Scientific Research Publishing

Theoretical Economics Letters, 2018, 8, 1793-1802
http://www.scirp.org/journal/tel
ISSN Online: 2162-2086
ISSN Print: 2162-2078

Regression Analyses of Income Inequality Indices

Johan Fellman

Hanken School of Economics, Helsinki, Finland
Email: fellman@hanken.fi

How to cite this paper: Fellman, J. (2018) Regression Analyses of Income Inequality Indices. *Theoretical Economics Letters*, **8**, 1793-1802.
https://doi.org/10.4236/tel.2018.810117

Received: April 11, 2018
Accepted: June 18, 2018
Published: June 21, 2018

Abstract

Scientists have analysed different methods for numerical estimation of Gini coefficients. Using Lorenz curves, various numerical integration attempts have been made to identify accurate estimates. Central alternative methods have been the trapezium, Simpson and Lagrange rules. They are all special cases of the Newton-Cotes methods. In this study, we approximate the Lorenz curve by polynomial regression models and integrate optimal regression models for numerical estimation of the Gini coefficient. The attempts are checked on theoretical Lorenz curves and on empirical Lorenz curves with known Gini indices. In all cases the proposed methods seem to be a good alternative to earlier methods presented in the literature.

Keywords

Gini Index, Income Distribution, Lorenz Curve, Regression Models, Trapezium Rule, Simpson Rule, Lagrange Rule, Newton-Cotes Method

1. Introduction

Income distributions are commonly unimodal and skew with a heavy right tail. Therefore, different skew models, such as the lognormal and the Pareto, have been proposed as suitable descriptions of income distributions, and the corresponding Lorenz curves have been obtained. These are usually applied in specific empirical situations. For general studies, more wide-ranging tools have been considered. In a long series of studies, different models and methods have been proposed. The target for them is to introduce inequality measures, such as Gini and Pietra indices, that are usable for comparisons of different distributions. Primary income data yield the most exact estimates of income inequality coefficients, but when the income distribution is unknown the use of Lorenz curves is common. In this article, we present a new regression model that approximates

the Lorenz curve by polynomial regression models and after integration of the optimal regression models one obtains numerical estimation of the Gini coefficient.

Income inequality indices. Consider the set of ordered points $\left(p, L\left(p\right)\right)$, where p is the cumulative proportion of the income-receiving units and the Lorenz curve, $L\left(p\right)$, is the corresponding cumulative proportion of income received when the units are arranged in ascending order of income. When Lorenz curves are compared, especially when they intersect, the comparisons are based on numerical indices. The most frequently used index is the Gini coefficient, G [1]. Using the Lorenz curves, this coefficient is the ratio of the area between the diagonal and the Lorenz curve and the whole area under the diagonal. The formula is

$$G = 1 - 2\int_0^1 L\left(p\right)dp. \tag{1}$$

This definition yields Gini coefficients satisfying the inequalities $0 < G < 1$. The higher the G value, the lower the Lorenz curve and the stronger the inequality. The reason for the popularity of the Gini coefficient is that it is easy to compute, being a ratio of two areas in Lorenz curve diagrams. The Gini coefficient allows direct comparison of the income of two income distributions, regardless of their sizes or patterns. The Gini does not capture where in the distribution the inequality occurs. As an additional result, two very different distributions of income, even if they have intersecting Lorenz curves, can have the same Gini index.

In many empirical situations, the income distribution $F(x)$ is given in grouped tables. If the mean or the total incomes in the groups are known, the cumulative distribution can be modified to a Lorenz curve, but the subintervals do not have constant length. Consequently, Simpson's rule is not applicable. One has to replace it with the trapezium rule or with Lagrange polynomials. The trapezium rule is a weak alternative because it yields positive bias for the area under the Lorenz curve and negative bias for the Gini coefficient.

As an application of these methods, Fellman [2] considered different Lorenz models: the Kakwani and Podder model, the generalized Pareto model analysed by Rasche *et al.* [3] and the Gupta model [4]. In addition, Rao and Tam [5] constructed a generalized Gupta model. Furthermore, Rao and Tam introduced a simplified version of the Rao-Tam model. Chotikapanich [6] defined an alternative Lorenz curve. The Pareto, Chotikapanich and Gupta models contain only one parameter. They are so simple that it is impossible to distinguish between the length of the range of the income distribution function and the Gini coefficient. With only one parameter to estimate, these distribution properties cannot be independently estimated. We pay special attention to these models and analyse them in more detail. Using Lorenz curves, various numerical integration attempts were made to determine the accuracy of the estimates. For example, Mettle *et al.* [7] considered Lorenz curves and estimated the Gini coefficient of in-

come by Newton-Cotes methods, and then compared the accuracy of these estimates for some (Ghanaian) data.

In this study, we review income analysis methods based on Lorenz curves. To test the proposed methods, the analyses are initially applied to theoretical models with known inequality indices. The empirical value of the method is based on analyses of real data in the literature with Gini indices of strong accuracy, and our obtained results are compared with earlier findings.

2. Methods

There are several different situations, and consequently, alternative analyses of Gini coefficients have to be performed. Common estimation alternatives are the use of the trapezium, Simpson and Lagrange rules. They are all special cases of the Newton-Cote method. A common property of these is that they split the $(0,1)$ interval into subintervals and approximate the Lorenz curve in such a way that the polynomials obtain the same values as the Lorenz curve at the end points of the subintervals.

When Lorenz curves are considered, the simplest situations are that they are defined for five quintiles or for ten deciles. In the first case, the most commonly used method is the trapezium rule. For Simpson's rule, the number of subintervals should be even and the intervals should have the same length. Consequently, the comparison of the results of different rules can be performed for Lorenz curves with deciles.

Our new attempt proposed here is to assume that the approximating function of $L(p)$ is a regression polynomial consisting of non-negative integer powers of the argument *p,* fitted to the values of the Lorenz curve. The optimal polynomial comes close to the Lorenz curve, but at no point obtains exactly the same value. Furthermore, the points of the Lorenz curves do not need to be equidistantly distributed.

Let the regression model be

$$\hat{L}(p) = \hat{\alpha} + \hat{\beta}_1 p + \hat{\beta}_2 p^2 + \ldots + \hat{\beta}_n p^n . \tag{2}$$

When one integrates the regression model over the interval $(0,1)$, one obtains the area under the Lorenz curve having the formulae

$$\int_0^1 \hat{L}(p)\,dp = \hat{\alpha} + \frac{1}{2}\hat{\beta}_1 + \frac{1}{3}\hat{\beta}_2 + \ldots + \frac{1}{n+1}\hat{\beta}_n \tag{3}$$

and

$$\hat{G} = 1 - 2\int_0^1 \hat{L}(p)\,dp = 1 - 2\left(\hat{\alpha} + \frac{1}{2}\hat{\beta}_1 + \frac{1}{3}\hat{\beta}_2 + \ldots + \frac{1}{n+1}\hat{\beta}_n\right). \tag{4}$$

3. Results

3.1. Theoretical Lorenz Curves

We apply our method on theoretical models in order to compare the obtained

Gini indices with theoretical ones. We follow the assumption that the polynomial is at most of six degree ($n \le 6$). This restriction is inposed by the maximum degree of the polynomial trend lines in the Excel system.

3.2. Pareto Model

The first one is the Pareto model, $F(x) = 1 - x^{-\alpha}$, with a finite mean, that is $\alpha > 1$. The Lorenz curve is $L(p) = 1 - (1-p)^{\frac{\alpha-1}{\alpha}}$, the mean is $\mu = \frac{1}{\alpha - 1}$ and $G = \frac{1}{2\alpha - 1}$. In this study, we assume that the parameter value is $\alpha = 4$. Hence, $\mu = \frac{1}{3}$ and $G = \frac{1}{2\alpha - 1} = 0.142857$. The Lorenz curve is presented in **Figure 1**. The estimated regression model is

$$\hat{L}(p) = 0.001462\alpha + 0.7056p + 0.28469p^2 - 0.33355p^4 + 0.34561p^6. \quad (5)$$

After integration, our method gives the observed value $\hat{G} = 0.141508$, which compared with the theoretical value given above indicates good agreement.

3.3. Chotikapanich Model

The Chotikapanich model [6] has the Lorenz curve

$$L(p) = \frac{e^{kp} - 1}{e^k - 1} \quad \text{for} \quad k > 0.$$

The Gini index is

$$G = \frac{(k-2)e^k + 2 + k}{k(e^k - 1)}. \text{ For given } \mu, \; F(x) = \frac{1}{k}\ln\left(\frac{x(e^k - 1)}{\mu k}\right) \text{ [8].}$$

In this study, we assume that $k = 5$. Hence, the Gini index is $G = 0.613567$. The Lorenz curve is presented in **Figure 2**.

The estimated regression model is

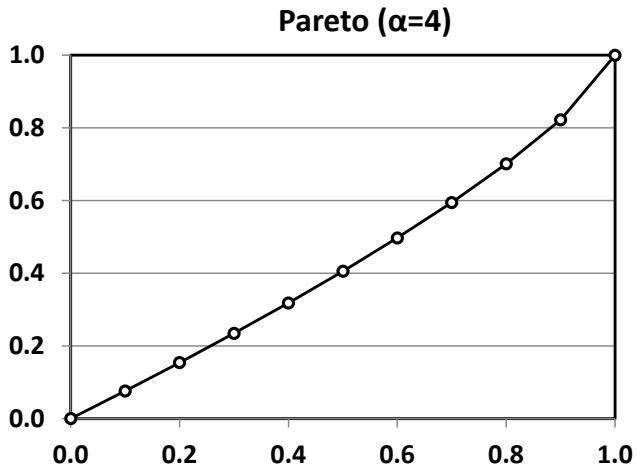

Figure 1. The Lorenz curve for a Pareto distribution with $\alpha = 4$ and $\hat{G} = 0.141508$.

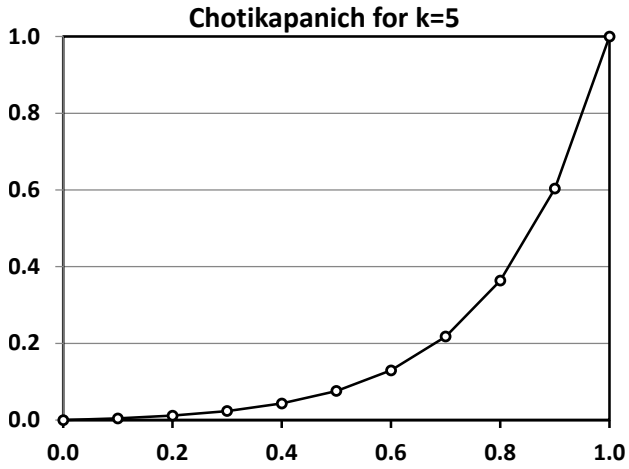

Figure 2. The Lorenz curve for a Chotikapanich distribution with $k=5$ and $\hat{G}=0.613462$.

$$\hat{L}(p)=-0.0000079+0.264809p^2+0.735051p^6. \quad (6)$$

The estimated Gini index based on a this polynomial is $\hat{G}=0.613462$, indicating good agreement with the theoretical value.

3.4. Gupta Model [4]

The Lorenz curve of the Gupta model is $L(p)=p\beta^{p-1}$ with the Gini coefficient

$$G=1-\frac{2}{\ln(\beta)}\left[1-\frac{\beta-1}{\beta\ln(\beta)}\right]. \quad (7)$$

Despite the Gupta model being relatively simple, the corresponding income distribution is not attainable. The explanation of this is that the variable p is included in the model both as a factor and exponent. If $\beta=5$, then the theoretical numerical value is $G=0.375021$. The Lorenz curve is presented in **Figure 3**. Our regression model is

$$\hat{L}(p)=-0.000189+0.211948p+0.235348p^2+0.459725p^3+0.09329p^6. \quad (8)$$

The estimated Gini index based on this polynomial is $\hat{G}=0.375015$, and the agreement with the theoretical value is acceptable.

The Lorenz curves for the Pareto, Chotikapanich and Gupta models are presented in deciles, and therefore, we can compare their regression results. In **Figure 4**, we compare their residuals. One observes that the residuals show stronger variations for Pareto and Chotikapanich than for Gupta. In fact, $SE=0.003144$ for Pareto, $SE=0.001521$ for Chotikapanich and $SE=0.000269$ for Gupta.

3.5. Empirical Data

The obtained results concerning theoretical models are acceptable, but in order to check the model the proposed method must also be applied on numerical

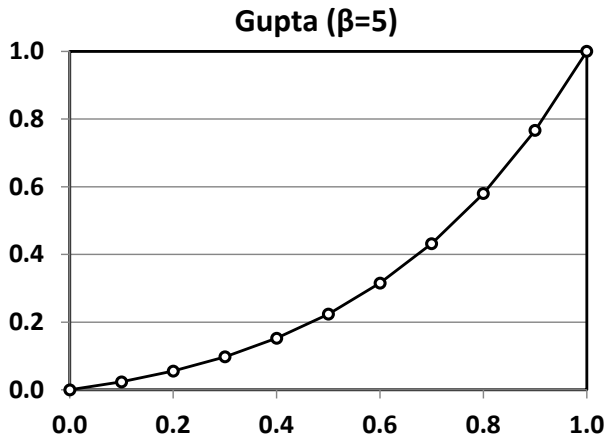

Figure 3. The Lorenz curve for a Gupta distribution with $\beta = 5$ and $\hat{G} = 0.375015$.

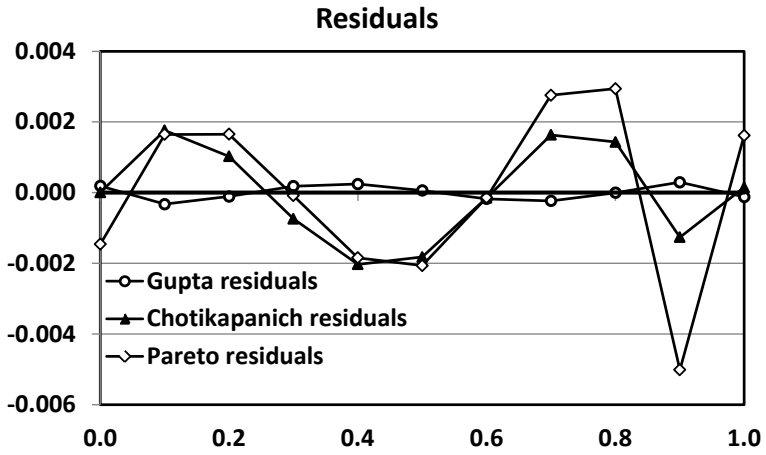

Figure 4. Comparison of the residuals for the Pareto, Chotikapanich and Gupta models. The differences between the models are discussed in the text.

empirical data. We choose from the literature empirical data for which G values have previously been estimated with good accuracy.

3.6. Ogwang Data

The Lorenz curve is based on the data given by Ogwang [9]. The data is household income in Israel, originally derived from the Family Expenditure Survey 1986/87 reported by Fishelson [10]. The data are presented as a Lorenz curve with several intervals. In this case, the subintervals are of different lengths. Consequently, one has no possibility to use Simpson's rule. The Lorenz curve is presented in **Figure 5**. Ogwang's analyses of Gini coefficients based on deciles yield the interval $0.3234 \le G < 0.3511$.

The estimated regression model is

$$\hat{L}(p) = 0.000822 + 0.144475p + 1.036686p^2 - 0.523957p^3 + 0.340210p^6. \quad (9)$$

The estimated Gini index based on the polynomial (9) is $\hat{G} = 0.3275$. This value is located well within the interval proposed by Ogwang.

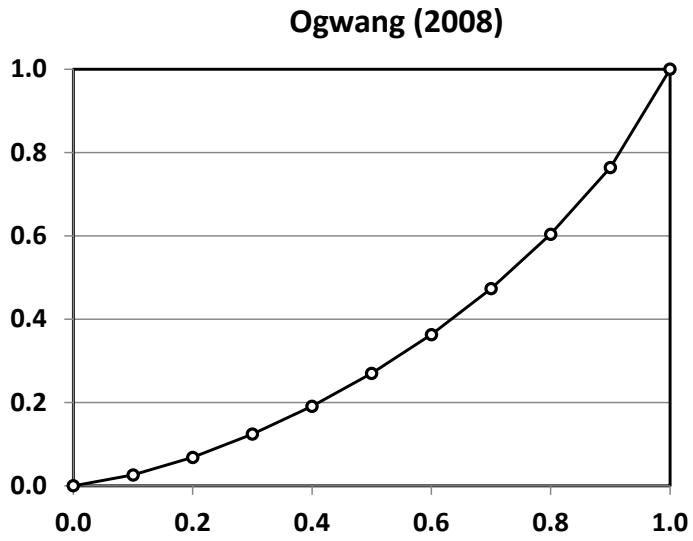

Figure 5. The Lorenz curve for the Ogwang data yielding $\hat{G} = 0.3275$.

3.7. Tepping's Data

Tepping estimated an accurate Gini coefficient from the Current Population Survey (CPS) data from 1968 [11]. The estimated Gini index was 0.4014. Gastwirth [12] tested Tepping's data, applied different methods and obtained interval estimates that were close to Tepping's estimate. In this study, we construct the Lorenz curve for Tepping's data and approximate the curve using our polynomial regression model. We obtain the optimal regression model

$$\hat{L}(p) = -0.0000079 + 0.264809p^2 + 0.735051p^6. \quad (10)$$

After integration, the Gini estimate is 0.4005. This finding is very close to Tepping's result (**Figure 6**).

3.8. Lorenzen Data

Lorenzen [13] presents information about the total distribution of income for households in Germany in 1973 in his "*Tabelle* 2". The Gini coefficient calculated by Lorenzen is based on data pooled in his "*Tabelle* 3", which yielded 0.30. Using Lorenzen's "*Tabelle* 3", Fellman [1] performed a comparison of the estimates obtained based on the trapezium rule, the Lagrange rule and the modified Golden method [14]. The estimated Gini coefficient according to the trapezium rule shows negative biases relative to Lorenzen's result, being 0.2920. The Lagrange interpolation yields the estimate 0.3486 and the modified Golden method 0.3002. The available empirical data are insufficient for a comparison of the accuracy of the methods or identification of the optimal one. In this study, we present the regression method. The Lorenz model of the data is presented in **Figure 7** and the optimal regression model is

$$\hat{L}(p) = 0.000913 + 0.160800p + 0.961946p^2 - 0.647743p^4 + 0.514366p^6. \quad (11)$$

However, we obtained the estimate $\hat{G} = 0.308212$.

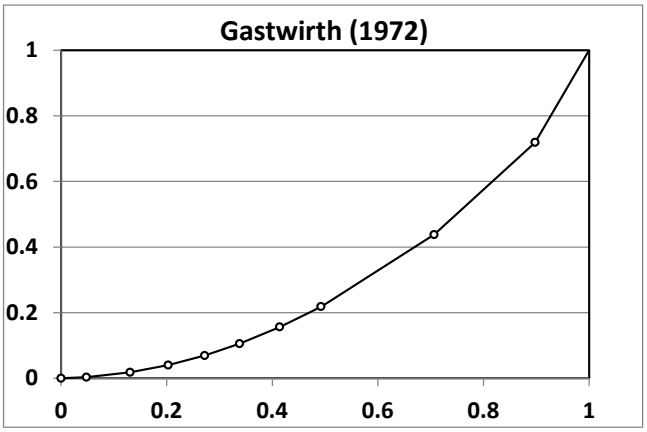

Figure 6. The Lorenz curve for the Tepping data yielding $G = 0.4005$.

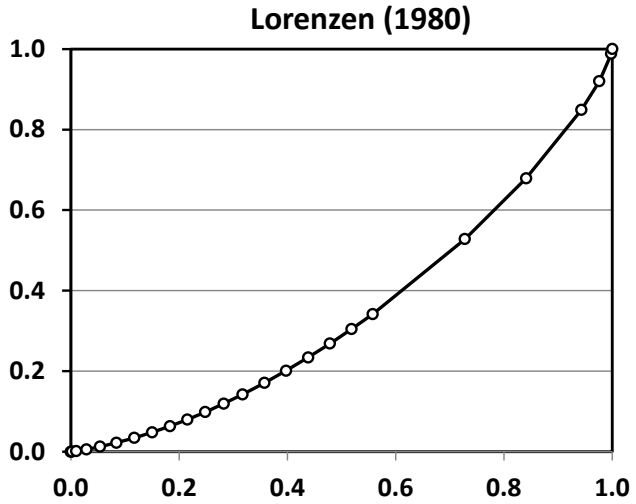

Figure 7. The Lorenz curve for the Lorenzen data $\hat{G} = 0.308212$.

In all cases the proposed methods seem to be a good alternative to earlier methods presented in the literature.

4. Discussion

The comparison between different estimation methods is in general difficult to perform. These difficulties are mainly caused by the fact that the true Gini coefficient is unknown, but sometimes, where more detailed studies have already resulted in accurate estimates, the comparisons are possible. Such comparison problems are eliminated if the numerical estimations are applied to theoretical distributions. Therefore, when one introduces a new method one must base it on theoretical Lorenz curves with known exact theoretical Gini indices [1].

The first model in this study is the Pareto model analysed by Rasche *et al.* [3], the second is the Chotikapanich model and the third the Gupta model [4]. The Gupta model, the simplified Rao-Tam and the Chotikapanich contain only one parameter. With only one parameter to estimate, the range and the Gini index cannot be independently estimated ([3] [4] [5] [6] [15] [16] [17]).

The step from the Lorenz curve to distribution function is more difficult than that from distribution function to the Lorenz curve. There is a difference between advanced and simple Lorenz models. Advanced models yield a better fit to data, but are difficult to connect to exact income distributions. Simple one-parameter models can more easily be associated with the corresponding income distribution, but when statistical analyses are performed the goodness of fit is often poor.

In order to perform comparisons between the estimated and theoretical Gini coefficients, Fellman [1] analysed classes of theoretical Lorenz curves with varying Gini coefficients. In this study, we compare Gini estimates for the Pareto, the simplified Rao-Tam and the Chotikapanich distributions.

Fellman [18] studied the Lorenz curves for the Pareto, Chotikapanich and Gupta models presented the Gini and Pietra indices for variable parameter values. He compared these indices and showed the relation between them.

Acknowledgements

This study was in part supported by a grant from the "Magnus Ehrnrooths Stiftelse" Foundation.

References

[1] Fellman, J. (2012) Estimation of Gini Coefficients Using Lorenz Curves. *Journal of Statistical and Econometric Methods*, **1**, 31-38.

[2] Fellman, J. (2012) Modelling Lorenz Curve. *Journal of Statistical and Econometric Methods*, **1**, 53-62.

[3] Rasche, R.H., Gaffney, J., Koo, A.Y.C. and Obst, N. (1980) Functional Forms for Estimating the Lorenz Curve. *Econometrica*, **48**, 1061-1062. https://doi.org/10.2307/1912948

[4] Gupta, M.R. (1984) Functional Form for Estimating the Lorenz Curve. *Econometrica*, **52**, 1313-1314. https://doi.org/10.2307/1911001

[5] Rao, U.L.G. and Tam, A.Y.-P. (1987) An Empirical Study of Selection and Estimation of Alternative Models of the Lorenz Curve. *Journal of Applied Statistics*, **14**, 275-280. https://doi.org/10.1080/02664768700000032

[6] Chotikapanich, D. (1993) A Comparison of Alternative Functional Forms for the Lorenz Curve. *Economics Letters*, **41**, 129-138. https://doi.org/10.1016/0165-1765(93)90186-G

[7] Mettle, F.O., Darkwah, K.A., Nortey, E.N.N. and Lotsi, C.A. (2016) An Estimation of the Gini Coefficient of Income Using Newton-Cotes Methods. CBAS Annual Science and Development Platform, University of Ghana, Volume 1.

[8] Fellman, J. (2012) Modelling Lorenz Curve. *Journal of Statistical and Econometric Methods,* **1**, 53-62.

[9] Ogwang, T. (2006) An Upper Bound of the Gini Index in the Absence of Mean Income Information. *Review of Income and Wealth*, **52**, 643-652. https://doi.org/10.1111/j.1475-4991.2006.00210.x

[10] Fishelson, G. (1994) Do the Intradeciles Inequalities Matter? A Note. *Empirical Economics*, **19**, 171-178. https://doi.org/10.1007/BF01205733

[11] Bureau of the Census (1967) Trends in the Income of Families and Persons in the United States, 1947-1964. Technical Paper No. 17, US Government Printing Office.

[12] Gastwirth, J.L. (1972) The Estimation of the Lorenz Curve and Gini Index. *The Review of Economics and Statistics,* **54**, 306-316. https://doi.org/10.2307/1937992

[13] Lorenzen, G. (1980) Was ist ein "echtes" Konzentrationsmass? *Algemeines Statistisches Archiv,* **4**, 390-400.

[14] Golden, J. (2008) A Simple Geometric Approach to Approximating the Gini Coefficient. *Journal of the Economic Education*, **39**, 68-77. https://doi.org/10.3200/JECE.39.1.68-77

[15] Ogwang, T. and Rao, U.L.G. (2000) Hybrid Models of the Lorenz Curve. *Economics Letters*, **69**, 39-44. https://doi.org/10.1016/S0165-1765(00)00274-3

[16] Cheong, K.S. (2002) An Empirical Comparison of Alternative Functional Forms for the Lorenz Curve. *Applied Economics Letters*, **9**, 171-176. https://doi.org/10.1080/13504850110054058

[17] Rohde, N. (2009) An Alternative Functional Form for Estimating the Lorenz Curve. *Economics Letters*, **105**, 61-63. https://doi.org/10.1016/j.econlet.2009.05.015

[18] Fellman, J. (2018) Income Inequality Measures. *Theoretical Economics Letters,* **8**, 557-574. https://doi.org/10.4236/tel.2018.83039