VATT-KESKUSTELUALOITTEITA
VATT-DISCUSSION PAPERS

59

NON-PARAMETRIC ESTIMATION OF LORENZ CURVES USING LOCALLY WEIGHTED REGRESSION

Ilpo Suoniemi

Valtion taloudellinen tutkimuskeskus
Government Institute for Economic Research
Helsinki 1994

ISBN 951-561-078-8
ISSN 0788-5016

Valtion taloudellinen tutkimuskeskus
Government Institute for Economic Research
Hämeentie 3, 00530 Helsinki

Painatuskeskus Pikapaino Opastinsilta
Helsinki 1994

SUONIEMI, ILPO: NON-PARAMETRIC ESTIMATION OF LORENZ CURVES USING LOCALLY WEIGHTED REGRESSION. Helsinki: VATT, Valtion taloudellinen tutkimuskeskus, 1994. (C, ISSN 0788-5016, No 59). ISBN 951-561-078-8.

ABSTRACT: In this study a non-parametric estimation method for the Lorenz and concentration curves is presented. The estimator is a collection of local fits each based on weighted least squares utilizing a set of nearby observations. The method is well suited for grouped data with only a few observations available. In the study parametric forms for Lorenz curves are compared. These are the generalized Gamma and Beta functions due to Kakwani and Podder and the elliptical form presented by Villaseñor & Arnold along with a simple form based on a fourth degree polynomial fit. An empirical example is presented in which the fits are compared on their ability to estimate the Gini coefficient and the location of the mean and mode of Finnish consumption distribution. The easily applicable polynomial fit is found to produce results on closely similar level of accuracy as the other methods which are specially constructed for the particular estimation exercise. This finding encourages the application of the former method in situations where no specially tailored functional forms are available, eg. in the case of concentration curve analysis.

KEY WORDS: Inequality; Lorenz Curves; Functional Forms; Non-parametric estimation.

SUONIEMI, ILPO: NON-PARAMETRIC ESTIMATION OF LORENZ CURVES USING LOCALLY WEIGHTED REGRESSION. Helsinki: VATT, Valtion taloudellinen tutkimuskeskus, 1994. (C, ISSN 0788-5016, No 59). ISBN 951-561-078-8.

TIIVISTELMÄ: Tutkimuksessa esitetään ei-parametrinen menetelmä Lorenzin ja keskittymiskäyrien estimoimiseksi. Käytetty estimaattori on kokoelma lokaaleja sovitteita, joista kukin
perustuu painotettuun pienimmän neliösumman estimointiin käyttäen lähiympäristön havaintoja. Menetelmä soveltuu hyvin myös sellaiseen luokiteltuun aineistoon, jossa on käytössä vain muutamia havaintoja. Tutkimuksessa vertaillaan Lorenzin käyrille esitettyjä parametrisia funktioita. Kakwanin ja Podderin esittämiä yleistettyjä gamma- ja betafunktioita verrataan Villaseñorin ja Arnoldin esittämään elliptiseen muotoon. Tarkastelun kohteena on myös suoraviivainen neljännen asteen polynomisovite. Empiirisessä sovelluksessa sovitteita verrataan siinä suhteessa, miten hyvin ne estimoivat Suomen kulutusjakauman Gini-kertoimen ja jakauman keskiarvon sekä moodin. Helpoiten sovellettava polynomisovite tuotti lähes yhtä tarkkoja tuloksia kuin muut menetelmät. Tämä havainto rohkaisee menetelmän soveltamista myös sellaisissa tilanteissa, joissa räätälöityjä funktiosovitteita ei ole. Näin esimerkiksi on laita keskittymiskäyriä analysoitaessa.

ASIASANAT: Eriarvoisuus; Lorenz käyrät; Funktioesitykset; Ei-parametrinen estimointi.

CONTENTS

1. INTRODUCTION

In the economic analysis of inequality and redistribution much effort has been spent in developing methods to describe income and consumption distributions. A popular approach adopted in the literature is to analyse inequality with the use of the more informative Lorenz and concentration curve techniques in connection with derived inequality indicators, generalized Gini and concentration coefficients (see eg. Atkinson (1970) and Gastwirth (1972)). The theoretical derivations do not, however, often take into account data imperfections likely to be encountered in practical empirical analysis. A very severe problem concerns the error free measurement of the various subcomponents of consumption and income. An influential line of work starting with Kakwani and Podder (1973) considers parametric proposals for a functional relationship for the Lorenz curve and its econometric estimation, in contrast to an alternative strategy of proposing a family of distributions as a descriptive model of the income distribution (see eg. McDonald (1984)).

In this paper a chosen functional relationship for the Lorenz curve is combined with a non-parametric local curve fitting method due to Cleveland (1979). Here, a curve is fitted locally to data using weighted regression with a set of nearby observation with less weight given to the more distant observations. The final fit is a collection of local fits. The method is well suited for fitting grouped data with information on only a few points on the Lorenz curve, eg. quintile data. The present paper introduces a simple fourth degree polynomial fit with endpoint restrictions for comparisons with the generalized gamma and beta functions (Kakwani & Podder, 1973 and 1976) and with the elliptical form introduced by Villaseñor & Arnold (1979). The fits are compared on their ability to correctly estimate the Gini coefficient, the mean and the mode from grouped data on Finnish consumption.

Encouraging results from the comparisons with other func-

tional forms suggest that the method based on a local polynomial fit has a lot of potential applications in situations where less demanding a priori restrictions are available for the fitted curve so as to inhibit the application of specially tailored functional forms. This is the case with the concentration curve analysis. Note that after smoothing out sample variability from the data, the estimated concentration curves may allow for more precise statements concerning marginal conditional stochastic dominance. This concept introduced by Slemrod and Yitzhaki (1987) in turn allows interesting statements to be made on social welfare with quite weak reliance on behaviourial assumptions. The author has previously used the method to analyze Finnish data on consumption expenditures and the utilization of public welfare services (Suoniemi, 1993).

The paper is organized as follows. Section 2 introduces the analytical framework and various functional forms for Lorenz curves. Section 3 presents the non-parametric estimation method used in the study. In section 4 an empirical application is presented. Here, the different functional forms are compared on their ability to correctly estimate the Gini coefficient, and the location of the mean and the mode in a case where grouped consumption data with differing numbers of groups are considered. The final section concludes the study.

2. FUNCTIONAL FORMS FOR LORENZ CURVES

In this section methods for parametric estimation of Lorenz curves are presented. The Lorenz (and concentration) curves are frequently used to represent and analyze the size distribution of income and expenditure, respectively. The **Lorenz curve**, L(p), is defined as the proportion of income which is earned by the least privileged p-fraction of consumers, i.e.

$$L_F(p) = \int_0^{F^{-1}(p)} y dF_y / M_y, \quad with\ p \in [0,1]. \qquad (1)$$

Above F is the cumulative distribution function of income, y, and M_y is the mean income. The **concentration curve** of expenditure on a commodity, x, is defined similarly as the proportion of aggregate consumption of x which is consumed by the least privileged p-fraction of consumers w.r.t. the distribution of y. More specifically, the integrand in (1) is replaced with x which is defined as a function of y, E(x|y), and M_y replaced with the mean expenditure M_x. The concentration curve is convex (concave) to the origin if the income elasticity of the commodity is positive (negative).

Several families of distributions have been proposed as models for the income distribution, see McDonald (1984). An alternative research strategy is to consider a functional specification for the relationship L(p). The following characterization of Lorenz curves has been attributed to Gaffney and Anstis (Villaseñor and Arnold (1989)).

Proposition 1 Assume L(p) is a continuous function on the interval [0,1] with the second derivative, L". The function L(p) is a Lorenz curve if and only if
$L(0) = 0$, $L(1) = 1$, $L'(0^+) > 0$, and $L''(x) > 0$
for all x in the open interval]0,1[.

An early proposal for an explicit parametric function was put forth by Kakwani and Podder (1973). They set

$$L(p) = p^{\delta}\exp\{-\eta(1-p)\}, \quad with\ 1 < \delta < 2,\ \eta > 0. \tag{2}$$

The curve defined by (2) will below be referred to as the generalized gamma function or the Kakwani-Podder I-form.

On the other hand, Villaseñor and Arnold (1989) noted that segments of ellipses provide a flexible family of Lorenz curves which perform "remarkably well in fitting the data". Start by defining the general quadratic form

$$ax^2 + bxy + cy^2 + dx + ey + f = 0. \tag{3}$$

This includes many curves (x,y) passing through the points (0,0) and (1,1) which satisfy the conditions of the preceding theorem and hence may be considered as Lorenz curves. The curve is a segment of an ellipse, a parabola or a hyperbola as $b^2 - 4ac < 0$, $= 0$, or > 0. In order of the curve (3) to pass through the end points (0,0) and (1,1) one must have

$$f = 0, \tag{4 a}$$

and

$$e = -(a + b + c + d). \tag{4 b}$$

If $c = 0$, the curve (x,y) represented by (3) collapses to a hyperbola. If $c \neq 0$, one may normalize and set $c = 1$ with no loss of generality. Taking into account the end point restriction (4 b) one may write (3) in the form

$$y(1-y) = a(x^2-y) + by(x-1) + d(x-y), \tag{5}$$

which is well suited for fitting the data. In this case the equation (3) is a quadratic equation in y and it has two roots

$$L(x) = \frac{\left[-(bx+e) \pm (\alpha x^2+\beta x+e^2)^{1/2}\right]}{2}, \tag{6}$$

where $e = -(a + b + d + 1)$, $\alpha = b^2 - 4a$, and $\beta = 2be - 4d$.

Villaseñor and Arnold (1989) note that the root obtained by setting the plus-sign in the middle of expression (6) corresponds either to a bathtub shaped or a monotone density function. Therefore, they have limited suitability as descriptions of say, income data. The class obtained under the minus sign contains hyperbolic ($\alpha > 0$), elliptical ($\alpha < 0$), and parabolic ($\alpha = 0$) Lorenz curves. Villaseñor and Arnold (1989) focus on elliptical Lorenz curves and present necessary and sufficient conditions on the parameters, for (6) to represent an elliptical Lorenz curve. Furthermore, they characterize the underlying density and the class of distributions with elliptical Lorenz curves. In the present paper the location of the mode and the mean relative to the median is examined as a summary way of describing the underlying density distribution.

Among the numerous other suggestions the modified Beta function developed by Kakwani and Podder (1976) seems to be the most influential. In this case the original domain of the argument, $[0,1]$, is first changed to $[0,\sqrt{2}]$ by a change of the coordinates $(p,L) \rightarrow 2^{-\frac{1}{2}}(p+L,p-L) = (\pi,\eta)$. Geometrically one now measures orthogonal distance of the original Lorenz curve from the egalitarian line. In the conventional Lorenz box-diagram this line is the diagonal through the unit square (figure 1). Next the transformed (π,η)-coordinates are connected by an application of a modified Beta function

$$\eta = a\pi^{\alpha}(\sqrt{2}-\pi)^{\beta}. \qquad (7)$$

In logarithmic form which is amendable for estimation the result is (given in original (p,L)-coordinates)

$$\log[2^{-\frac{1}{2}}(p-L(p))] = \log a + \alpha\log[2^{-\frac{1}{2}}(p+L(p))] + \beta\log[\sqrt{2}-2^{-\frac{1}{2}}(p+L(p))], \qquad (8)$$

The above expression will be referred to as the generalized

beta function or the Kakwani-Podder II-form.[1]

Alternative specifications have been proposed by Baylock and Smallwood (1982) who use the Box-Cox transformation to extend the Kakwani-Podder II functional form while simultaneously allowing for heteroscedastic errors in fitting the curve and by Rasche et al. (1980). These proposals are, however, not considered in detail here but a very simple form is considered instead.

Here, one utilizes a fourth degree polynomial

$$y = ax^4 + bx^3 + cx^2 + dx + f. \tag{9}$$

with the end point restrictions

$$f = 0,$$

and

$$1 = a + b + c + d.$$

Taking into account the end point restrictions one may solve for d and write (9) in the form

$$x - y = a(x-x^4) + b(x-x^3) + c(x-x^2). \tag{10}$$

which can be readily applied in estimating the curve.

The following results are useful for examining whether a given Lorenz curve is appropriate for representing an income distribution. First differentiate (1) to get

$$L_F'(p) = F^{-1}(p)/M_y, \quad with\ p \in [0,1]. \tag{11}$$

Therefore,

---

[1] The Kakwani-Podder function has singularity at the end points. This, however, has been found to produce no material effect in empirical applications (Kakwani, 1980). Above the function is given in the original form (Kakwani & Podder, 1976). Note, however, that the extension of the domain which produces the various $\checkmark$ -terms has no material effect on the subsequent curve fitting.

$$L_F'(p) \leq 1 \iff F^{-1}(p) \leq M_y, \quad \forall\, p \in [0,1]. \tag{12}$$

The above formula shows how the location of the mean, L'(p) = 1, is determined if the Lorenz curve of the distribution is known. By further differentiation one obtains (Villaseñor and Arnold (1989))

Proposition 2 If L"(p) exists and is positive (almost) everywhere in an open interval $]p_1, p_2[$ then the corresponding distribution F has a finite positive density in the interval $]M_y L(x_1^+), M_y L(x_2^-)[$ and the density is given by

$$f(y) = 1/M_y L''(F(y)). \tag{13}$$

Finally, one notes by differentiating (13) that the location of the modes are obtained as the points of singularity of the third derivative, L''', of the Lorenz curve.

The **Gini concentration coefficient** is widely used as a summary measure of the extent of inequality. It is calculated as twice the area between the forty-five degree line and the Lorenz curve, using either of the two formula

$$G(y) = 1 - 2\int_0^1 L_F(p)\,dp \tag{14}$$

$$= 2\,Cov(y,\ F(y))\ /M_y, \tag{15}$$

where M_y is the mean of the variable y. Subsequently, in the paper the functional forms of the Lorenz curve are evaluated by comparing how successful they are in estimating the Gini coefficient, and the location of the mean and the mode.

3. LOCALLY WEIGHTED REGRESSION

Consider a random sample of size n, (X_i, ω_i) with (possibly) unequal sampling weights ω_i, i = 1,...,n. For notational convenience let the weights be such that $\sum \omega_i = 1$, ignoring the estimate on the size of the finite sampling population. Let $(X_{(i)}, \omega_{(i)})$, i = 1,.., n, be the (ascending) order statistics of X with the corresponding weights attached. The frequency distribution function is given by

$$F_n(x) = \sum_1^{i-1} \omega_{(j)} + \frac{1}{2}\omega_{(i)} + \frac{1}{2}\left[\frac{\omega_{(i+1)}+\omega_{(i)}}{X_{(i+1)}-X_{(i)}}\right](x - X_{(i)}), \qquad (16)$$

where i is the largest integer with $X_{(i)} \leq x$.

The corresponding non-parametric estimate of the Lorenz curve is given similarly by, $L_n(0) = 0$, $L_n(1) = 1$, and for $p \in]0,1[$

$$L_n(p) = \frac{\sum_1^{i-1} \omega_{(j)}X_{(j)} + \frac{1}{2}\omega_{(i)}X_{(i)} + \frac{1}{2}\left[\frac{\omega_{(i+1)}X_{(i+1)}+\omega_{(i)}X_{(i)}}{X_{(i+1)}-X_{(i)}}\right](x-X_{(i)})}{\sum_1^n \omega_i X_i}, \qquad (17)$$

where i is the largest integer with $X_{(i)} \leq x$, and $p = F_n(x)$.[2]

The Lorenz curves are estimated in this study by using a non-parametric local fitting method based on an idea due to Cleveland (1979). The raw data consists of a frequency Lorenz curve (F_n, L_n) which is based on the frequency distribution function of the data and is given in a parametric form by (16) and (17). The estimation procedure used here is a smoothing filter applied on the frequency Lorenz curve.

---

[2] In the case of ties one orders the observations that are affected w.r.t. the weights (in decreasing order) to guarantee succesively increasing slopes on the piecewise linear Lorenz curve connecting the points given by (16) and (17).

This filter has two elements. First, a parametric function which is locally fitted to the data near a given point, say t using weighted least squares. Second, a weight function, or a kernel, that assigns less weight to more distant observations. Therefore points that are close to t play a large role in the determination of the fit $\mu(t)$ while points far away have a lesser role.

In this study the following specific kernel function is used, for points t_i, i = 1,..., n,

$$w_t(t_i) = \phi\left[\frac{t_i - t}{\gamma_t}\right], \quad \text{if } |t_i - t| < 2.5\gamma_t, \quad \text{and} \tag{18}$$

$$= 0, \quad \text{if } |t_i - t| \geq 2.5\gamma_t, \tag{19}$$

where ϕ is the Gaussian function, $\phi(u) = \exp\{-u^2/2\}$.

In estimation one selects the bandwidth locally as to guarantee a sufficient number of observations, say 2k+1 that are effectively involved in the local fit. This is done by centering the data on t_j and setting γ so that $|t_j - u| = 2.5\gamma$, at the 2k'th nearest neighbour of t_j. In considering the asymptotic properties of the non-parametric fitting procedure one will select the bandwidth so as to guarantee that the number of effective observations tends locally to infinity while simultaneously the bandwidth gets to zero, as the sample size goes to infinity.[3]

Consider using a local base of polynomial functions $\{1, t-u, \ldots, (t-u)^m\}$ in the local fit of data on (t,y) near a given data point t. Let $T_i = (1, t-t_i, \ldots, (t-t_i)^m)$ denote

---

[3] In the procedure used in the present application the choice is done automatically depending on the total number of points in the sample, n, so that the number of observations, 2k +1, effective at the local fit gets to infinity at order $n^{2/3}$ (see Table 1). By ordering the observations by t the computational burden is considerably diminished. Estimation procedures that incorporate all the above features are written using GAUSS™ (Aptech Systems) programming language and are available from the author on request.

the m+1 -dimensional column vector associated with the observation i, i = 1,..., n, and T_u be defined similarly. The estimated local parameters are given by

$$b_n(t) = \left[\sum_1^n w_t(t_j)\, T_j T_j^T\right]^{-1} \left[\sum_1^n w_t(t_j)\, y_j T_j\right], \qquad (20)$$

with the weights given by (16) and (17). The local fit at u near data point t is given by

$$\mu_n(u,t) = b_n(t)^T T_u = T_u^T \left[\sum_1^n w_t(t_j)\, T_j T_j^T\right]^{-1} \left[\sum_1^n w_t(t_j)\, y_j T_j\right]. \qquad (21)$$

At the point u = t, T_u = (1, 0,..., 0) and the local fit $\mu_n(t,t)$ is given directly by the parameter estimate of the constant term in the local fit. Similarly

$$\left[\frac{\partial^k \mu_n(u,t)}{\partial u^k}\right]_{u=t} = k!\,[b_n(t)]_k, \qquad (22)$$

i.e. the k'th coefficient of the Taylor expansion of the fit μ at point t is given by the k'th component of the vector $b_n(t)$, k = 1,..,m.

To arrive at the (local) smoothing filter interpretation one can write the fit as a weighted mean of original observations

$$\mu_n(u,t) = \sum_1^n K_j(u,t,\gamma_t)\, y_j, \qquad (23)$$

where

$$K_j(u,t,\gamma_t) = w_t(t_j)\, T_u^T \left[\sum_1^n w_t(t_i)\, T_i T_i^T\right]^{-1} T_j\,. \qquad (24)$$

Generally the estimation procedure can be seen as a member in the class of non-parametric regression. Another and widely used non-parametric regression procedure is smoothing splines (see Eubank, 1988). Splines have been popular because they are the solution to an intuitively appealing mathematical optimization criterion. But, on the other hand, they optimize on a global criterion and are not generally local. Secondly, since splines are results of optimization, it may be difficult to determine how they operate on data. In particular, it is considerably more difficult to determine the effective bandwidth of a spline estimate at a given point whereas in the above case this is straightforward to do.[4] In large data sets, a case less relevant to the application examined in the present paper, the computational requirements of splines are substantial.

In this study local fits are estimates using a fourth degree polynomial fit (10), the elliptic form (5), and the generalized gamma (2) and beta functions (8). The Gini concentration coefficients are calculated analytically by piecewise integration of the locally smoothed curve. In the case of the frequency Lorenz curve, the area between the forty-five degree line and the Lorenz curve can be calculated either directly or by using the covariance formula (13). These methods give values with no discrepancies at the reported level of accuracy.

---

[4] This is particularly important near the extreme points of the data. Here, a local fit can often guarantee a better fit.

5. AN APPLICATION AND COMPARISONS

Locally weighted regression is used to produce smoothed Lorenz curves using both the elliptical and a simple fourth degree polynomial fits of the data and the corresponding forms due to Kakwani and Podder (1976). The data are drawn from the Finnish 1985 Consumer Expenditure Survey, collected by the Central Statistical Office of Finland, for general information see Tilastokeskus (1987).

The goodness of fit of the various functional forms are compared by utilizing data in tabulated form with diminishing degrees of information (number of income groups). Here the complete microdata for 1985 (with 8200 obs.) are used along with data on fractiles of one part per 1000th parts, percentage data, decile and finally quintile data.

In table 1 the relative success of the methods in estimating the Gini coefficient for total consumption data is shown. The first estimate is based on the area under the Lorenz curve calculated directly from the frequency curve based on grouped data. The other columns are obtained by analytically integrating the locally smoothed curves fitted on the same data (details available on request).

The last two columns give information on the characteristics of the local regression involved in the estimation. The first of these gives the number of observations effectively used in each local fit, and the last column gives the number of observations in the data used for estimation, eg. 10 in the case of decile data (the third row). Note that the local fit used in the case of the quintile data (the fourth row) uses all the observations but naturally with greatly varying weights across the local fits. Last row gives the results for the whole sample available.

A notable feature of the results is that all locally fitted parametric curves produce results that are superior to simple formula using frequency data in cases of tabulated

data at conventional levels of availability.[5] The accuracy in estimating the Gini-coefficient is remarkable even in the case where only quintile data is available. To some degree this is due the nice behaviour of the underlying specific distribution.

Table 1: Estimation of the Gini-coefficients.

Gini coefficients					Number of observations	
Frequency distr.	Elliptic curve	Fourth deg. polynomial	Kakwani Podder I	Kakwani Podder II	2K+1	N
0.3328	0.3327	0.3328	0.3331	0.3328	199	1000
0.3327	0.3328	0.3329	0.3325	0.3329	43	100
0.3289	0.3327	0.3331	0.3282	0.3329	9	10
0.3200	0.3323	0.3327	0.3285	0.3330	5	5
0.3328	0.3327	0.3328	0.3329	0.3320	813	8200

An additional property of the local fit concerns its ability to estimate the density of the underlying data. On this point the simple polynomial regression is a priori on a less solid ground than methods that incorporate the necessary restrictions directly in the functional form. An examination of the estimated parameters reveals that the parameter estimates fulfil the constraints implied by proposition 2.[6]

In table 2 the estimates of the mean are given for the fits employed[7]

---

[5] Some allowance in comparisons should be made for the fact that the Kakwani-Podder I -form has locally one estimable parameter less than the other functions.

[6] In fact one may run into some minor but annoying local difficulties in some fits covering small segments of the interval [0,1] if relatively disaggregated data are used (the first and the last rows) in conjunction with greatly varying sample weights across the original observations.

[7] When local estimation methods are employed, the mean (and the mode) are estimated in the following way. For each local interval the corresponding local estimate of say mean is made. If the estimate lies within the relevant interval it is chosen, otherwise, it is discarded.

Table 2: Estimation of the Mean.

Fractile of the mean				Number of observations	
Elliptic curve	Fourth deg. polynomial	Kakwani Podder I	Kakwani Podder II	2K+1	N
0.5790	0.5798	0.5775	0.5791	199	1000
0.5815	0.5863	0.5750	0.5821	43	100
0.5861	0.6040	0.5650	0.5911	9	10
0.5865	0.6098	0.5604	0.5915	5	5

Examination of the total sample gives the estimate of 0.5771 for the location of the mean of the distribution. In this respect the fourth degree polynomial fit is slightly inferior to the other fits when tabulated data with only few groups are available.

Table 3: Estimation of the Mode.

Fractile of the mean				Number of observations	
Elliptic curve	Fourth deg. polynomial	Kakwani Podder I	Kakwani Podder II	2K+1	N
0.2952	0.2711	0.1549	0.0842	199	1000
0.2736	0.2503	0.1566	0.0852	43	100
0.2456	0.2561	0.1552	0.1177	9	10
0.2478	0.2645	0.1509	0.1215	5	5

In table 3 the estimates of the mode are given. In this case the results vary a lot with the elliptic fit and the fourth degree polynomial fit giving comparable values. This may be partly due to the fact that these two forms imply global unimodality with a simple linear formula of parameters characterizing the modal value (Appendix). In the case of the Kakwani-Podder forms the corresponding equations are more complicated giving some possibility for bimodality.[8]

---

[8] In the case of the generalized gamma function, KP-form I, the modal values are given by the roots of a third degree polynomial (Appendix). It turned out, however, that only one of these is the proper choice for the mode in the case considered here. Further details on the calculations are available on request.

It should be noted that although unimodality is a common feature of distributions that are frequently chosen to describe, say income data, (McDonald, 1984) it is a convenient assumption and not necessarily a property of the actual data. In the above case examination of the frequency density distribution of the total sample gives indication of multiple local maxima. If one decreases the number of classes the number of candidates for a modal values decrease. A global maximum value of the density seems to emergence at around the value 0.28 which is quite near the values given by the methods corresponding to the first two columns of table 3. In contrast, the methods proposed by Kakwani & Podder give markedly different values. This probably illustrates the sensitivity of results in trying to estimate the third derivative of a curve, although figure 2 shows that all four curves fitted to quintile data are remarkably close to each other both in actual values and in terms of the derivative of the curve.[9]

The final point to be made concerns the performance of the simple polynomial fit. This is on a similar level of accuracy as the other curves specially constructed for this particular estimation case. This is probably due to the local estimation method which is particularly successful in our application (figure 2).

Furthermore, in the simple polynomial case the Gini concentration coefficient and the mean and modal values of the distribution can be estimated with considerably less computational burden than in the other cases considered in this paper. The above comforting results make the polynomial regression a particularly appealing method in situations where a priori restrictions on the functional form are less demanding or even non-existent. This is true for instance in

---

[9] A slight caveat concerns the rather tedious and intransparent formulae and calculations needed to produce the modal point in the case of the generalized beta function, KP-form II (see the Appendix). Although the calculations have been repeatedly checked and double checked it is difficult to rule out a possible error definitely.

the case of the concentration curves where one is only able to set the end point restrictions on the fitted curve.

5. CONCLUSION

In this paper locally weighted regression has been applied to produce a non-parametric estimator of the Lorenz curve. Some influential proposals for a parametric form of a Lorenz curve are compared with a simple polynomial function on their ability to fit grouped data and correctly estimate the Gini coefficient and the mean and the mode of the distribution. In these respects the simple polynomial method has a similar level of performance with the rivals that incorporate the necessary restrictions directly in the functional form.

The positive and encouraging implications of this study are the following. The method based on a local polynomial fit has many potential applications in situations where less demanding a priori restrictions are available for the curve to be fitted so as to inhibit the application of specially tailored functional forms. This is the case with concentration curve analysis. Here smoothing out sample variability from the raw data and using estimated concentration curves for inferential purposes may allow for more precise statements concerning marginal conditional stochastic dominance (Slemrod and Yitzhaki, 1987) between, say expenditures on various consumption categories. This, in turn, allows one to make interesting statements on social welfare with quite weak reliance on behaviourial assumptions. The author has previously applied this method to Finnish data to examine consumption together with the utilization of public welfare services (Suoniemi, 1993).

REFERENCES

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

Baylock, J.L. and D. M. Smallwood (1982). "Analysis of Income and Food Expenditure Distributions: A Flexible Approach." **The Review of Economics and Statistics**, 64, 104-109.

Cleveland, W.S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." **Journal of the American Statistical Association**, 74, 829-836.

Eubank, R.L. (1988). "**Spline Smoothing and Nonparametric Regression.**" Marcel Dekker, Inc.: New York.

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

Kakwani, N.C. (1980). "Applications of Lorenz curves in economic analysis." **Econometrica**, 48, 1064-1066.

Kakwani, N.C. and N. Podder (1973). "On the Estimation of the Lorenz Curves from Grouped Observations." **International Economic Review**, 14, 278-291.

Kakwani, N.C. and N. Podder (1976). "Efficient Estimation of the Lorenz Curve and associated Inequality Measures from Grouped Observations." **Econometrica**, 44, 137-148.

McDonald, J. B. (1984). "Some Generalized Functions for the Size Distribution of Income." **Econometrica**, 52, 647-663.

Rasche, R.H., J. Gaffney, A.Y.C. Koo, and N. Obst (1980). "Functional Forms for Estimating the Lorenz Curve." **Econometrica**, 48, 1061-1063.

Slemrod, J., and S. Yitzhaki (1987). "Welfare dominance: an application to commodity taxation." **NBER Working Paper**, 2451.

Suoniemi, I. (1993). "Public Welfare Services and Inequality: Introduction to methodology and some examples with the 1985 Finnish Household Expenditure Survey data." VATT Discussion Paper 45.

Tilastokeskus (1987). "Finnish Household Expenditure Survey 1985. User Guide." (in Finnish). **Tilastokeskus.**, Helsinki.

Villaseñor, J.A. and B.C. Arnold (1989. "Elliptical Lorenz Curves." **Journal of Econometrics**, 40, 327-338.

APPENDIX

The location of a mode of the underlying distribution is characterized by the condition, $L'''(p) = 0$. Here it is shown what the corresponding condition is in the (π,η) - representation of the Lorenz curve introduced by Kakwani & Podder (1976).[10] Start by defining

$$\begin{bmatrix} \eta \\ \pi \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} L \\ p \end{bmatrix}. \qquad \textbf{(A 1)}$$

A straightforward differentiation gives

$$\frac{\partial \eta}{\partial \pi} = \frac{1 - L'}{1 + L'}, \qquad \textbf{(A 2)}$$

noting that $d\eta/dp = 2^{-1/2}(1 - L')$ and $d\pi/dp = 2^{-1/2}(1 + L')$.

Incidentally (A 2) shows that the mean ($L' = 1$) corresponds to the point where $\partial\eta/\partial\pi = 0$. This occurs at the point

$$\pi = \frac{\sqrt{2}\,\alpha}{\alpha + \beta}. \qquad \textbf{(A 3)}$$

Similarly,

$$\frac{\partial^2 \eta}{\partial \pi^2} = -\frac{2\sqrt{2}\,L'}{(1 + L')^3}. \qquad \textbf{(A 4)}$$

Furthermore,

$$\frac{\partial^3 \eta}{\partial \pi^3} = \frac{-4L'''}{(1 + L')^4} + \frac{12\,(L'')^2}{(1 + L')^5}. \qquad \textbf{(A 5)}$$

Finally, one can conclude that the condition, $L'''(p) = 0$, implies that

$$\eta'''\,(1 + \eta') = 3\,(\eta'')^2. \qquad \textbf{(A 6)}$$

---

[10] The equations for the first two derivatives, (A2) and (A4), were obtained already by Kakwani & Podder, in terms of the underlying variable y. The derivation in terms of p and the other formulae seem to be new.

Considering the other functional forms the polynomial fit gives the unique modal point at

$$p = -\frac{b}{4a}, \qquad \textbf{(A 7 i)}$$

The mean is one of the roots of a trinomial

$$4ap^3 + 3bp^2 + 2cp + d = 0. \qquad \textbf{(A 7 ii)}$$

The elliptic fit gives the mean as a root of

$$(b+2)(\alpha p^2+\beta p+e^2)^{\frac{1}{2}} \pm \frac{1}{2}(2\alpha p+\beta) = 0, \qquad \textbf{(A 8 i)}$$

where the sign depends on which of the branches is selected.

The unique mode is simply given by

$$p = -\frac{\beta}{2\alpha}, \qquad \textbf{(A 8 ii)}$$

The generalized gamma function by Kakwani & Podder (1976) gives the mean as a root of

$$(\delta-1)\log p + \eta(p-1) + \log(\eta p+\delta) = 0, \qquad \textbf{(A 9 i)}$$

and an estimate of the mode as a root of

$$(\eta p+\delta)^3 - 3\delta(\eta p+\delta) + 2\delta = 0. \qquad \textbf{(A 9 ii)}$$

Figure 1: Lorenz curve

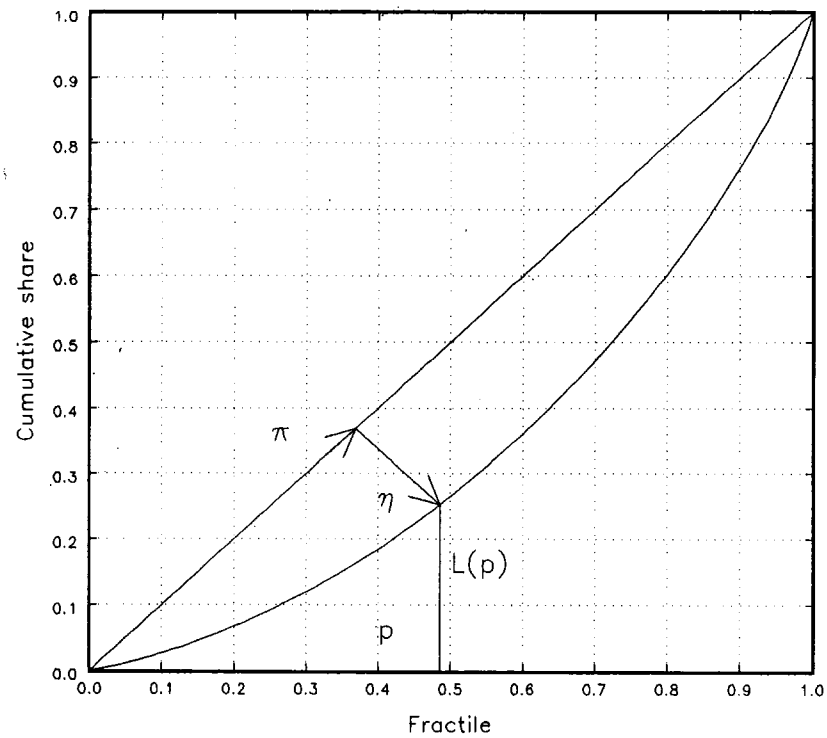

Figure 2: Lorenz curves fitted on Quintile data.