



MEJORANDO LAS ESTIMACIONES DE DESIGUALDAD DE INGRESOS: EL PAPEL DE LA IMPUTACIÓN MÚLTIPLE EN LA PRECISIÓN DEL ÍNDICE DE GINI

Dpto. Métodos Cuantitativos para la Economía y la Empresa.

Universidad de Granada

AMOR-PULIDO, RAÚL
MOYA-FERNÁNDEZ, PABLO J.

RESUMEN

La desigualdad de ingresos es un tema de gran interés tanto para el público general como para la comunidad científica. Una de las medidas más utilizadas para cuantificarla es el índice de Gini, ampliamente empleado por organismos oficiales de estadística, como EUROSTAT o el INE, en la elaboración de informes y estadísticas sobre ingresos y condiciones de vida. Dado que el índice de Gini es una medida poblacional generalmente desconocida, su estimación se basa en los datos recogidos a través de encuestas.

Sin embargo, las encuestas presentan con frecuencia el problema de datos faltantes, los cuales pueden deberse a omisiones involuntarias (por parte del encuestado o del investigador) o a negativas deliberadas a proporcionar cierta información. La presencia de datos incompletos es, por tanto, una característica habitual en este tipo de estudios. Para abordar esta limitación, es común recurrir a métodos de imputación, siendo la imputación única una de las estrategias más empleadas.

En este trabajo se analiza el comportamiento empírico de diversos métodos tradicionales de imputación única aplicados a la estimación del índice de Gini. Los resultados evidencian que este tipo de imputación tiende a subestimar la varianza, lo que conduce a intervalos de confianza artificialmente estrechos y con coberturas empíricas deficientes. Ante esta limitación, se propone el uso de métodos de imputación múltiple, cuya aplicación se explora y se compara con las técnicas tradicionales, poniendo en evidencia sus ventajas en términos de precisión y validez inferencial.

ÍNDICE DE GINI

Definición: Sea Y una variable aleatoria continua no negativa, el Índice de Gini se puede definir como:

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

Donde $\mu_Y = E[Y] = \int_0^\infty y dF_Y(x)$ es la media de Y y $F_Y(y)$ es la función de distribución.

G es una medida de desigualdad que toma valores entre 0 y 1; siendo 0 una situación de igualdad total y 1 lo contrario

METODOLOGÍA

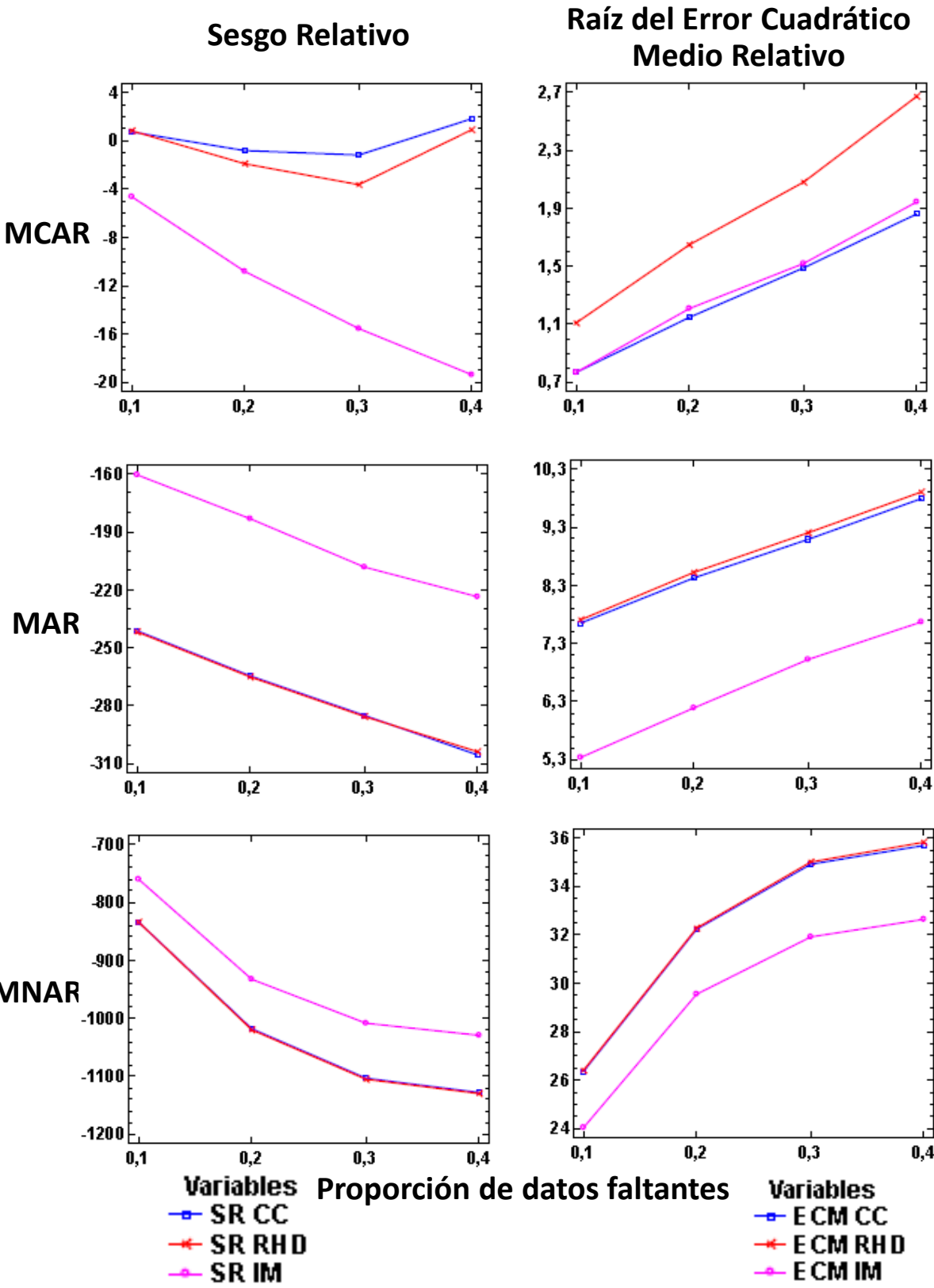
Se ha realizado un estudio de simulación Montecarlo para evaluar diferentes métodos de imputación en la estimación del índice de Gini con 1000 réplicas.

Por simplificación, se consideraron 3 variables (eqIncome, age, hy040n) de la eusilc de la librería leaken de R como una muestra seleccionada de una población infinita. En cada réplica se generaban una proporción de valores faltantes en la variable eqIncome con los diferentes mecanismos de respuesta MCAR, MAR y MNAR y se consideraban varios métodos de imputación para estimar el índice de Gini. Los estimadores se evaluaron en términos de sesgo relativo y la raíz del error cuadrático medio relativo en tanto por 10000 para proporciones de datos faltantes entre 0,10 y 0,40.

Se consideraron los siguientes métodos:

- Se consideraron los siguientes métodos:
1. Análisis de casos completos (CC)
 2. Random Hot Deck (RHD)
 3. Imputación mediante Regresión (Reg)
 4. Imputación múltiple usando random forest (IM)

RESULTADOS



CONCLUSIÓN

- A medida que aumenta la proporción de datos faltantes, aumenta tanto el Sesgo Relativo como el Error Cuadrático Medio Relativo.
- El método de imputación mediante Regresión no se ha incluido en las gráficas porque presenta mucho peores resultados que los otros y distorsiona la gráfica.
- Los métodos de imputación funcionan peor bajo el mecanismo de respuesta MNAR.
- En general, el método que mejores resultados proporciona es el de Imputación múltiple.
- Como próximos pasos en la investigación:
 - Estudios de Simulación Monte Carlo con más repeticiones para estabilizar resultados.
 - Considerar números más elevados de “m” y de “maxit” (m = número de conjunto de datos completos en imputación múltiple; maxit = número de iteraciones del algoritmo).
 - Ampliar el número de muestras para realizar el estudio.
 - Evaluar otros escenarios que incluyan conjuntos de datos con más variables, datos faltantes en más de una variable, poblaciones finitas,...
 - Evaluar intervalos de confianza construidos por diferentes métodos..