesgo.

Otro modo de proceder consiste en estimar antes la transformación logit de $m(x)$:

$$g\left(x\right) = log \frac{m\left(x\right)}{1 - m\left(x\right)} \qquad [4]$$

mediante maximización de la log-verosimilitud local lineal ponderada:

$$\begin{aligned} L\left(x\right) &= \Sigma_i log \left\{ \left[m\left(x_i\right) \right]^{z_i} \left[1 - m\left(x_i\right) \right]^{1-z_i} \right\} K_h \left(x_i - x \right) d_i^{-1} = \\ &= \Sigma_i \left\{ z_i g\left(x_i\right) - log \left[1 + e^{g(x_i)} \right] \right\} K_h \left(x_i - x \right) d_i^{-1} \end{aligned} \qquad [5]$$

Si la solución de esta maximización es $\hat{g}(x)$, nuestra solución para estimar m(x) es su transformada logit inversa:

$$\hat{m}\left(x\right) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} \qquad [6]$$

4.2. Estimador de una función de regresión, generalizando el método de *Successive sampling (SS)*, Gile, (2011) para una proporción

El método anterior basado en Volz y Heckathorn supone un muestreo con reemplazamiento, esto es, se permite que un individuo muestral reclute dos veces a un mismo individuo. Sin embargo, el método SS evita la posible inclusión repetida de individuos muestrales; es decir, realiza el muestreo sin reemplazamiento en el vector de los grados **d** = (d_i).

En tal caso, el estimador SS de una proporción pasa a ser:

$$\widehat{\mu_{SS}} = \frac{\Sigma_{i=1}^{n} z_i \,/\, \hat{\pi}_i \left(\boldsymbol{d} \right)}{\Sigma_{i=1}^{n} 1 \,/\, \hat{\pi}_i \left(\boldsymbol{d} \right)} \qquad [7]$$

Es de destacar que, con respecto al estimador VH, este estimador SS evita el sesgo debido a la fracción muestral, cuando esta es grande.

Este estimador lo podemos generalizar al caso de una función de regresión en la siguiente forma:

Si z_i (x) es una variable binaria sobre covariable x, y deseamos estimar m (x) = E [$\{Z|x\}$], podemos emplear el estimador $\widehat{m(x)} = \alpha_0$ que es la solución en α de:

$$min\ \Sigma\left\{z_i - \alpha - \beta\left(x_i - x\right)^2\right\}^2 K_h\left(x_i - x\right)\hat{\pi}_i\left(\boldsymbol{d}\right)^{-1} \qquad [8]$$

donde $\hat{\pi}_i\left(\boldsymbol{d}\right)$ es el estimador del vector de probabilidades sobre los grados y $\hat{\pi}_i\left(\boldsymbol{d}\right)^{-1}$ es la ponderación extra para evitar el sesgo (ver Cristóbal y Alcalá, 2000).

4.3. Estimador de una función de regresión, generalizando el método de Salganik y Heckathorn (2004) para una proporción

Es conocido en la amplia literatura de casos empíricos sobre la aplicación de muestreo *RDS* para estimar una proporción de elementos de una población oculta que tienen una determinada característica (por ejemplo, una determinada enfermedad) que los dos estimadores anteriores VH y *SS* para estimar una proporción presentan un inconveniente de que se comportan mal en presencia de homofilia, o tendencia de los individuos con atributos similares para conectarse entre sí.

Hay otro procedimiento de estimación de una proporción, introducido por Salganik y Heckathorn (2004), que es menos sensible a la homofilia y también se comporta mejor cuando la muestra inicial es fuertemente no representativa de toda la población. Este estimador SH se basa en la información de los enlaces para estimar la media, y considera que cada enlace tiene la misma probabilidad de ser incluido en la muestra. Por lo tanto es un estimador que corrige el sesgo por separado en cada grupo según los individuos tengan la característica analizada o no.

Concretamente, el estimador SH para la proporción de individuos que presentan una característica definida por la variable dicotómica *Z*, viene definido en la siguiente forma:

Sea el total en la población de nodos de tipo $z = 0$ y N_1 el total en la poblaciónde nodos de tipo $z = 1$. Representemos con:

$T01 : \frac{1}{N_0}$ [número de enlaces desde nodos tipo *0* hasta nodos tipo *1*].

y con

$D_0 = \frac{1}{N_0}$ [suma de todos los grados de nodos tipo *z = 0*].

Si $C_{01} = \frac{T_{01}}{D_0}$, entonces:

$$\mu = \frac{1}{N}\sum_{k=1}^{n} z_k = \frac{N_1}{N_0 + N_1} = \frac{T_{01}}{T_{01} + T_{10}} = \frac{D_0 C_{01}}{D_0 C_{01} + D_1 C_{10}} \quad [9]$$

Con lo que el estimador SH viene definido mediante:

$$\hat{\mu}_{SH} = \frac{\hat{D}_0 \hat{C}_{01}}{\hat{D}_0 \hat{C}_{01} + \hat{D}_1 \hat{C}_{10}} \quad [10]$$

donde:

$$\hat{C}_{k,1-k} = \frac{R_{k,1-k}}{R_{k,1-k} + R_{k,k}}, \textit{(k = 0,1)}$$

que es un estimador de $C_{k,1\text{-}k}$ sin corrección de sesgo, siendo $R_{k,1\text{-}k}$ el número de reclutamientos desde nodos de tipo *k* a nodos de tipo *1-k*, y

$$\hat{D}_k = \frac{\sum_{i:z_i=k} d_i / d_i}{\sum_{i:z_i=k} 1 / d_i} = \frac{n_k}{\sum_{i:z_i=k} 1 / d_i}, \textit{(k = 0,1)} \quad [11]$$

es un estimador D_k de que corrige el sesgo que es proporcional al grado, en cada grupo por separado.

Definimos ahora una generalización del estimador SH a la función de regresión. Denominamos E (x, δ) a los elementos de la muestra con la covariable x en *[x-δ, x+δ]* y llamamos $\widehat{C}_{k,1\text{-}k}$ (x,h) = $\widehat{C}_{k,1\text{-}k}$, los calculados según el estimador SH en E (x, δ).

De manera análoga, llamamos $\widehat{D}_0$ (x, h), $\widehat{D}_1$ (x, h), a los estimadores calculados en E (x, δ), según el estimador lineal local (o el estimador de máxima verosimilitud local) ponderado con d_i^{-1}, para D_0 y D_1, definidos en el proceso SH anterior.

Finalmente, estimamos la función de regresión mediante:

$$\hat{\mu}_{SH}(x) = \frac{\hat{D}_0(x,\delta)\hat{C}_{01}(x,\delta)}{\hat{D}_0(x,\delta)\hat{C}_{01}(x,\delta) + \hat{D}_1(x,\delta)\hat{C}_{10}(x,\delta)}, \text{ (δ, parámetro de suavizado)} \quad [12]$$

Este estimador así construido se comporta bien bajo reclutamiento diferencial por el modo en que se ha construido. Adicionalmente, la interpretación de sus dis-

continuidades pueden dar información importante acerca de cómo se comporta la población según distintos grupos disjuntos determinados por la variable explicativa x (por ejemplo, en una población de PSH se pueden encontrar diferentes grupos de edad en los que el desempleo se comporta de manera similar dentro de cada uno).

A modo de conclusión, me gustaría terminar señalando que se ha pretendido difundir una metodología no muy utilizada en las diversas áreas de la Economía pero que resulta imprescindible cuando no es posible realizar un muestreo probabilístico basado en un marco muestral. Dicha metodología está fundamentada en un muestreo de tipo *RDS,* para el que he realizado la propuesta de tres nuevos estimadores de una función de regresión, que sirve para explicar la influencia de ciertas variables sobre diversas características en una población oculta. Se ha utilizado el caso ilustrativo de explicar distintas tendencias de características como probabilidad de impagos, tiempo en desempleo, esperanza de vida, consumo de drogas, etc. para distintas franjas de edad en la población de personas sin hogar.

BIBLIOGRAFÍA

Broadhead, R. S. y Heckathorn, D. (1994). AIDS prevention outreach among injection drug users: Agency problems and new approaches. *Social Problems,* 41, pp. 473-495.

Cabrera, P. (2012). *Estudio personas sin techo.* Zaragoza 2012. Zaragoza: Cruz Roja.

Cristóbal, J. A. y Alcalá, J. T. (2000). Nonparametric regression estimators for length biased data. *Journal of Statistical Planning and Inference,* 89, pp.145-168.

European Observatory on Homelessness. Red de FEANTSA, Federación Europea de Organizaciones Nacionales que Trabajan con Personas Sin Hogar. http://www.feantsaresearch.org

Gile, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association,* 106, pp. 135-146.

Goodman, L. A. (1961). Snowball Sampling. *Annals of Mathematical Statistics,* 32, pp. 148–170.

Informe del Ministerio de Sanidad, Servicios Sociales e Igualdad (2016). Estrategia Nacional Integral para Personas sin Hogar 2015-2020. http://publicacionesoficiales.boe.es.

Killworth, P D. y Bernard, H. R. (1978). The reversal small world experiment. *Social Networks,* 1, pp. 159-192.

Ott, Q. O., Gile, K. J., Harrison, M. T., Johnston, L. G. y Hogan, J. W. (2019). Reduced bias for respondent-driven sampling accounting for non-uniform edge sampling probabilities in people who inject drugs in Mauritius. *Applied Statistics, Series C,* 68 (5), pp. 1411-1429.

Salganik, M. J. y Heckathorn, D. D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology,* 34, pp. 193-239

Thompson, S. K. y Frank, O. (2000).Model-Based Estimation with Linktracing Sampling Designs. *Survey Methodology,* 26(1), pp. 87–98.

Volz, E. y Heckathorn, D. D. (2008). Probability Based Estimation Theory for Respondent Driven Sampling. *Journal of Official Statistics,* 24, pp.79–97.

Wejnert, C. (2009). An empirical test of respondent-driven sampling: point estimates, variance, degree measures, and out-of-equilibrium data. *Sociological Methodology,* 39(1), pp. 73-116.