Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Propuesta de mejora en la medición del bienestar económico

Tesis de Licenciatura en Ciencias de Datos

Lautaro Lasorsa

Director: Rodrigo Castro
Codirector: Walter Sosa Escudero
Buenos Aires, Argentina, 2024

Índice general

1. MEDICIÓN DEL BIENESTAR ECÓNOMICO

En este cápitulo se presenta la problematica a tratar en la tésis, se motiva su estudio y se comparán las formas actuales de medirlo con las alternativas propuestas.

1.1. Definición y motivación

1.1.1. Definición

En este trabajo definiremos el bienestar económico (BE) de un individuo como la utilidad que obtiene de sus ingresos. Es decir, el valor que puede obtener de los recursos economicos, principalmente monetarios, de los que dispone.

Sea W el nivel de ingresos de un individuo, modelamos la utilidad que obtiene de esos ingresos como $U = log(W)$. Entonces, $BE = U(W) = log(W)$

Es importante tener presente que si bien al definirlo como $BE = log(W)$ estamos dando una definición constructiva que nos permite calcularlo a partir de los datos recabados, esta es una aproximación que realizamos al concepto en base al comportamiento que esperamos que tenga:

- **Es creciente**: Esperamos que un mayor ingreso cause un mayor bienestar económico en el inviduo que lo recibe. Formalmente, $W_1 \geq W_2 \rightarrow U(W_1) \geq U(W_2)$
- **Ley de utilidades marginales decrecientes**: Esperamos que, a partir de un valor W_1, la utilidad marginal del ingreso sea decreciente. Formalmente:

$$\exists W_1 / \forall W \geq W_1, \epsilon > 0, U(W + \epsilon) - U(W) > U(W + 2 * \epsilon) - U(W + \epsilon)$$

Este es un comportamiento analogo a la especificación y la implementación de una función en programación, donde las expectativas que tenemos sobre el indicador (su definición conceptual) cumplen el rol de la especificación y la definición objetivamente medible cumple el rol de la implementación.

1.1.2. Motivación

La correcta medición del bienestar económico tiene diversas motivaciones, como por ejemplo:

- Es un objeto de los objetos de estudio elementales de las ciencias económicas.
- Permite medir el impacto de las politicas publicas.
- Es un insumo para otras disciplinas y puede utilizarse como predictora de otras variables de interes.

De estas motivaciones puede deducirse que tiene tanto un valor intrinseco (por si mismo) como un valor instrumental, y que en conjunto vuelven a la correcta medición de este concepto un asunto de interes.

Notar que al ser un concepto al cuál aproximamos con una definición constructiva, hay 2 aspectos independientes:

- La calidad de la implementación que realicemos, es decir, qué tan bien captura la definición objetivamente medible las expectativas que tenemos sobre el concepto.
- La calidad de la medición de esta implementación.

De estos dos puntos, el presente trabajo propone posibles mejoras sobre el segundo basandose en la definición empirica ya explicada. Sin embargo, al estudiar la correlación entre el bienestar económico y otras variables, ambos puntos tendrán impacto aunque solo el segundo se estudie explicitamente.

1.2. Medición actual

1.2.1. Metodología

La medición actual del BE, utilizada por ejemplo por el **Índice de Desarrollo Humano** [1] (HDI por sus siglas en ingles, Human Development Index), es mediante el logaritmo del promedio de los íngresos.

En el caso puntual del HDI utiliza como insumo el GNI (Gross Domestic Income, Ingreso Nacional Bruto) a Paridad de Poder Adquisitivo (PPA, PPP en inglés). Es decir:

$$HDI_{income} = log(GNI\ per\ cpita)$$

Generalizandolo, dada una población de N individuos, cuyos ingresos son $X_1, X_2, \ldots, X_N$, la medición actual del bienestar ecónomico es:

$$BE(X) = log(\frac{1}{N} * \sum_{i=1}^{N} X_i)$$

1.2.2. Ventajas

La principal ventaja de este metodo es a nivel logistico, dado que calcula el BE en base a una colección de datos que pueden ser calculados o estimados de forma independiente:

- **GNI PPP**: El ingreso total de los individuos del país, deflactado a la paridad de poder adquisitivo. Inclusive puede ser posible calcular de forma separada el GNI nomial y el deflactor de PPP, pero también hay metodologías que apuntan directamente al valor GNI PPP utilizando cantidades en vez del valor monterio de los bienes y servicios.
- **Población**: La población de esa economía durante ese año. Usualmente se utiliza la cantidad de población a mitad del año.

Al ser estas metricas (GNI, deflactor, población) ya calculadas por tener interes en sí mismas, no supone una dificultad adicional obtener esta metrica derivada. Adicionalmente, es una metodología confiable ya que su confianza se deriva de la confianza que tenemos en la capacidad de medir las metricas base.

Además, las metricas base que se utilizán para calcular esta metrica tienen la ventaja adicional de ser valores aditivos. Es decir, cada uno de estos puede ser medido de forma independiente en subdivisiones de la economía a estudiar (por ejemplo las provincias de un

país, y dentro de estas los municipios) y luego simplemente sumar las cantidades medidas por cada subdivisión para obtener la cantidad correspondiente al total de la economía.

A su vez, cómo el GNI se pueden descomponer en distintos componentes, estos componentes pueden calcularse independientemente.

Se puede hacer esto utilizando que $GNI = GDP + EX_{net}$, donde:

- GDP (Gross Domestic Product, Producto Interior Bruto) es la producción total de bienes y servicios en el país
- EX_{net} es el neto de pagos y transferencisa (salvo importaciones y exportaciones, que se incluyen en el GDP) hacia el exterior (donde el signo positivo índica que a la economía ingreso más dinero del que salió)

Y a su vez podemos descomponer al principal sumando, el GDP, en sus componentes de dos maneras distintas, utilizando la Perspectiva de los Gastos y la Perspectiva de los ingresos.

Perspectiva de Gastos

$$GDP = C + I + G + (X - M)$$

Donde:

- C = consumo
- I = Inversión
- G = Gasto del gobierno
- X = Exportaciones
- M = Importaciones

Persectiva de los ingresos

$$GDP = S + B + R + I + II + D - SU$$

Donde:

- S = Salarios
- B = Beneficios empresariales
- R = Rentas (alquileres)
- I = Intereses
- II = Impuestos Indirectos (ejemplo: IVA)
- D = Depreciación de bienes de cápital
- SU = Subsidios

El tener estos dos enfoques permite:

- Descomponer el calculo del GNI en muchas variables chicas que se pueden medir de forma especializada y tienen interes en si mismo. Por tanto, se vuelve en si mismo una estadistica derivada de otras.
- Al tener 2 metodos independientes de calcular el GDP, que es el factor más importante en el computo del GNI, es posible detectar y corregir errores e inconsistencias en las mediciones.

En sintesis, la medición actualmente utilizada supone una gran cantidad de ventajas en matería logistica y de ser consecuencia de otra bateria de mediciones con interes en si mismas.

1.2.3. Desventajas

Hay un primer problema que podemos observar en esta metodología y es que el GNI incluye también los ingresos empresariales y del gobierno, mientras que el BE lo definimos a nivel de individuos y sus ingresos personales. Sin embargo, como las empresas y gobieros pueden utilizar estos recursos para proveer bienes y servicios a las personas (que a estas no se les imputan dentro de sus ingresos personales, por ejemplo la educación pública no arancelada), este alejamiento de la definición que dimos originalmente puede permitir capturar mejor el bienestar económico de una sociedad.

Sin embargo, hay otro problema que en principio es más importante a tener en cuenta. La utilidad marginal decreciente aplica en los ingresos de cada individuo y no en los ingresos agregados de la sociedad. Por tanto, es importante aplicar la función de utilidad (U, en teste caso log) a los ingresos de cada individuo y no al promedio de los ingresos. Y lo importante es que **el logaritmo del promedio no es el promedio de los logaritmos**

Es esta última desventaja la que este trabajo busca subsanar proponiendo una medición alternativa, y a su vez evaluar que mejoras ofrece dicha alternativa frente a la metodogía actual.

1.3. Propuesta de mejora: Medición ideal

1.3.1. Metodología

Dada una población X de N individuos cuyos ingresos son $X_1, X_2, \ldots, X_N$, definimos el Bienestar Económico (promedio) de esa población como:

$$BE(X) = \frac{1}{N} * \sum_{i=1}^{N} log(X_i) = log(\sqrt[N]{\prod_{i=1}^{N} X_i})$$

1.3.2. Ventajas

Es la medición exacta de la definición que propusimos de Bienestar Económico. Por tanto, es la mejor aproximación a dicho concepto. Lo que tiene especial sentido si pensamos que los ingresos son un instrumento para obtener bienestar y no el bienestar en si mismo, y eso es lo que se busca reflejar al aplicarles una función de utilidad distinta de la identidad.

Adicionalmente, al contemplar que un aumento en los ingresos de individuos que ya tienen altos ingresos les genera un menor beneficio marginal que aumentar en la misma

cantidad absoluta los ingresos de individuos de menores ingresos, esta metrica refleja le eficiencia de la distribución del ingreso en una sociedad.

De esta forma, esta metrica premia simultaneamente un aumento de la productividad de una ecónomia como una distribución más eficiente de la misma.

1.3.3. Desventajas

La principal desventaja de esta metodología es la contracara de la ventaja de la metodología actual, la logistica. En este caso llega al punto de la infactibilidad, puesto que para poder realizar el calculo propuesto es necesario conocer los ingresos de cada uno de los individuos, algo que es logisticamente imposible.

1.4. Propuesta de mejora: Mediciones aproximadas

1.4.1. Metodología

Debido a la imposibilidad factica de aplicar la metodología ideal propuesta en este trabajo, es necesario buscar aproximaciones con los datos disponibles.

De esta forma podemos introducir el concepto de **Granularidad** de una medición, G.

Sea una población X de N individuos, cuyos ingresos son $X_1, X_2, \ldots, X_N$, con $X_i \leq X_{i+1} \forall 1 \leq i < N$, y sea la granularidad G, tal que $N = G * T$, definimos $BE_G(X)$ como:

$$BE_G(X) = \frac{1}{G} * \sum_{i=0}^{G-1} log(\frac{1}{T} \sum_{j=i*T}^{(i+1)*T-1} X_j)$$

De esta definición se pueden reescribir la medición actual, utilizando el promedio de los ingresos de los individuos, como BE_1 y la medición ideal antes propuesta como BE_N. Si en lugar del promedio de los ingresos de los individuos, utilizamos el GNI (que también incluye ingresos de las empresas y del gobierno), lo llamaremos BE_{GNI}

1.4.2. Ventajas

Esta metodología tiene, parcialmente, las ventajas tanto de la metodología ideal propuesta como de la metodología actual:

- $\exists G > 1$ para el cuál la metodología es logisitcamente factible. Puede calcularse en base a encuestas de ingresos como la EPH (Encuesta Permanente de Hogares).
- Modera más que BE_1 el impacto de los outliers (ultra ricos) en la medición del bienestar ecónomico, ya que solo impactan de forma líneal en el último de los G grupos en los que se divide la población.

1.4.3. Desventajas

Tiene la desventaja de ser una solución de compromiso, y justamente parte del interes de este trabajo es ver, para valores logisticamente factibles de G, cuál es la diferencia entre BE_G y BE_N.

2. TÉCNICAS A UTILIZAR

En este cápitulo se explicarán las técnicas que se utilizarán a lo largo del trabajo para analizar los datos, tanto sinteticos como reales, y obtener conclusiones de los mismos.

2.1. Correlación de Pearson

El coeficiente de correlación de Pearson se utiliza para medir la similitud líneal entre 2 variables. Se define como:

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X * \sigma_Y}$$

Donde $Cov(X,Y)$ es la covarianza entre ambas variables, y σ_X y σ_Y son los desvios estandar de las mismas. Este indicador tiene varias propiedades interesantes:

- Si planteamos el modelo $Y = \alpha * X + \beta$, para los valores $\alpha*$ y $\beta*$ que minimizan el MSE (Mean Square Error):

  - $$R^2 = \frac{\sum_{i=1}^{N}(Y_i - X_i * \alpha * + \beta)^2}{\sum_{i=1}^{N}(Y_i - \mu_Y)^2}$$
  - $$R^2 = (r(X,Y))^2$$

  Es decir, captura la capacidad predictiva (en sentido estadistico) del modelo líneal $Y = \alpha * X + \beta$

- Podemos tratar a L^2 (variables aleatorias con segundo momento finito) como un espacio vectorial euclideo donde:

  - $Cov(X,Y)$ es el producto interno entre X e Y
  - $Var(X) = Cov(X,X)$ es la norma cuadrada de X. Analogamente, $std(X) = \sigma_X = ||X||$
  - $r(X,Y) = \frac{<X,Y>}{||X||*||Y||}$ es el coseno del angulo entre X e Y.

  Esto reafirma su interpretación como medida de similitud entre las variables.

2.2. Correlaciones no lineales

Notar que la correlación de Pearson solo captura relaciones líneales entre las variables, y si quieremos capturar otras posibles relaciones (por ejemplo, $Y = log(X)$) corresponde utilizar otras metricas, como son la correlación de Spearman y la τ de Kendall.

2.2.1. Correlación de Spearman

Sea $rank_X[x]$ la función que dado un individuo nos dice su índice en la población ordenada,

$$Spearman(X,Y) = 1 - \frac{6}{N*(N^2-1)} * \sum_{i=1}^{N} (rank_X[X_i] - rank_Y[Y_i])^2$$

2.2.2. τ de Kendall

$$\tau(X,Y) = \frac{2}{N*(N-1)} * (P_C - P_D)$$

Donde

- $P_C = ||\{(i,j)|i<j \wedge ((X_i<X_j) \wedge (Y_i<Y_j) \vee (X_i>X_j) \wedge (Y_i>Y_j))\}||$ son los pares que tienen la misma relación ordinal en X y en Y.
- $P_D = ||\{(i,j)|i<j \wedge ((X_i<X_j) \wedge (Y_i>Y_j) \vee (X_i>X_j) \wedge (Y_i<Y_j))\}||$ son los pares que tienen distinta relación ordinal en X que en Y.

Una forma alternativa de escribirla es:

$$\tau(X,Y) = \frac{2}{N*(N-1)} * \sum_{i=1}^{N} * \sum_{j=1}^{i-1} (signo(X_i - X_j) * signo(Y_i - Y_j))$$

2.3. Bootstrap

Bootstrap es una tecnica que consiste en utilizar los datos de una muestra para estimar la distribución de la misma, en base a esa estimación generar datos sinteticos y estudiar sobre esos datos sinteticos la distribución del indicador de interes.

En el caso de este trabajo se utilizará bootstrap no parametrico de la siguiente forma:

Dada una muestra M de N individuos $M_1, M_2, \ldots, M_N$, una nueva muestra se generará tomando N elementos de M con reposición (es decir, al seleccionar un elemento de M para incluirlo en nuestra nueva muestra, no lo eliminamos de M), y aplicamos el índicador (por ejemplo, la correlación entre 2 dimensiones de los individuos) a la población simulada.

3. DATOS SINTETICOS

En este cápitulo se utilizán datos simulados para explorar caracteristicas de los índicadores (existentes y propuestos) en sí mismos.

3.1. Generación

Para la generación de los datos se modela la distribución del ingreso en una sociedad como $logNorm(\mu, \sigma^2)$, basandonos en (Gibrat, 1931)[2]. Es decir, sea X una población con N individuos cuyos ingresos son $X_1, X_2, \ldots, X_N$, las variables X_i son IID (independientes e identicamente distribuidas) y $X_i \sim LN(\mu, \sigma^2)$.

Recordar que

$$X \sim LN(\mu, \sigma^2) \iff log(X) \sim N(\mu, \sigma^2)$$

Notar que bajo esta distribución:

- El bienestar económico tiene distribución normal $N(\mu, \sigma^2)$
- $BE_N(X)$ es un estimador insesgado de μ

La metodología consistio en lo siguiente:

- Genarar poblaciones con $N = 1{,}000{,}000$ individuos cada una.
- Ordenar a los individuos de la población, en orden creciente de ingresos.
- Para cada población X^i y para divisor G de N, calcular $BE_G(X^i)$
- Almacenar para posteriro uso los valores $BE^i_j = BE_{G_j}(X^i)$

Notar que todas las observaciones y todas las poblaciones generadas son independientes entre si.

Como modificar μ es lo mismo que multiplicar a todos los X_i por $e^{\Delta\mu}$, utilizaremos $\mu = 0$ para la generación de datos sinteticos.

Respecto del valor de σ^2, se generaron 2 datasets:

- **Datos** $LN(0, 1)$: Un dataset donde $X_i \sim LN(0, 1)$, para el que se simularon contiene 20,000 poblaciones
- **Datos** $LN(0, \sigma^2)$: Un dataset donde se toman diversos valores de σ^2, en el cuál se simularon 1,000 poblaciones para cada valor entero de σ^2 entre 1 y 10, generando 10,000 poblaciones en total.

La generación de datos se paralelizo utilizando GPU mediante CUDA[3].

3.2. Datos $LN(0,1)$

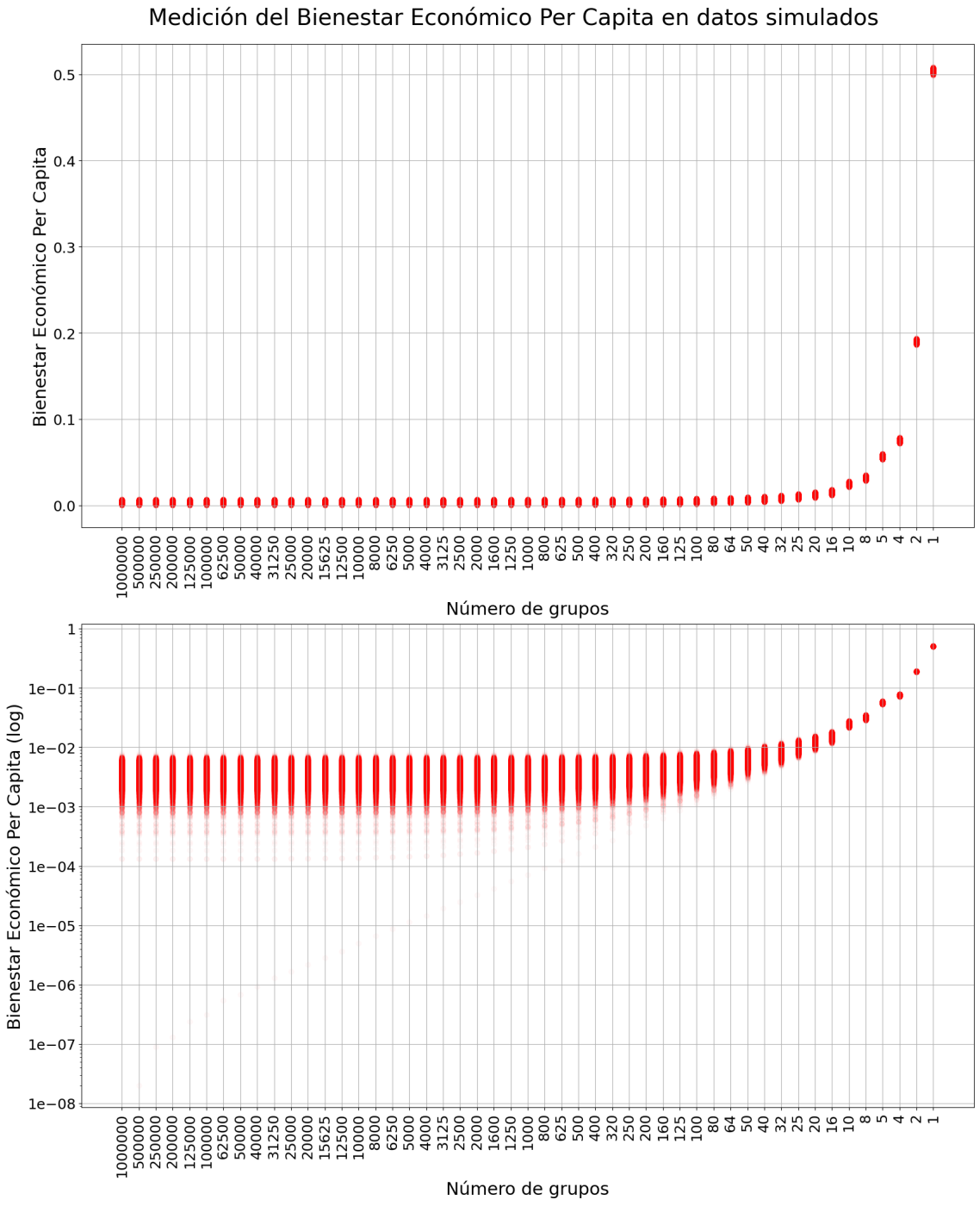

Fig. 3.1: Distribución de BE^i_j, donde el eje X es la granularidad (cantidad de grupos) de la medición. Para facilitar la visualización, a cada elemento se le resto el mínimo de todo el dataset

Como puede verse en 3.1, al tener más de 200 grupos es dificil distinguir las distintas distribuciones, inclusive en escala logaritmica. A su vez, se nota que cuando la granularidad

es baja, cada aumento de granularidad acerca significativamente BE_{G_j} a BE_N

Esto nos permite formular la primera de las hipotesis del presente trabajo:

hyp This is my first hypothesis.

3.3. Datos $LN(0, \sigma^2)$

En esta sección se analizán los datos generados utilizando una distribución $LogNormal(0, \sigma^2)$ para distintos valores de σ^2. Trabajar con distintos valores de σ^2 permite una mayor variedad y riqueza de análisis.

3.3.1. Comportamiento condicional

Se estudia el comportamiento condicional a σ^2, tratandolo como 10 poblaciones independientes y estudiandolas.

3.3.2. Interacción con granularidad

Se estudia como se comportan las mediciones con distintas granularidades al modificar σ^2

3.3.3. Comportamiento no condicional

Se trabajará con toda la muestra sin condicionar por σ^2, es decir tratandola como una unica población heterogenea.

Bibliografía

[1] United Nations Development Programme. Technical notes: Human development report 2023/24. Technical report, United Nations Development Programme (UNDP), 2023. Accessed: 2024-10-11, https://hdr.undp.org/sites/default/files/2023-24_HDR/hdr2023-24_technical_notes.pdf.

[2] R. Gibrat. *Les inégalités économiques*. Sirey, 1931.

[3] Lautaro Lasorsa. Códigos para la generación de datos simulados, 2024. https://gist.github.com/LautaroLasorsa/739dae8ec5f9041f243150183070815a.