Economics Bulletin

Volume 41, Issue 4

Income Inequality and Health: New Methodology and an Application

Jaesang Sung
Georgia State University

Qihua Qiu
Augusta University

James Marton
Georgia State University

Abstract

Prior studies propose a way to express the Gini index of income inequality as a function of the ratio of mean to median household income under the assumption that individual income follows the lognormal distribution. This allows for easy and precise construction of annual US income inequality indices at different levels of geography. In this paper, we are the first to express the Atkinson index in a similar manner. We also contribute to the literature by expressing both indices under the assumption that individual income follows the Pareto distribution. We merge these indices into an individual level dataset consisting of the 2001-2012 annual editions of the U.S. Behavioral Risk Factor Surveillance System at the state and county level. In an application, we find preliminary evidence that greater income inequality negatively affects overall self-reported health.

We would like to thank Tom Mroz, Charles Courtemanche, Rusty Tchernis, and Pierre Nguimkeu for their comments. Any errors are, of course, our own.

Citation: Jaesang Sung and Qihua Qiu and James Marton, (2021) "Income Inequality and Health: New Methodology and an Application", *Economics Bulletin*, Vol. 41 No. 4 pp. 2676-2689.

Contact: Jaesang Sung - jaesang27@gmail.com, Qihua Qiu - qqiu@augusta.edu, James Marton - marton@gsu.edu.

Submitted: January 02, 2021. **Published:** December 29, 2021.

1. Introduction and Background

Prior studies propose a way to express the Gini index of income inequality as a function of the ratio of mean to median household income under the assumption that individual income follows the lognormal distribution (Aitchison and Brown, 1957; Crow and Shimizu, 1987). This approach allows for easy and precise calculation of such annual US income inequality measures at different levels of geography. In this paper, we are the first to express the Atkinson index in the same manner. We also contribute to the literature by expressing both indices under the assumption that individual income follows the Pareto distribution.[1]

The absolute income hypothesis (AIH) suggests that increasing income improves health outcomes. The AIH is well-established and supported by prior literature (Bechtel et al., 2012; Lorgelly and Lindley, 2008). The notion that everyone's health in society is reduced when there is more income inequality is known as the income inequality hypothesis (IIH) (Subramanian and Kawachi, 2004; Lynch et al., 2004; Wilkinson and Pickett, 2006; Gravelle and Sutton, 2009). Kaplan et al. (1996) and Kennedy et al. (1996) suggest that under-investment in human and social capital such as education and medical care caused by inequality is a potential mechanism behind the significant relationship between greater inequality and worse health outcomes. Kawachi and Kennedy (1997) also argue that lower inequality promotes social integration that is closely associated with individual well-being.

The ideal dataset to test the IIH would be an individual level dataset containing information about income and health measured frequently over time and space, as well as geographic identifiers. However, no such dataset exists and no annual income inequality measures at different levels of geography are available in the United States. The US Census provides the American Community Survey (ACS) 1-year and 5-year estimates of the state- and county-level Gini index only after 2010. [2, 3]

Mellor and Milyo (2002) use data from the 1995-1999 Current Population Survey (CPS) to calculate annual inequality measures at different levels of geography and investigate the effect of income inequality on individual health status. However, concurrent work published in the same year, Blakely et al. (2002), criticizes Mellor and Milyo (2002) for their use of the CPS for income inequality analysis. Blakely et al. (2002) argue that CPS data are not precise enough to calculate inequality measures at local (i.e. sub-state) levels of geography and goes on to suggest that decennial Census data would produce more precise income inequality measures due to larger sample sizes.

The literature has subsequently focused on the use of individual income data from the decennial Census to construct regional income inequality measures. Employing 1990 decennial Census data or 2000 decennial Census data or both, Blakely et al. (2002) and Lopez (2004) show support for the IIH, whereas Chang and Christakis (2005) and Chen and Crawford (2012) find no

[1] The earliest references regarding modeling the distribution of income are Pareto (1896) and Gibrat (1931) who proposed the Pareto distribution and the lognormal distribution respectively.

[2] Please refer to https://data.census.gov/cedsci/table?q=income%20inequality&tid=ACSDT1Y2019.B19083.

[3] Meanwhile, our suggested methodology, which will be explained in detail in the next section, enables us to construct both the Gini index and the Atkinson index going as far back as 1991. This is critical in that recent literature (such as Piketty et al., 2017) find that US income inequality strikingly increased during the period including the 1990s and 2000s.

effects or mixed effects of inequality on health status and health behaviors. However, a general limitation associated with all of these studies using decennial Census data is the inability to estimate time varying effects of inequality. This is due to the fact that Census data is only produced every ten years.[4] In summary, use of either the CPS or the decennial Census has serious limitations in that they do not allow for the construction of income inequality measures from a dataset with a large sample size that measures income frequently over time (Kopczuk et al., 2010).

Alternatively, Kelly (2000) and Brush (2007) apply a new methodology derived by Aitchison and Brown (1957) and Crow and Shimizu (1987) to construct the annual Gini index as a function of the ratio of mean to median household income under the assumption that US individual income follows the lognormal distribution. Kelly (2000) and Brush (2007) compute annual mean and median household income at both the state and county level using various Federal data sources and then plug them into the derived equation. Both papers examine the relationship between income inequality and crime rather than health.

In this paper we apply the same methodology that Kelly (2000) and Brush (2007) used to address the issues raised by using either the CPS or the decennial Census to estimate the effect of income inequality on health. Furthermore, we contribute to the income inequality literature by applying this methodology to express the Atkinson index, an alternative measure of income inequality, as a function of the ratio of mean to median household income. This is because Kennedy et al. (1996), Weich et al. (2002) and Laporte (2002) suggest that the estimated effect of income inequality on health differs with respect to the choice of income inequality measure. Finally, we also contribute to the literature by deriving both the Gini and Atkinson index under the alternate assumption that individual income follows the Pareto distribution. This is significant in that prior studies such as Piketty (2013) and Sommeiller and Price (2016) model the distribution of US income using the Pareto distribution.

We construct an individual level dataset by combining the 2001-2012 annual editions of the U.S. Behavioral Risk Factor Surveillance System (BRFSS) at the state and county level with annual regional inequality measures computed by our suggested methodologies. In an application, we find preliminary evidence that greater income inequality negatively affects self-reported overall health in the United States.

2. Methods

2.1 Review of the Gini Index under the Lognormal Income Distribution Assumption

Equation (1) describes the Gini index (Sen, 1973), the most commonly used measure of income inequality in the literature:

[4] Alternative approaches to calculating an annual inequality measure include using individual tax filing data from the Internal Revenue Service (IRS) or income data from the ACS. IRS data provide large sample sizes and annual data but access to the data is limited and censored below a threshold level of income. The ACS provides both state and sub-state geographic identifiers via Public Use Microdata Areas (PUMAs). However, the ACS has not provided income data every year. In addition, use of PUMAs as a measure of sub-state geography creates challenges as the PUMA definitions can change over time, unlike other measures of sub-state geography such as counties.

$$G = \frac{1}{n}\left[n + 1 - 2\left(\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i}\right)\right] \quad for\ \mathrm{y}_1 \le y_2 \le \cdots \le y_n \tag{1}$$

Here y_i represents the income of individual (or household) *i* and *n* represents the total number of individuals (or households) being considered. The Gini index varies from 0 (complete income equality) to 1 (complete income inequality). Crow and Shimizu (1987) show that the Gini index can be derived as a function of mean and median household income, under the assumption that individual income is log-normally distributed. This version of the Gini index is represented by equation (2):

$$G = 2\Phi\left(\sqrt{\ln\left(\frac{\bar{y}}{\ddot{y}}\right)}\right) - 1 \tag{2}$$

Here $\Phi(.)$ is the cumulative density function of standard normal distribution. $\bar{y}$ and $\ddot{y}$ are mean and median household income, respectively, of all the households living in the reference group.

2.2 Our Derivation of the Atkinson Index under the Lognormal Income Distribution Assumption

The Atkinson (1970) index measures the social utility that can be gained by total redistribution from current income distribution to equality. The Atkinson index ranges from 0 (complete equality) to 1 (complete inequality). It is given as:

$$A = \begin{cases} 1 - \frac{1}{\bar{y}}\left(\frac{1}{n}\sum_{i=1}^{n} y_i^{1-\varepsilon}\right)^{\frac{1}{1-\varepsilon}} & for\ 0 \le \varepsilon \ne 1 \\ 1 - \frac{1}{\bar{y}}\left(\prod_{i=1}^{n} y_i\right)^{\frac{1}{n}} & for\ \varepsilon = 1 \end{cases} \tag{3}$$

Here ε is the "inequality aversion parameter." The higher the value of ε is, the higher level of the society's aversion toward inequality, and the more gain by redistribution from inequality to equality. In practice, ε values of 0.5, 1, or 2 are used commonly (Lorgelly and Lindley, 2008; Bechtel et al., 2012). Under the lognormal income distribution assumption such that: $y_i \sim \ln N(\mu, \sigma^2)$, the Atkinson index can be expressed as (Lubrano, 2013):

$$A = 1 - e^{-\frac{1}{2}\sigma^2\varepsilon} \tag{4}$$

For individual household income that follows lognormal distribution, the mean and median household income are $\bar{y} = e^{\mu+\frac{1}{2}\sigma^2}$ and $\ddot{y} = e^{\mu}$. So we solve for σ^2:

$$\sigma^2 = 2\ln\left(\frac{\bar{y}}{\ddot{y}}\right) \tag{5}$$

Plugging equation (5) into equation (4), we can derive the Atkinson index under lognormal income distribution as:

$$A = 1 - \left(\frac{\ddot{y}}{\bar{y}}\right)^{\varepsilon} \tag{6}$$

2.3 Our Derivation of the Gini Index and Atkinson Index under the Pareto Income Distribution Assumption

Assume individual household income, y_i, follows the Pareto distribution within the reference group such that: $y_i \sim Pareto\ (y_m, \alpha)$. Here $y_m > 0$ denotes the minimum possible value of y_i. The positive parameter α is the Pareto index when the Pareto distribution is used to model the distribution of wealth. Under the Pareto income distribution assumption, the Gini index and the Atkinson index can be expressed as equation (7) and equation (8) respectively (Lubrano, 2013):

$$G = \frac{1}{2\alpha - 1} \tag{7}$$

$$A = 1 - \frac{(\alpha - 1)\alpha^{\frac{1}{1-\varepsilon}}}{\alpha[\alpha + \varepsilon - 1]^{\frac{1}{1-\varepsilon}}} \tag{8}$$

Sung et al. (2020) solve the Pareto index, α, as a function of mean and median household income such that: [5]

$$\alpha = \frac{\ln 2}{\ln 2 + W(\theta)},\ \ W(\theta) \approx \theta - \theta^2 + \frac{3}{2}\theta^3 - \frac{8}{3}\theta^4 + \frac{125}{24}\theta^5,\ \ \theta = -\frac{\ln 2}{2} * \frac{\ddot{y}}{\bar{y}} \tag{9}$$

Here $W(\theta)$ is the Lambert W function expressed as a Taylor series that can be approximated using the Lagrange inversion theorem. Plugging equation (9) into equations (7) and (8), we can derive the Gini index and Atkinson index (ε=0.5, 1, 2) under Pareto income distribution assumption as:

$$G = \frac{\ln 2 + W(\theta)}{\ln 2 - W(\theta)} \tag{10}$$

[5] Sung et al. (2020) derive the Yitzhaki index of relative deprivation as a function of mean and median household income under the assumption that US individual income follows the lognormal or Pareto distribution.

$$A = \begin{cases} \left(\dfrac{\ln 2 + W(\theta)}{\ln 2 - W(\theta)}\right)^2 & for\ \varepsilon = 0.5 \\ \dfrac{\ln 2 + W(\theta)}{\ln 2} & for\ \varepsilon = 1 \\ \left(\dfrac{\ln 2 + W(\theta)}{\ln 2}\right)^2 & for\ \varepsilon = 2 \end{cases} \tag{11}$$

Thus, we can calculate the Gini index and Atkinson index for any time period and level of geography for which we have mean and median household income, regardless of the availability of individual income data, under either the lognormal or Pareto income distribution assumption. [6]

2.4 Calculation of Annual Regional Mean and Median Household Income

In order to implement this methodology, we compute annual mean and median household income at both the state and county level using various Federal data sources following Sung et al. (2020), a study focusing on relative deprivation rather than income inequality. We obtain regional median household income data from the Census Small Area Income and Poverty Estimate. We calculate regional mean household income by multiplying regional mean personal income from the Bureau of Economic Analysis by regional mean household size from the ACS.[7]

The BRFSS provides both individual income data and health data. However, there are two reasons that we use external aggregate annual income data via the Federal sources to calculate regional mean and median household income. First, the accuracy of these external estimates of income is arguably stronger than estimates of income derived from a smaller individual survey such as the BRFSS or the CPS, especially at sub-state levels of geography such as county. Second, the ability to do both state and sub-state level analysis is important when estimating the impact of income inequality, as the literature has debated the appropriate level of geography to determine an individual's reference group. In particular, we do not believe the BRFSS and the CPS provide a sufficient number of observations to credibly calculate the Gini index or the Atkinson index at the county level. For example, in 2007, Camp county in Texas has only 6 individuals sampled in the BRFSS. It is very difficult to believe that the index calculated using such a small sample size could represent income inequality of the whole population in a county.[8] This trade-off between accurate income estimates and the frequency of such estimates is why some papers in the literature use Decennial Census data (strong on accuracy, weak on frequency) and others use survey data such as the CPS (weaker on accuracy, stronger on frequency). Our approach using annual aggregate income and household size data from external

[6] As an extension, we derive in the appendix Gini indices as a function of socioeconomic subgroup mean and median household incomes under a mixture of lognormal distributions or a mixture of Pareto distributions.
[7] We use a linear interpolation to approximate mean household size in the years that the ACS data is not available.
[8] Besides that, the BRFSS income data are reported in ranges and are right truncated: less than US$ 10,000, 10,000-14,999, 15,000-19,999, 20,000-24,999, 25,000-34,999, 35,000-49,999, 50,000-74,999, and 75,000 or above. Therefore, mean household income calculated using BRFSS income data is underestimated.

Federal sources allows us to have the "best of both worlds" in terms of annual calculation of the index and with accurate income data.

2.5 Econometric Model

Using the annual Gini index and Atkinson Index at different levels of geography calculated by our suggested methodologies, we examine the effect of income inequality on individuals' overall health by estimating linear models specified as in equation (12) below:

$$H_{ist} = \alpha + \beta_1 I_{st} + X_{ist}\beta_2 + \beta_3 U_{st} + \delta_s + \lambda_t + \varepsilon_{ist} \quad (12)$$

where H_{ist} represents the self-reported overall health of person *i* in region (state or county) *s* at time *t*. We estimate linear probability models (LPMs) for two binary outcomes: reporting "excellent" health or not, and reporting "fair or poor" health or not.[9] Our primary independent variable of interest is denoted by I_{st}. It represents the Gini index and the Atkinson index for region *s* at time *t*.

We also include a vector of demographic characteristics denoted by X_{ist}. It consists of logarithmic group average household income stratified by age, gender and education in each state- or county-year (Ruhm, 2005),[10] age dummies, race dummies, education dummies, a gender dummy, a marital status dummy, and an indicator for respondents participating via cell phone only.[11] The regional unemployment rate in region *s* at time *t* is denoted by U_{st}. In addition, δ_s controls for regional fixed effects and λ_t controls for year-month fixed effects. Finally, ε_{ist} is the idiosyncratic error term. We cluster heteroskedasticity-robust standard errors at the state or county level, respectively, according to the level of reference groups.

3. Results

Table I presents our estimated effect of income inequality on individuals' self-reported health status. Overall the results, regardless of income distribution assumption, suggest that income inequality negatively affects self-reported health with the highest degree of statistical significance coming from counties, in terms of our geographic measures. The coefficients imply that a one standard deviation increase in the state income inequality measures reduces the probability of reporting excellent health by 1.96 to 2.00 percent.[12] A one standard deviation increase in the county income inequality measures reduces the probability of reporting excellent

[9] In the county-level analysis incorporating 2,342 county fixed effects with 144 year-month fixed effects, use of an LPM significantly reduces the computational burden relative to a nonlinear model such as probit or logit. As a robustness check, we compare the results from LPM and nonlinear models at the state level and they are very similar. The results using nonlinear models at the state level are available upon request.

[10] Ruhm (2005) conducts state-level analysis alone using the BRFSS, while we conduct both state and county level analysis.

[11] This "cell phone only" indicator has been available in the BRFSS since 2011, indicating whether the respondent exclusively use their cell phone to participate. Individuals who only use their cell phone could have different characteristics than others in the survey (Barbaresco et al., 2015; Courtemanche et al., 2018).

[12] For example, the estimated coefficient of the Gini index in column (1) is -0.147, implying that a one standard deviation increase in the state Gini index reduces the probability of reporting excellent health by 0.408 percentage points or 1.98 percent relative to the 20.65 percent average probability of reporting excellent health.

health by 1.58 to 1.67 percent, and increases the probability of reporting fair or poor health by 2.08 to 2.33 percent.

Table I. Effect of Income Inequality Indices on Health Outcomes

	Lognormal				Pareto			
	"Excellent"		"Fair or Poor"		"Excellent"		"Fair or Poor"	
Explanatory Variables	**State (1)**	**County (2)**	**State (3)**	**County (4)**	**State (5)**	**County (6)**	**State (7)**	**County (8)**
Gini	-0.147*	-0.066*	0.100	0.070***	-0.111*	-0.051*	0.073	0.054***
	(0.084)	(0.034)	(0.080)	(0.027)	(0.065)	(0.026)	(0.060)	(0.021)
Percent Change [a]	**-1.98%**	**-1.61%**	**1.69%**	**2.23%**	**-2.00%**	**-1.67%**	**1.66%**	**2.25%**
Atkinson (ε=0.5)	-0.142	-0.065*	0.087	0.072***	-0.104	-0.047*	0.060	0.053**
	(0.088)	(0.035)	(0.078)	(0.028)	(0.067)	(0.027)	(0.058)	(0.021)
Percent Change [a]	**-2.00%**	**-1.61%**	**1.54%**	**2.33%**	**-2.00%**	**-1.59%**	**1.45%**	**2.29%**
Atkinson (ε=1)	-0.101*	-0.046*	0.067	0.049***	-0.127*	-0.057**	0.090	0.059***
	(0.059)	(0.024)	(0.055)	(0.019)	(0.071)	(0.029)	(0.069)	(0.02)
Percent Change [a]	**-1.99%**	**-1.63%**	**1.66%**	**2.28%**	**-1.96%**	**-1.64%**	**1.75%**	**2.14%**
Atkinson (ε=2)	-0.097*	-0.044**	0.072	0.044***	-0.094*	-0.043*	0.062	0.046***
	(0.052)	(0.022)	(0.053)	(0.017)	(0.055)	(0.022)	(0.051)	(0.018)
Percent Change [a]	**-1.92%**	**-1.58%**	**1.80%**	**2.08%**	**-2.00%**	**-1.66%**	**1.67%**	**2.23%**

Robust standard errors clustered at the state or county level are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).
a. Percent changes from a one standard deviation change in the respective index.

Table II reports our estimated effects of logarithmic group average income on self-reported health, which strongly support the absolute income hypothesis that higher absolute income promotes health, thus being consistent with previous literature. It is worth noting that these estimated effects are completely robust across specifications using different inequality measures (i.e., Gini, Atkinson) and income distribution assumptions (i.e., Lognormal, Pareto), which is in line with prior literature suggesting that absolute income effects are not sensitive to various inequality measures (Lorgelly and Lindley, 2008).

Table II. Effect of Logarithmic Group Average Income on Health Outcomes

	"Excellent"		"Fair or Poor"	
	State	**County**	**State**	**County**
Explanatory Variable	**(1)**	**(2)**	**(3)**	**(4)**
Ln (Group Average Income)	0.049***	0.036***	-0.080***	-0.061***
	(0.009)	(0.002)	(0.006)	(0.002)

Robust standard errors clustered at the state or county level are in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

4. Discussion

In this paper we extend the income inequality literature by expressing both the Gini and Atkinson index as a function of the ratio of the mean to median household income. We also model individual income as following either the lognormal or Pareto distribution when using each index. This approach produces precise annual Gini and Atkinson indices at different levels of geography, thus solving the sample size problem in the literature by incorporating externally calculated inequality measures. In an application, we find preliminary evidence that greater income inequality negatively affects overall self-reported health in the United States.

Interest within the academic literature and the popular press on the consequences of income inequality has exploded in recent years. This is no doubt driven by the dramatic widening of the income distribution in both the United States as well as many other developed countries over the past 30 years (Piketty et al., 2017; Piketty and Saez, 2003; Boustan et al., 2013). In light of this, we believe our suggested approach will be of help to those interested in investigating the effect of increases in income inequality across a variety of research fields.

One potential limitation of our work is that estimating the effect of income inequality on health is challenging because the causal relationship might also run from health to inequality. However, our application focuses on a period in the U.S. with tremendous changes in economic conditions (including the economic boom from 2002 to 2006 and the Great Recession from 2006 to 2009), which one may argue could exogenously affect income inequality. We believe, therefore, such exogenous changes in inequality might reduce this concern to some degree.

References

Aitchison, J. and J.A.C. Brown (1957) *The Lognormal Distribution*, Cambridge University Press.

Atkinson, A.B. (1970) "On the measurement of inequality" *Journal of Economic Theory* **2**(3), 244-263.

Barbaresco, S., Courtemanche, C.J., and Y. Qi (2015) "Impacts of the Affordable Care Act dependent coverage provision on health-related outcomes of young adults" *Journal of Health Economics* **40**, 54-68.

Bechtel, L., Lordan, G., and D.S.P. Rao (2012) "Income inequality and mental health—empirical evidence from Australia" *Health Economics* **21**, 4-17.

Blakely, T.A., Lochner, K., I. Kawachi (2002) "Metropolitan area income inequality and self-rated health—a multi-level study." *Social Science & Medicine* **54**(1), 65-77.

Boustan, L., Ferreira, F., Winkler, H., E.M. Zolt (2013) "The effect of rising income inequality on taxation and public expenditures: Evidence from US municipalities and school districts, 1970–2000" *The Review of Economics and Statistics* **95**(4), 1291-1302.

Brush, J. (2007) "Does income inequality lead to more crime? A comparison of cross-sectional and time-series analyses of United States counties" *Economics Letters* **96**(2), 264-268.

Chang, V.W. and N.A. Christakis (2005) "Income inequality and weight status in US metropolitan areas" *Social Science & Medicine* **61**(1), 83-96.

Chen, Z. and C.A.G. Crawford (2012) "The role of geographic scale in testing the income inequality hypothesis as an explanation of health disparities." *Social Science & Medicine* **75**(6), 1022-1031.

Courtemanche, C., Marton, J., Ukert, B., Yelowitz, A., and D. Zapata (2018) "Early effects of the Affordable Care Act on health care access, risky health behaviors, and self-assessed health" *Southern Economic Journal* **84**(3), 660-691.

Crow, E.L. and K. Shimizu (1987) *Lognormal Distributions*, New York: Marcel Dekker.

Gibrat, R. (1931) *Les Inégalités économiques*, Librairie du Recueil Sirey, Paris.

Gravelle, H. and M. Sutton (2009) "Income, relative income, and self-reported health in Britain 1979–2000" *Health Economics* **18**(2), 125-145.

Kaplan, G.A, Pamuk, E.R., Lynch, J.W., Cohen, R.D., and J.L. Balfour (1996) "Inequality in income and mortality in the United States: analysis of mortality and potential pathways" *Bmj* **312**(7037), 999-1003.

Kelly, M. (2000) "Inequality and crime" *The Review of Economics and Statistics* **82**(4), 530-539.

Kennedy, B.P., Kawachi, I., D. Prothrow-Stith (1996) "Income distribution and mortality: cross sectional ecological study of the Robin Hood index in the United States". *The BMJ* **312**(7037), 1004-1007.

Kopczuk, W., Saez, E., and J. Song (2010) "Earnings inequality and mobility in the United States: evidence from social security data since 1937" *Quarterly Journal of Economics* **125**(1), 91-128.

Laporte, A. (2002) "A note on the use of a single inequality index in testing the effect of income distribution on mortality" *Social Science & Medicine* **55**(9), 1561-1570.

Lopez, R. (2004) "Income inequality and self-rated health in US metropolitan areas: a multi-level analysis" *Social Science & Medicine* **59**(12), 2409-2419.

Lorgelly, P.K. and J. Lindley (2008) "What is the relationship between income inequality and health? Evidence from the BHPS" *Health Economics* **17**(2), 249-265.

Lubrano, M. (2013) "The econometrics of inequality and poverty. Lecture 4: Lorenz curves, the Gini coefficient and parametric distributions" http://www.vcharite.univ-mrs.fr/PP/lubrano/poverty.htm. Accessed 1 August 2019.

Lynch, J., Smith, G.D., Harper, S.A., M. Hillemeier, N. Ross, G.A. Kaplan, and M. Wolfson (2004) "Is income inequality a determinant of population health? Part 1. A systematic review" *The Milbank Quarterly* **82**(1), 5-99.

Mellor, J.M. and J. Milyo (2002) "Income inequality and health status in the United States: evidence from the current population survey" *The Journal of Human Resources* **37**(3), 510-539.

Modalsli, J. (2015) "Inequality in the very long run: inferring inequality from data on social groups" *Journal of Economic Inequality* **13**(2), 225–247.

Pareto, V. (1896) *Cours d'economie politique*, Lausanne. (Vol. 1). F. Rouge.

Piketty, T., Saez, E., and G. Zucman (2017) "Distributional national accounts: methods and estimates for the United States" *Quarterly Journal of Economics* **133**(2), 553-609.

Piketty, T. (2013) *Capital in the twenty-first century*, Harvard University Press.

Piketty, T. and E. Saez (2003) "Income inequality in the United States, 1913–1998" *Quarterly Journal of Economics* **118**(1), 1-41.

Ruhm, C.J. (2005) "Healthy living in hard times" *Journal of Health Economics* **24**(2), 341-363.

Sarabia, J.M., Castillo, E., Pascual, M., and M. Sarabia (2005). "Mixture Lorenz curves" *Economics Letters* **89**(1), 89-94.

Sen, A. (1973) *On economic inequality*, Oxford University Press.

Sommeiller, E., Price, M., and E. Wazeter (2016). "Income inequality in the U.S. by state, metropolitan area, and county" *Economic Policy Institute*, 2.

Subramanian, S.V. and I. Kawachi (2004) "Income inequality and health: what have we learned so far?" *Epidemiologic Reviews* **26**(1), 78-91.

Sung, J., Qiu, Q., and J. Marton (2020) “Relative deprivation: a new derivation and application” *Applied Economics Letters*, 1-4.

Weich, S., Lewis, G., and S.P. Jenkins (2002) “Income inequality and self-rated health in Britain” *Journal of Epidemiology & Community Health* **56**(6), 436-441.

Wilkinson, R.G. and K.E. Pickett (2006) “Income inequality and population health: a review and explanation of the evidence” *Social Science & Medicine* **62**(7), 1768-1784.

Young, A. (2011) The Gini Coefficient for a Mixture of Ln-Normal Populations. The London School of Economics and Political Science, London, UK.

Appendix: Derivation for Mixture Gini Indices (Lognormal, Pareto)

In this appendix, we first briefly review the Gini index from a mixture of lognormal distribution done by prior studies. We then extend this work by expressing the lognormal mixture Gini index as a function of subgroup mean and median household incomes. Following that, we derive a mixture Gini index from Pareto distribution that can be expressed as a function of subgroup mean and median household incomes. Due to the limitation on data availability, we do not have data on mean and median household incomes of socioeconomic subgroups at the state- or county-level from reliable Federal data sources for an application of our derived mixture Gini indices. Future studies can apply our mixture Gini indices if such aggregate income data become available.

A1. Review and extension of Gini Index from a mixture of lognormal distributions

Prior studies have derived the Gini index from a mixture of lognormal distributions as follows (Young, 2011; Modalsli, 2015):

$$G = \sum_{i=1}^{M}\sum_{j=1}^{M} \frac{p_i p_j \bar{y}_i}{\bar{y}} \left(2\Phi\left(\frac{\mu_i - \mu_j + 0.5\sigma_i^2 + 0.5\sigma_j^2}{\sqrt{\sigma_i^2 + \sigma_j^2}} \right) \right) - 1 \tag{A1}$$

where M is the total number of socioeconomic subgroups, and i and j denote subgroups (rather than individuals). $\bar{y}_i$ is the mean household income of subgroup i, and $y \sim \ln N(\mu_i, \sigma_i^2)$ in every subgroup i.

We extend equation (A1) to a function of subgroup mean to median household incomes as equation (A2):

$$G = \sum_{i=1}^{M}\sum_{j=1}^{M} \frac{p_i p_j \bar{y}_i}{\bar{y}} \left(2\Phi\left(\frac{\ln\left(\frac{\bar{y}_i \bar{y}_j}{\ddot{y}_j^{\,2}}\right)}{\sqrt{2\ln\left(\frac{\bar{y}_i \bar{y}_j}{\ddot{y}_i \ddot{y}_j}\right)}} \right) \right) - 1 \tag{A2}$$

where $\bar{y}_j$ is the mean household income of subgroup j, and $\ddot{y}_i$ ($\ddot{y}_j$) is the median household income of subgroup i (j).

A2. Our derivation of Gini Index from a Mixture of Pareto distributions

We derive the Gini index following a mixture of Pareto distribution and express it as a function of subgroup mean and median household incomes.

For household income that follows Pareto distribution $y \sim Pareto\ (y_m, \alpha_i)$ in each subgroup i, we have the probability density function and cumulative density function as

$$f_i(y) = \frac{\alpha_i y_m^{\alpha_i}}{y^{\alpha_i+1}}, \qquad F_i(y) = 1 - \left(\frac{y_m}{y}\right)^{\alpha_i}$$

Therefore, the probability density function of the total population with M subgroups is:

$$f(y) = \sum_{i=1}^{M} p_i f_i(y) = \sum_{i=1}^{M} p_i \frac{\alpha_i y_m^{\alpha_i}}{y^{\alpha_i+1}}$$

Following that, the cumulative density function of the total population with M subgroups:

$$F(y) = \int_{y_m}^{y} f(u)du = \int_{y_m}^{y} \sum_{i=1}^{M} p_i f_i(u)\, du = \sum_{i=1}^{M} p_i \int_{y_m}^{y} f_i(u)du = \sum_{i=1}^{M} p_i F_i(y)$$

We have known the Lorenz Curve under the Pareto distribution in each subgroup i is:

$$L_i(F_i) = 1 - (1 - F_i)^{1-\frac{1}{\alpha_i}}$$

The Lorenz Curve under a mixture of Pareto distributions in the total population with M subgroups can be expressed as a weighted mean as well (Sarabia et al., 2005):

$$L(F) = \sum_{i=1}^{M} p_i L_i(F_i) = \sum_{i=1}^{M} p_i \left(1 - \left(1 - F_i(y)\right)^{1-\frac{1}{\alpha_i}}\right)$$

Therefore, we can calculate the Gini index under a mixture of Pareto distributions as:

$$\begin{aligned}
G &= 1 - 2\int_0^1 L(F)dF \\
&= 1 - 2\sum_{i=1}^{M} p_i \int_0^1 \left(1 - \left(1 - F_i(u)\right)^{1-\frac{1}{\alpha_i}}\right) d\left(\sum_{i=1}^{M} p_j F_j(u)\right) \\
&= 1 - 2\sum_{i=1}^{M}\sum_{j=1}^{M} p_i p_j \int_0^1 \left(1 - \left(1 - F_i(u)\right)^{1-\frac{1}{\alpha_i}}\right) dF_j(u) \\
&= 1 - 2\sum_{i=1}^{M}\sum_{j=1}^{M} p_i p_j \int_{y_m}^{\infty} \left(1 - \left(\frac{y_m}{u}\right)^{\alpha_i-1}\right) d\left(1 - \left(\frac{y_m}{u}\right)^{\alpha_j}\right) \\
&= 1 - 2\sum_{i=1}^{M}\sum_{j=1}^{M} p_i p_j \int_{y_m}^{\infty} \left(\left(\frac{y_m}{u}\right)^{\alpha_i-1}\right) d\left(\left(\frac{y_m}{u}\right)^{\alpha_j}\right) \\
&= 1 - 2\sum_{i=1}^{M}\sum_{j=1}^{M} p_i p_j\, {y_m}^{\alpha_i-1} {y_m}^{\alpha_j} \int_{y_m}^{\infty} u^{1-\alpha_i} du^{-\alpha_j} \\
&= 1 - 2\sum_{i=1}^{M}\sum_{j=1}^{M} p_i p_j\, {y_m}^{\alpha_i-1} {y_m}^{\alpha_j} (-\alpha_j) \int_{y_m}^{\infty} u^{-\alpha_i-\alpha_j} du
\end{aligned}$$

Solving the above equation, we obtain:

$$G = 1 - 2\sum_{i=1}^{M}\sum_{j=1}^{M} p_i p_j \frac{\alpha_i}{1-\alpha_i-\alpha_j} \tag{A3}$$

We rename i as j and j as i for symmetry and we get:

$$G = 1 - 2\sum_{i=1}^{M}\sum_{j=1}^{M} p_i p_j \frac{\alpha_j}{1-\alpha_i-\alpha_j} \tag{A4}$$

Averaging equation (A3) and (A4), we finish the derivation and obtain the Gini index under a mixture of Pareto distributions as:

$$G = 1 - \sum_{i=1}^{M}\sum_{j=1}^{M} p_i p_j \frac{\alpha_i+\alpha_j}{1-\alpha_i-\alpha_j} \tag{A5}$$

where for each subgroup $i(j)$,

$$\alpha_i = \frac{\ln 2}{\ln 2 + W(\theta_i)},\ W(\theta_i) \approx \theta_i - {\theta_i}^2 + \frac{3}{2}{\theta_i}^3 - \frac{8}{3}{\theta_i}^4 + \frac{125}{24}{\theta_i}^5,\ \theta_i = -\frac{\ln 2}{2} * \frac{\ddot{y}_i}{\bar{y}_i}$$