Inferencia para el índice de Gini bajo imputación múltiple: varianza e intervalos de confianza con bootstrap

Pablo José Moya Fernández[1]; Juan Francisco Muñoz Rosas[2]; Raúl Amor Pulido[3]

[1]Facultad de Ciencias Económicas y Empresariales/Métodos Cuantitativos para la Economía y la Empresa/Universidad de Granada. Correo-e: pjmoyafernandez@ugr.es

[2]Facultad de Ciencias Económicas y Empresariales/Métodos Cuantitativos para la Economía y la Empresa/Universidad de Granada. Correo-e: jfmunoz@ugr.es

[3]Facultad de Ciencias Económicas y Empresariales/Métodos Cuantitativos para la Economía y la Empresa/Universidad de Granada. Correo-e: ramor@ugr.es

Resumen

La desigualdad económica es un fenómeno con implicaciones relevantes para el análisis político, en la medida en que condiciona la distribución de recursos, oportunidades y resultados sociales. Su estudio empírico requiere medidas fiables y procedimientos de inferencia que permitan cuantificar adecuadamente la incertidumbre asociada a su estimación. En este contexto, el índice de Gini es el indicador más utilizado, pero su estimación a partir de encuestas de ingresos plantea desafíos metodológicos derivados de la presencia de datos faltantes y de la naturaleza asimétrica de las distribuciones de ingreso.

Este trabajo presenta una contribución metodológica orientada a la inferencia del índice de Gini bajo el supuesto de datos faltantes al azar (MAR), utilizando imputación múltiple. En particular, se aborda la estimación de la varianza del estimador y la construcción de intervalos de confianza, comparando dos estrategias para combinar imputación múltiple y remuestreo: (i) bootstrap seguido de imputación múltiple y (ii) imputación múltiple seguido de bootstrap.

Palabras clave: Desigualdad, datos faltantes, distribuciones asimétricas

1. Introducción

La medición de la desigualdad constituye una cuestión central en el análisis económico y social y tiene implicaciones directas para el diagnóstico, el diseño y la evaluación de políticas públicas. Entre los indicadores disponibles, el índice de Gini continúa siendo una de las medidas de desigualdad más utilizadas. Su popularidad se explica, entre otros motivos, por su interpretación sencilla y por la posibilidad de comparar distribuciones expresadas en escalas diferentes (Gini, 1912; Giorgi y Gigliarano, 2017). Sin embargo, la utilidad de cualquier indicador de desigualdad depende de que su estimación sea suficientemente fiable. En aplicaciones basadas en encuestas, esta fiabilidad puede verse afectada por el sesgo muestral del propio estimador, por la forma de la distribución de ingresos y, de manera particularmente importante, por la presencia de datos faltantes.

La falta de respuesta es habitual en las encuestas y puede ser especialmente relevante cuando se solicita información sensible, como los ingresos o la riqueza. Es conveniente distinguir entre falta de respuesta de unidad, cuando no se obtiene información de una unidad seleccionada, y falta de respuesta de ítem, cuando el individuo participa en la encuesta pero no responde a una o varias preguntas. Esta segunda situación es la que interesa en el presente trabajo. Su tratamiento suele realizarse mediante procedimientos de imputación, cuyo objetivo es aprovechar la información observada y evitar que el análisis dependa exclusivamente de los casos completos (Little y Rubin, 2019; Van Buuren, 2018).

Las consecuencias de los datos faltantes no se limitan a una reducción del tamaño muestral. Si quienes responden y quienes no responden difieren sistemáticamente en la variable de interés, puede aparecer sesgo por falta de respuesta. Además, la pérdida de información puede incrementar la variabilidad de los estimadores y reducir la precisión de la inferencia. A ello se añade un problema especialmente relevante para esta investigación: un tratamiento inadecuado de los valores faltantes puede producir estimaciones incorrectas de la varianza y, en consecuencia, intervalos de confianza con coberturas alejadas de su nivel nominal.

El efecto de la falta de respuesta depende del mecanismo que la genera. Rubin (1976) distingue entre datos faltantes completamente al azar (MCAR), datos faltantes al azar (MAR) y datos faltantes no al azar (MNAR). En el mecanismo MAR, que constituye el marco de esta investigación, la probabilidad de que falte un valor puede depender de información observada, pero no del valor no observado una vez se condiciona por dicha información. Esta hipótesis permite utilizar variables auxiliares observadas para modelizar el proceso de imputación y constituye uno de los supuestos de referencia para la aplicación de la imputación múltiple.

Álvarez-Verdejo et al. (2021) estudiaron específicamente el efecto de los datos faltantes sobre la estimación puntual y por intervalos del índice de Gini mediante un estudio de simulación Monte Carlo que contemplaba numerosos escenarios. Su análisis comparó diferentes métodos de imputación simple y distintos procedimientos para construir intervalos de confianza bajo varios

mecanismos de falta de respuesta, proporciones de datos faltantes, niveles de desigualdad y grados de asociación entre la variable de interés y una variable auxiliar. Los resultados mostraron que la falta de respuesta puede generar sesgos relevantes y deteriorar sustancialmente la cobertura de los intervalos de confianza. También pusieron de manifiesto que el uso de información auxiliar puede mejorar el comportamiento de la estimación cuando la relación entre la variable auxiliar y la variable de interés es elevada.

La imputación simple, no obstante, presenta una limitación importante para la inferencia. Cada valor faltante se sustituye por una única realización y, si posteriormente se analiza el conjunto completado como si todos sus valores hubieran sido observados, no se representa adecuadamente la incertidumbre asociada al proceso de imputación. La imputación múltiple fue desarrollada precisamente para incorporar esa fuente adicional de incertidumbre. En lugar de generar un único conjunto completado, se construyen varios conjuntos de datos plausibles, cada uno de los cuales se analiza por separado. Las estimaciones se combinan posteriormente mediante las reglas de Rubin, que distinguen entre la variabilidad dentro de las imputaciones y la variabilidad entre imputaciones (Rubin, 1976; Little y Rubin, 2019).

La extensión desde imputación simple a imputación múltiple constituye, por tanto, una continuación natural del trabajo de Álvarez-Verdejo et al. (2021). Sin embargo, el problema inferencial no queda completamente resuelto por el mero uso de imputación múltiple. Las distribuciones de ingresos suelen ser marcadamente asimétricas y el índice de Gini es un estadístico no lineal. En este contexto, los procedimientos de remuestreo ofrecen una vía flexible para aproximar la distribución muestral de un estimador sin imponer una forma paramétrica específica. El bootstrap constituye una de las herramientas más utilizadas con esta finalidad (Efron y Tibshirani, 1993) y se ha aplicado también a la inferencia del índice de Gini (Davidson, 2009; Giorgi y Gigliarano, 2017).

Cuando existen simultáneamente datos faltantes y una distribución muestral que se desea aproximar mediante bootstrap, surge una decisión metodológica adicional: el orden en el que deben combinarse la imputación múltiple y el remuestreo. Brand et al. (2019), en un contexto de datos de coste-efectividad caracterizados también por asimetría y falta de respuesta, muestran que ambas técnicas pueden anidarse en órdenes diferentes y que su combinación no garantiza automáticamente una inferencia válida. En particular, distinguen entre procedimientos en los que primero se generan muestras bootstrap y dentro de cada una se realiza la imputación, y procedimientos en los que primero se generan los conjuntos imputados y posteriormente se aplica el bootstrap a cada uno de ellos. El orden de anidamiento afecta a la forma en la que se representa la incertidumbre procedente del muestreo y de la imputación.

El objetivo de este trabajo es trasladar este problema al estudio de la desigualdad y desarrollar la inferencia del índice de Gini en presencia de datos faltantes bajo un mecanismo MAR. En concreto, se comparan dos estrategias para combinar imputación múltiple y bootstrap: (i) bootstrap seguido de imputación múltiple y (ii) imputación múltiple seguida de bootstrap. El

interés se centra en la estimación de la varianza del estimador y en la construcción de intervalos de confianza mediante bootstrap percentil, con el propósito de evaluar qué procedimiento representa de forma más adecuada la incertidumbre total asociada a la estimación del índice de Gini.

La investigación se encuentra actualmente en una fase metodológica. La comparación de ambas estrategias se realizará mediante simulación Monte Carlo, manteniendo una estructura próxima a la empleada por Álvarez-Verdejo et al. (2021), pero centrada en el mecanismo MAR, trabajando con poblaciones teóricas infinitas y sustituyendo la imputación simple por imputación múltiple. La evaluación prestará especial atención a la estimación de la varianza, la cobertura empírica y la amplitud de los intervalos de confianza. En consecuencia, esta contribución presenta el planteamiento metodológico y el diseño de evaluación que servirán de base para el análisis posterior de los resultados de simulación.

2. Marco metodológico

2.1. Índice de Gini y estimador puntual

Sea Y una variable aleatoria continua no negativa, con función de distribución $F_Y(y)$ y función de densidad $f(y)$. El índice de Gini se define como

$$G = \frac{1}{2\mu_Y} \int_0^{+\infty} \int_0^{+\infty} |x - y| \; dF_Y(x)\, dF_Y(y).$$

donde la media de Y viene dada por

$$\mu_Y = E[Y] = \int_0^{+\infty} y\, f(y)\, dy = \int_0^{+\infty} y \; dF_Y(y).$$

El índice de Gini toma valores comprendidos entre 0 y 1. El valor G = 0 representa una situación de igualdad perfecta, mientras que valores progresivamente mayores indican una mayor concentración de los ingresos y, por tanto, un mayor nivel de desigualdad. El valor 1 constituye el límite superior de la medida. Esta escala facilita la interpretación y la comparación entre distribuciones, aunque el índice presenta limitaciones conocidas, entre ellas una sensibilidad diferente a cambios producidos en distintas zonas de la distribución (Giorgi y Gigliarano, 2017).

Para una población infinita, sea S el conjunto de índices de una muestra de tamaño n. Denotamos por y_i, $i \in S$, las observaciones de variables aleatorias no negativas con la misma distribución que Y, y por $y_{(i)}$ sus valores ordenados de forma no decreciente. La media muestral se define como

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

A partir de estas cantidades, un estimador habitual del índice de Gini es

$$\hat{G} = \frac{2}{\overline{y}n^2} \sum_{i \in S} i\, y_{(i)} - \frac{n+1}{n}.$$

Este estimador puede presentar sesgo en muestras pequeñas (Deltas, 2003). Por ello, se emplea su versión corregida por sesgo:

$$\hat{G}^{bc} = \frac{2}{\overline{y}n(n-1)} \sum_{i \in S} i\, y_{(i)} - \frac{n+1}{n-1}.$$

La relación entre ambos estimadores es

$$\hat{G}^{bc} = \frac{n}{n-1} \hat{G}.$$

En consecuencia, a lo largo del trabajo la estimación del índice de Gini se realizará con la versión corregida por sesgo, $\hat{G}^{bc}$. La corrección reduce el sesgo muestral, aunque no garantiza su eliminación completa para cualquier distribución y tamaño muestral (Deltas, 2003; Davidson, 2009; Álvarez-Verdejo et al., 2021).

2.2. Datos faltantes y mecanismo MAR

Los datos faltantes constituyen un problema habitual en las encuestas y pueden afectar tanto a la estimación puntual como a la inferencia estadística. Cuando la ausencia de información no se produce de forma completamente aleatoria, los individuos con información observada pueden diferir sistemáticamente de aquellos para los que determinados valores no están disponibles, generando sesgo por falta de respuesta. Además, la pérdida de información puede incrementar la variabilidad de los estimadores y reducir su precisión. Estas cuestiones son especialmente relevantes cuando la variable de interés es el ingreso, dado que se trata de una variable sensible y, al mismo tiempo, fundamental para la medición de la desigualdad.

En este trabajo se considera un problema de no respuesta parcial o item non-response: los valores faltantes aparecen únicamente en la variable de interés Y, mientras que una variable auxiliar X, relacionada con Y, se encuentra completamente observada. Para cada observación i, sea R_i el indicador de respuesta, definido como

$$R_i = \begin{cases} 1, & si\ Y_i\ está\ observado \\ 0, & si\ Y_i\ está\ ausente \end{cases}$$

La naturaleza del proceso que determina qué valores son observados se describe mediante el mecanismo de datos faltantes. Siguiendo la clasificación de Rubin (1976), un mecanismo es *Missing At Random* (MAR) cuando, condicionando por la información observada, la probabilidad de ausencia no depende de los valores no observados. En el contexto considerado en este trabajo,

donde X está completamente observada y los valores faltantes se producen únicamente en Y, el supuesto MAR puede expresarse como

$$P(R_{\mathrm{i}} = 1 \mid Y_{\mathrm{i}}, X_{\mathrm{i}}) \;=\; P(R_{\mathrm{i}} \;=\; 1 \mid X_{\mathrm{i}})$$

Por tanto, una vez condicionamos por X_{i}, la probabilidad de que Y_{i} sea observado no depende adicionalmente de su valor no observado. Conviene señalar que MAR no implica que la falta de respuesta sea completamente aleatoria. A diferencia del mecanismo MCAR, permite que la probabilidad de respuesta varíe sistemáticamente en función de características observadas. Lo que excluye es una dependencia adicional respecto a los valores faltantes una vez considerada la información observada relevante (Rubin, 1976; Little y Rubin, 2019).

La disponibilidad de información auxiliar relacionada con la variable incompleta desempeña, por tanto, un papel central bajo MAR. Incorporar X al modelo de imputación permite utilizar las diferencias observadas entre unidades para generar valores plausibles de Y y preservar, en la medida en que el modelo de imputación esté adecuadamente especificado, la relación existente entre ambas variables. Esta es la base sobre la que se introduce la imputación múltiple en la sección siguiente.

2.3. Imputación múltiple mediante regresión con perturbación aleatoria

La imputación constituye una estrategia habitual para tratar problemas de no respuesta parcial, sustituyendo los valores faltantes por valores obtenidos a partir de la información observada. En la imputación simple, cada valor faltante se sustituye una única vez y se obtiene, por tanto, un solo conjunto de datos completado. Aunque este procedimiento permite recuperar una base de datos completa, no representa directamente la incertidumbre asociada al hecho de que los valores imputados son desconocidos y han debido ser estimados.

La imputación múltiple extiende este planteamiento sustituyendo cada valor faltante por M valores plausibles, lo que da lugar a M conjuntos de datos completados. Estos conjuntos no deben ser idénticos: las diferencias entre ellos pretenden representar la incertidumbre asociada a los valores no observados. Cada conjunto se analiza por separado y las estimaciones resultantes se combinan posteriormente, teniendo en cuenta tanto la variabilidad dentro de cada conjunto completado como la variabilidad existente entre las distintas imputaciones (Rubin, 1976; Van Buuren, 2018).

En este trabajo se emplea imputación múltiple mediante ecuaciones encadenadas (MICE), utilizando un modelo de regresión lineal con perturbación aleatoria. Dado que Y es la única variable incompleta y X está completamente observada, el modelo de imputación se basa en la distribución condicional de Y dada X. De forma esquemática, para una observación i con Y_{i} faltante, el valor generado en la imputación m puede representarse como

$$Y_i^{(m)} = \beta_0^{(m)} + \beta_1^{(m)} X_i + \varepsilon_i^{(m)}, \quad m = 1, \ldots, M.$$

La perturbación aleatoria evita que los valores imputados coincidan simplemente con las predicciones puntuales de una regresión. Además, el procedimiento de imputación múltiple incorpora incertidumbre sobre los parámetros del modelo de imputación, de modo que las distintas bases completadas representan realizaciones plausibles de los valores faltantes. Esta variación entre imputaciones es esencial para reflejar la incertidumbre adicional introducida por la presencia de datos no observados (Van Buuren, 2018).

En cada conjunto completado se calcula la versión corregida por sesgo del índice de Gini definida en la sección 2.1. Para simplificar la notación, en las expresiones siguientes se utiliza $\hat{G}_m$ para denotar la estimación obtenida en el conjunto completado m. La estimación puntual combinada mediante imputación múltiple viene dada por

$$\overline{G}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{G}_m.$$

La incertidumbre derivada de los valores faltantes queda reflejada en la variabilidad de las estimaciones obtenidas en los diferentes conjuntos completados. La varianza entre imputaciones se estima mediante

$$B_M = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{G}_m - \overline{G}_M \right)^2.$$

En la formulación general de la imputación múltiple, si U_m representa una estimación de la varianza de $\hat{G}_m$ condicionada al conjunto completado m, la varianza media dentro de las imputaciones es

$$\overline{U}_M = \frac{1}{M} \sum_{m=1}^{M} U_m,$$

y la varianza total de acuerdo con las reglas de Rubin viene dada por

$$T_M = \overline{U}_M + \left(1 + \frac{1}{M}\right) B_M.$$

De este modo, la inferencia mediante imputación múltiple combina la incertidumbre existente dentro de cada conjunto completado con la variabilidad adicional generada por los valores faltantes. En el caso del índice de Gini, sin embargo, la obtención de U_m no es trivial, especialmente debido al carácter no lineal del estimador y a la asimetría habitual de las distribuciones de ingresos. Esta cuestión motiva la utilización del bootstrap y conduce al problema central del trabajo: analizar cómo cambia la inferencia según el orden en que se combinan bootstrap e imputación múltiple.

3. Inferencia para el índice de Gini bajo imputación múltiple

La estimación de la incertidumbre debe recoger dos fuentes distintas de variabilidad. La primera es la variabilidad muestral: aun si todos los datos estuvieran observados, diferentes muestras procedentes de la misma población producirían diferentes valores de $\hat{G}$. La segunda es la incertidumbre debida a los datos faltantes: para una misma muestra incompleta pueden generarse distintos valores plausibles en el proceso de imputación. La combinación de bootstrap e imputación múltiple pretende representar conjuntamente ambas fuentes de incertidumbre.

Brand et al. (2019) distinguen entre dos órdenes de anidamiento. En su terminología, BM corresponde al bootstrap como bucle exterior y a la imputación como bucle interior, mientras que MB sitúa la imputación en el bucle exterior y el bootstrap dentro de cada conjunto completado. En este trabajo se trasladan ambos esquemas al índice de Gini y se utiliza el intervalo bootstrap percentil para la inferencia por intervalos. En ambos procedimientos, el estimador del Gini utilizado en cada conjunto completo es la versión corregida por sesgo definida en la sección 2.1. Para aligerar la notación, en las expresiones de esta sección se mantiene $\hat{G}$ sin el superíndice bc. La estimación puntual de referencia es la obtenida aplicando M imputaciones a la muestra incompleta original; la diferencia entre las estrategias se concentra en la forma de estimar la varianza y construir el intervalo de confianza.

3.1. Bootstrap seguido de imputación múltiple (B-IM)

En el primer procedimiento, el bootstrap constituye el bucle exterior y la imputación múltiple el bucle interior. A partir de la muestra incompleta original se generan B muestras bootstrap mediante muestreo aleatorio simple con reemplazamiento de las n unidades. El remuestreo se realiza antes de la imputación, por lo que cada muestra bootstrap conserva conjuntamente la variable auxiliar, el indicador de respuesta y los valores observados o faltantes correspondientes a las unidades seleccionadas.

Dentro de cada muestra bootstrap b, $b = 1, \ldots, B$, se realiza posteriormente la imputación múltiple, generándose M conjuntos de datos completados. Sea $\hat{G}_{bm}^{\text{boot}}$ la estimación del índice de Gini obtenida en la imputación m de la muestra bootstrap b. Las M estimaciones correspondientes a esa muestra se combinan mediante

$$\overline{G}_b^{\text{boot}} = \frac{1}{M} \sum_{m=1}^{M} \hat{G}_{bm}^{\text{boot}}.$$

Por tanto, cada muestra bootstrap produce una única estimación combinada, $\overline{G}_b^{\text{boot}}$, que incorpora el proceso de imputación múltiple realizado dentro de ella. Repitiendo el procedimiento para las B muestras bootstrap se obtiene la distribución empírica formada por $\overline{G}_1^{\text{boot}}, \ldots, \overline{G}_B^{\text{boot}}$. Su media es

$$\overline{G}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \overline{G}_b^{\text{boot}},$$

y la varianza del estimador combinado se estima como

$$\hat{V}_{\text{B-IM}} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\overline{G}_b^{\text{boot}} - \overline{G}_{\text{boot}} \right)^2.$$

Para construir el intervalo percentil de nivel $1-\alpha$, sean $q_{\alpha/2}$ y $q_{1-\alpha/2}$ los cuantiles correspondientes de la distribución empírica de $\overline{G}_b^{\text{boot}}$. El intervalo se define como

$$IC_{\text{B-IM}} = \left[q_{\alpha/2} \left(\overline{G}_b^{\text{boot}} \right),\ q_{1-\alpha/2} \left(\overline{G}_b^{\text{boot}} \right) \right].$$

Este esquema corresponde al procedimiento BM_p considerado por Brand et al. (2019): el bootstrap se aplica a los datos todavía incompletos y, dentro de cada muestra bootstrap, se lleva a cabo un nuevo proceso de imputación múltiple. De este modo, la distribución utilizada para la inferencia se obtiene a partir de B estimaciones combinadas del Gini, cada una de las cuales resume las M imputaciones generadas dentro de la correspondiente muestra bootstrap.

3.2. Imputación múltiple seguida de bootstrap (IM-B)

En el segundo procedimiento se invierte el orden: la imputación múltiple constituye el bucle exterior y el bootstrap el bucle interior. En primer lugar, a partir de la muestra incompleta original se generan M conjuntos de datos completados mediante imputación múltiple. En cada conjunto completado m, $m = 1, \ldots, M$, se calcula la estimación $\hat{G}_m$. A continuación, dentro de

cada uno de estos conjuntos se generan B muestras bootstrap y se obtienen las estimaciones $\hat{G}_{mb}^{\text{boot}}$, con $b = 1, \dots, B$.

Para cada imputación m, la media de las estimaciones bootstrap es

$$\overline{G}_m^{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} \hat{G}_{mb}^{\text{boot}},$$

y la varianza bootstrap dentro de esa imputación se estima mediante

$$U_m^{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B} \left(\hat{G}_{mb}^{\text{boot}} - \overline{G}_m^{\text{boot}}\right)^2.$$

La media de estas varianzas proporciona el componente de variabilidad dentro de las imputaciones,

$$\overline{U}_{\text{boot}} = \frac{1}{M}\sum_{m=1}^{M} U_m^{\text{boot}}.$$

La variabilidad entre imputaciones se estima, como antes, a partir de las estimaciones obtenidas en los conjuntos completados originales,

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M} \left(\hat{G}_m - \overline{G}_M\right)^2.$$

En esta estrategia se propone combinar ambas fuentes de variación mediante la regla habitual de imputación múltiple, utilizando la varianza bootstrap como estimación de la varianza dentro de cada imputación:

$$\hat{V}_{\text{IM-B}} = \overline{U}_{\text{boot}} + \left(1 + \frac{1}{M}\right) B_M.$$

Esta expresión constituye la adaptación al índice de Gini del principio de combinación de Rubin cuando la variabilidad dentro de cada conjunto completado se estima mediante bootstrap; no debe interpretarse como una fórmula específica propuesta por Brand et al. (2019).

Para construir el intervalo percentil, en cada imputación m se calculan los cuantiles $q_{m,\alpha/2}$ y $q_{m,1-\alpha/2}$ de sus B estimaciones bootstrap. Siguiendo el esquema MB_p analizado por Brand et al. (2019), los límites de los intervalos obtenidos en las distintas imputaciones se combinan mediante su promedio:

$$L_{\text{IM-B}} = \frac{1}{M}\sum_{m=1}^{M} q_{m,\alpha/2},$$

$$U_{\mathrm{IM-B}} = \frac{1}{M}\sum_{m=1}^{M} q_{m,1-\alpha/2},$$

de modo que

$$IC_{\mathrm{IM-B}} = [L_{\mathrm{IM-B}},\ U_{\mathrm{IM-B}}].$$

Este procedimiento corresponde al bootstrap anidado dentro de la imputación múltiple. Brand et al. (2019) muestran que, en su aplicación, el comportamiento de este esquema percentil difiere del obtenido cuando el orden se invierte. Precisamente por ello, no se presupone que los dos procedimientos presenten propiedades equivalentes para el índice de Gini: su comportamiento debe evaluarse de forma específica mediante simulación.

4. Diseño de evaluación mediante simulación Monte Carlo

Para evaluar las propiedades inferenciales de los procedimientos B-IM e IM-B se plantea un experimento de simulación Monte Carlo. El diseño mantiene, en la medida de lo posible, la lógica empleada por Álvarez-Verdejo et al. (2021), aunque restringe el análisis al mecanismo MAR, considera una población teórica infinita y sustituye la imputación simple por los dos esquemas de imputación múltiple y bootstrap descritos anteriormente. Esta continuidad permite comparar el nuevo planteamiento con el análisis previo manteniendo una estructura de simulación similar.

En cada una de las $R = 1000$ repeticiones Monte Carlo se genera una muestra aleatoria independiente de tamaño $n = 100$ procedente de una distribución lognormal. Siguiendo la parametrización del estudio previo, se fija el parámetro de localización de la distribución lognormal en $\mu = 5$ y se consideran dos valores del parámetro de dispersión, $\sigma \in \{0.36, 1.19\}$. Para una variable lognormal, el índice de Gini depende únicamente de σ y viene dado por

$$G = 2\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1,$$

donde Φ es la función de distribución de una normal estándar. Los valores anteriores generan aproximadamente $G = 0.2$ y $G = 0.6$, respectivamente, por lo que permiten analizar escenarios de desigualdad baja y elevada.

La variable auxiliar se genera siguiendo Álvarez-Verdejo et al. (2021) mediante

$$X = Y + e,$$

donde e es una perturbación normal independiente cuya desviación estándar se selecciona para que la correlación entre Y y X sea $\rho \in \{0.5, 0.95\}$. Se analizan así situaciones en las que la

información auxiliar presenta una asociación moderada con los ingresos y situaciones en las que su capacidad predictiva es muy elevada.

En cada muestra se generan valores faltantes únicamente en Y bajo un mecanismo MAR. La variable auxiliar X permanece completamente observada y determina la información disponible para modelizar tanto el mecanismo de falta de respuesta como el proceso posterior de imputación.

$$p \in \{0.1,0.2,0.3,0.4\}.$$

La combinación de dos niveles de desigualdad, dos niveles de correlación y cuatro proporciones de falta de respuesta genera un total de 16 escenarios MAR.

En cada escenario y repetición Monte Carlo se aplicarán los dos procedimientos descritos en la sección anterior. La imputación múltiple se realizará con $M = 5$ conjuntos completados mediante regresión con perturbación aleatoria. En cada procedimiento se utilizarán $B = 1000$ muestras bootstrap y el índice de Gini se calculará mediante la versión corregida por sesgo definida en la sección 2.1. Los dos procedimientos se evaluarán, por tanto, con los mismos valores de M y B.

La comparación se centrará principalmente en la estimación de los intervalos de confianza, los indicadores principales serán la tasa de cobertura empírica y la amplitud media. Si L_r y U_r representan los límites inferior y superior del intervalo obtenido en la réplica r, la tasa de cobertura es

$$CR = \frac{1}{R}\sum_{r=1}^{R} I\,(L_r \leq G \leq U_r),$$

y la amplitud media viene dada por

$$W = \frac{1}{R}\sum_{r=1}^{R}(U_r - L_r).$$

Los intervalos se construirán con un nivel nominal del 95%. Por tanto, un procedimiento adecuado debería proporcionar valores de cobertura próximos a 0.95 sin generar intervalos innecesariamente amplios. Además, dado que el orden de anidamiento modifica de forma sustancial el número de procesos de imputación que deben ejecutarse, se registrará también el coste computacional como criterio complementario de comparación, aspecto que Brand et al. (2019) muestran que puede ser relevante en la práctica.

Como medidas complementarias para evaluar la estimación puntual obtenida mediante imputación múltiple se conservarán el sesgo relativo y el error cuadrático medio relativo utilizados por Álvarez-Verdejo et al. (2021):

$$RB = \frac{E(\hat{G}) - G}{G},$$

$$RRMSE = \frac{\sqrt{MSE(\hat{G})}}{G},$$

donde las correspondientes aproximaciones Monte Carlo se obtienen mediante

$$E(\hat{G}) = \frac{1}{R}\sum_{r=1}^{R} \hat{G}^{(r)},$$

$$MSE(\hat{G}) = \frac{1}{R}\sum_{r=1}^{R} (\hat{G}^{(r)} - G)^2.$$

Dado que la estimación puntual de referencia es la misma en los dos órdenes de anidamiento, estas medidas se utilizarán como control del comportamiento del estimador bajo imputación múltiple y no como criterio principal para distinguir entre B-IM e IM-B.

5. Conclusiones y líneas de trabajo

Este trabajo plantea una extensión metodológica del análisis desarrollado por Álvarez-Verdejo et al. (2021), sustituyendo la imputación simple por imputación múltiple para estudiar la inferencia del índice de Gini en presencia de datos faltantes bajo un mecanismo MAR. La contribución se centra en un aspecto concreto: cómo combinar la incertidumbre derivada del muestreo con la incertidumbre introducida por el proceso de imputación cuando se emplea bootstrap para estimar la varianza y construir intervalos de confianza.

Siguiendo la distinción analizada por Brand et al. (2019), se consideran dos estrategias que difieren en el orden de anidamiento: bootstrap seguido de imputación múltiple (B-IM) e imputación múltiple seguida de bootstrap (IM-B). Aunque ambas utilizan los mismos componentes, el orden en que se aplican da lugar a procedimientos inferenciales diferentes. En B-IM, el remuestreo se realiza sobre los datos incompletos y cada muestra bootstrap incorpora un nuevo proceso de imputación; en IM-B, el bootstrap se aplica dentro de los conjuntos previamente completados y las distintas fuentes de variabilidad se combinan posteriormente. El experimento Monte Carlo permitirá evaluar si estas diferencias se traducen en comportamientos distintos en la estimación de la varianza y en la cobertura y amplitud de los intervalos de confianza para el índice de Gini.

La siguiente fase de la investigación consiste en completar el estudio Monte Carlo descrito en la sección anterior y analizar comparativamente ambos procedimientos bajo los distintos escenarios considerados. Al encontrarse las simulaciones actualmente en ejecución, esta

contribución no presenta resultados empíricos ni anticipa la superioridad de ninguna de las dos estrategias. Su objetivo en esta etapa es establecer el marco metodológico y el diseño de evaluación que permitirán estudiar sus propiedades inferenciales de forma sistemática.

La utilización de una única especificación de imputación múltiple permite concentrar el análisis en el objeto principal del estudio: el efecto del orden de combinación entre imputación múltiple y bootstrap sobre la inferencia del índice de Gini. Una vez evaluado este aspecto, el análisis podrá extenderse a otros métodos de imputación múltiple, como *predictive mean matching*, empleado por Brand et al. (2019), y estudiar la sensibilidad de los resultados al número de imputaciones y a otras variantes de intervalos bootstrap. Otras extensiones incluyen la consideración de diferentes tamaños muestrales y el análisis de mecanismos de falta de respuesta distintos de MAR.

Referencias

Álvarez-Verdejo, E., Moya-Fernández, P. J. y Muñoz-Rosas, J. F. (2021). Single Imputation Methods and Confidence Intervals for the Gini Index. *Mathematics*, 9, 3252. https://doi.org/10.3390/math9243252

Brand, J., van Buuren, S., le Cessie, S. y van den Hout, W. B. (2019). Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. *Statistics in Medicine*, 38, 210-220. https://doi.org/10.1002/sim.7956

Davidson, R. (2009). Reliable inference for the Gini index. *Journal of Econometrics*, 150, 30-40.

Deltas, G. (2003). The small-sample bias of the Gini coefficient: Results and implications for empirical research. *Review of Economics and Statistics*, 85(1), 226-234.

Efron, B. y Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Gini, C. (1912). Variabilità e mutabilità. En E. Pizetti (Ed.), *Memorie di metodologica statistica*. Libreria Eredi Virgilio Veschi.

Giorgi, G. M. y Gigliarano, C. (2017). The Gini concentration index: A review of the inference literature. *Journal of Economic Surveys*, 31, 1130-1148.

Little, R. J. y Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3.ª ed.). John Wiley & Sons.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC Press.