Técnicas Estadísticas para las Ciencias Sociales

Vanesa Jordá

Máster Oficial en Economía: Instrumentos del Análisis Económico

Curso 2020-2021

Objetivos

- Estimación puntual de medidas de desigualdad con información completa
- Entender las limitaciones de los datos
- Conocer estrategias de estimación que nos permitan superar dichas limitaciones
- Estimación de medidas de desigualdad con información limitada a nivel nacional
- Conocer diferentes especificaciones para modelizar distribuciones de renta
- Bondad del ajuste
- Estimación de medidas de desigualdad a nivel regional

Desigualdad de renta: definición y conceptos

- La desigualdad, la pobreza y el crecimiento, así como los vínculos entre estos conceptos han recibido una gran atención en el ámbito económico, político y social.
- Estas variables juegan un papel crucial para evaluar el nuevo contexto económico y social de las profundas transformaciones de la realidad económica sufridas en las últimas décadas:
  - El crecimiento exponencial de China en los últimos 20 años.
  - La relación entre la globalización y la desigualdad.
  - La crisis financiera y sus consecuencias para la economía real.

Caso práctico I: Cálculo empírico de medidas de desigualdad

Práctica 1

El siguiente conjunto de datos proporciona información sobre la renta de una muestra de nueve individuos (en miles de dólares) en dos localidades distintas:
Localidad A: 13, 25, 15, 7, 12, 38, 42, 53, 7.
Localidad B: 4, 23, 36, 18, 39, 20, 9, 45, 12.
Representar las curvas de Lorenz de las localidades A y B en un solo gráfico.

Caso práctico I: Cálculo empírico de medidas de desigualdad

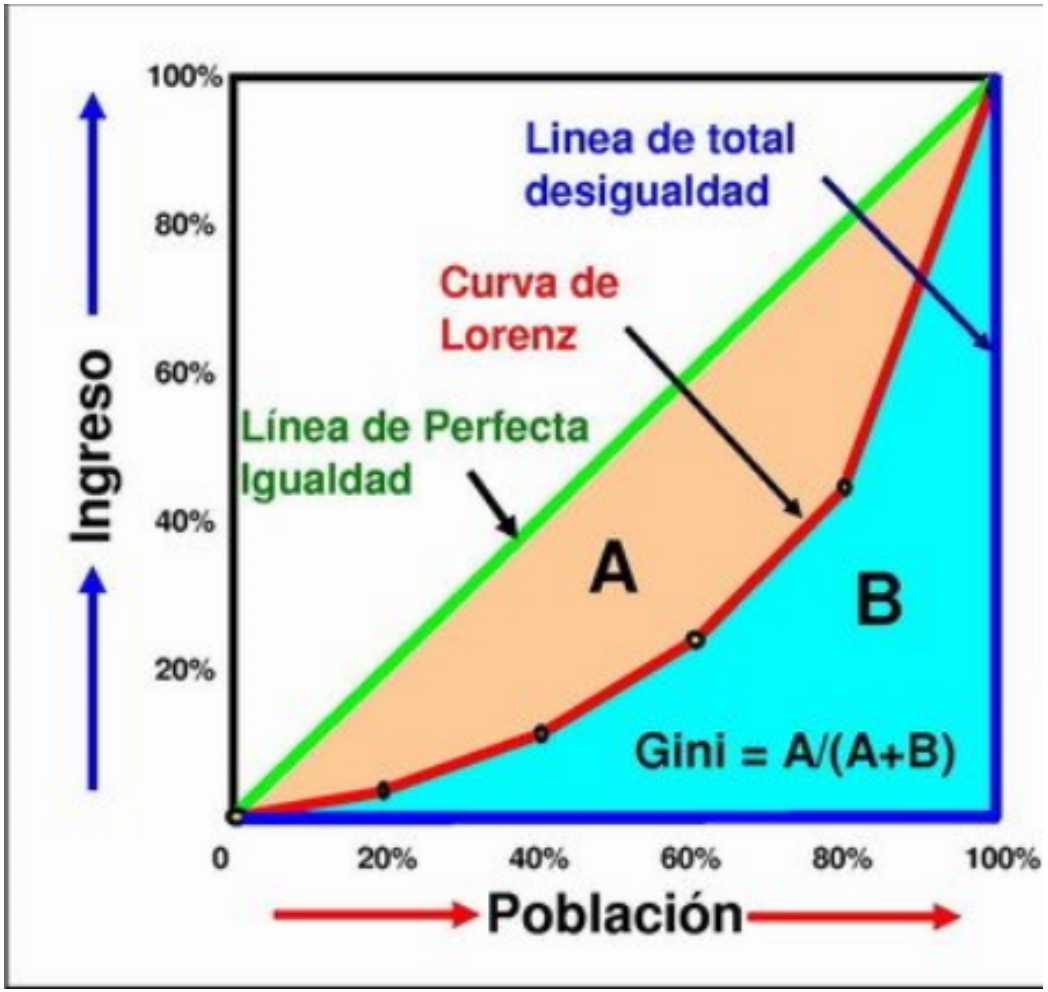
- Para estimar medidas de desigualad tenemos, en este caso, microdatos sobre la renta de los individuos.
- Curva de Lorenz: representa en el eje de ordenadas los porcentajes de renta asociados a los porcentajes de población que se muestran en eje de abscisas.
- Sea $x_1, ..., x_n$ una muestra aleatoria simple, se define la curva de Lorenz como:

$$Lc(p = j/n) = \frac{1}{n} \sum_{i=1}^{j} x_{(i)},$$

donde p es lo que se conoce como función de distribución empírica y $x_{(i)}$ representa la observación i-ésima de la muestra ordenada.

$$Pr(X \leq x) = \frac{\text{numero de valores en la muestra} \leq x}{n} = j/n$$

La curva de Lorenz



La curva de Lorenz

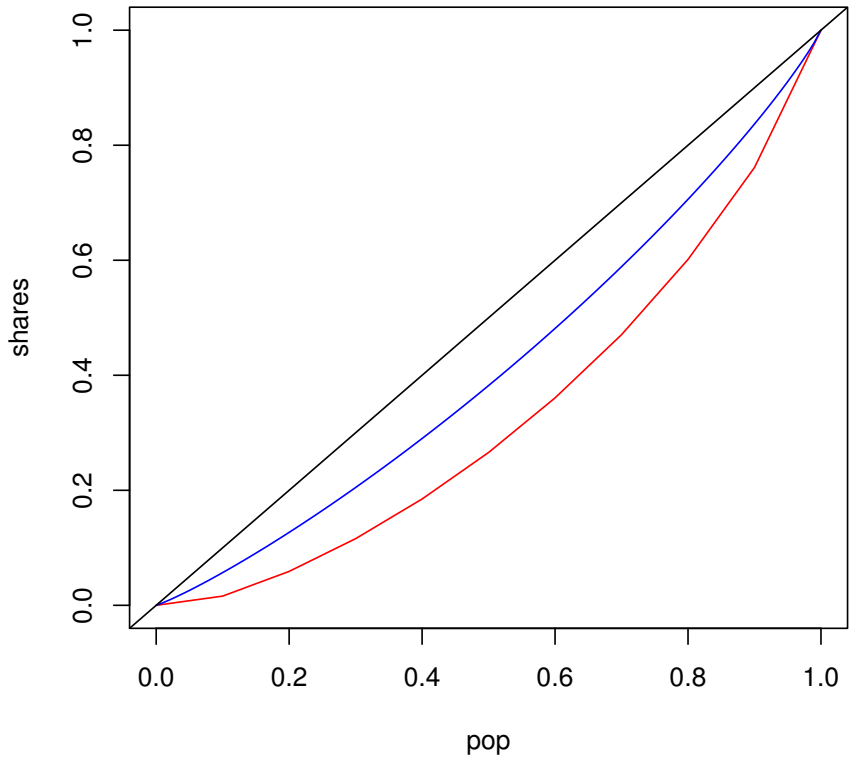


Figure: Dominancia en el sentido de Lorenz

La curva de Lorenz

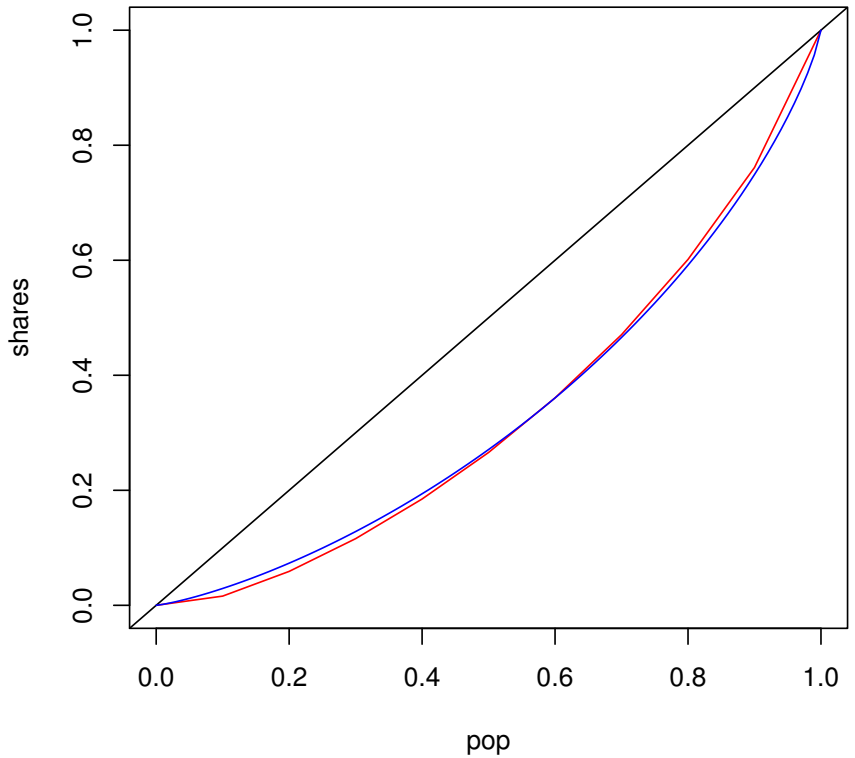


Figure: Curvas de Lorenz que se cortan

Medidas de desigualdad relativas

- Índice de Gini

$$G(x) = \frac{1}{2n^2\bar{x}} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|,$$

donde $\bar{x}$ representa el ingreso medio.

- Índice de Atkinson:

$$A_\epsilon(x) = 1 - \left[\frac{1}{n} \sum_{i=1}^{n} \left[\frac{x_i}{\bar{x}} \right]^{1-\epsilon} \right]^{\frac{1}{1-\epsilon}}, \epsilon \neq 1$$

$$A_1(x) = 1 - \frac{1}{\bar{x}} \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

donde $\epsilon > 0$ es el parámetro de aversión a la desigualdad. Cuanto mayor es este parámetro mayor es el peso que reciben, en este caso, los individuos más pobres.

Medidas de desigualdad relativas

- Índice de entropía generalizada

$$E_\theta(x) = \frac{1}{\theta^2 - \theta}\left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i}{\bar{x}}\right]^\theta - 1\right], \theta \neq 0, 1,$$

$$E_1(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{\bar{x}} log\left(\frac{x_i}{\bar{x}}\right); E_0 = \frac{1}{n}\sum_{i=1}^{n} log\left(\frac{\bar{x}}{x_i}\right),$$

  donde θ es un parámetro que mide la sensibilidad de esta medida en distintas partes de la distribución.
- El caso límite $\theta = 1$ se conoce como índice de Theil (Theil, 1967) y asigna la misma importancia a todas las observaciones.
- Si $\theta < 1$ la medida es más sensible a las diferencias entre los individuos más pobres.
- Si $\theta > 1$ la medida es más sensible a las diferencias entre los individuos más ricos.
- El caso límite $\theta = 0$ se conoce como mean log deviation (MLD) uno de los casos límite.

Medidas de desigualdad relativas

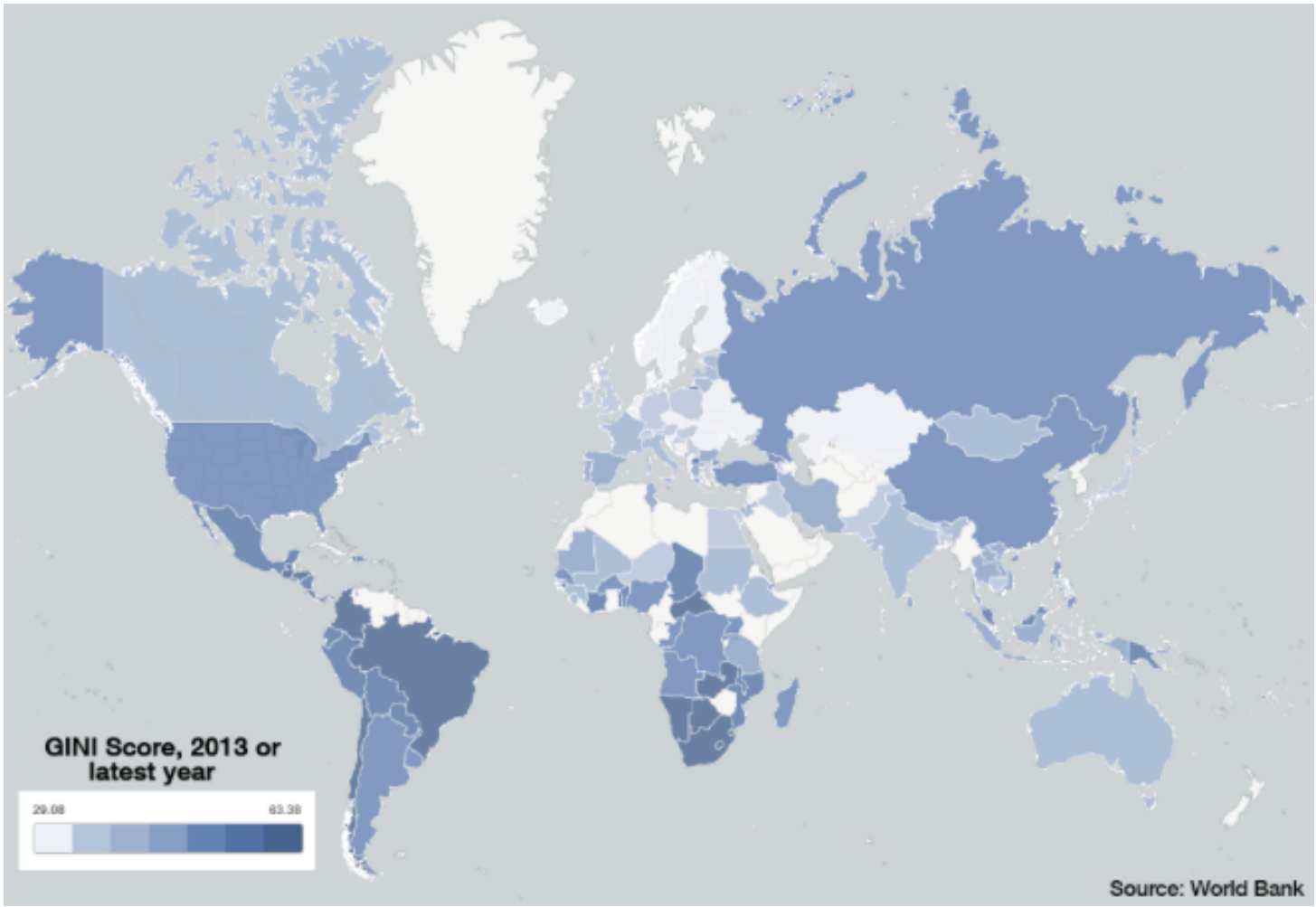
- Las medidas de desigualdad relativas son congruentes con el orden de Lorenz. En caso de que exista dominancia de Lorenz, la curva de Lorenz más cercana a la recta de equidistribución siempre será declarada como más equitativa por todas las medidas de desigualdad anteriores.
- Estas medidas de desigualdad son invariantes ante cambios de escala. Si cambiamos la unidad de medida de la variable ingreso de miles de euros a euros, el valor de las medidas de desigualdad no cambia.

Caso práctico I: Cálculo empírico de medidas de desigualdad

Práctica 2

Con las muestras sobre la renta individual de las localidades A y B:
a) Calcular el índice de Gini, el índice de Atkinson ($\epsilon = 2$), el índice de Theil y el MLD.
b) Comparar el efecto sobre la desigualdad de un impuesto fijo sobre la renta de 1000 dólares frente a un impuesto del 10%.

Índice de Gini en diferentes países del mundo



Estimación de la desigualdad con información limitada

- Hasta ahora hemos aprendido a calcular medidas de desigualdad cuando los micro-datos están disponibles.
- En determinados casos, especialmente para países en desarrollo no tenemos acceso a datos individuales de renta.
- La única información disponible suele presentarse en forma de datos agrupados.
- World Income Inequality Database (WIID) la base de datos agrupados de renta más extensa en términos de cobertura geográfica y temporal.

Ejemplo: USA, 2012 (WIID v3.4, US Census Bureau - Current Population Survey)

% Población	20%	40%	60%	80%	100%
% Renta	3.2%	8.3%	14.4%	23%	51%

Ejemplo: USA (2012)

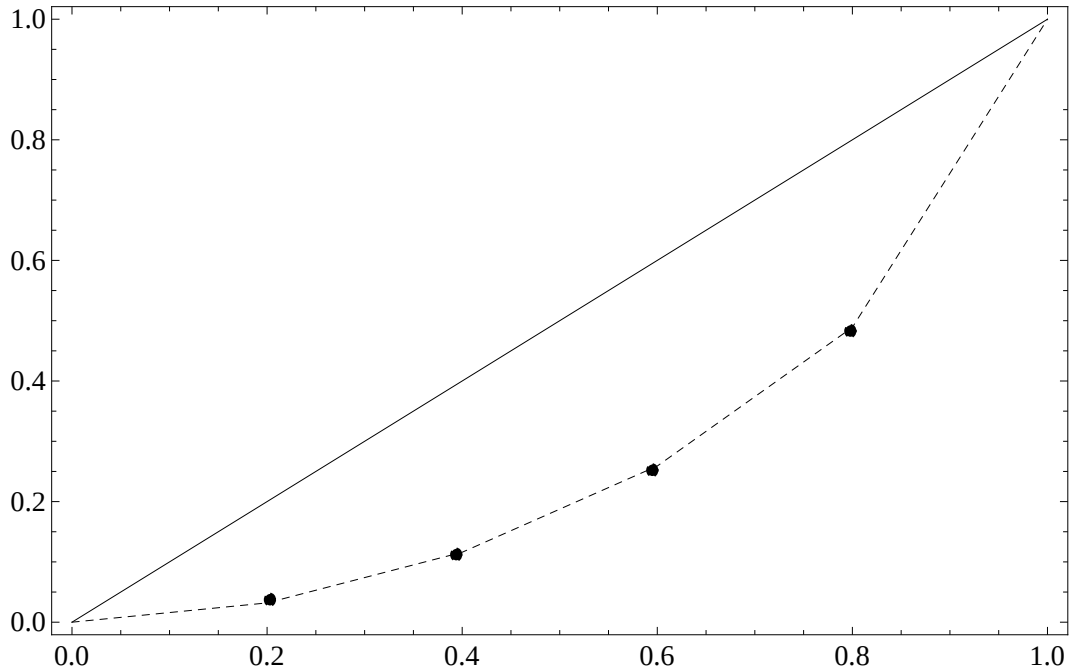
Práctica 3

Calcular el indice de Gini en Estados Unidos para el año 2012 y representar gráficamente la curva de Lorenz.

Un poco de notación...

- Sea **x** una muestra aleatoria simple de observaciones $i.i.d.$ de una distribución continua $f(x;\boldsymbol{\theta})$ definida en el soporte $H=[0,\infty)$, donde $\boldsymbol{\theta}\in\boldsymbol{\Theta}\subseteq\mathbb{R}^k$, siendo $\boldsymbol{\Theta}$ el espacio paramétrico.
- Supón que H se divide en J intervalos mutuamente excluyentes $H_j=(h_{j-1},h_j], j=1,\ldots,J$.
- Denotamos por $c_j=\sum_{i=1}^{N}1_{(h_{j-1},h_j]}(x_i)x_i/\sum_{i=1}^{N}x_i, j=1,\ldots,J$ a la proporción de ingreso total que tiene los individuos de la muestra pertenecientes al grupo$j$-ésimo y su proporción acumulada por $s_j=\sum_{k=1}^{j}c_k$.
- Sea $p_j=\sum_{i=1}^{N}1_{(h_{j-1},h_j]}(x_i)/N, j=1,\ldots,J$ la frecuencia de la muestra **x** en el intervalo j-ésimo y $u_j=\sum_{k=1}^{j}p_k$ la frecuencia acumulada.
- Por tanto las proporciones de ingreso $(s_j, j=1,\ldots,J)$ son puntos de la curva de Lorenz en el eje de ordenadas correspondientes con la abscisa $u_j, j=1,\ldots,J$.

Ejemplo: USA (2012)



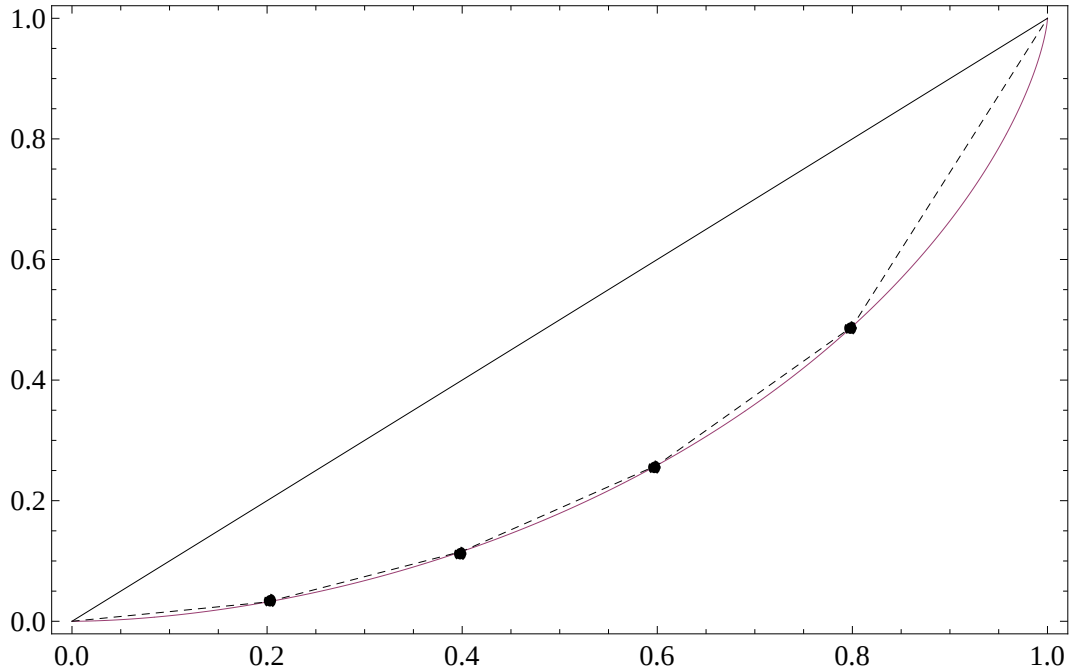
Ejemplo: USA (2012)

- Dada la estructura de los datos, no podemos emplear la fórmula del Gini presentada anteriormente.
- Cuando los datos están agrupados aproximamos el índice de Gini mediante la siguiente fórmula:

$$G(s_k, u_k) \approx 1 - \sum_{k=1}^{K} (s_k + s_{k-1})(u_k - u_{k-1}).$$

donde $u_k, k = 1, ..., K$ representa el porcentaje (acumulado) de población del grupo k-ésimo, y $s_k, k = 1, ..., K$ el porcentaje (acumulado) de renta de dicho grupo.

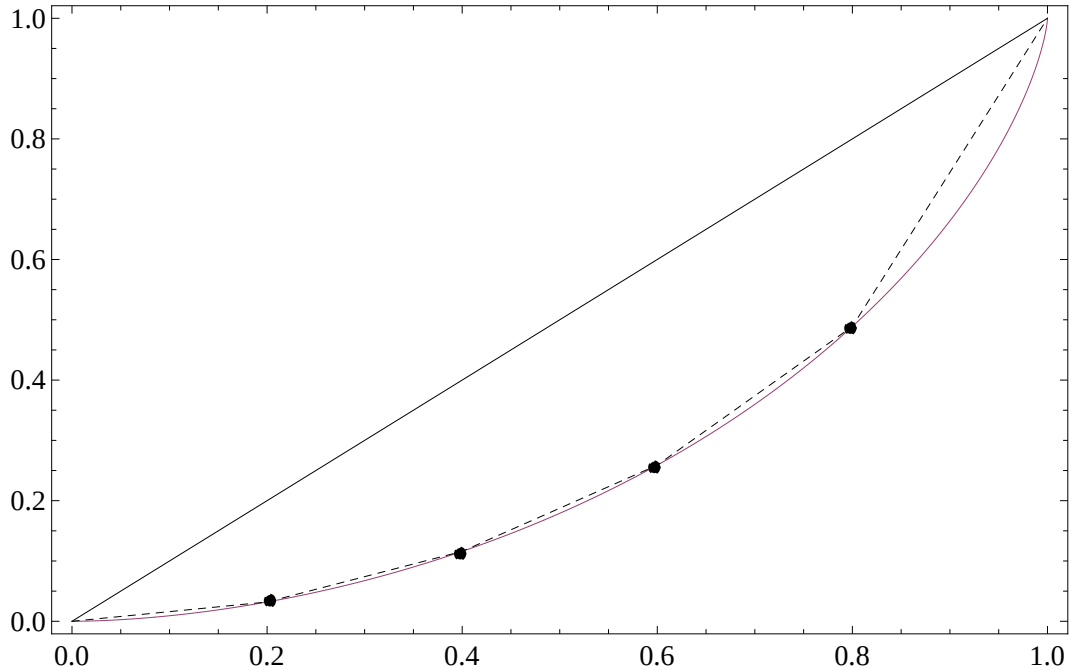
Example: USA (2012)



Estimación sesgada: supone que todos los individuos del grupo tienen la misma renta

- Índice de Gini de la encuesta: 0.477
- Índice de Gini empírico:0.441 (-0.036)
- Índice de Gini paramétrico:0.473 (-0.004)

Example: USA (2012)

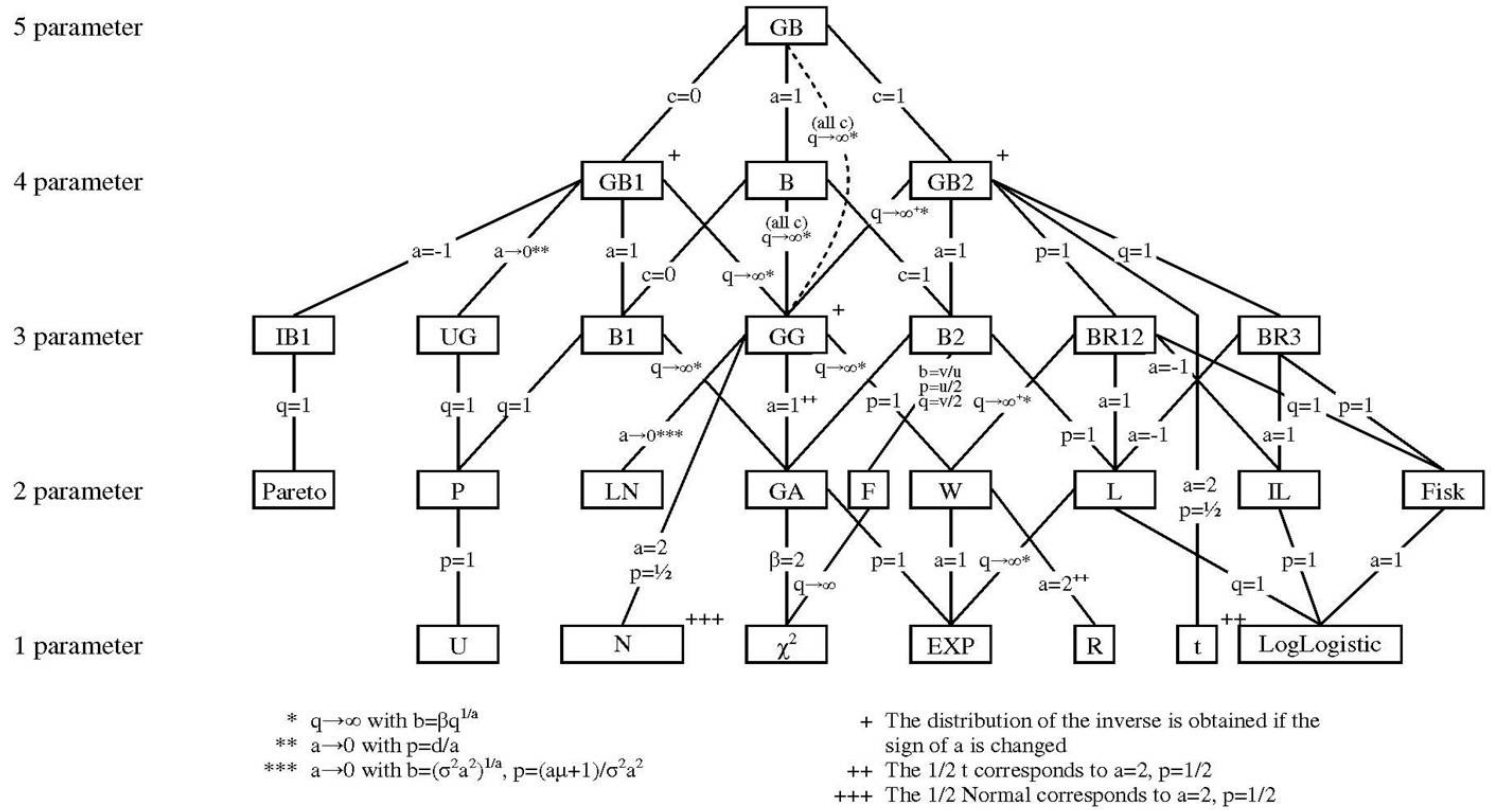


Estimación sesgada: supone que todos los individuos del grupo tienen la misma renta

- Índice de Gini de la encuesta: 0.477
- Índice de Gini empírico:0.441 (-0.036)
- Índice de Gini paramétrico:0.473 (-0.004)

Modelos paramétricos de distribución de la renta

Fuente: McDonald(1984)



Modelos paramétricos de distribución de la renta

Fuente: Jordá et al.(2018)

Table 1: Cumulative distribution function, kth moment distribution, kth moment and Gini index for a selection of distributions of the GB2 family

Distribution	CDF	kth moment distribution	$E(X^k)$	Gini Index
GB2	$B\left(\frac{(x/b)^a}{1+(x/b)^a};p,q\right)$	$GB2\left(a,p+\frac{k}{a},q-\frac{k}{a}\right)$	$\frac{b^k B(p+\frac{k}{a},q-\frac{k}{a})}{B(p,q)},\ q>k/a$	see Eq. (7)
Second kind beta	$B\left(\frac{x/b}{1+x/b};p,q\right)$	$B2(p+k,q-k)$	$\frac{b^k B(p+k,q-k)}{B(p,q)},\ q>k$	$\frac{2B(2p,2q-1)}{pB^2(p,q)},\ q>1.$
Singh-Maddala	$1-\left(1+\left(\frac{x}{b}\right)^a\right)^{-q}$	$GB2\left(a,1+\frac{k}{a},q-\frac{k}{a}\right)$	$\frac{b^k\Gamma(1+\frac{k}{a})\Gamma(q-\frac{k}{a})}{\Gamma(q)},\ q>k/a$	$1-\frac{\Gamma(q)\Gamma(2q-\frac{1}{a})}{\Gamma(q-\frac{1}{a})\Gamma(2q)},\ q>1/a.$
Dagum	$\left(1+\left(\frac{x}{b}\right)^{-a}\right)^{-p}$	$GB2\left(a,p+\frac{k}{a},1-\frac{k}{a}\right)$	$\frac{b^k\Gamma(p+\frac{k}{a})\Gamma(1-\frac{k}{a})}{\Gamma(p)},\ k/a<1$	$\frac{\Gamma(p)\Gamma(2p+\frac{1}{a})}{\Gamma(2p)\Gamma(p+\frac{1}{a})}-1,\ a>1.$
Lognormal	$\Phi\left(\frac{\log x-\mu}{\sigma}\right)$	$LN(\mu+k\sigma^2,\sigma)$	$\exp\left(k\mu+k^2\sigma^2/2\right)$	$2\Phi\left(\frac{\sigma}{\sqrt{2}}\right)-1.$
Fisk	$1-\left(1+\left(\frac{x}{b}\right)^a\right)^{-1}$	$GB2\left(a,1+\frac{k}{a},1-\frac{k}{a}\right),\ k/a<1$	$b^k\Gamma(1+k)\Gamma(1-k),\ k<1$	$\frac{1}{a},\ a>1.$
Weibull	$1-e^{-(x/b)^a}$	$GG\left(a,1+\frac{k}{a}\right)$	$b^k\Gamma\left(1+\frac{k}{a}\right)$	$1-\frac{1}{2^{1/a}},\ a>1.$

Source: Arnold and Sarabia (2018), Kleiber and Kotz (2003) and McDonald (1984).
Note: $B(x;a,b)$ denotes the incomplete beta function. The existence of kth moment distribution, defined as $F_{(k)}(x) = (\int_o^x t^k dF(t))/(\int_0^\infty t^k dF(t)), x>0$, requires the same constraints about the parameters than the kth moment and $E(X^k)<\infty$.

Método de estimación: Mínimos cuadrados no lineales

- Sea X una variable aleatoria con función de distribución $F(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}$ y curva de Lorenz $L(u;\boldsymbol{\theta})$, la estrategia de estimación se plantea del siguiente modo:

$$\min_{\boldsymbol{\theta}} \sum_{j=1}^{J-1} (L(u_j;\boldsymbol{\theta}) - s_j)^2,$$

$L(u_j;\boldsymbol{\theta})$ representa la forma función de la curva de Lorenz de la distribución paramétrica empleada para modelizar la variable renta. Para la distribución GB2 se expresa como:

$$L_{GB2}(u;a,p,q) = B\left(B^{-1}(u;p,q); p+\frac{1}{a}, q-\frac{1}{a}\right), \quad 0 \leq u \leq 1,$$

donde $q > 1/a$ y $B^{-1}(x;p,q)$ es la inversa de la función beta ratio y $B(x,a,b)$ es la función beta incompleta.

Método de estimación

- La curva de Lorenz es independiente de la escala, por tanto con este método sólo podemos estimar los parámetros de forma a, p, q.
- Esto no es una limitación si nuestro interés reside en estimar medidas de desigualdad relativas ya que los parámetros de escala no juegan ningún papel.
- La ventaja de este método es que reduce la dimensionalidad del problema de optimización.
- Para estimar el parámetro de escala (b) resolvemos la siguiente ecuación:

$$\bar{X} = \mu(\hat{a}, b, \hat{p}, \hat{q}),$$

donde $\bar{X}$ es la media muestral y $\mu(a, b, p, q) = \int_{\mathbb{R}_+} x f(x; a, b, p, q) dx$ es la esperanza de la distribución paramétrica.

Práctica

Práctica 4

Con los datos agrupados de estados Unidos:
a) Ajustar las distribuciones GB2, Beta 2, Dagum y Singh-Maddala.
b) Calcular los índices de Gini estimados para cada modelo.
c) Calcular el índice de Atkinson ($\epsilon = 1.25$), el índice de Theil y el MLD.

Bondad del ajuste

- Las medidas de bondad del ajuste comparan los puntos observados de la curva de Lorenz con los teóricos que derivan de las distribuciones paramétricas:

  1. Desviación cuadrática media

$$sse = \frac{1}{J-1}\sum_{j=1}^{J-1}(s_j - \hat{L}(u_j;\boldsymbol{\theta}))^2$$

  2. Versión adaptada del criterio de información de Akaike

$$aic = e^{2K/(J-1)}\frac{1}{J-1}\sum_{j=1}^{J-1}(s_j - \hat{L}(u_j;\boldsymbol{\theta}))^2$$

- Los estadísticos de bondad del ajuste nos permiten determinar el mejor modelo de entre los prepuestos pero en ningún caso validan la especificación paramétrica considerada.

Ejemplo: USA (2012)

Práctica 5

Determinar qué modelo de los ajustados en la Práctica 4 presenta mayor bondad del ajuste empleando los estadísticos anteriores.

Desigualdad a nivel regional

Práctica 6

Calcular la desigualdad de renta en Norte América en el año 2012 empleando la varianza y las medidas de entropía generalizada con $\theta = 0, 1, 1.5$.

Desigualdad a nivel regional

- A partir de medidas de desigualdad nacionales es posible construir la desigualdad de una determinada región.
- Vamos a construir varias medidas de desigualdad de renta en Norte América.
- Para ello necesitamos datos de Canadá en el año 2012.

% Población	10	20	30	40	50	60	70	80	90	100
% Renta	2.2	4	5.2	6.5	7.9	9.2	10.7	12.6	15.4	26.1

Descomposición de las medidas de entropía generalizada

- Las medidas de entropía generalizada admiten una descomposición aditiva en términos de la desigualad entre los países y dentro de los países.

$$GE_T(X;\theta) = GE_W(X;\theta) + GE_B(X;\theta)$$

- Para descomponer las medidas de entropía generalizada se aplican las siguientes fórmulas, siendo la desigualdad total la suma de ambos componentes:

$$GE_W(X;\theta) = \sum_{i=1}^{N} \lambda_i^{1-\theta} s_i^{\theta} GE_i(\theta),$$

$$GE_B(X;\theta) = \frac{1}{\theta(\theta-1)} \left(\sum_{i=1}^{N} \lambda_i \left(\frac{\mu_i}{\mu} \right)^{\theta} - 1 \right),$$

donde $\lambda_i, i = 1, ..., N$, representa el porcentaje de población del país i, s_i porcentaje de renta media del país i sobre la renta total: $s_i = \frac{\lambda_i \mu_i}{\mu} = \frac{\lambda_i \mu_i}{\sum_{i=1}^{N} \lambda_i \mu_i}$ y $GE_i(X;\theta)$ es el valor de la medida de entropía generalizada en el país i.

Descomposición de las medidas de entropía generalizada

- Para los casos especiales de estas medidas, el índice de Theil ($\theta = 1$) y el MLD ($\theta = 0$), la descomposición viene dada, respectivamente, por las siguientes fórmulas:

$$T_W(X) = \sum_{i=1}^{N} s_i T_i;\ T_B = \sum_{i=1}^{N} s_i log\left(\frac{\mu_i}{\mu}\right),$$

$$L_W(X) = \sum_{i=1}^{N} \lambda_i L_i;\ L_B = \sum_{i=1}^{N} \lambda_i log\left(\frac{\mu}{\mu_i}\right),$$

donde $T_i(X)$ and $L_i(X)$ son, respectivamente, los índices de Theil y MLD del país i.

Otros ejemplos de estimación con información limitada

- Estimación de elasticidades de pobreza (Bresson, 2009)
- Estimación de desigualdad de educación (Jordá y Alonso, 2017)

Referencias

- Bresson, F. (2009). On the estimation of growth and inequality elasticities of poverty with grouped data. Review of Income and Wealth, 55(2), 266-302.
- Cowell, F. (2011). Measuring Inequality. Oxford University Press, Oxford.
- Jordá, V., Alonso, J. M. (2017). New estimates on educational attainment using a continuous approach (1970?2010). World Development, 90, 281-293.
- Jorda, V., Sarabia, J. M., and Jäntti, M. (2018). Estimation of income inequality from grouped data. arXiv preprint arXiv:1808.09831.
- Kleiber, C., and Kotz, S. (2003). Statistical size distributions in economics and actuarial sciences (Vol. 470). John Wiley and Sons.
- Lakner, C., and Milanovic, B. (2016). Global income distribution: from the fall of the Berlin Wall to the Great Recession.The World Bank Economic Review, 30(2), 203?232.
- McDonald, J. B. (1984). Some generalized functions for the size distribution of income. Econometrica: Journal of the Econometric Society, 647-663.
- Niño-Zarazúa, M., Roope, L., and Tarp, F. (2017). Global inequality: Relatively lower, absolutely higher. Review of Income and Wealth, 63(4), 661-684.
- Wasserman, Larry (2003). *All of statistics*. New York: Springer.