Modelling Inequality with a Single Parameter[1]

J. M. Henle,[2] N. J. Horton, and S. J. Jakus
Smith College

Abstract *We argue that the Lorenz curve for income is well-modelled by a member of the one-parameter family of functions:*

$$\{y = (1 - (1 - r)^k)^{\frac{1}{k}}\}.$$

We justify this statement with data from the Luxembourg Income Study. The family of curves arises from a dynamic model of income growth, in which the parameter k has a direct economic interpretation.

0. Introduction

The unequal distribution of a resource is captured in all its variety by the Lorenz curve which charts, given the rank r $(0 \leq r \leq 1)$ of an individual (based on the individual's level of the resource in a given population), the proportion $L(r)$ of the resource belonging to all those of lower rank.

In theory, the Lorenz curve, and hence inequality in a society, is a multifaceted phenomenon. The curve is subject only to the constraints that it passes through $(0, 0)$ and $(1, 1)$ and that its derivative is non-decreasing. In practice, however, real Lorenz curves appear to follow a very distinct pattern and in nearly every case is well-modelled by a member of a one-parameter family of curves, the Lamé curves:

$$\{y = (1 - (1 - r)^k)^{\frac{1}{k}}\}.$$

In section 1 we introduce our family of curves and use it to model Lorenz curves for a number of countries and years, chiefly for income data.

In section 2 we develop two economic models based on "trickle-up" theories. Both yield Lamé curves.

In section 3 we consider a number of reality checks on our model and its consequences.

[1] The authors would like to thank Peter Lambert for a careful reading and many helpful suggestions.

[2] The first author would like to dedicate his part in this paper to his father ([He]).

In section 4 we pose a few questions.

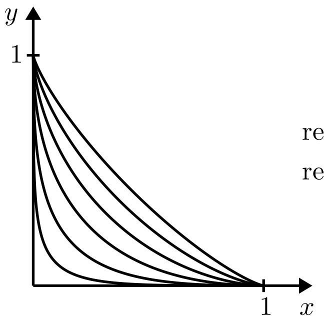
For the most part, we will restrict our attention to inequality of income. The Luxembourg Income Study (LIS) provides excellent income data for many countries and many years. Data on the distribution of wealth are less reliable or comparable.

We will use r to denote the rank ($0 \leq r \leq 1$) of a family in terms of its income and $I(r)$ for the income of a family at rank r. We will use N for the number of families and A for the aggregrate income of all families (i.e. $A = N \int_0^1 I(x)\,dx$). The Lorenz curve, $L(r)$, is the fraction of income earned by families of rank $\leq r$, that is, $L(r) = \frac{\int_0^r I(x)\,dx}{\int_0^1 I(x)\,dx} = \frac{N}{A} \int_0^r I(x)\,dx$. Alternatively, we can write: $I(r) = \frac{A}{N} L'(r)$. For background on this and inequality in general, see [L].

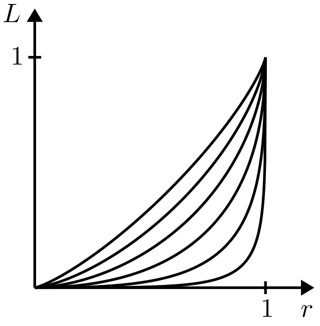
1. Modelling the Lorenz Curve

The curves, $\{x^k + y^k = 1\}$ are a special case of the more general Lamé curves, $\{\left(\frac{x}{a}\right)^k + \left(\frac{y}{b}\right)^k = 1\}$. These were studied by the Danish engineer and designer Piet Hein and have been called, when $k > 1$, "superellipses."

Our family of functions results from taking $a = b = 1$ and $k < 1$,



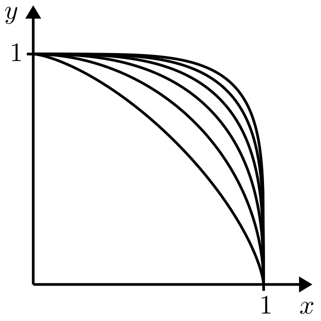
replacing x with $1 - r$ and
replacing y with L,



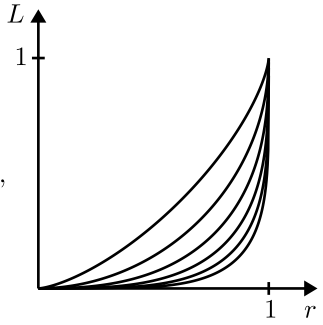
giving us

$$L(r) = (1 - (1 - r)^k)^{1/k}.$$

A second family can be formed by taking $a = b = 1$ and $k > 1$,



replacing x with r and
replacing y with $1 - L$,



giving us

$$L(r) = 1 - (1 - r^k)^{1/k}.$$

Others have proposed more elaborate versions of this family. Sarabia, Castillo, and Slottje [SCS], for example, explore among other families

$$L(r) = (1 - (1 - r^k)^{\gamma}.$$

Necessarily, because of the additional freedom, they achieve a better fit. McDonald [McD] catalogued a hierarchy of probability models (ranging from one to four parameters) for the size distribution of income. We are struck, however, by how good a fit is possible without that freedom. In addition, for many situations where data are limited (i.e. estimates are available only at the decile level), it is less clear that the additional flexibility introduced with more than one parameter is worth the cost in potential overfitting and increased complexity of interpretation.

We tested our family on 89 sets of data from LIS, each set consisting of decile data for a country and year, specifically, Austria (4 years), Australia (4 years), Belgium (4 years), Canada (8 years), Denmark (4 years), Finland (4 years), France (4 years), Germany (8 years), Ireland (4 years), Israel (4 Years), Italy (3 years), Mexico (6 years), Netherlands (4 years), Norway (4 years), Taiwan (4 years), Sweden (7 years), United Kingdom (8 years), and the United States (5 years). In addition, we took two sets for the United States used by [S] from [R] which in addition to decile points included values at $r = .91, .92, \ldots .99$. The years considered ranged from 1967 to 2000; median: 1991. The nonlinear regression function in *Stata* version 9.1 was used for estimation.

The results for the 91 observations are impressive:

Variable	Mean	Std. Dev.	Min.	Max.
root mean square error (MSE)	.0043318	.0032447	.0004838	.0200169
mean absolute deviation	.003455	.0026484	.0003716	.0153033
maximum absolute deviation	.0074414	.0060083	.0007958	.0446752

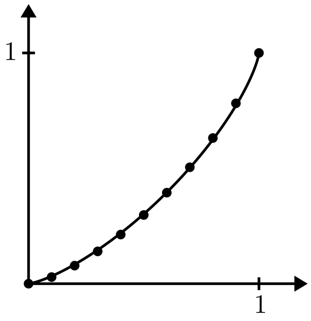
The average root mean square error for the models overall was 0.0043. The maximum absolute deviation of the predicted value from any observed value was

0.045, and the largest MSE for any country/year combination was 0.020 (Italy in 1991 for both).[3] The largest MSE for any other country/year combination (not including US Sarabia) was 0.010 (US 1991) with corresponding maximum absolute deviation of 0.018.

35.2% of the models fit for each country/year combinations yielded a max absolute deviation of less than 0.005; 80.2% were always within 0.01 of the observed value.

The proportion of variance accounted for by the single parameter model was quite high (all R^2 values ≥ 0.998). While the addition of a second parameter may lead to a statistically significant better fit, it is less clear whether this is of practical significance.

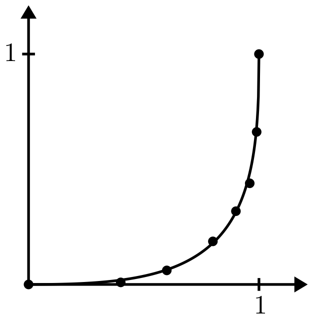
Below is a typical example, LIS data for Canada in 1997, with the graph of $y = (1-(1-r)^{.752})^{1/.752}$.



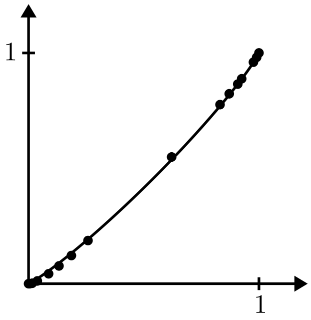
The results from the one parameter model explain 99.99% of the variability, with mean absolute deviation of 0.0046 and maximal deviation of 0.0098. A plot of the residuals indicates that while these deviations are of relatively small magnitude, the primary lack of fit is due to the symmetry assumption of the one-parameter model. The two parameter model of Sarabia and colleagues provides an even better fit (mean absolute deviation of 0.0003 and maximal deviation of 0.0008) but at the cost of potentially overfitting the data, and with less readily interpretable parameters.

Wealth, which is more unequally distributed than income, was also well modelled by members of this family of curves. Shown below is wealth data for the United States in 1983 from [W] with $y = (1-(1-r)^{.417})^{1/.417}$.

[3] The numbers for Italy in 1991 are suspect. The first decile is negative (-.01), the only example in all the data from LIS. In addition, the first decile was positive five years earlier (+.03) and positive again four years later (+.02).



Education, which is more equally distributed than income, again fits the pattern. Below is data for years of educational attainment among U.S. citizens 15 years and older, modelled by $y = (1 - (1 - r)^{.8862})^{1/.8862}$.



The data, from the U.S. Census Bureau [U], was broken down into enough categories to yield 16 points on the Lorenz curve (see Appendix A).

2. Modelling the Redistribution of Wealth

One justification for our one-parameter family is its success in matching real Lorenz curves. We have a second justification. The family of curves is the solution to a simple dynamic model of income growth which we present here.

We start by viewing I and L as functions of two variables, rank and time. We imagine income rising or falling for each family, which in turn affects the Lorenz curve.

We also adopt a sort of “trickle-up” theory. This theory posits that families earn money off families of lower rank. The wealthier a family is, the faster its income will grow. In other words, we assume that $\frac{\partial I}{\partial t}$ is related to $\frac{\frac{A}{N}L}{N(1-r)} = \frac{AL}{N^2(1-r)}$. $\frac{A}{N}L$ is the aggregate income of families of lower rank. $N(1 - r)$ is the number of families of higher rank (a family at rank r must share development rights on poorer families with all richer families).

What is the relation? It seems reasonable to assume that $\frac{AL}{N^2(1-r)}$, and $\frac{\partial I}{\partial t}$ are simultaneously zero or simultaneously non-zero. Thus we can allow for considerable possibilities by assuming that their logarithms satisfy a linear equation. This leads to

$$\log\left(\frac{\partial I}{\partial t}\right) = B\log\left(\frac{AL}{N^2(1-r)}\right) + C,$$

or,

$$\frac{\partial I}{\partial t} = e^C\left(\frac{AL}{N^2(1-r)}\right)^B,$$

for some constants B and C.

Now, for a small interval of time Δt, we have $\Delta I = e^C\left(\frac{AL}{N^2(1-r)}\right)^B \Delta t$. For a fixed rank r, we have:

$$L + \Delta L = \frac{\int_0^r (I+\Delta I)\,dx}{\int_0^1 (I+\Delta I)\,dx} = \frac{\frac{A}{N}L + \int_0^r e^C\left(\frac{AL}{N^2(1-x)}\right)^B \Delta t\,dx}{\frac{A}{N} + \int_0^1 e^C\left(\frac{AL}{N^2(1-x)}\right)^B \Delta t\,dx}.$$

We are interested in shape of L in the steady-state, that is, when $\Delta L = 0$. This reduces the equation to:

$$L\int_0^1\left(\frac{L}{1-x}\right)^B dx = \int_0^r\left(\frac{L}{1-x}\right)^B dx.$$

The integral, $\int_0^1\left(\frac{L}{1-x}\right)^B dx$ is a constant; we will call it H. Taking the derivative of both sides with respect to r, we have:

$$H\frac{dL}{dr} = \left(\frac{L}{1-r}\right)^B$$

We can solve this equation by separation of variables:

$$\begin{array}{rcl} H\int L^{-B}dL &=& \int(1-r)^{-B}dr \\ \frac{H}{1-B}L^{1-B} &=& \frac{-1}{1-B}(1-r)^{1-B} + F \\ HL^{1-B} &=& -(1-r)^{1-B} + F(1-B). \end{array}$$

If we relabel $k = 1-B$, this simplifies to

$$HL^k + (1-r)^k = Fk.$$

In practice, $k > 0$. Substituting the points $r = 1$, $L = 1$ and $r = 0$, $L = 0$, gives us that $H = 1$ and $F = \frac{1}{k}$ and we are left with

$$L^k + (1-r)^k = 1, \quad \text{or,} \quad L = (1-(1-r)^k)^{1/k}.$$

We can attempt a corresponding “trickle-down” theory by assuming that $\frac{\partial I}{\partial t}$ depends on $\frac{\frac{A}{N}(1-L)}{Nr}$—a family at rank r developing, with those of lower rank (Nr), the wealth of those of higher rank ($\frac{A}{N}(1-L)$). From

$$\frac{\partial I}{\partial t} = e^C \left(\frac{A(1-L)}{N^2 r} \right)^B$$

we derive a second family of Lamé curves:

$$L = 1 - (1 - r^k)^{1/k}.$$

But in this case, the constant $k = 1 - B$ is greater than 1, meaning B is negative. In other words, we are left with another trickle-up theory, which one might describe as a dollar in the hands of someone at rank r sharing development rights on families of lower rank with all the dollars in the hands of those of higher rank.

Both familes model the real Lorenz curves well. It would be difficult to argue that one is signficantly better than the other.[4]

3. Checks and Balances

The success of the Lamé curves suggests that there is something fundamentally one-dimensional about inequality. That is a radical hypothesis that should be treated with caution. We explore the hypothesis and its ramifications here.

A. Lamé curves, as solutions to a differential equation, do not cross. But Lorenz curves do cross. Kakwani [K] reports that in a collection of Lorenz curves 21% of the pairs intersected. Does this falsify our hypothesis?

We don’t believe it does. Consider what we might find if the Lorenz curve for a particular time and place were computed from two independently collected data sets. The curves would follow the same basic arc but would vary up and down. The two would almost certainly cross several times. For countries whose Lorenz curves are close, it doesn’t seem surprising that they would cross.

B. We have just defended the hypothesis by appealing to possible errors or random variation in the data. But the data are also the basis for our argument. Is that a difficulty?

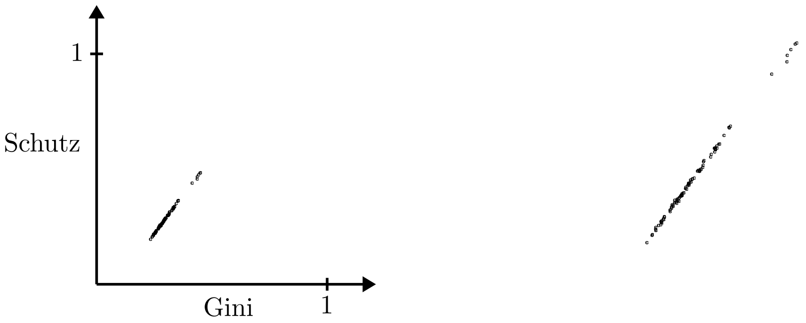
The LIS data, we understand, is the gold-standard for income data, yet we did experience some difficulties with it. The program supplied for computing

[4] These trickle-up theories suggest two additional theories, one in which dollars develop dollars (with $\frac{\partial I}{\partial t}$ proportional to $\frac{1-L}{L}$) and one in which people develop people (with $\frac{\partial I}{\partial t}$ proportional to $\frac{1-r}{r}$). Both of these result in one-parameter families, but the families lack closed-form expressions because the differential equations can’t be solved analytically. The MSEs associated with these approaches are of the same order of magnitude as those associated with the other approaches.

deciles, for example, had a bug. Even after dealing with that, we found at least one set of numbers that raised suspicions.[5] But unless the data have systemic biases, it seems a reasonable source on which to base our models. Note that our confidence in the data does have limits. We have four different one-parameter families which all model Lorenz curves well, but we don't feel we can distinguish among them.

C. If Lorenz curves were fundamentally Lamé curves, then all monotonic measures of inequality would be equivalent in the sense that knowing one measure gives you all the others. Suppose, for example, we knew the Gini coefficient g of a Lorenz curve. Then we could find the unique Lamé curve with Gini coefficient g. From that we could compute the Schutz index. Conversely, given the Schutz index, we could recover the Gini coefficient. Further, if the computations of two measures are continuous, then plotting the measures against each other should result in a connected curve.

Indeed, that seems to be the case. Here is the plot for Gini vs. Schutz. The graph on the right is a magnification.



Another measure of inequality is suggested by the trickle-up theory, the exponent B in the partial differential equation,

$$\frac{\partial I}{\partial t} = e^C \left(\frac{AL}{N^2(1-r)} \right)^B .$$

Since $B = 1 - k$, B can be determined from the best-fitting model of the Lorenz curve from the family $L(r) = \{(1-(1-r)^k)^{1/k}\}$. We could call this measure the "sensitivity factor" since it reflects how sensitive income growth for an individual is to the incomes of others. Perfect equality occurs when sensitivity is zero ($B = 0$, $k = 1$):

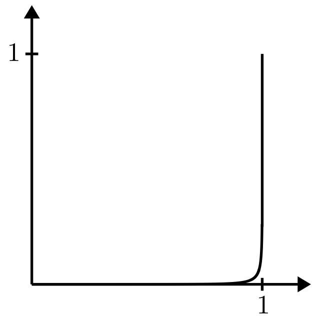
$$L(r) = 1 - (1-r)^1)^{1/1} = r.$$

In that case, all incomes grow at the same absolute rate: $\left(\frac{\partial I}{\partial t} = e^C\right)$. At steady-state, where the Lorenz curve doesn't change, incomes can still grow, but they

[5] Italy, 1991, as mentioned in the previous section.

must all grow proportionally. The only way incomes can grow at the same absolute rate and the same proportional rate is if they are all equal.

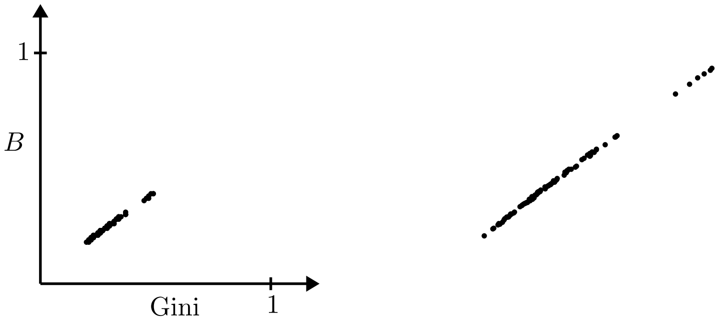
Similarly, as k approaches 0, the Lamé curves approach absolute inequality.



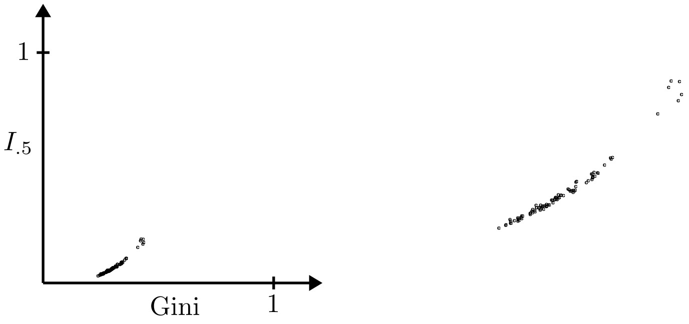
$$L(r) = (1-(1-r)^{.2})^{1/.2}$$

This reflects a growth rate directly proportional to what we might call the “opportunity for development,” $\frac{AL}{N^2(1-r)}$.

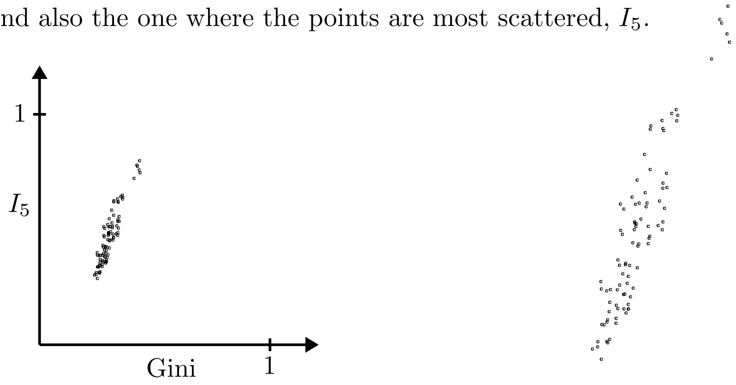
Compared to other measures, the sensitivity factor is, perhaps, less abstract and more directly meaningful. It also tracks well with the Gini coefficient.



D. James Harvey explores the relationship between the Gini coefficient and several of the Atkinson indices I_r [Ha]. His plots show large scattering which would seem to refute our hypothesis. We computed for the LIS data two Atkinson indices, one where the relationship is well-behaved in Harvey's paper, $I_{.5}$,



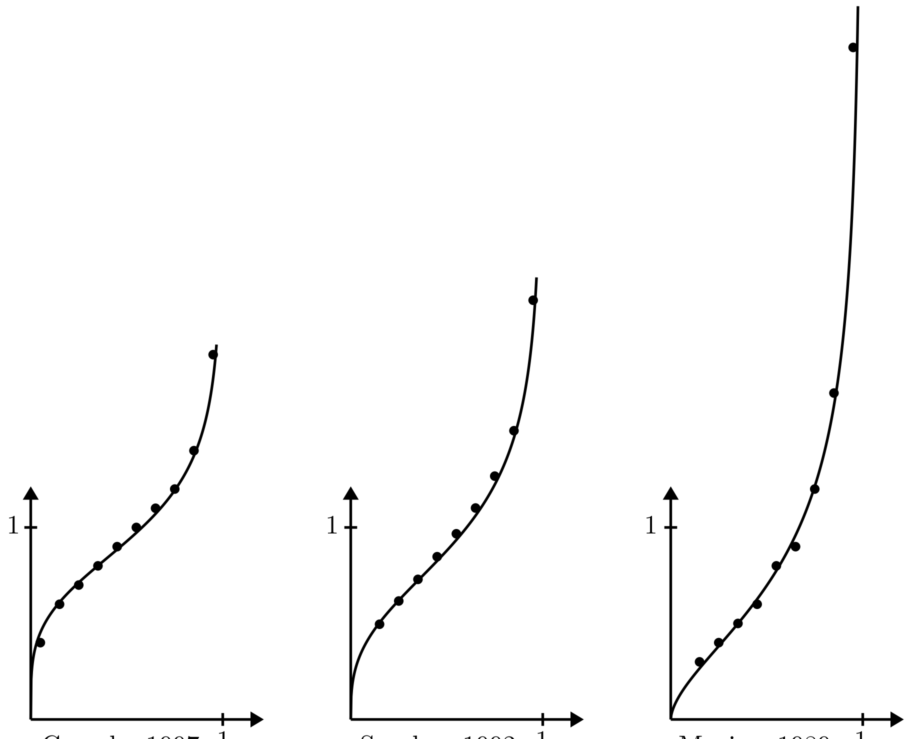
and also the one where the points are most scattered, I_5.



The last isn't clean, but it's more organized and linear than in [Ha]. The higher subscript tends to exaggerate differences at low end of the Lorenz curve. An error, for example, of ϵ in the calculation of $L(.1)$ can change I_5 by more than 10ϵ.

E. Finally, we are modelling a curve that is confined to a small space, a curve that must go from $(0,0)$ to $(1,1)$ with a constantly increasing derivative. Under those circumstances, modelling closely with a carefully chosen family of curves may seem unspectacular.

We considered this and thought to test how well the derivatives of our curves matched the derivatives of the Lorenz curves. This is a significantly greater challenge, since the derivative is theoretically unbounded. The following graphs show the derivatives of the Lamé curves for countries with varying degrees of inequality. The points are the difference quotients formed from consecutive decile data.



4. Questions

We have presented evidence that Lorenz curves for income taken at different times in different countries are well-modelled by curves from a one-parameter family of functions. Of course, additional parameters produce better fits. Modern economies are subject to countless disturbances which must vary the Lorenz curve in local but significant ways.

But the closeness of the approximations produced by a single parameter implies that the distribution of income in a society is largely characterized by a single number. This raises some related questions.

1. Is there a single economic variable that drives inequality?

2. What are the ways in which the sensitivity factor can be changed through economic policy?

3. What does the success of trickle-up theories have to say about how governments should stimulate the economy?

4. Can $B = 1 - k$ be seen as a measure of the efficiency of an economy? If so, does this suggest an explicit trade-off between efficiency and equality?

5. Inequality in the United States decreased from 1950 to 1970 ([He]) and increased from 1979 to 2000 (the Gini coefficient increased steadily from .301 to .368). Can the framework of this paper help explain these trends?

6. The relationship between the Gini coefficient and the sensitivity factor appears almost linear. Does this mean that the Gini coefficient has a concrete interpretation? That is, does the Gini coefficient tell us something definite about the relation between the rate of growth of one's income and the income of those who earn less?

7. We were not able to distinguish among the four models of income growth $(\frac{L}{1-r}, \frac{r}{1-L}, \frac{L}{1-L}, \frac{r}{1-r})$. Is there a way of determining which is best?

Appendix A On the Education Data

We found a Lorenz curve for educational attainment using data from the U.S. Census Bureau. In the table below, we have noted the number of years we attached to each category.

none (years: 0)	1st-4th (years: 2.5)	5th-6th (years: 5.5)	7th-8th (years: 7.5)	9th (years: 9)	10th (years: 10)	11th (years: 11)	H.S. grad (years: 12)	Some college (years: 13)	Associate degree (years: 14)	Bachelor's degree (years: 16)	Master's degree (years: 17)	Professional Degree (years: 18)	Doctorate Degree (years: 22)
887	2091	3911	9039	8212	9795	12993	66210	38315	13998	3090	9295	2586	1869

Educational Attainment of Persons 15 Years Old and Over
(all races, both sexes, in thousands)

The data gave us (with $(0,0)$ and $(1,1)$) 16 points on the Lorenz curve.

Appendix B Computational Details

For interest, we report here on the techniques we used to compute (a) the Gini coefficient and (b) the Schutz index.

(a) We computed the Gini coefficient from quintile data using a Newton-Cotes formula.

Given the value of a function f at three values, a, $a + .5(b-a)$, b, Simpson's Rule approximates the integral of f on $[a,b]$ by integrating the quadratic passing through the three points. Given the value of f at more points the Newton-Cotes formulae find more accurate approximations by integrating polynomials of higher degree. The particular formula we used (appropriate for the six points given by quintile data) approximates $\int_a^b f(x)\,dx$ by $\frac{95}{288}f(a) + \frac{125}{96}f(a+.2(b-a)) + \frac{125}{144}f(a+.4(b-a)) + \frac{125}{144}f(a+.6(b-a)) + \frac{125}{96}f(a+.8(b-a)) + \frac{95}{288}f(b)$.

(b) The Schutz index is the greatest distance between the Lorenz curve and the straight line from the origin to $(1,1)$. The difficulty is determining this given only decile data for the Lorenz curve.

A little calculus tells us that the point where this distance is greatest is where $L'(r) = 1$. For most Lorenz curves, that comes when r is between .6 and .7. We then approximate L with the cubic passing through the points, $(.5, L(.5))$,

$(.6, L(.6))$, $(.7, L(.7))$, $(.8, L(.8))$ and use this to find a such that $L'(a) = 1$. and then to evaluate $a - L(a)$ (the Schutz index).

References

[A] Atkinson, A. B., "On the Measure of Inequality," *Journal of Economic Theory*, Vol. 2, PP. 244-263, 1970.

[AB] Atkinson, A. B. and Bourguignon, F., editors, *Handbook of Income Distribution, Volume 1*, North-Holland, 1998.

[C] Champernowne, D. G., "A Comparison of Measures of Inequality of Income Distribution," *The Economic Journal*, vol. 84, No. 336, pp. 787-816 (1974).

[F] Figini, Paolo, "Measures, Equivalence Scales and Adjustment for Household Size and Composition," Technical Paper No. 98/8, Trinity Economic Paper Series, Trinity College, Dublin.

[Ha] Harvey, James, "A note on the 'natural rate of subjective inequality' hypothesis and the approximate relationship between the Gini coefficient and the Atkinson index," *Journal of Public Economics* vol. 89, 1021-1025, 2005.

[He] Henle, Peter, "Exploring the distribution of earned income," *Monthly Labor Review*, December, 1972 16-27.

[K] Kakwani, Nanak, "Welfare Ranking of Income Distributions,", *Advances in Econometrics*, vol. 3, pp. 191-213.

[L] Lambert, Peter, *The Distribution and Redistribution of Income*, Manchester University Press, third edition, 2001.

[McD] McDonald, James, "Some Generalized Functions for the Size Distribution of Income," *Ecnometrica*, vol. 52, No. 3, May, 1984, pp. 647-664.

[R] Ryu, H., Slottje, D., "Two flexible functional forms for approximating the Lorenz curve," *Journal of Econometrics* 72 (1996) 251-274.

[SCS] Sarabia, J.-M., Castillo, Enrique, and Slottje, Daniel J., "An ordered family of Lorenz curves," *Journal of Econometrics* 91 (1999) 43-60.

[S] Silber, Jacques, editor, *Handbook of Income Inequality Measurement* (Recent Economic Thought Series, 71), Kluwer Academic Publishers, 2001.

[T] Thiel, H., *Economics and Information Theory*, Rand McNally & Co., 1967.

[U] U.S. Census Bureau, "Educational Attainment in the United State: March 1998 (Update)."

[W] Wolff, Edward N., "Recent Trends in Wealth Ownership, 1983-1998," Working Paper No. 300, Jerome Levy Economics Institute.