JOURNAL OF ECONOMIC DEVELOPMENT
Volume 19, Number 2, December 1994

Gini Coefficient, Lorenz Curves, and Lorenz Dominance Effect:
An Application to Korean Income Distribution Data*

Hakchung Choo**
and
Hang Keun Ryu

A unique Gini coefficient can be obtained from a given Lorenz curve, but transformation from the given Gini coefficient to the Lorenz curve is not unique. A class of Lorenz curves will produce the same Gini coefficient so that measuring income inequality with a Gini coefficient and interpreting its result should be done with caution. In this paper, by expanding the Lorenz curve in a power series expansion, we show that knowledge of the Gini coefficient is equivalent to information of the zeroth order integration of the Lorenz curve. Therefore, the difference in conveyed information between the Gini coefficient and the Lorenz curve is that the Lorenz curve gives information of first, second, and higher order integration of the Lorenz curve while the Gini coefficient does not. If we expand the Lorenz curve in a Legendre series, then knowledge of the Gini coefficient is equivalent to the zeroth order coefficient of the Legendre series. If extra information is known, such as the first order integration of the Lorenz curve, then the ambiguity in constructing the Lorenz curve can be partially removed. We apply our results to measure income inequality in Korea for 1988, and discuss inaccurate interpretation associated with the Gini coefficient.

I. Introduction

There are two basic issues in studies on income distribution. The first is search for causes of income inequality. Observed size distribu-

[*] Comment of the referee was very helpful in revising this paper.

[**] Korea Development Institute and Dept. of Economics, Chung-Ang University, respectively.

tion of income depends on individual capability and on the social configuration of any given economy. It might be thought that measurable proxies for capability should be normally distributed, while clearly empirical income graduations are J-shaped distributions with non-zero skewness and kurtosis, cf. Sahota (1978), Creedy (1985), and Ryu and Slottje (1993) for discussion of this work. The second issue is related with construction of a Lorenz curve and search for the functional form of the Lorenz curve and a most widely used measure of income inequality.

In this paper, we focus on the second issue. In principle, if there are T income samples independently observed from an identical distribution, then there are various way to find the underlying probability density function. See for example, Nadayara's (1965) non-parametric density estimation, orthonormal basis estimation method (Prakasa Rao (1983)), maximum entropy estimation method (Zellner and Highfield (1988) and Ryu (1993)). Once the density function is approximated, we no longer need any more information. We can derive a Lorenz curve, a Gini coefficient, or any other income inequality measure using the approximated density function.

However, the tradition is not to derive the underlying density function, but to represent income inequality with a single summary measure. As a result, many different income distributions might yield the same summary measure. For example, when two Lorenz curves intersect, it is not clear that improvement in the summary measure is improvement in income inequality. It is possible that the lowest quintile of a society with higher Gini coefficient can receive more accumulated income compared to the lowest quintile of another society with lower Gini coefficient. Comparing these two societies, improvement in Gini coefficient may or may not be a desirable change for the economy.

We shall introduce a power series expansion for the Lorenz curve because the required convexity condition of the Lorenz curve can be established with ease, and graphical comparison of various Lorenz curves are relatively easy. As an alternative way to expand the Lorenz curve in a series, we can introduce a Legendre polynomial series. The Lorenz curve is decomposed into several orthogonal Legendre components, of which the zeroth order coefficient corresponds to the Gini coefficient. The first order component corresponds to redistribution of income between the individuals which is orthogonal to the lower order component. The Gini coefficient specifies the area under the Lorenz curve so that it can be considered as a normalization constant. However the mean of Lorenz curve, which can be derived from the Lorenz curve, describes the degree of fat tailedness of J shaped curve. A related study can be found in Ryu and Slottje (1992b).

In section two we discuss the theoretical part of this paper. We introduce the power series expansion and the Legendre series expansion for the Lorenz curves. We discuss the relationship between the Lorenz curve and Gini coefficient. In section three we actually perform experiments using measured family income observations of Korea. It is followed by a concluding remarks.

II. The Theory

There are two fundamental issues related with parametric representation of the Lorenz curve. Why should we care for the parametric form of the Lorenz curve when the empirical Lorenz curve already exist? By approximating the empirical Lorenz curve with some parametric functional form and by introducing the estimated parameters into the approximated form, the best we can hope is that the approximated Lorenz curve will be very closely located to the empirical Lorenz curve. However, the justification for using a parametric functional form comes from parsimony of representation. Suppose we have a large number of sample observations, and we want to report the Lorenz curve to another person. One possible way is to report all the values at each point. This is very difficult but accurate way. Another way is to represent the Lorenz curve with a simple parametric functional form and report the parameter values. The trade off is that we are losing some information by approximating the empirical Lorenz curve with certain functional form. However the positive aspect is that we can easily reproduce the reported result.

The second way requires a good simple parametric functional form to approximate the true unknown Lorenz curve. Basmann, Hayes, Johnson, and Slottje (1990) established a nonlinear functional form while Ryu and Slottje (1992a) suggested expansion of the income function in an exponential series so that the Lorenz curve can be represented as an integration of this income function. However, in this paper, we directly represent the Lorenz curve in a power series expansion or in a Legendre series expansion because we intend to demonstrate some relationship between the Gini coefficient and the Lorenz curve. See Choo (1982) for related discussion.

A. Power Series Expansion of the Lorenz Curve

An N^{th} order power series expansion is defined as

$$(2.1) \quad L_N(z) = a_1 + a_1 z + a_2 z^2 + \cdots + a_N z^N$$

where $z = (0, 1)$ is a population variable. Since the true Lorenz curve is a continuous smooth function of z, $L_N(z)$ converges in L^2 to the unknown true Lorenz curve when N goes to infinity and the parameters were chosen properly. Here we are restricting the functional form of the Lorenz curve to a class of the power series, but such restriction will not produce any finite departure because a power series expansion is a complete set. See Appendix for definition of completeness.

Let us define the following power integration of the Lorenz curve.

$$(2.2) \quad \mu_n \equiv \int_0^1 z^n L_N(z)dz = \int_0^1 z^n [a_0 + a_1 z + a_2 z^2 + \cdots + a_N z^N]dz$$

If $L_N(x)$ is known, we can compute $\mu_0,\ldots,\mu_N,\ldots$ and alternatively if $\mu_0,\ldots, \mu_N$ is given we can compute $a_0, a_1,\ldots, a_N$.

Now let us explain the relationship between the Gini coefficient and the Lorenz curve. Once the Lorenz curve is known, computation of the Gini coefficient is immediate.

$$(2.3) \quad \text{Gini} \equiv 1 - 2\int_0^1 L(z)dz \doteq 1 - 2\int_0^1 L_N(z)dz = 1 - 2\mu_0$$

where L(z) is the true unknown Lorenz curve and $L_N(z)$ is the N^{th} order approximation. As the size of series becomes large, the difference between L(z) and $L_N(z)$ becomes negligible. The knowledge of Gini is equivalent to the knowledge of the zeroth integration of the Lorenz curve, in this case it is important to note that no more information is provided by the Lorenz curve so that a class of Lorenz curves corresponding to different values of $\mu_1,\ldots, \mu_N$ will produce the same Gini coefficient. Since the Lorenz curve is a convex function, and we defined μ_n with (2.2), there are certain rules connecting the values of μ_0, $\mu_1,\ldots,\mu_N$.

However, search for the restricting conditions in $\mu_0, \mu_1,\ldots, \mu_N$ space seems to be difficult. In the following, we impose a sufficient condition for convexity of the Lorenz curve in the parameter space.

$L_N(z)$ is a covex function if $a_0 \geqslant 0,\ldots,a_N \geqslant 0$

We provide several examples assuming the Gini coefficient is 0.25. In Example 4, we relax this restriction with $0 \leqslant \text{Gini} \leqslant 0.5$.

Example 1: Suppose $L_1(z) = a_0 + a_1 z$ and we do not impose the boundary conditions $L_1(0) = 0$ and $L_1(1) = 1$. One extreme case is 25% of the population has no income and the remaining 75% of the population has equal share of the total income. This is plotted in Fig. 1a. Another extreme case is when 75% of the total income is uniformly distributed to everyone and the remaining 25% goes to the richest individual. This is plotted in Fig. 1b. From the knowledge of the Gini coefficient, we can not distinguish Fig. 1a from Fig. 1b. However, the first moment of Lorenz curve $\mu_1 \equiv \int_0^1 zL(z)dz$ will be bigger for the distribution of Fig. 1b.

Example 2: Suppose $L_2(z) = a_0 + a_1 z + a_2 z^2$ and we impose the boundary conditions $L_2(0) = 0$ and $L_2(1) = 1$ and convexity conditions, $a_1,\ a_2 \geqslant 0$. We have three equations and three variables. Thus the Lorenz curve is

$$L_2(z) = 0.25z + 0.75z^2$$

This curve is plotted in Fig. 1c. We assumed the Gini coefficient $G = 0.25$ and derived a convex $L_2(z)$, but the Gini coefficient should not be too large if we want to derive a convex Lorenz curve. For example, if $G = 0.5$, a convex Lorenz curve can not be established which satisfies both the boundary and convexity conditions. From $L(0) = 0$, we get $a_0 = 0$, from $L(1) = 1$, we get $a_1 + a_2 = 1$. From $L'(z) = a_1 + 2a_2 z > 0$, we get $L'(0) = a_1 > 0$. Also note $L''(z) = 2a_2 > 0$. From the definition of $\mu_0 = \int_0^1 L(z)dz = a_1/2 + a_2/3 = (1 - G)/2$. We get $0 \leqslant a_2 (= 3\text{Gini}) \leqslant 1$. Then Gini coefficient should be less than one third if the Lorenz curve is to satisfy the convexity condition and the given functional form.

Examples 3: $L_3(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3$. Now impose the boundary conditions $L_3(0) = 0$ and $L_3(1) = 1$, the Gini restriction (2.3) and convexity conditions, $a_1,\ a_2,\ a_3 \geqslant 0$. We have three equations and four variables. Let $\text{Gini} = 0.25$.

$$L_3(0) = a_0 = 0$$

$$L_3(1) = a_1 + a_2 + a_3 = 1$$

$$\mu_0 = \int_0^1 L_3(z)dz = \frac{a_1}{2} + \frac{a_2}{3} + \frac{a_3}{4}$$

$$\mu_0 = \frac{1 - \text{Gini}}{2} = \frac{3}{8}$$

Removing a_3 and imposing positivity of a_1 and a_2,

$$0.25 \leqslant a_1 \leqslant 0.5 \qquad a_2 = 1.5 - 3a_1 \qquad a_3 = 1 - a_1 - a_2.$$

If $a_1 = 0.25$, then a $_2 = 0.75$ and $a_3 = 0$. If $a_1 = 0.5$, then $a_2 = 0$ and $a_3 = 0.5$. This is plotted in Fig. 1d ($L(z) = 0.5z + 0.5z^3$).

Example 4: Let $0 \leqslant G \leqslant 0.5$ and $L(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3$ with boundary conditions $L(0) = 0$ and $L(1) = 1$ and convexity conditions, a_1, a_2, $a_3 \geqslant 0$. We have three equations and five variables (a_0, a_1, a_2, a_3, G).

$$L(0) = a_0 = 0$$

$$L(1) = a_1 + a_2 + a_3 = 1$$

$$\int_0^1 L(z)dz = \frac{a_1}{2} + \frac{a_2}{3} + \frac{a_3}{4} = (1 - G)/2$$

Therefore

$$1 - 3G \leqslant a_1 \leqslant 1 - 2G, \qquad a_2 = 3(1 - a_1 - 2G),$$

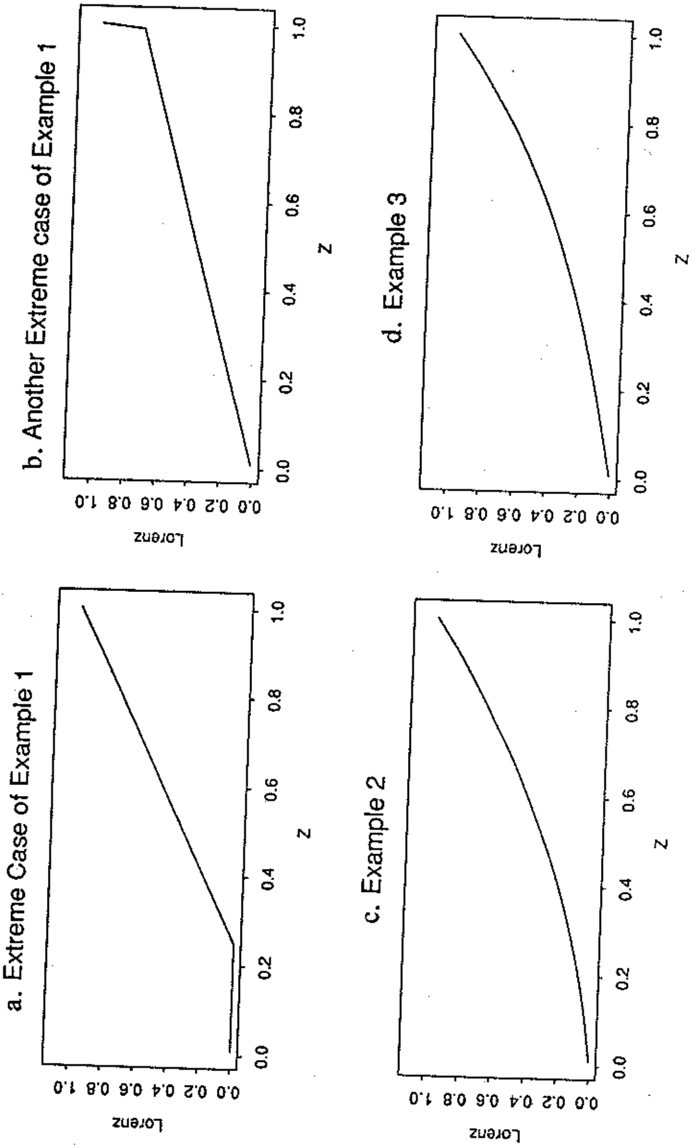
$$a_3 = 1 - a_1 - a_2$$

Plotting the corresponding Lorenz curves will be difficult because we have two degree of freedom (five variables with three equations).

In this subsection, we have expanded the Lorenz curve with a polynomial series. We have seen that information contained with the Gini coefficient is equivalent to the zeroth order moment of the Lorenz curve.

However, it is important to note that we are facing a potentially crucial problem in this method. Suppose we are given with a large number of sample observations of individual income. From this, we can calculate the empirical Lorenz curve and the Gini coefficient. However, if we approximate the above empirical Lorenz curve with a lower order polynomial series, for example, $L_3(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3$, the Gini coefficient derived by $1 - 2\int_0^1 L_3(z)$ will be different from the sample estimated Gini coefficient. Since L_3 is just an approximation of the true unknown Lorenz curve, and that $\int L_3(z)dz \neq \int L(z)\,dz$. The justification for using a complete set is that when the size of series is very large, the difference between the approximated function and the unknown true function can be made as small as possible. But for a

Figure 1
Graph of the Lorenz Curves


a. Extreme Case of Example 1

b. Another Extreme case of Example 1

c. Example 2

d. Example 3

finite sample and finite series expansion, approximation can be very rough.

Furthermore, when we increase the size of the series, the estimated parameters will fluctuate so that the estimated Gini coefficient $= 1-2\int_0^1 L_N(z)dz$ will also fluctuate. Therefore, our claim that we can approximate the Lorenz curve with a polynomial series expansion and the Gini coefficient can be obtained from the zeroth order integral may not be very meaningful if the calculated Gini coefficient depends on the size of series expansion of $L_N(z)$. In the following subsection, we introduce a Legendre series where the orthogonality condition of the given sequences will provide stable parameter estimation and stable Gini values when we increase the size of series.

B. Legendre Polynomial Series Expansion of the Lorenz Curve

The definition of the Legendre polynomial series is will explained in Arfken (1985), but a brief review for the concept of the orthonormal basis and completeness is provided in the Appendix. Originally, the Legendre series is defined on a domain $-1 \leqslant x \leqslant +1$, but for convenience of calculation, we made linear transformation for the original Lorenz curves with $0 \leqslant z \leqslant 1$ such that we have

(2.4)
$$\begin{aligned} P_0(z) &= 1 \\ P_1(z) &= \sqrt{3}(2z-1) \\ P_2(z) &= \sqrt{5}(6z^2-6z+1) \\ P_3(z) &= \sqrt{7}(20z^3-30z^2+12z-1) \\ P_4(z) &= \sqrt{9}(70z^4-140z^3+90z^2-20z+1) \end{aligned}$$

These functions are plotted in Fig. 2a – Fig. 2e. The orthonormality condition of the Legendre condition means

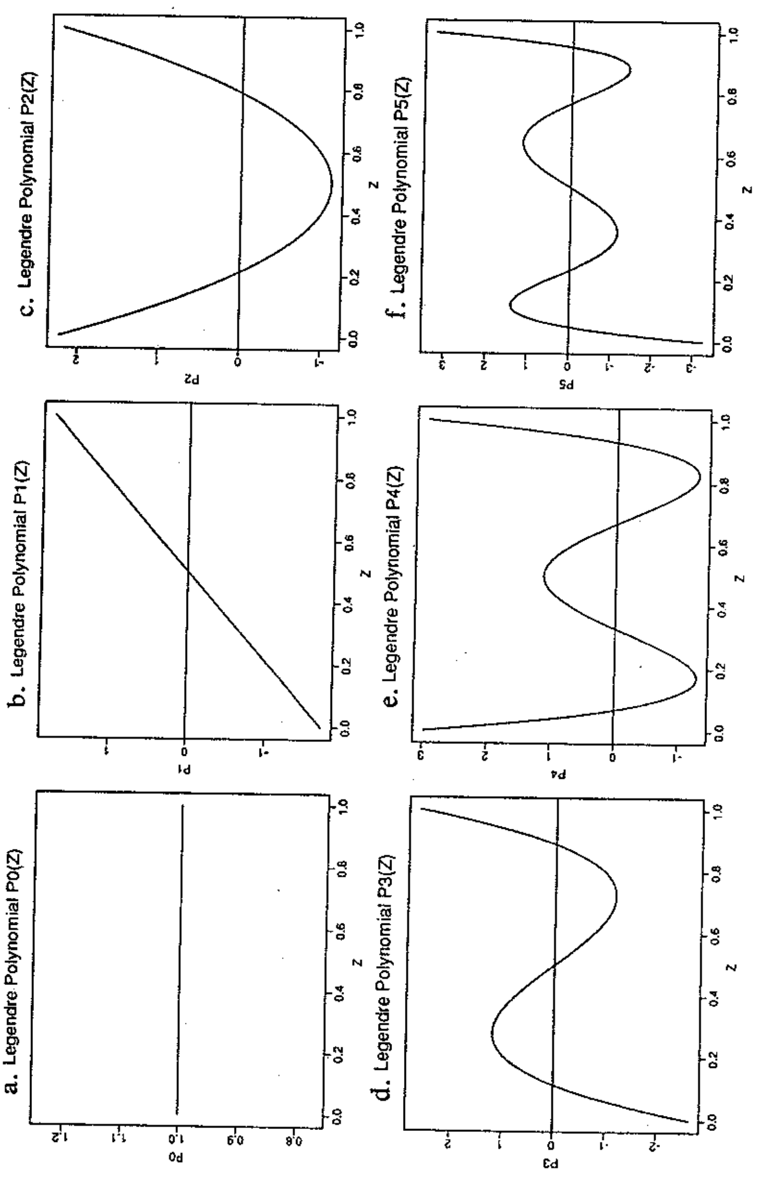
(2.5)
$$\int_0^1 P_m(z)P_n(z)dz = \delta_{mn}$$

where $\delta_{mn} = 1$ if $m = n$ and zero otherwise. The boundary conditions of the Legendre series are

$$\begin{aligned} P_0(0) &= 1 & P_0(1) &= 1 \\ P_1(0) &= -\sqrt{3} & P_1(1) &= \sqrt{3} \\ P_2(0) &= \sqrt{5} & P_2(1) &= \sqrt{5} \\ P_3(0) &= -\sqrt{7} & P_3(1) &= \sqrt{7} \end{aligned}$$

Figure 2

Graph of the Legendre Functions



Now expand the Lorenz curve with a Legendre series,

$$(2.6) \quad R_N(z) = a_0P_0(z) + a_1P_1(z) + a_2P_2(z) + \cdots + a_NP_N(z)$$

The Gini coefficient determines the zeroth coefficient, a_0.

$$(2.7) \quad a_0 = (1 - G)/2 \text{ because } \int_0^1 L(z)dz \doteq \int_0^1 R_N(z)dz = a_0$$

There is a big difference between (2.3) and (2.7). In (2.3), the calculated Gini coefficient changes its value when we increase the size of series expansion N, but not in (2.7). The orthogonality condition of the Legendre series (2.5) provides such convenience so that the zeroth decomposition of the Lorenz curve with the Legendre series will provide equivalent information to the Gini coefficient.

Once the Gini coefficient is given, a primitive Lorenz curve can be obtained by imposing the two boundary conditions $R(0) = 0$ and $R(1) = 1$, then

$$R_2(z) = a_0P_0(z) + a_1P_1(z) + a_2P_2(z)$$

can be computed using the boundary conditions

$$a_0 = (1 - G)/2, \quad a_1 = \frac{1}{2\sqrt{3}}, \quad \text{and } a_2 = \frac{G}{2\sqrt{5}}$$

$$(2.8) \quad R_2(z) = (\frac{1-G}{2})P_0(z) + \frac{1}{2\sqrt{3}} P_1(z) + \frac{G}{2\sqrt{5}} P_2(z)$$

or introducing the Legendre functions stated in (2.4).

$$R_2(z) = 3Gz^2 + (1 - 3G)z$$

Therefore $R_2(z)$ is a convez function if $0 \leqslant G \leqslant 1/3$. This is the same restriction stated in example 2.

Now let us introduce a third order Legendre series for the Lorenz curve. For given G, the basic functional form for our Lorenz curve is

$$(2.9) \quad R_3(z) = a_0P_0(z) + a_1P_1(z) + a_2P_2(z) + a_3P_3(z)$$
$$= (\frac{1 - G}{2})P_0(z) + a_1P_1(z) + \frac{G}{2\sqrt{5}} P_2(z) + \frac{(0.5 - \sqrt{3}a_1)}{7} P_3(z)$$

Since we have four variables a_0, a_1, a_2, a_3 and three equation

$a_0=(1-G)/2$, $R_3(0)=0$, and $R_3(1)=1$, we represented (2.9) as a function of a_1.

We have described the Lorenz curve for the given information of the Gini coefficient and the boundary conditions. Though we have derived the second order Legendre series (2.8) and the third order Legendre series (2.9), we can not go to any higher order Legendre series unless more information about the shape of the Lorenz curve is given.

C. Parametric Approximation of Empirical Lorenz Curve

In the subsection, we discuss how to approximate the empirical Lorenz curve with a Legendre series. Here we assume knowledge of the empirical Lorenz curve at all points but we shall not impose the boundary condition for convenience of exposition.

$$L(z)=a_0P_0(z)+a_1P_1(z)+a_2P_2(z)+\cdots+a_NP_N(z)+\varepsilon \tag{2.10}$$

There are other well known Lorenz curves. Basmann, Hayes, Johnson, and Slottje (1990) approximated the Lorenz curve with the following nonlinear form,

$$L(z)=z^{az+b}\exp[-g(1-z^2)-h(1-z)+\mu]$$

Ryu and Slottje (1992a) applied Gastwirth's (1971) definition of Lorenz curve. Suppose there are T observed incomes, $\{x_1,\ldots,x_T\}$, assumed to be independently drawn from a distribution F(x). Let $z\equiv F(x)=\int_0^x f(x')dx'$ and we define the inverse distribution function as $F^{-1}(w)=\inf_x(x;\ F(x)\geqslant w)$. The Lorenz curve is defined as

$$L(z)\equiv\frac{1}{\mu}\int_0^z F^{-1}(w)dw \text{ for } z\in(0,1)$$

where $\mu=\int_X xf(x)dx$. As a particular case, Ryu and Slottje (1992a) introduced an exponential series for the inverse function.

$$\hat{L}(z)=\frac{1}{\hat{\mu}}\int_0^Z \exp[\hat{\beta}_0+\hat{\beta}_1 w+\hat{\beta}_2 w^2+\hat{\beta}_3 w^3]dw \text{ for } z\in(0,1)$$

As an alternative way to represent the Lorenz curve, this paper introduced a series expansion for the Lorenz curve stated in (2.10). To estimate the parameters of (2.10), define a matrix

$$(2.11)\qquad X=\begin{bmatrix} P_0(z_1) & P_1(z_1) & \cdots & P_N(z_1)\\ P_0(z_2) & P_1(z_2) & \cdots & P_N(z_2)\\ \vdots & \vdots & \ddots & \vdots\\ P_0(z_T) & P_1(z_T) & \cdots & P_N(z_T)\end{bmatrix}$$

where $z_1, z_2,\ldots,z_T = 1/T, 2/T,\ldots, 1$. The orthogonality condition (2.5) produces

$$\frac{1}{T}(X'X)_{mn}=\frac{1}{T}\sum_{t=1}^{T}P_m(z_t)P_n(z_t)\doteq\int_0^1 P_m(z)P_n(z)dz=\delta_{mn}$$

where $\doteq$ means approximately equal to and δ_{mn} is zero if $m\neq n$ and one if $m=n$. Therefore, the least squares parameter estimation method of (2.10) will be

$$(2.12)\qquad \hat{a}_n=(X'X)^{-1}X\hat{L}(z)\doteq\frac{1}{T}\sum_{t=1}^{T}P_n(z_t)\hat{L}(z_t)$$

where $\hat{L}(z_t)$ is the empirical Lorenz curve.

It is important to note that parameter estimation does not depend on the size of the series expansion. The parameters estimated by (2.12) is uniquely determined by the given empirical Lorez curve and the chosen Legendre series. We are decomposing the Lorenz curve in a sequence of Legendre series and each term will project out its component from the Lorenz curve and the orthogonality condition guarantees independence of each projection. As a particular example of this orthogonality condition, we have indicated that the Gini coefficient which can be obtained from $a_0=(1-G)/2$ does not depend on the size of the chosen series.

To summarize this section, we expanded the Lorenz curve in a simple polynomial series and derived the relationship between the Gini coefficient and the Lorenz curve. However, such relationship depended on the size of the chosen series so that we introduced a Legendre series where estimated parameters does not depend on the size of the series. To derive a Lorenz curve from the given Gini coefficient, we can not include many terms inside the series because there is no way to estimate the parameters. However, if we have access to the empirical Lorenz curve, then we can approximate the unknown Lorenz curve with a Legendre series which asymptotically converges to the unknown true Lorenz curve as we increase the size of the series.

III. Application

We want to compare the performance of various approaches. We applied Korean cross sectional family income data for 1988. The Korea Development Institute (KDI) surveyed total household income for 5,111 families (excluding Cheju island). Among these families, 4,613 heads of family agreed to meet specially trained surveyors and answered specific items. Others refused to be visited. We also divided these family incomes into 100 income classes because it is customary not to use full sample data as it involves various complicated problems. See Basmann et. al. (1990) for more details.

In the first column in Table 1, z represents the population index. If z=0.2, it represents the 20th poorest percentile out of the 100. Similarly, if z=0.9, it represents the 10th richest percentile out of the 100. The second column is the actual empirical Lorenz curve, but in the next two columns, the Lorenz curves are estimated using the ordinary least squares (OLS) method and the orthonormal basis (ONB) method.

$$(2.10)\quad L(z) = a_0P_0(z) + a_1P_1(z) + a_2P_2(z) + \cdots + a_NP_N(z) + \varepsilon$$

In the OLS method, the parameters are estimated by

$$(2.12)\quad \hat{a}_n = (X'X)^{-1}X\hat{L}(z)$$

In the ONB method, the parameters are estimated by

$$\hat{a}_n = \frac{1}{T}\sum_{t=1}^{T} P_n(z_t)L(z_t)$$

The difference in the parameter estimation methods produced little difference in the estimated values. Both OLS and ONB methods seems to produce good approximation. However, some departure is observed at the tail area, z=0.95 and z=1.

The Table 2, we report the result of the Lorenz curve derived by the following approach.

$$(2.8)\quad R_2(z) = [\frac{1-G}{2}]P_0(z) + \frac{1}{2\sqrt{3}}P_1(z) + \frac{G}{2\sqrt{5}}P_2(z)$$
$$= 3Gz^2 + (1-3G)z$$

If we introduced sample Gini (0.4016) which is bigger than one third,

Table 1

COMPARISON OF LORENZ CURVES BASED ON EMPIRICAL METHOD, OLS METHOD, AND ONB METHOD

z^a	Lorenz[b]	*OLSL*[c]	ONBL[d]
0.05	0.0039	0.0054	0.0055
0.10	0.0138	0.0109	0.0121
0.15	0.0290	0.0243	0.0263
0.20	0.0484	0.0439	0.0464
0.25	0.0695	0.0683	0.0710
0.30	0.0942	0.0961	0.0990
0.35	0.1223	0.1267	0.1297
0.40	0.1542	0.1593	0.1625
0.45	0.1892	0.1939	0.1972
0.50	0.2281	0.2304	0.2339
0.55	0.2701	0.2692	0.2731
0.60	0.3135	0.3111	0.3154
0.65	0.3608	0.3569	0.3619
0.70	0.4135	0.4081	0.4138
0.75	0.4718	0.4662	0.4728
0.80	0.5378	0.5331	0.5408
0.85	0.6090	0.6111	0.6199
0.90	0.6921	0.7027	0.7127
0.95	0.7961	0.8107	0.8220
1.00	1.0000	0.9383	0.9508

Notes: [a] The population index is represented by z. If $z=0.2$, it represents the 20th poorest percentile out of the 100.

[b] Lorenz means the empirical Lorenz curve.

[c] For the Lorenz curve $L(z)=a_0P_0(z)+a_1P_1(z)+\cdots+a_4P_4(z)+\epsilon$, the parameters are estimated by the OLS method.

[d] For the Lorenz curve $L(z)=a_0P_0(z)+a_1P_1(z)+\cdots+a_4P_4(z)+\epsilon$, the parameters are estimated by the ONB method, $\hat{a}_n=\frac{1}{T}\Sigma_{t=1}^{T}P_n(z_t)\hat{L}(z_t)$.

the Lorenz curve (2.8) will not be convex. Therefore, the only way to make (2.8) a convex function is to introduce Gini$=0.333$. Since the performance of R_2 is not very good, we increase flexibility. We extend our series to the third order.

$$(2.9)\qquad R_3(z)=[\frac{1-G}{2}]P_0(z)+a_1P_1(z)+\frac{G}{2\sqrt{5}}P_2(z)+\frac{(0.5-\sqrt{3}a_1)}{7}P_3(z)$$

Table 2

COMPARISON OF EMPIRICAL LORENZ CURVES WITH THOSE DERIVED BY THE SECOND ORDER (R_2) AND THIRD ORDER LEGENDRE SERIES

z	Lorenz	ONBL	$R_2{}^a$	$R_3{}^b$
0.05	0.0039	0.0055	0.0025	0.0185
0.10	0.0138	0.0121	0.0100	0.0357
0.15	0.0290	0.0263	0.0225	0.0503
0.20	0.0484	0.0464	0.0400	0.0652
0.25	0.0695	0.0710	0.0625	0.0807
0.30	0.0942	0.0990	0.0900	0.0977
0.35	0.1223	0.1297	0.1225	0.1171
0.40	0.1542	0.1625	0.1600	0.1398
0.45	0.1892	0.1972	0.2025	0.1667
0.50	0.2281	0.2339	0.2500	0.1988
0.55	0.2701	0.2731	0.3025	0.2368
0.60	0.3135	0.3154	0.3600	0.2818
0.65	0.3608	0.3619	0.4225	0.3347
0.70	0.4135	0.4138	0.4900	0.3962
0.75	0.4718	0.4728	0.5625	0.4674
0.80	0.5373	0.5408	0.6400	0.5492
0.85	0.6090	0.6199	0.7225	0.6425
0.90	0.6921	0.7127	0.8100	0.7481
0.95	0.7961	0.8220	0.9025	0.8669
1.00	1.0000	0.9508	1.0000	1.0000

Notes: [a] The Lorenz curve R_2 is derived by using the following functinal form.

$$(2.8)\quad R_2(z) = [\frac{1-G}{2}]P_0(z) + \frac{1}{2\sqrt{3}}P_1(z) + \frac{G}{2\sqrt{5}}P_2(z)$$

To impose convexity, we need Gini coefficient be smaller than or equal to 0.333. Terefore, we introduced this value rather than observed value of 4.016.

[b] The Lorenz curve R_3 is derived by the following functional form.

$$R_3(z) = a_0P_0(z) + a_1P_1(z) + a_2P_2(z) + a_3P_3(z)$$

$$(2.9)\quad = [\frac{1-G}{2}]P_0(z) + a_1P_1(z) + \frac{G}{2\sqrt{5}}P_2(z) + \frac{(0.5-\sqrt{3a}_1}{\sqrt{7}}P_3(z)$$

This approach applies extra information of $\hat{a}_1$ compared to the R_2 approach. We find that performance of R_3 is quite satisfactory because extra information of $\hat{a}_1$ removes most of the uncertainty involved in the derivation of the Lorenz curve. We used OLS method to estimate a_1 in (2.9).

In the following we compare the performance of three approaches, ONB method, R_2 expansion, and R_3 expansion using the sum of the squared residual criteria.

$$\text{RSS(ONB)} = \sum_{t=1}^{T}[\hat{L}(z_t) - L_{ONB}(z_t)]^2 = 0.0100$$

$$\text{RSS(R2)} = \sum_{t=1}^{T}[\hat{L}(z_t) - L_{R_2}(z_t)]^2 = 0.3608$$

$$\text{RSS(R3)} = \sum_{t=1}^{T}[\hat{L}(z_t) - L_{R_3}(z_t)]^2 = 0.0849$$

where $\hat{L}(z)$ is the empirical Lorenz curve. We see R_2 is based only upon one parameter $\hat{a}_0$(i.e., Gini), R_3 is based upon two parameters $\hat{a}_0$(i.e., Gini) and $\hat{a}_1$, and the ONB is based upon a_0(i.e., Gini), a_1, a_2, a_3. As a result, the sum of the squared residual decreased as we increase knowledge of parameters.

VI. Conclusion

Our objective has been to establish a simple relationship between the Gini coefficient and the underlying true Lorenz curve. When we expanded the Lorenz curve in a Legendre series, its zeroth order integration corresponded to knowledge of the Gini coefficient. The higher order integration explains the income redistribution process between income groups. The zeroth order integration of the Lorenz curve is equivalent to the normalization constraint so that the area under the curve be constant. The first order component describes the mean of the Lorenz curve so that skewness and fat tailedness of the J curve is described by this component, related study can be found in Ryu and Slottje (1992b). They derived the probability density function (i.e., the share function), which can be defined as the normalized income density function, with an exponential Legendre series. The higher order coefficient corresponded to the redistribution of income between the groups.

Though there exists a relationship between the Gini coefficient and the underlying true Lorenz curve, such relationship will not hold when we replace the truce Lorenz curve with an approximated function. Furthermore, the estimated parameters will change when we increase the size of the polynomial series. However, by introducing an orthonormal polynomial series in the Lorenz curve, the estimated parameters were stable with respect to the size of the series. Therefore the Gini coefficient could be established as a function of the approximated Lorenz curve by expanding the true unknown Lorenz curve in a Legendre series.

Appendix

Review of Mathematical Concepts of Completeness, Orthonormality, and Basis

Let us review the concepts of completeness, orthonormality, and basis. Let (X, B, μ) be a measure space where X is a compact space, the set B is a Borel-sigma algebra, and μ is the Lebesgue measure. Define $L^2(X)$ to be the set of all Borel measurable real valued functions f whose squares are integrable on X, i.e., $\int_X |f(x)|^2 dx < \infty$. The set $L^2(X)$ is a normed linear space with norm of $||f(x)|| = [\int_X |f(x)|^2 dx]^{1/2}$. There exists in $L^2(X)$ a countable set of elements, which is everywhere dense in the sense that every element of $L^2(X)$ can be approximated arbitrarily closely by elements belonging to this set. In other words, $L^2(X)$ is separable.

An orthonormal sequence satisfies

$$\int_X P_n(x) P_m(x) dx = \delta_{nm}, \quad n,m, = 0,\ 1,\ 2,\ldots$$

where $d_{nm} = 1$ if n = m and zero otherwise.

The orthogonal sequence $\{P_n\}$ in the space $L^2(X)$ is called *complete* if there is no element $f \neq 0$ of $L^2(X)$ which is orthogonal to all the elements P_n. In other words, for a complete system of equalities

$$\int_X f(x) P_n(x) dx = 0 \ (n = 0,1,2,\ldots)$$

and for $f(x) \in L^2(X)$, it follows that $f(x) = 0$ for almost all $x \in X$.

Suppose we have a countable set of elements in a sequence: g_0, g_1, $g_2, \ldots, g_n, \ldots$. We can then construct an orthonormal sequence by the Gram-Schmidt procedure. From the given sequence we delete the function 0 (if it occurs) and all those functions which can be expressed as linear combinations of the preceding ones. When proper normalization is performed, the resulting sequence $\{P_0(x), P_1(x), P_2(x),\ldots\}$ is orthonormal.

$$P_0(x) = g_0(x) / ||g_0(x)||$$

$$Q_n(x) = g_n(x) - \sum_{i=0}^{n-1} P_i(x) \int_X g_n(x)P_i(x)dx, \quad n \geq 1$$

$$P_n(x) = Q_n(x) / ||Q_n(x)||$$

If the orthonormal sequence $\{P_0(x), P_1(x), P_2(x),\ldots\}$ satisfies the completeness condition, it is called as an orthonormal basis.

We provide two examples of sequences and their corresponding ONBS.

1) Trigonometric expansion: Suppose we have a sequence $\{1, \cos x, \sin x,\ldots, \cos nx, \sin nx,\ldots\}$ for $x \in [-\pi, +\pi]$. Since they are orthogonal, normalization will produce a complete ONB:

$$P_0(x) = \frac{1}{\sqrt{2\pi}}, \; P_1(x) = \frac{\cos x}{\sqrt{\pi}}, \; P_2(x) = \frac{\sin x}{\sqrt{\pi}}, \; P_3(x) = \frac{\cos 2x}{\sqrt{\pi}},$$

$$P_4(x) = \frac{\sin 2x}{\sqrt{\pi}}, \ldots$$

2) Legendre expansion: Suppose we have a sequence $\{1, x, x^2,\ldots\}$ for $x \epsilon [-1, +1]$, then the Gram-Schmidt orthogonalization will produce a complete ONB:

$$P_0(x) = \frac{1}{\sqrt{2}}, \; P_1(x) = \sqrt{\frac{3}{2}}\, x, \; P_2(x) = \sqrt{\frac{5}{8}}(3x^2 - 1), P_3(x) =$$

$$\sqrt{\frac{7}{8}}(5x^3 - 3x), \ldots$$

When we change the domain of x to [0,1], we get the Legendre polynomials stated in section 2.2.

References

Arfken, G., *Mathematical Methods for Physicists,* Academic Press, 1985.

Basmann, R.K., Hayes, K., Johnson, J. and D. Slottje, "A General Function Form for Approximating the Lorenz Curve," *Journal of Econometrics,* 43, 1990, 77-90.

Choo, H., *Income Distribution and Its Determinants in Korea,* Appendix 1, in Choo, H. (ed.), *On the*

Measure of Income Inequality, the Korean Development Institute, 1982.

Creedy, J., *Dynamics of Income Distribution,* Basil Blackwell, Oxford, 1985.

Gastwirth, J., "A General Definition of the Lorenz Curve," *Econometrica,* 39, 1971, 1037-1039.

Nadaraya, E., "On Non-Parametric Estimation of Density Functions and Regression Curves," *Theory Prob. Applic.,* 10, 1965, 186-190.

Prakasa Rao, B., *Nonparametric Functional Estimation,* Academic Press, Orlando, 1983.

Ryu, H., "Maximum Entropy Estimation of Density and Regression Functions," *the Journal of Econometrics,* 56, 1993, 397-440.

Ryu, H. and D. Slottje, "Two Flexible Functional Form Approaches for Approximating the Lorenz Curve," Working Paper, Southern Methodist University, 1992a.

——— and ———, "Another Perspective on Recent Changes in the U.S. Income Distribution: An Index Space Representation," Working Paper, Southern Methodist University, 1992b.

Sahota, G., "Theories of Personal Income Distribution: A Survey," *Journal of Economic Literature,* 16, 1, 1978, 1-55.

Zellner, A. and R. Highfield, "Calculation of Maximum Entropy Distributions and Approximation of Marginal Posterior Distributions," *Journal of Econometrics,* 37, 1988, 195-209.