Published in The American Statistician 66, issue 4, 225-231, 2012
which should be used for any reference to this work

Histogram-Based Interpolation of the Lorenz Curve and Gini Index for Grouped Data

Yves TILLÉ and Matti LANGEL

In grouped data, the estimation of the Lorenz curve without taking into account the within-class variability leads to an overestimation of the curve and an underestimation of the Gini index. We propose a new strictly convex estimator of the Lorenz curve derived from a linear interpolation-based approximation of the cumulative distribution function. Integrating the Lorenz curve, a correction can be derived for the Gini index that takes the intraclass variability into account.

KEY WORDS: Binning; Class intervals; Income distribution; Inequality.

1. INTRODUCTION

According to the U.S. Census Bureau (DeNavas-Walt, Proctor, and Smith 2011), in 2010, the shares of aggregate incomes are distributed as follows. The poorest 20% of households earn 3.3% of the total aggregate income, while the poorest 80% of households earn 49.8% of the total income. The evolution of the shares is presented in Table 1. The curve that associates each proportion α of poorest to this share of income is called the Lorenz (1905) curve. This curve is the basic tool for the computation of several indices of inequality of income and wealth, the most well-known and broadly used of which is the Gini index (Gini 1914).

The Gini concentration index can be derived from the Lorenz curve. Larger Gini coefficients mean greater inequality of income. The Gini index is standardized in the sense where it varies between 0 and 1. Value 0 occurs when all the households have the same income, while value 1 occurs when only one household earns all the available income. The Gini index for the household income distribution in the United States in 2010 was estimated to be 0.469 (DeNavas-Walt, Proctor, and Smith 2011).

Although the Gini index is mostly used to measure the unequal allocation of income, its area of application is very wide, ranging from computer science to ecology or industrial concentration. Thus, both the Lorenz curve and the Gini index have generated an impressive amount of literature (for a review, see, e.g., Giorgi 1990, 1999; Cowell 2000; Xu 2004; Langel and Tillé 2013).

Inside this literature, a large set of articles is devoted to how bounds may be defined for the Lorenz curve and the Gini index when estimated from grouped data. Deriving a *lower (upper)* bound for the Lorenz curve generally allows for the derivation of an *upper (lower)* bound for the Gini index. The Lorenz curve and the Gini index can be estimated by assuming that all the incomes are equal within the groups. This reductive hypothesis leads to overestimation of the Lorenz curve (and an underestimation of the Gini index) and thus an upper bound for the Lorenz curve (Gastwirth 1972; Mehran 1975).

Different approaches exist for obtaining lower bounds of the Lorenz curve and, by extension, upper bounds of the Gini index. As shown by Ogwang (2006), proposed upper bounds for the Gini index differ mostly in the grouping information they require, such as the limits of each interval and the mean income of each class. Moreover, some approaches make assumptions on the shape of the underlying density function.

Gastwirth (1972) proposed distribution-free bounds for the Lorenz curve and the Gini index that involve a grouping-correction term (see also Fuller 1979; Gastwirth, Nayak, and Krieger 1986; Giorgi and Pallini 1986, 1987; Ogwang and Wang 2004), as well as improved bounds at an additional cost of only weak assumptions on the distribution function. The equivalence between different approaches is shown by Ogwang (2003). Mehran (1975) and Silber (1990) suggested approaches that do not require knowledge of the mean income or of the limits of the grouping intervals, while Krieger (1979) derived bounds for the Lorenz curve and the Gini index under the assumption that the probability density function is unimodal.

The Lorenz curve can also be directly modeled (Kakwani and Podder 1976) and estimated by modeling the distribution of the variable of interest (e.g., using Pareto, beta, log-normal, generalized-beta distributions). A review of expressions of the Lorenz curve and the Gini index for a variety of parametric distributions can be found in Giorgi and Nadarajah (2010). Income distributions, however, seem to have modeling-resistant distributions. Schader and Schmid (1994) showed that most parametric Lorenz curves produce unreliable estimates in empirical applications. Finally, Cowell and Mehta (1982) used a split-histogram technique to estimate various inequality measures, while Gastwirth and Glauberman (1976) and Schrag and Krämer (1993) had proposed to estimate the Lorenz curve using a Hermite interpolator. The latter method seems to give reasonable estimates for the Gini index.

Yves Tillé is Professor (E-mail: *yves.tille@unine.ch*) and Matti Langel is doctoral assistant (E-mail: *matti.langel@unine.ch*), Institute of Statistics, Faculty of Economics, University of Neuchâtel, Neuchâtel, Switzerland. The authors thank Monique Graf for her comments that helped improve the article, as well as the Editor, the Associate Editor, and an anonymous referee for insightful suggestions.

Table 1. Shares of aggregate income earned by poorest fractions of households (source: DeNavas-Walt, Proctor, and Smith 2011)

Year	20%	40%	60%	80%
2010	3.3	11.8	26.4	49.8
2000	3.6	12.5	27.3	50.3
1990	3.8	13.4	29.3	53.3
1980	4.2	14.4	31.2	55.9

Unfortunately, the underlying estimated Lorenz curve is not always convex.

In this article, after reviewing definitions, we use a very simple procedure to construct a realistic quadratic interpolation of the Lorenz curve. The histogram is used as the estimator of density. We then derive an estimator of the cumulative distribution function, a quantile function, and a quadratic interpolation for the Lorenz curve. Using this convex estimator of the Lorenz curve, we derive an estimator of the Gini index that contains a correction for the within-class variability.

Note that estimating the density via a histogram depends heavily on the number and length of classes of the available grouped data. Moreover, this issue is accentuated by skewness, which is a common feature of income distributions. The choice of the binning procedure when producing a histogram is a well-known issue (Sturges 1926; Doane 1976; Scott 1979) and a particular case of the bandwidth problem in nonparametric statistics. As it is clear that, for example, larger bins result in loss of information; the quality of estimation of the Lorenz curve or the Gini index is affected by the choice of class intervals. However, this issue is not addressed in this article because the binning procedure is viewed here as more of a constraint than a choice. Indeed, for confidentiality reasons, income data are often only available in grouped form with a given number of classes and given interval bounds.

2. LORENZ CURVE AND GINI INDEX IN INFINITE POPULATION

The Lorenz curve is the share of total income earned by the $100\alpha\%$ poorest. For example, the statement that the poorest 40% of households in the United States earned altogether 11.8% of the total income in 2010 is simply evaluating the Lorenz curve for the household income distribution of the United States at $\alpha = 0.4$.

More formally, let X be a positive and continuous income random variable with a strictly increasing cumulative distribution function $F(.)$. We assume that the expectation μ and the variance σ^2 exist. The quantile function is the inverse of $F(.)$:

$$Q(\alpha) = F^{-1}(\alpha), \alpha \in (0, 1).$$

The interpretation of the quantile is that $100\alpha\%$ of the population has an income less than or equal to $Q(\alpha)$ and $100(1-\alpha)\%$ of the population has an income larger than or equal to $Q(\alpha)$.

The Lorenz curve is defined by the ratio

$$L(\alpha) = \frac{\int_0^{Q(\alpha)} xdF(x)}{\int_0^{\infty} xdF(x)}. \quad (1)$$

It is the share of average income earned by the $100\alpha\%$ poorest, which in a finite population is also the total income of the $100\alpha\%$ poorest divided by the total income of the entire population. By posing $x = Q(p)$ (and thus $F(x) = p$), we can also define the Lorenz curve as

$$L(\alpha) = \frac{\int_0^{\alpha} Q(p)dp}{\int_0^{1} Q(p)dp}.$$

The numerator of the Lorenz curve is the incomplete (or partial) first moment evaluated at the quantity $Q(\alpha)$. The denominator of the Lorenz curve in (1) is equal to the expectation of X:

$$\int_0^1 Q(p)dp = \int_0^{\infty} xdF(x) = \mu.$$

When $F(.)$ is strictly increasing, the Lorenz curve is strictly convex, because its derivative is the quantile function divided by the mean, which is a strictly increasing function. Indeed,

$$\frac{dL(\alpha)}{d\alpha} = \frac{Q(\alpha)}{\mu}. \quad (2)$$

Figure 1 gives an example of the graphical representation of the Lorenz curve for a uniform and a lognormal distribution. It is commonly plotted together with the diagonal line, because, when everyone has the same income, the Lorenz curve is equal to the diagonal line. The diagonal line is thus the benchmark for the perfect equality of income.

The Lorenz curve is always below the diagonal line. The further the Lorenz curve is from the diagonal line, the greater is the level of inequality. The area between the curve and the line thus serves as a measure of inequality. Doubling this area directly gives the well-known Gini index that can also be defined

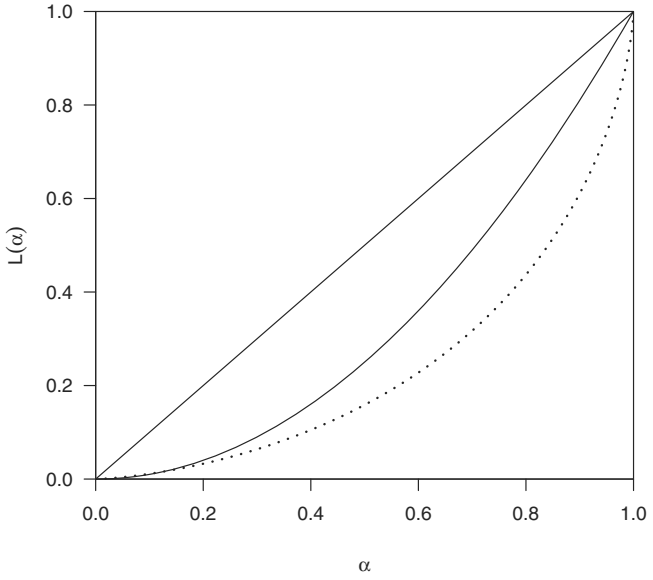

Figure 1. Lorenz curve for a uniform distribution ($a = 0, b$) and for a lognormal distribution (dotted curve) with parameters ($\mu = 0, \sigma^2 = 1$).

by means of the cumulative distribution function

$$G = 1 - 2\int_0^1 L(\alpha)d\alpha = \frac{\int_0^\infty \int_0^\infty |x-y|dF(x)dF(y)}{2\mu}. \quad (3)$$

The value of the Gini index lies between 0 (perfect equality) and 1 (perfect inequality). Perfect inequality means that only one person has a positive income and all the others have a null income. In this case, the Lorenz curve would be 0 and jump to 1 at $\alpha = 1$.

If X follows a known parametric distribution, the Lorenz curve and the Gini index can be derived as functions of these parameters (for a review, see Sarabia 2008; Giorgi and Nadarajah 2010). For example,

- if $X \sim \text{Uniform}(a, b)$:

$$L(\alpha) = \frac{2a\alpha + (b-a)\alpha^2}{a+b},$$
$$G = \frac{b-a}{3(a+b)},$$

- if $X \sim \text{Lognormal}(\mu, \sigma^2)$:

$$L(\alpha) = \Phi[(\Phi^{-1}(\alpha) - \sigma],$$
$$G = 2\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1,$$

where $\Phi(.)$ denotes the cumulative distribution function of the standard normal distribution. Thus, the Gini index for a Uniform $(a = 0, b)$ distribution is equal to 1/3 for any value of $b > 0$, while it is approximately equal to 0.52 for a Lognormal $(\mu, \sigma^2 = 1)$ distribution for any value of μ.

The Gini index and the Lorenz curve are invariant under scale changes because they are ratios between two quantities that linearly depend on the scale. These measures do not depend on whether the units are dollars or cents or euros. However, these measures are not invariant under location changes. This means, for instance, that if all the incomes are increased by a constant, the Gini index decreases, while if all the incomes are increased by the same proportion, the Gini index remains unchanged.

3. GROUPED DISTRIBUTION

Table 2 contains a small example of grouped income data of the U.S. Current Population Survey 2010 (U.S. Census Bureau 2011). The data consist of $n = 118{,}683$ observations grouped in $J = 4$ classes. We have advisedly regrouped the data in a small number of classes to better put the differences between the methods into evidence. We arbitrarily put the upper bound of the last class to 500,000. The mean computed by using the centers of classes is

$$\bar{x} = \sum_{j=1}^{J} f_j x_j^C = 72{,}904.9.$$

The following notations are used in Table 2 and throughout the article:

- J is the number of classes,
- n is the sample size,

Table 2. Grouped distribution of incomes of the U.S. Current Population Survey 2010

x_j^-	x_j^+	n_j	f_j	F_j	x_j^C	ℓ_j	L_j
0	49,999	59,831	0.50	0.50	24,999.5	50,000	0.17
50,000	99,999	34,618	0.29	0.80	74,999.5	50,000	0.47
100,000	199,999	19,607	0.17	0.96	149,999.5	100,000	0.81
200,000	500,000	4627	0.04	1.00	349,999.5	300,000	1.00

- x_j^- is the lower bound of class $j = 1, \ldots, J$,
- x_j^+ is the upper bound of class $j = 1, \ldots, J$, with $x_j^+ = x_{j+1}^-$,
- $x_j^C = (x_j^- + x_j^+)/2$ is the center of class $j = 1, \ldots, J$,
- n_j is the frequency of class $j = 1, \ldots, J$,
- $f_j = n_j/n$ is the relative frequency of class $j = 1, \ldots, J$,
- $F_j = \sum_{k=1}^{j} f_k$ is the cumulative relative frequency of class $j = 1, \ldots, J$ (with $F_0 = 0$ and $F_J = 1$),
- $\ell_j = x_j^+ - x_j^-$ is the length of class j,
- $\bar{x} = \sum_{j=1}^{J} f_j x_j^C$ is the mean computed by using the centers of classes, and
- $p_j = \int_{x_j^-}^{x_j^+} f(x)dx$ is the probability of interval $[x_j^-, x_j^+]$ in the population, where $f(x) = dF(x)/dx$ is the density of the underlying population.

In a grouped distribution, a common estimator of the Lorenz curve (see, e.g., Gastwirth 1972; Pyatt, Chen, and Fei 1980; Lerman and Yitzhaki 1989; Deltas 2003; Wodon and Yitzhaki 2003) is the continuous piecewise linear function connecting the points (F_j, L_j), where $L_0 = 0$, $L_J = 1$, and

$$L_j = \frac{1}{\bar{x}} \sum_{k=1}^{j} f_k x_k^C.$$

The empirical curve is thus defined by

$$\widehat{L}_1(\alpha) = \frac{1}{\bar{x}} \sum_{j=1}^{J} f_j x_j^C H\left(\frac{\alpha - F_{j-1}}{f_j}\right), \quad (4)$$

where $H(.)$ is the cumulative distribution function of a uniform [0, 1] distribution:

$$H(x) = \begin{cases} 0, & \text{if } x < 0, \\ x, & \text{if } 0 \le x < 1, \\ 1, & \text{if } x \ge 1. \end{cases}$$

Figure 2 shows a plot of the Lorenz curve $\widehat{L}_1(\alpha)$ for the data of Table 2.

This estimator derives from taking the following step function $\widehat{\widehat{F}}(.)$ as an estimator of the cumulative distribution function F:

$$\widehat{\widehat{F}}(x) = \begin{cases} 0, & \text{if } x < x_1^-, \\ F_j, & \text{if } x_j^- \le x < x_j^+,\ j = 1, \ldots, J, \\ 1, & \text{if } x \ge x_J^-. \end{cases} \quad (5)$$

Estimator $\widehat{L}_1(\alpha)$ is known to overestimate the Lorenz curve and is besides used by Mehran (1975) as an upper bound for $L(\alpha)$. If we apply definition (3) on (4), we obtain after some algebra

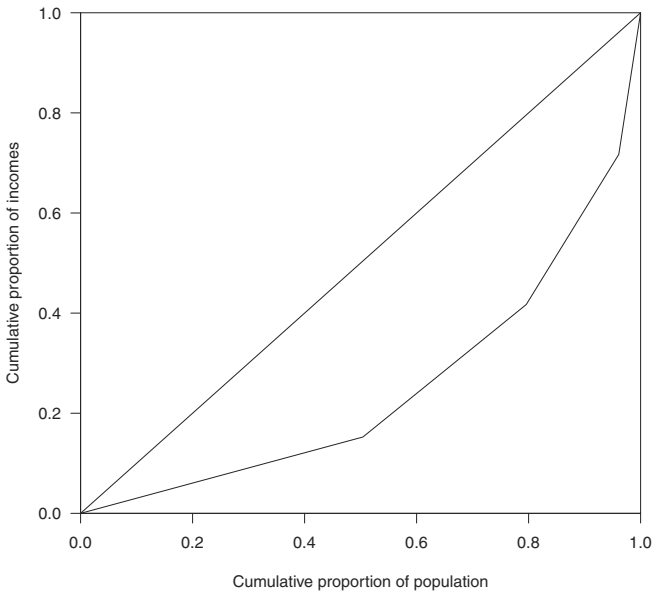

Figure 2. Estimated Lorenz curve $\widehat{L}_1(\alpha)$ for the dataset presented in Table 2.

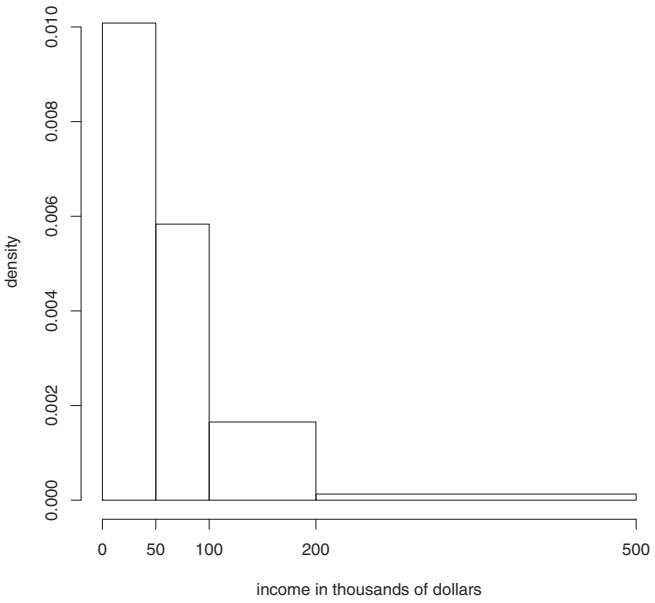

Figure 3. Histogram of the dataset presented in Table 2.

an estimator of the Gini index:

$$\widehat{G}_1 = \frac{1}{2\bar{x}} \sum_{j=1}^{J} \sum_{k=1}^{J} f_j f_k \left| x_j^C - x_k^C \right|,$$

which underestimates the population value because the within-class inequality is not accounted for. This estimate of the Gini index is $\widehat{G}_1 = 0.4414$.

4. ESTIMATION OF THE DENSITY, MEAN, AND VARIANCE

In grouped data, the most usual estimator of the density function is the histogram that is thus defined as

$$\widehat{f}(x) = \begin{cases} 0, & \text{if } x < x_1^-, \\ f_j/\ell_j, & \text{if } x_j^- \le x < x_j^+, \quad j = 1, \ldots, J, \\ 0, & \text{if } x \ge x_J^+. \end{cases}$$

Figure 3 shows a plot of the histogram or empirical density function $\widehat{f}$ for the data of Table 2. Integrating the estimated density distribution, we obtain an estimator of the cumulative distribution function by linear interpolation in the classes:

$$\begin{aligned} \widehat{F}(x) &= \int_0^x \widehat{f}(z)dz \\ &= \begin{cases} 0, & \text{if } x < x_1^-, \\ F_{j-1} + f_j \frac{x - x_j^-}{\ell_j}, & \text{if } x_j^- \le x < x_j^+, \quad j = 1, \ldots, J, \\ 1, & \text{if } x \ge x_J^+. \end{cases} \end{aligned} \tag{6}$$

While $\widehat{\widehat{F}}(.)$ in (5) is a step function, $\widehat{F}(x)$ is piecewise linear. Using the estimated density or the estimated cumulative distribution function, estimators of moments can be derived. The estimator of the mean is

$$\begin{aligned} \widehat{\mu} &= \int_{x_1^-}^{x_J^+} x d\widehat{F}(x) = \sum_{j=1}^{J} \int_{x_j^-}^{x_j^+} x d\widehat{F}(x) \\ &= \sum_{j=1}^{J} \frac{f_j}{\ell_j} \ell_j x_j^C = \sum_{j=1}^{J} f_j x_j^C. \end{aligned}$$

The estimator of the second moment is

$$\begin{aligned} \widehat{\mu}_2 &= \int_{x_1^-}^{x_J^+} x^2 d\widehat{F}(x) = \sum_{j=1}^{J} \int_{x_j^-}^{x_j^+} x^2 d\widehat{F}(x) = \sum_{j=1}^{J} \frac{f_j}{\ell_j} \int_{x_j^-}^{x_j^+} x^2 dx \\ &= \sum_{j=1}^{J} \frac{f_j}{\ell_j} \frac{\left[\left(x_j^+\right)^3 - \left(x_j^-\right)^3 \right]}{3} \\ &= \frac{1}{3} \sum_{j=1}^{J} f_j \left[\left(x_j^+\right)^2 + x_j^+ x_j^- + \left(x_j^-\right)^2 \right] \\ &= \sum_{j=1}^{J} f_j \left(x_j^C\right)^2 + \frac{1}{12} \sum_{j=1}^{J} f_j \ell_j^2, \end{aligned}$$

and enables us to propose an estimator of the variance

$$\widehat{\sigma}^2 = \widehat{\mu}_2 - \widehat{\mu}^2 = \sum_{j=1}^{J} f_j \left(x_j^C - \bar{x}\right)^2 + \frac{1}{12} \sum_{j=1}^{J} f_j \ell_j^2. \tag{7}$$

Note that the use of the density $\widehat{f}(.)$ automatically provides an estimator of the variance that includes a correction for the within-class variability. This estimator is the sum of the variance between the centers of classes and of the variance within the classes. Under simple random sampling with replacement, the alternate estimator

$$\widehat{\sigma}_A^2 = \frac{n}{n-1} \sum_{j=1}^{J} f_j \left(x_j^C - \bar{x}\right)^2 + \frac{1}{12} \sum_{j=1}^{J} f_j \ell_j^2$$

is, however, preferable because it unbiasedly estimates

$$\tilde{\sigma}^2 = \sum_{j=1}^{J} p_j \left(x_j^C - \mu\right)^2 + \frac{1}{12}\sum_{j=1}^{J} p_j \ell_j^2.$$

Given the loss of information that results from grouping the data, σ^2 cannot be unbiasedly estimated. However, part of the intraclass variability is recovered when using the proposed estimators. Moreover, if the distribution inside each class in the population is uniform, $\widehat{\sigma}_A^2$ is an unbiased estimator of σ^2. Relaxing this assumption, estimator $\widehat{\sigma}_A^2$ unfortunately remains biased with respect to σ^2, but its bias only depends on $\tilde{\sigma}^2 - \sigma^2$, that is, on the distribution of the variable within the classes. In the next section, the same histogram-based approach is used to propose estimators of the Lorenz curve and of the Gini index, which take the within-class variability into account.

5. ESTIMATION OF THE LORENZ CURVE AND THE GINI INDEX

Since the quantile function is the inverse of the cumulative distribution function, an estimator of the quantile can be defined as given below using expression (6). First, let α be the order of the quantile and k be the index of the class such that $F_{k-1} \leq \alpha < F_k$. For a grouped distribution, the quantile function can then be estimated by

$$\widehat{Q}(\alpha) = x_k^- + \frac{\alpha - F_{k-1}}{f_k}\left(x_k^+ - x_k^-\right),$$

with $\widehat{Q}(0) = x_1^-$ and $\widehat{Q}(1) = x_J^+$. By using this definition, we obtain an estimated Lorenz curve. Since the quantile function is piecewise linear, expression (2) directly implies that the Lorenz curve is piecewise quadratic.

$$\begin{aligned}
\widehat{L}_2(\alpha) &= \frac{1}{\bar{x}}\int_0^{\alpha} \widehat{Q}(p)dp \\
&= \frac{1}{\bar{x}}\left\{\sum_{j=1}^{k-1}\int_{F_{j-1}}^{F_j} \widehat{Q}(p)dp + \int_{F_{k-1}}^{\alpha} \widehat{Q}(p)dp\right\} \\
&= \frac{1}{\bar{x}}\left\{\sum_{j=1}^{k-1}\left[p x_j^- + \frac{\frac{p^2}{2} - pF_{j-1}}{f_j}\left(x_j^+ - x_j^-\right)\right]_{F_{j-1}}^{F_j}\right. \\
&\quad \left. + \left[p x_k^- + \frac{\frac{p^2}{2} - pF_{k-1}}{f_k}\left(x_k^+ - x_k^-\right)\right]_{F_{k-1}}^{\alpha}\right\} \\
&= \frac{1}{\bar{x}}\left\{\sum_{j=1}^{k-1}\left[f_j x_j^- + \frac{f_j}{2}\left(x_j^+ - x_j^-\right)\right]\right. \\
&\quad \left. + \left[(\alpha - F_{k-1})x_k^- + \frac{(\alpha - F_{k-1})^2}{2f_k}\left(x_k^+ - x_k^-\right)\right]\right\} \\
&= \frac{1}{\bar{x}}\left\{\sum_{j=1}^{k-1} f_j x_j^C + \left[(\alpha - F_{k-1})x_k^-\right.\right. \\
&\quad \left.\left. + \frac{(\alpha - F_{k-1})^2}{2f_k}\left(x_k^+ - x_k^-\right)\right]\right\}.
\end{aligned}$$

In particular, we have $\widehat{L}_2(0) = 0$ and $\widehat{L}_2(1) = 1$. Now, we can compute

$$\begin{aligned}
\int_{F_{k-1}}^{F_k} \widehat{L}_2(\alpha)d\alpha &= \frac{1}{\bar{x}}\left[f_k \sum_{j=1}^{k-1} f_j x_j^C + \frac{1}{6}f_k^2\left(2x_k^- + x_k^+\right)\right] \\
&= \frac{1}{\bar{x}} f_k \left[\sum_{j=1}^{k} f_j x_j^C - \frac{1}{6}f_k\left(x_k^- + 2x_k^+\right)\right].
\end{aligned}$$

Thus, the area under the estimated Lorenz curve is

$$\begin{aligned}
\int_0^1 \widehat{L}_2(\alpha)d\alpha &= \sum_{k=1}^{J}\frac{1}{\bar{x}} f_k \left[\sum_{j=1}^{k} f_j x_j^C - \frac{1}{6}f_k\left(x_k^- + 2x_k^+\right)\right] \\
&= \frac{1}{\bar{x}}\left[\left(\bar{x} - \sum_{j=1}^{J} f_j x_j^C F_j + \sum_{j=1}^{J} f_j^2 x_j^C\right)\right. \\
&\quad \left. - \frac{\sum_{j=1}^{J} f_j^2 x_j^C}{2} - \sum_{j=1}^{J}\frac{1}{6}f_j^2\frac{\ell_j}{2}\right] \\
&= \frac{1}{\bar{x}}\left[\left(\bar{x} - \sum_{j=1}^{J} f_j x_j^C F_j\right) + \frac{\sum_{j=1}^{J} f_j^2 x_j^C}{2}\right. \\
&\quad \left. - \sum_{j=1}^{J}\frac{1}{6}f_j^2\frac{\ell_j}{2}\right].
\end{aligned}$$

Finally, the estimated Gini index is derived from the above result and expression (3):

$$\begin{aligned}
\widehat{G}_2 &= 1 - 2\int_0^1 \widehat{L}_2(\alpha)d\alpha \\
&= 1 - 2\frac{1}{\bar{x}}\left[\left(\bar{x} - \sum_{j=1}^{J} f_j x_j^C F_j\right) + \frac{\sum_{j=1}^{J} f_j^2 x_j^C}{2}\right. \\
&\quad \left. - \sum_{j=1}^{J}\frac{1}{6}f_j^2\frac{\ell_j}{2}\right] \\
&= -1 + \frac{1}{\bar{x}}2\sum_{j=1}^{J} f_j x_j^C F_j - \frac{1}{\bar{x}}\sum_{j=1}^{J} f_j^2 x_j^C + \frac{1}{\bar{x}}\sum_{j=1}^{J} f_j^2\frac{\ell_j}{6} \\
&= \frac{1}{2\bar{x}}\sum_{j=1}^{J}\sum_{k=1}^{J} f_j f_k\left|x_j^C - x_k^C\right| + \frac{1}{\bar{x}}\sum_{j=1}^{J}\frac{f_j^2\ell_j}{6}.
\end{aligned}$$

Estimator $\widehat{G}_2$ includes a correction term that depends on the within-class variability. Note that for a uniform continuous random variable between x_j^- and x_j^+, the Gini index is equal to $\ell_j/(6x_j^C)$. The second term thus depends on the Gini index measured within the classes. Under simple random sampling with replacement, the alternate estimator

$$\widehat{G}_A = \frac{1}{2\bar{x}}\frac{n}{n-1}\sum_{j=1}^{J}\sum_{k=1}^{J} f_j f_k\left|x_j^C - x_k^C\right| + \frac{1}{\bar{x}}\sum_{j=1}^{J}\frac{\left(nf_j^2 - f_j\right)\ell_j}{6(n-1)}$$

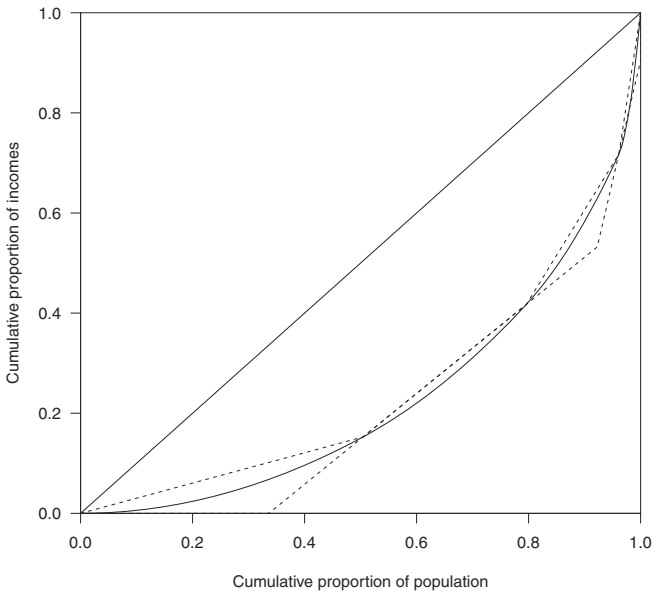

Figure 4. Estimated Lorenz curve $\widehat{L}_2(\alpha)$ and the two Mehran bounds (dotted lines).

is, however, preferable because $\bar{x}\widehat{G}_A$ unbiasedly estimates the population parameter

$$\frac{1}{2}\sum_{j=1}^{J} p_j p_k \left|x_j^C - x_k^C\right| + \sum_{j=1}^{J} \frac{p_j^2 \ell_j}{6},$$

as $\mathrm{E}(f_j^2) = p_i/n + p_i^2(n-1)/n$ and $\mathrm{E}(f_j f_k) = p_j p_j(n-1)/n$. Estimator $\widehat{G}_A$ unfortunately remains biased. This results from the fact that $\widehat{G}_A$ is a ratio with the random variable $\bar{x}$ at the denominator, and that it also depends on the distribution of the variable within the classes. Note that the numerator of the Gini index is unbiasedly estimated by $\bar{x}\widehat{G}_A$ under the assumption that the population inside each class is uniformly distributed.

Figure 4 shows the interpolation of the estimated Lorenz curve $\widehat{L}_2(\alpha)$ that lies between the two Mehran bounds. Note that the upper Mehran bound for the Lorenz curve is equal to $\widehat{L}_1(\alpha)$. The three Gini estimators for our dataset give $\widehat{G}_1 = 0.4414$, $\widehat{G}_2 = 0.4874$, and $\widehat{G}_A = 0.4874$. Gastwirth's (1972) upper bound for the Gini index is

$$\widehat{G}_U = \widehat{G}_1 + \frac{1}{\bar{x}}\sum_{j=1}^{J} \frac{f_j^2\left(x_j^- - x_j^C\right)\left(x_j^C - x_j^+\right)}{\ell_j} = 0.5565.$$

The difference between the estimates is thus far from being negligible. The estimators proposed in this article have the advantages of being intuitive, easy to compute, and directly coherent with the histogram as a density function. Moreover, the estimator of the Lorenz curve is strictly convex, which is in line with its infinite population definition.

REFERENCES

Cowell, F. A. (2000), "Measurement of Inequality," in *Handbook of Income Distribution* Vol. 1, eds. A. B. Atkinson and F. Bourguignon, Amsterdam: Elsevier, pp. 87–166. [225]

Cowell, F. A., and Mehta, F. (1982), "The Estimation and Interpolation of Inequality Measures," *Review of Economic Studies*, 49, 273–290. [225]

Deltas, G. (2003), "The Small-Sample Bias of the Gini Coefficient: Results and Implications for Empirical Research," *The Review of Economics and Statistics*, 85, 226–234. [227]

DeNavas-Walt, C., Proctor, B. D., and Smith, J. C. (2011), "Income, Poverty, and Health Insurance Coverage in the United States: 2010," Current Population Reports, Census Bureau, Washington, DC. [225]

Doane, D. P. (1976), "Aesthetic Frequency Classifications," *The American Statistician*, 30, 181–183. [226]

Fuller, M. (1979), "The Estimation of Gini Coefficients From Grouped Data: Upper and Lower Bounds," *Economics Letters*, 3, 187–192. [225]

Gastwirth, J. L. (1972), "The Estimation of the Lorenz Curve and Gini Index," *The Review of Economics and Statistics*, 54, 306–316. [225,227]

Gastwirth, J. L., and Glauberman, M. (1976), "The Interpolation of the Lorenz Curve and Gini Index From Grouped Data," *Econometrica*, 44, 479–483. [225]

Gastwirth, J. L., Nayak, T. K., and Krieger, A. M. (1986), "Large Sample Theory for the Bounds on the Gini and Related Indices of Inequality Estimated From Grouped Data," *Journal of Business & Economic Statistics*, 4, 269–273. [225]

Gini, C. (1914), "Sulla Misura della Concertrazione e della Variabilità dei Caratteri," *Atti del R. Istituto Veneto di Scienze, Lettere e Arti*, 73, 1203–1248. [225]

Giorgi, G. M. (1990), "Bibliographic Portrait of the Gini Concentration Ratio," *Metron*, 48, 183–221. [225]

——— (1999), "Income Inequality Measurement: The Statistical Approach," in *Handbook of Income Inequality Measurement*, ed. J. Silber, Norwell, MA: Kluwer Academic Publishers, pp. 245–267. [225]

Giorgi, G. M., and Nadarajah, S. (2010), "Bonferroni and Gini Indices for Various Parametric Families of Distributions," *Metron*, 68, 23–46. [225,227]

Giorgi, G. M., and Pallini, A. (1986), "Di Talune Soglie Inferiori e Superiori del Rapporto di Concentrazione," *Metron*, 44, 377–390. [225]

——— (1987), "About a General Method for the Lower and the Upper Distribution-Free Bounds on Gini's Concentration Ratio From Grouped Data," *Statistica (Bologna)*, 47, 171–184. [225]

Kakwani, N. C., and Podder, N. (1976), "Efficient Estimation of the Lorenz Curve and Associated Inequality Measures From Grouped Observations," *Econometrica*, 44, 137–148. [225]

Krieger, A. M. (1979), "Bounding Moments, the Gini Index and Lorenz Curve From Grouped Data for Unimodal Density Functions," *Journal of the American Statistical Association*, 74, 375–378. [225]

Langel, M., and Tillé, Y. (2013), "Variance Estimation of the Gini Index: Revisiting a Result Several Times Published," *Journal of the Royal Statistical Society*, Series A, 176, DOI: 10.1111/j.1467-985X.2012.01048.x. [225]

Lerman, R. I., and Yitzhaki, S. (1989), "Improving the Accuracy of Estimates of Gini Coefficients," *Journal of Econometrics*, 42, 43–47. [227]

Lorenz, M. O. (1905), "Methods of Measuring the Concentration of Wealth," *Publications of the American Statistical Association*, 9, 209–219. [225]

Mehran, F. (1975), "Bounds on the Gini Index Based on Observed Points of the Lorenz Curve," *Journal of the American Statistical Association*, 70, 64–66. [225,227]

Ogwang, T. (2003), "Bounds of the Gini Index Using Sparse Information on Mean Incomes," *Review of Income and Wealth*, 49, 415–423. [225]

——— (2006), "An Upper Bound of the Gini Index in the Absence of Mean Income Information," *Review of Income and Wealth*, 52, 643–652. [225]

Ogwang, T., and Wang, B. (2004), "A Modification of Silber's Algorithm to Derive Bounds on Gini's Concentration Ratio From Grouped Observations," *Statistica (Bologna)*, 64, 697–706. [225]

Pyatt, G., Chen, C.-N., and Fei, J. (1980), "The Distribution of Income by Factor Components," *The Quarterly Journal of Economics*, 95, 451–473. [227]

Sarabia, J. M. (2008), "Parametric Lorenz Curves: Models and Applications," in *Modeling Income Distributions and Lorenz Curves*, ed. D. Chotikapanich, New York: Springer, pp. 167–190. [227]

Schader, M., and Schmid, F. (1994), "Fitting Parametric Lorenz Curves to Grouped Income Distributions: A Critical Note," *Empirical Economics*, 19, 361–370. [225]

Schrag, H., and Krämer, W. (1993), "A Simple Necessary and Sufficient Condition for the Convexity of Interpolated Lorenz Curves," *Statistica (Bologna)*, 53, 167–170. [225]

Scott, D. (1979), "On Optimal and Data-Based Histograms," *Biometrika*, 66, 605–610. [226]

Silber, J. (1990), "On a New Algorithm to Derive Bounds on Gini's Concentration Ratio From Grouped Observations," *Statistica (Bologna)*, 50, 215–220. [225]

Sturges, H. (1926), "The Choice of a Class-Interval," *Journal of the American Statistical Association*, 21, 65–66. [226]

U.S. Census Bureau. (2011), *Current Population Survey, Annual Social and Economic (ASEC) Supplement*, Washington, DC: Author. [227]

Wodon, Q., and Yitzhaki, S. (2003), "The Effect of Using Grouped Data on the Estimation of the Gini Income Elasticity," *Economics Letters*, 78, 153–159. [227]

Xu, K. (2004), "How Has the Literature on Gini's Index Evolved in the Past 80 Years?" Working Papers Archive, Department of Economics, Dalhousie University. [225]