Journal of Statistical and Econometric Methods, Vol.10, No.1, 2021, 35-53
ISSN: 2241-0384 (print), 2241-0376 (online)
https://doi.org/10.47260/jsem/1013
Scientific Press International Limited

Empirical Analyses of Income: Finland (2009) and Australia (1967-1968)

Johan Fellman[1]

Abstract

Analyses of income data are often based on assumptions concerning theoretical distributions. In this study, we apply statistical analyses, but ignore specific distribution models. The main income data sets considered in this study are taxable income in Finland (2009) and household income in Australia (1967-1968). Our intention is to compare statistical analyses performed without assumptions of the theoretical models with earlier results based on specific models. We have presented the central objects, probability density function, cumulative distribution function, the Lorenz curve, the derivative of the Lorenz curve, the Gini index and the Pietra index. The trapezium rule, Simpson´s rule, the regression model and the difference quotients yield comparable results for the Finnish data, but for the Australian data the differences are marked. For the Australian data, the discrepancies are caused by limited data.

JEL classification numbers: D31, D63, E64.
Keywords: Cumulative distribution function, Probability density function, Mean, quantiles, Lorenz curve, Gini coefficient, Pietra index, Robin Hood index, Trapezium rule, Simpson´s rule, Regression models, Difference quotients, Derivative of Lorenz curve

[1] Hanken School of Economics.

Article Info: *Received:* September 29, 2019. *Revised:* January 21, 2020.
Published online: March 19, 2021.

1. Introduction

Analyses of income data are often based on assumptions concerning theoretical distributions. In this study, we apply statistical analyses, but ignore specific models concerning the income distributions. The income data sets in this study are taxable income in Finland (2009) and household income in Australia (1967-1968). Our intention is to compare statistical analyses performed without assumptions concerning theoretical income models with earlier results based on specific models. The central idea is that assumptions concerning distribution models yield only approximate results with inaccuracies comparable to results based on model-free statistical estimations. We have presented the central objects, probability density function (PDF), cumulative distribution function (CDF), the Lorenz curve, the derivative of the Lorenz curve, the Gini index and the Pietra index. The trapezium rule, Simpson´s rule, the regression model and the difference quotients yield comparable results for the Finnish data, but for the Australian data the differences are marked. For the Australian data, the discrepancies are caused by limited data.

2. Methods

We use the following notations. Let *X* be the income variable, let $F(x)$ be the cumulative distribution function (CDF), let $f(x)$ be the probability density function (PDF), let

$$\mu = \int_0^\infty xf(x)dx \tag{1}$$

be the mean of *X* and let x_p be the *p* quantile, that is $F(x_p) = p$. Then the Lorenz curve is

$$L(p) = \frac{1}{\mu}\int_0^{x_p} xf(x)dx. \tag{2}$$

The Lorenz curve (Figure 1) has the following general properties:

i. $L(p)$ is monotone increasing,

ii. $L(p) \le p$,

iii. $L(p)$ is convex,

iv. $L(0) = 0$ and $L(1) = 1$

The Lorenz curve $L(p)$ is convex because the income share of the poor is less than their proportion of the population. The higher the Lorenz curve the lesser the inequality in the income distribution.

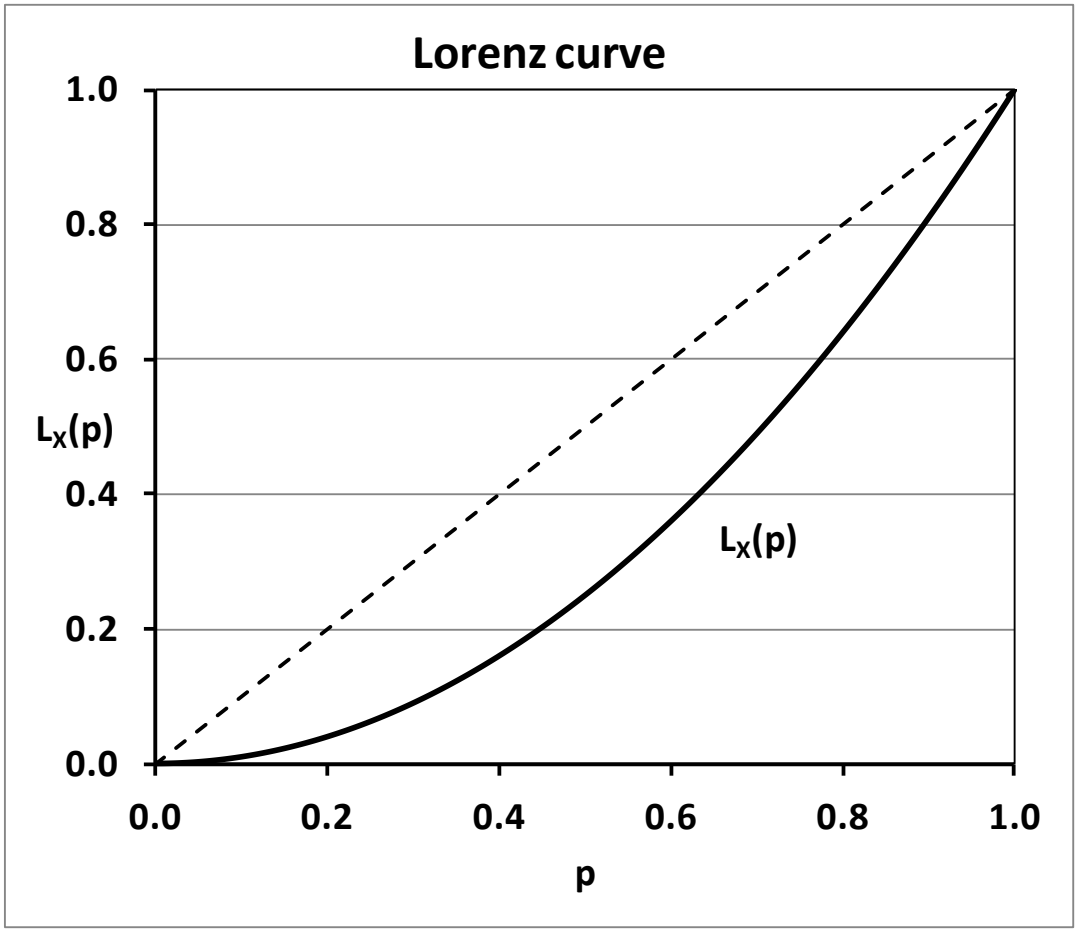

Figure 1: A sketch of a Lorenz curve $L(p)$ [1].

As explained in [1], the Gini coefficient, *G*, is defined as the ratio of the area between the diagonal and the Lorenz curve and the area of the whole triangle under the diagonal. Primary income data yield the most accurate estimates of the Gini coefficient. However, the estimation are often based on tables with grouped data or on Lorenz curves. Fellman [2] analysed the estimation of Gini coefficients using Lorenz curves. In empirical situations, the income distribution $F(x)$ is often given in grouped tables. If the number of observations and the mean of *X* or if the total incomes in the groups are known, the distribution can be considered as a Lorenz curve, but the subintervals are usually not of constant length. This is the case of the Australian data considered below.

We intend to perform statistical analyses without any assumptions about the theoretical PDF, CDF and Lorenz distributions and only base the calculations on the general statistical formulae.

There are several different situations and, consequently, alternative analyses of Gini coefficients have to be performed. When Lorenz curves are considered, the simplest

situation is that they are defined for five quintiles or for 10 deciles. In the first case, the most commonly used method is the trapezium rule. The trapezium rule generates numerical problems because every trapezium yields a positive bias to the estimated area under the Lorenz curve, and since the biases accumulate and no elimination of biases can be performed, the estimated Gini coefficient based on the trapezium rule always has a negative bias. Simpson´s rule is better fitted to the Lorenz curve, but demands an even number of subintervals of the same length. This means, for example, that Lorenz curves with 10 deciles are suitable. Consequently, the comparison of different rules can be performed for Lorenz curves with deciles.

Compared with the trapezium rule, Simpson´s rule gives more accurate approximations. As stressed above, Simpson´s rule demands two restrictions; the number of subintervals must be even and the subintervals of equal length. In order to apply Simpson´s rule, the subintervals must be grouped two by two. Each doubled subinterval has three values. The area under this part of the Lorenz curve is estimated such that a parabola obtaining the same values approximates the Lorenz curve. Simpson´s rule obviously yields exact results for quadratic curves, but, in fact, exactness also holds for cubic curves.

Fellman [3] presented a new attempt proposing 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 Lorenz curves are increasing and convex functions of *p* and the powers of *p* are also increasing and convex, and hence, such polynomials are suitable approximations of $L(p)$. In general, the optimal polynomial comes close to the Lorenz curve, but obtains at no point exactly the same value. Furthermore, the points of the Lorenz curves do not need to be equidistantly distributed. Let the obtained optimal regression model be

$$\hat{L}(p) = \hat{\alpha} + \hat{\beta}_1 p + \hat{\beta}_2 p^2 + \hat{\beta}_3 p^3 + ... + \hat{\beta}_n p^n . \tag{3}$$

When one integrates the regression model over the interval $(0,1)$, the area under the regression model is the formula

$$\int_0^1 \hat{L}(p)dp = \hat{\alpha} + \hat{\beta}_1 \frac{1}{2} + \hat{\beta}_2 \frac{1}{3} + \hat{\beta}_3 \frac{1}{4} + ... + \hat{\beta}_n \frac{1}{n+1}, \tag{4}$$

and consequently,

$$\hat{G} \approx 1 - 2\int_0^1 \hat{L}(p)dp = 1 - 2\left(\hat{\alpha} + \hat{\beta}_1 \frac{1}{2} + \hat{\beta}_2 \frac{1}{3} + \hat{\beta}_3 \frac{1}{4} + ... + \hat{\beta}_n \frac{1}{n+1} \right). \tag{5}$$

The obtained regression model can also be used when one wants to estimate the derivative of the Lorenz curve $L(p)$. The derivative of the regression model is

$$\hat{L}'(p) = \hat{\beta}_1 + 2\hat{\beta}_2 p + 3\hat{\beta}_3 p^2 + ... + n\hat{\beta}_n p^{n-1} . \tag{6}$$

If the Lorenz curve is differentiable, the derivatives have the following properties. Consider $L(p) = \frac{1}{\mu}\int_{x}^{x_p} xf(x)dx$, $F(x_p) = p$ and the frequency function $f(x)$. When we differentiate the equation $F(x_p) = p$ we obtain $\frac{dF(x_p)}{dp} = 1$ and

$$f(x_p)\frac{dx_p}{dp} = 1.$$

Consequently,

$$f(x_p)\frac{dx_p}{dp} = 1$$

and

$$\frac{dx_p}{dp} = \frac{1}{f(x_p)}.$$

If we differentiate

$$L'(p) = \frac{dL(p)}{dp} = \frac{1}{\mu}\frac{d\int_{0}^{x_p} xf(x)dx}{dx_p}\frac{dx_p}{dp} = \frac{1}{\mu}x_p,$$

then

$$L'(p) = \frac{x_p}{\mu} \tag{7}$$

and the mean μ is

$$\mu = \frac{x_p}{L'(p)}. \tag{8}$$

The Methods section is based on results obtained from our empirical data. Figure 2 presents the Lorenz curve for the Finnish data. In Figure 3, we apply (8) and present the mean income μ and x_p as a function of *p*. The derivative $L'(p)$ is presented in Figure 4 and is based on (7) and (8). Hence, the numerical data in Figures 2 to 4 are based on the Finnish data set.

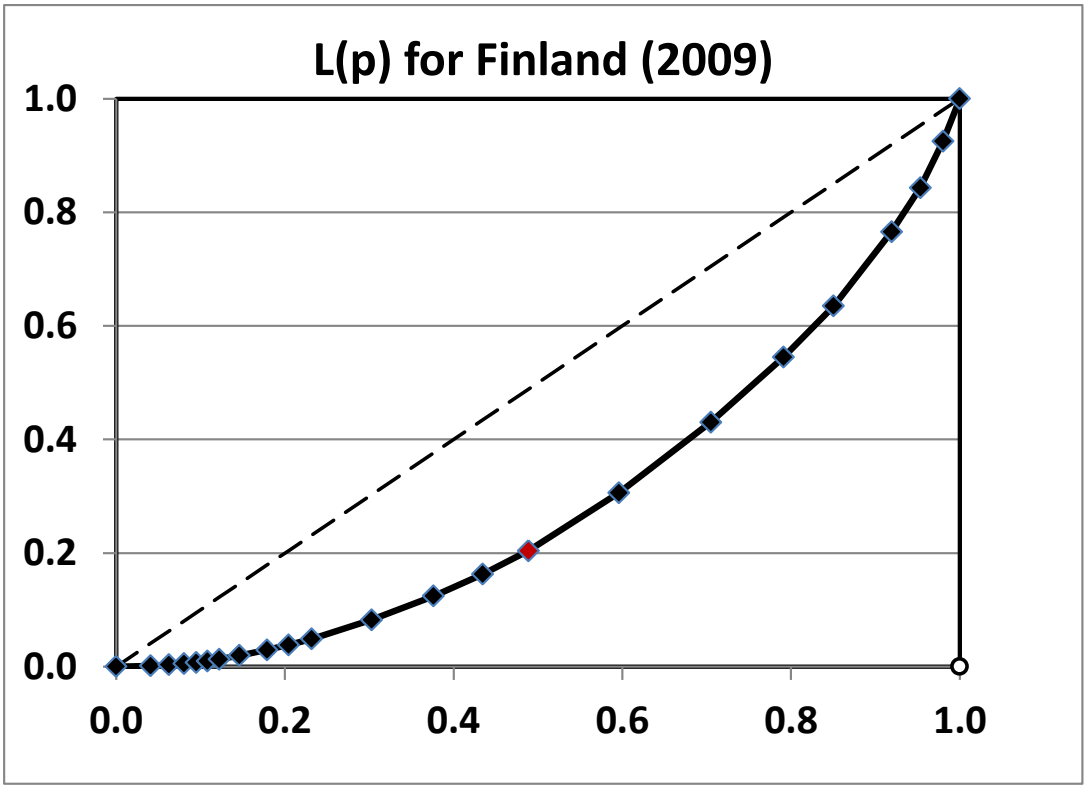

Figure 2: Lorenz curve for the Finnish data

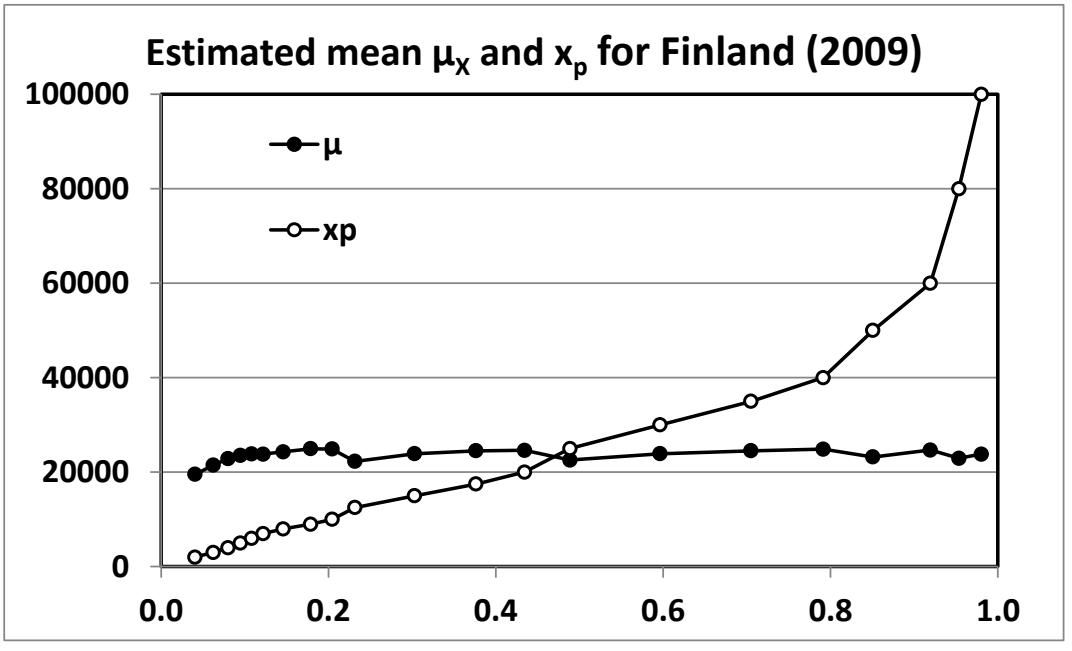

Figure 3: Estimated μ and x_p as functions of *p* for the Finnish data

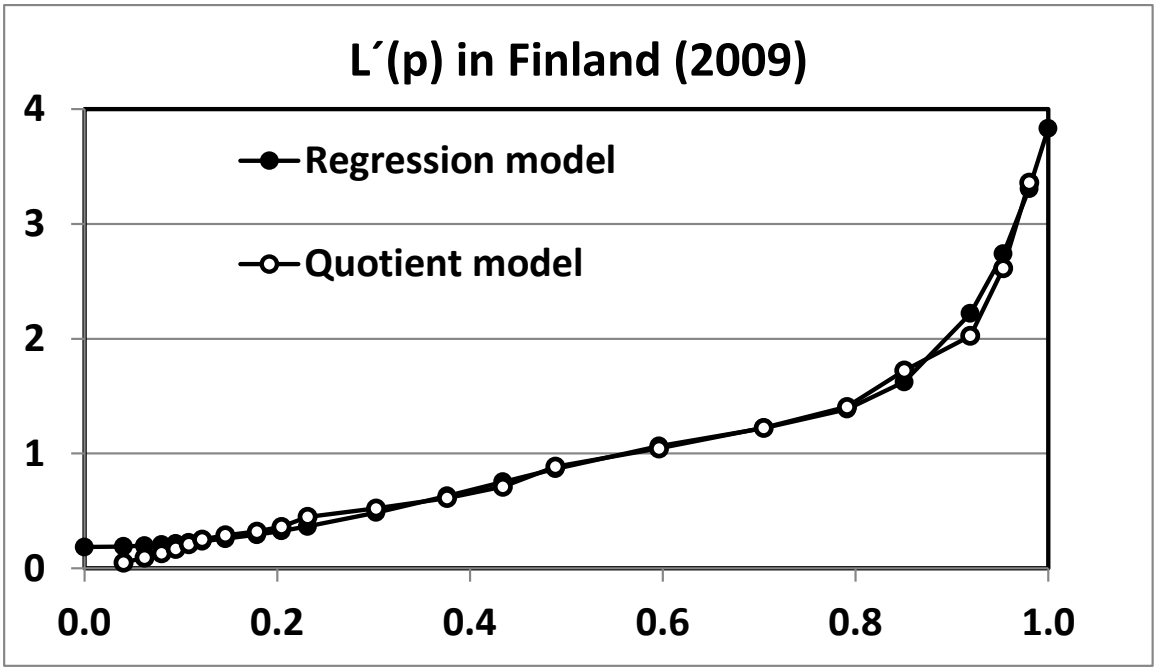

Figure 4: Derivates estimated by the difference quotients and by the regression models are included. The comparison between the findings is acceptable

The difference between the diagonal and the Lorenz curve has the properties

$$D = p - L(p), \tag{9}$$

$$\frac{dD}{dp} = 1 - L'(p) = 1 - \frac{x_p}{\mu},$$

$$\frac{d^2D}{dp^2} = -L''(p) = -\frac{1}{\mu}\frac{dx_p}{dp} = -\frac{1}{\mu f(x)} < 0.$$

The maximum of *D* implies $1 - \frac{x_p}{\mu} = 0$, that is, $x_p = \mu$.

For $x_p = \mu$, $L'(p) = 1$, and at the point $p_\mu = F(\mu)$, the tangent is parallel to the line of perfect equality. This is also the point at which the vertical distance between the Lorenz curve and the egalitarian line attains its maximum $P = D_{\max} = p_\mu - L(p_\mu)$. This maximum is defined as the Pietra index [4]. According to this definition, $0 < P < p_\mu < 1$. The lower bound is obtained when there is a total income equality, that is, the Lorenz curve coincides with the diagonal. The upper bound can be obtained when the curve converges towards the lower right corner. The Pietra index can be interpreted as the income of the rich ($p > p_\mu$) that should be redistributed to the poor ($p < p_\mu$) in order to obtain total income equality.

Therefore, the index is sometimes named the Robin Hood index.
An alternative definition of the Pietra index has also been given. It can be defined as twice the area of the largest triangle inscribed in the area between the Lorenz curve and the diagonal line [4]. In Figure 5, one observes that the triangle obtains its maximum when the corner lies on the Lorenz curve where the tangent is parallel to the diagonal. The height of the triangle is $h = \frac{P}{\sqrt{2}}$, and the base is the diagonal $b = \sqrt{2}$. The double of the area is

$$2area = 2\frac{h\sqrt{2}}{2} = \frac{2P\sqrt{2}}{2\sqrt{2}} = P .$$

Figure 5: The Lorenz curve and the geometric interpretations of the Pietra index [1].

In comparison, the *P* index is twice the area of the inscribed triangle and the Gini coefficient is twice the whole area between the diagonal and the Lorenz curve, and hence, $G > P$.

Compare two income variables *X* and *Y*. If $P_X < P_Y$, then the distribution $F_X(x)$ measured by the Pietra index has lower inequality than the distribution $F_Y(y)$, and we say that $F_X(x)$ *Pietra dominates* $F_Y(y)$. We denote this relation $F_X(x) \succ_P F_Y(y)$. For the Lorenz curves in Figure 1.1.3 in [1], $P_1 = 0.2500$ and $P_2 = 0.2940$. According to the relation between the two Pietra indices, $L_1(p)$ is more unequal than $L_2(p)$ [5]

3. Materials

Table 1: Taxable income receivers in Finland in 2009

Annual income (€)	**Income recipients (n)**
-1000	182281
1000 - 2000	96836
2000 - 3000	80056
3000 - 4000	65800
4000 - 5000	59595
5000 - 6000	62171
6000 - 7000	107558
7000 - 8000	146526
8000 - 9000	114602
9000 - 10000	121555
10000 - 12500	319042
12500 - 15000	329083
15000 - 17500	259979
17500 - 20000	243284
20000 - 25000	481753
25000 - 30000	487376
30000 - 35000	385672
35000 - 40000	266075
40000 - 50000	307810
50000 - 60000	152714
60000 - 80000	120327
80000 -	88488
All	4478583

Our data consist of the income distribution reported for Finland (2009) given in Table 1 and the Australian data given in Table 2 [6]. Our intentions are to consider the estimations of the central concepts based on empirical data. Our main difficulty is that our data for Finland do not contain exact information on individual incomes. We have only grouped data according to the incomes of the income receivers. For such an interval, we only know the number of the receivers and the minimum and the maximum of income individuals of the group. Consider group number *i*. Let the number of observations be n_i, the lower limit be a_i and the upper limit be b_i. The total income is restricted to the interval $\left(\sum n_i a_i, \sum n_i b_i\right)$. The total income is $\sum n_i y_i$, where $a_i \le y_i \le b_i$. The y_i values are unknown, but plausible estimates of total incomes are one of the alternatives $\sum n_i a_i$, $\sum \frac{1}{2} n_i (a_i + b_i)$ or $\sum n_i b_i$. We perform our statistical analyses based on the assumption that $y_i = b_i$. This assumption indicates that no income is underestimated.

Table 2: Australian household income data for 1967-1968

Income	Number	Mean
Below $1000	310	674.39
$1000 - $2000	552	1426.1
$2000 - $3000	1007	2545.79
$3000 - $4000	1193	3469.35
$4000 - $5000	884	4470.33
$5000 - $6000	608	5446.6
$6000 - $7000	314	6460.93
$7000 - $8000	222	7459.14
$8000 - $9000	128	8456.66
$9000 - $11000	112	9788.38
$11000 and over	110	15617.69

4. Results

In Figure 6, we present the distribution of the income for Finland (2009). In this figure, we assume that within every income interval the income is equal to the upper limit. In Figure 7, the cumulative income distribution is presented.

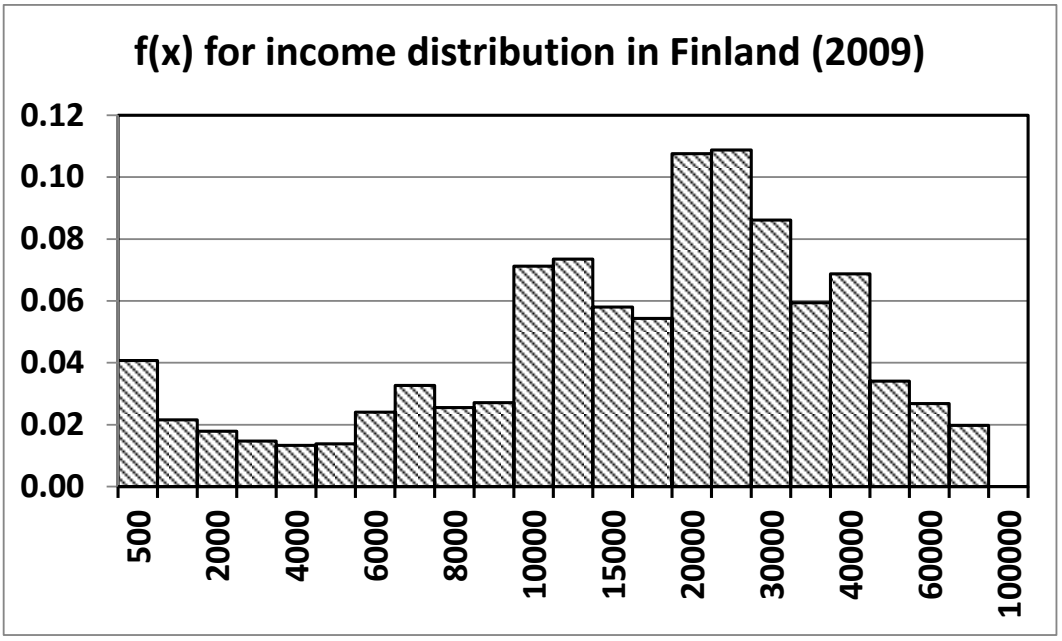

Figure 6: Distribution of the income for Finland (2009)

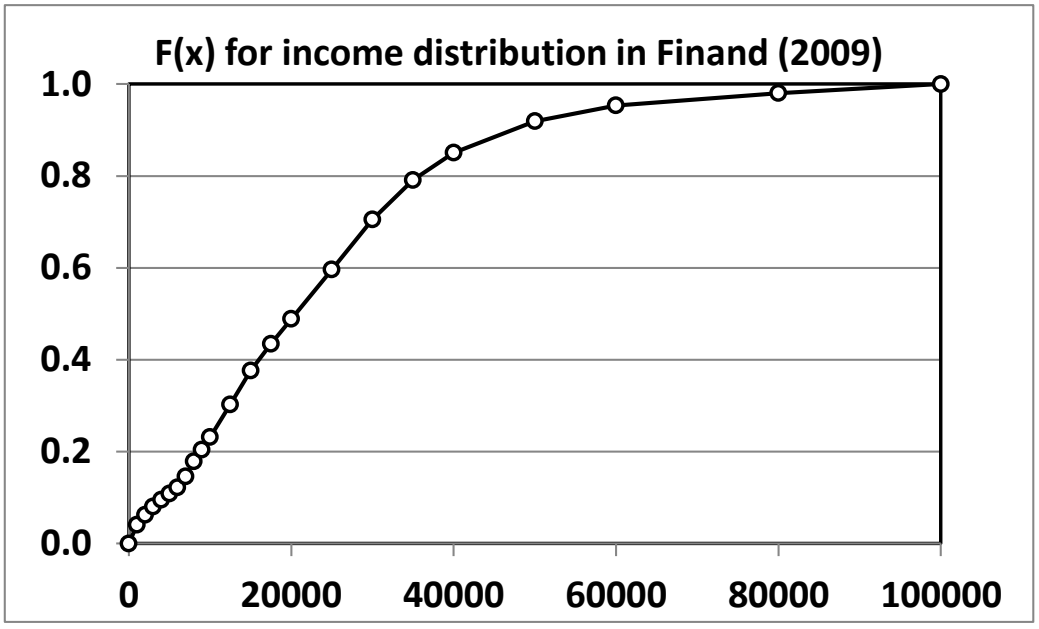

Figure 7: Cumulative income distribution of the income for Finland (2009)

We used already our empirical data when we described the theory in the Methods section. Especially we presented the Lorenz curve for the Finnish data in Figure 2. The Gini index based on the trapezium method is $G = 0.4056$. The regression method yields the slightly higher value 0.4081 [3]. This is in good agreement with the theoretical result that the trapezium rule always yields too low results.
The derivative of $L(p)$ is estimated by approximations based on the difference quotients and on the regression models. The derivative of $L(p)$ as a function of p is given in Figure 4. A good agreement can be observed.
In Figure 3, we used (7) and (8) and presented μ and x_p as functions of p.
In Figure 8, the Lorenz curve and the difference $p - L(p)$ are presented. The maximum of the difference $p - L(p)$ yields $p = 0.596198$, $L(p) = 0.306249$ and $p - L(p) = 0.28995$. Hence, the Robin Hood index is $P = 0.28995$. Note that $G = 0.4081$ is larger than $P = 0.28995$.

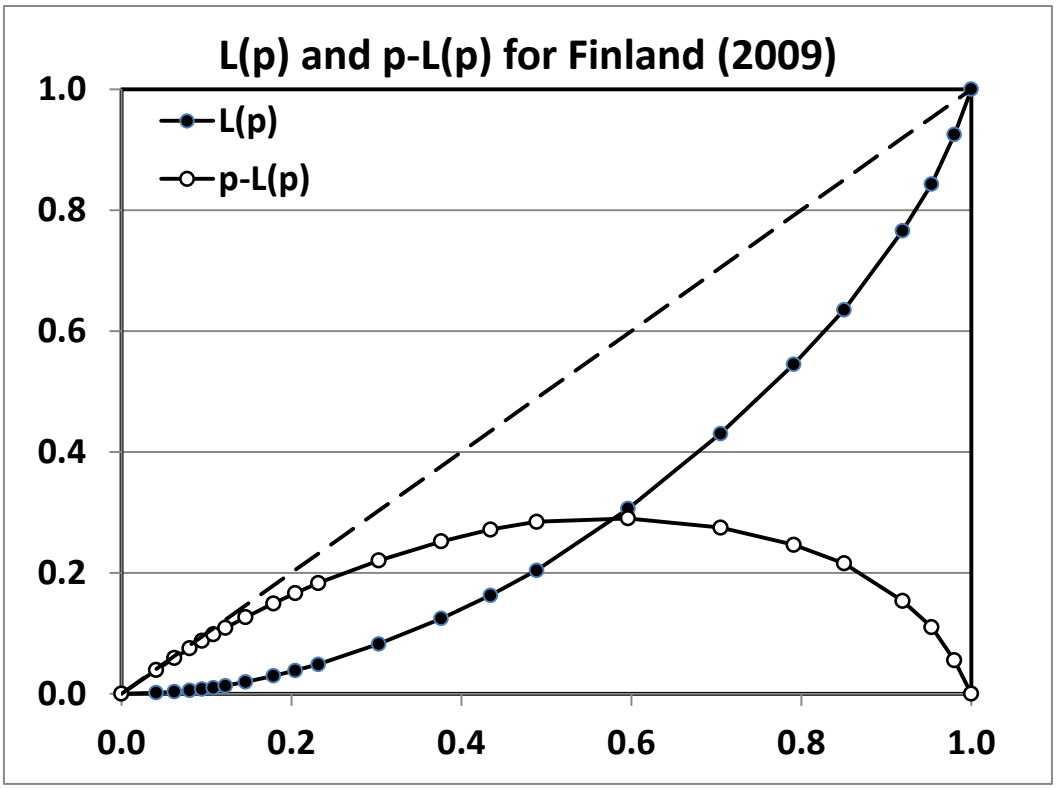

Figure 8: The Lorenz curve and the difference $D = p - L(p)$ as functions of *p*. The maximum $D = 0.28995$ is obtained for $p = 0.596198$

Australia. In general, we analyse the Australian data following the same steps as the analyses of the Finnish data. The Australia results are given in the figures below.

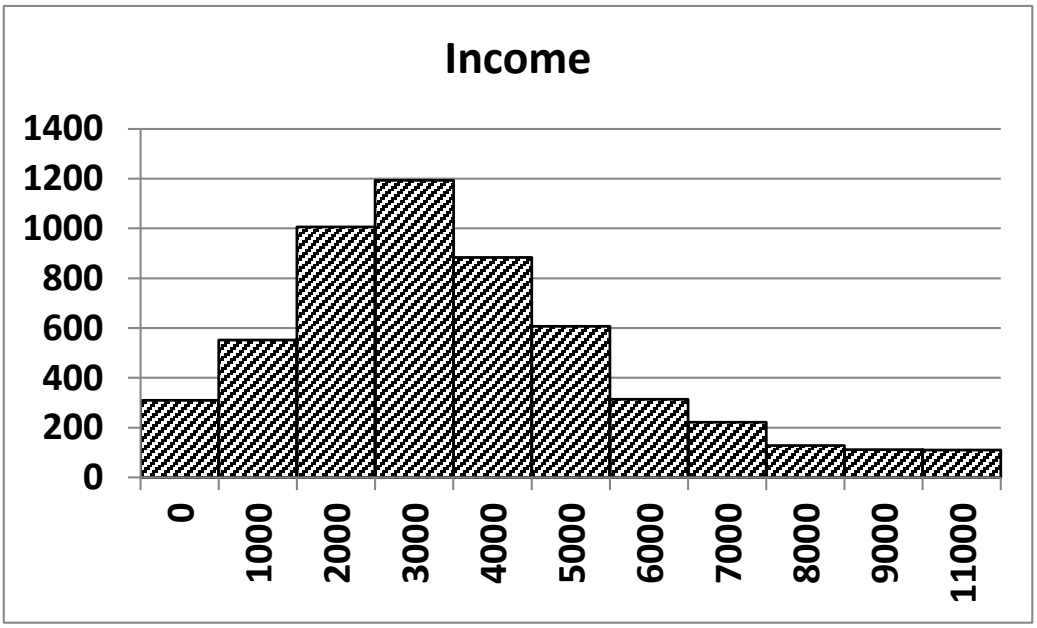

Figure 9: Distribution of the income in Australia (1967-1968)

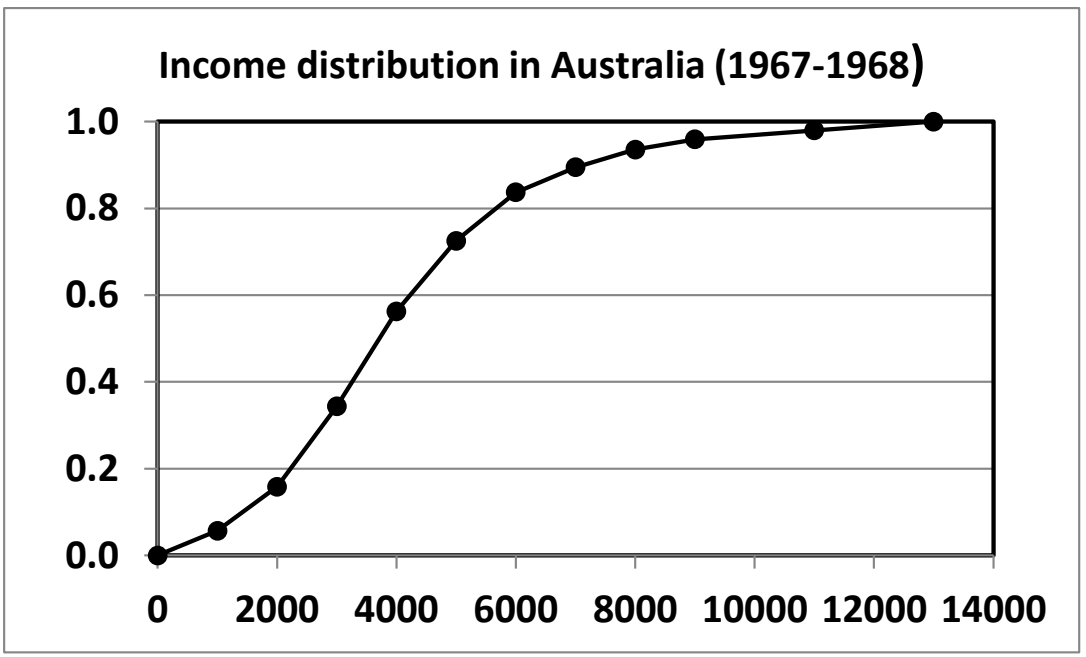

Figure 10: Cumulative income of the income in Australia (1967-1968)

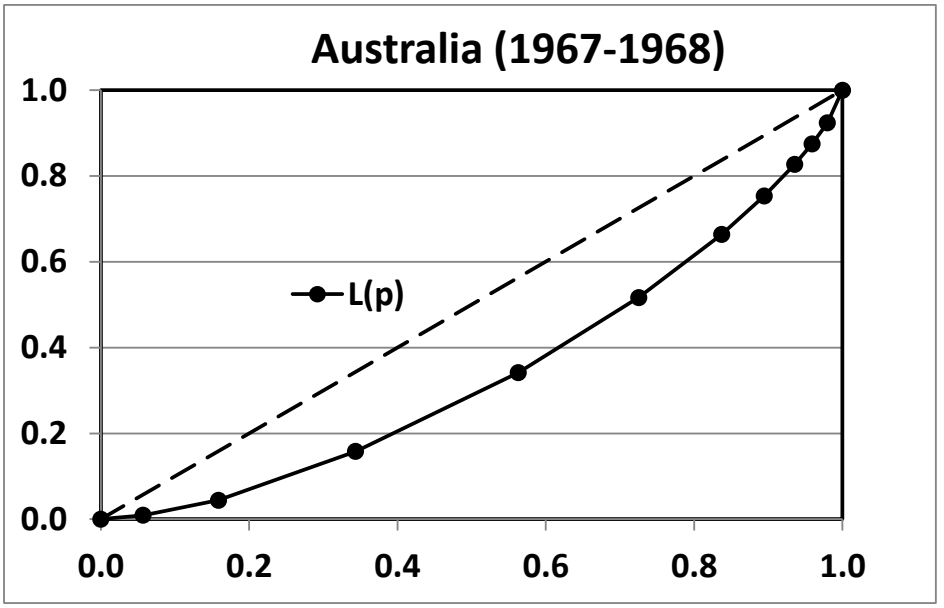

Figure 11: The Lorenz curve $L(p)$ in Australia (1967-1968)

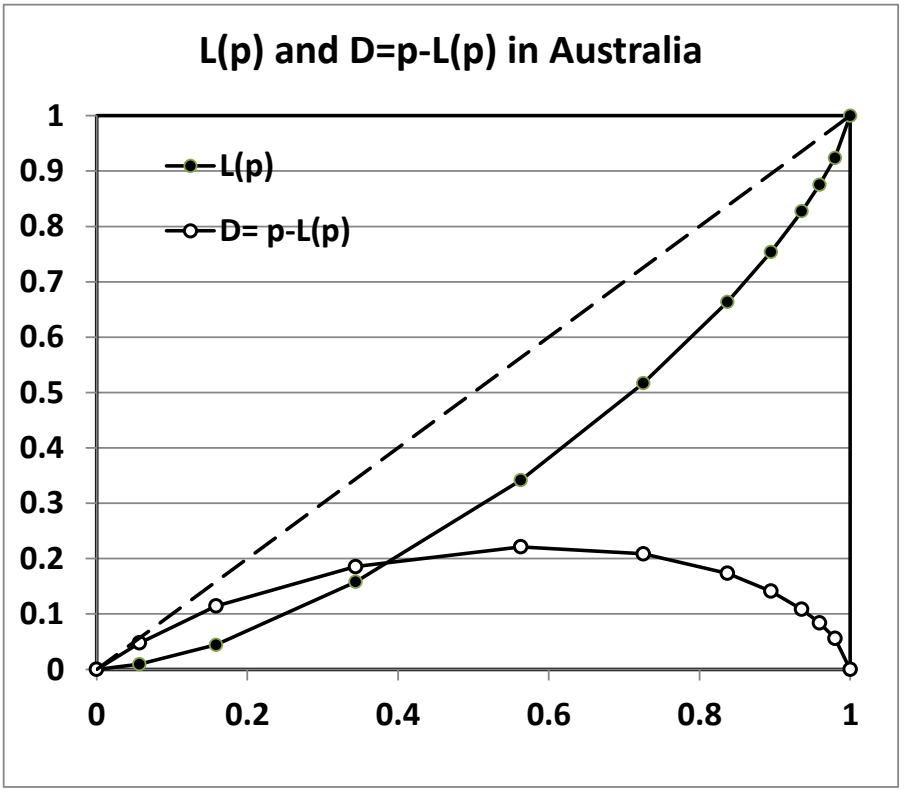

Figure 12: The Lorenz curve $L(p)$ and $D = p - L(p)$ as functions of p in Australia (1967-1968). The maximum of $D = 0.2214$ is obtained for $p = 0.5629$.

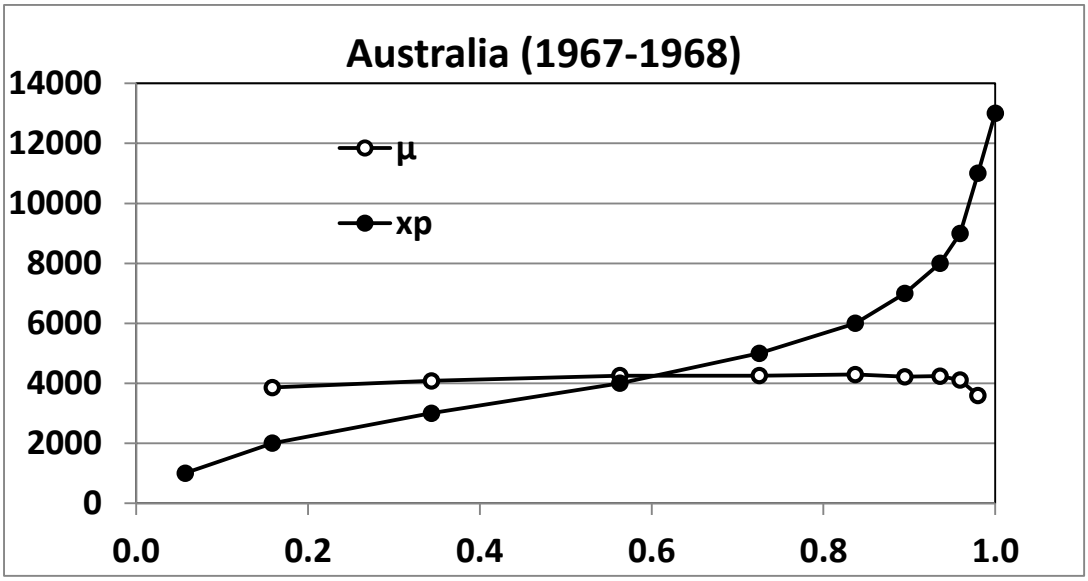

Figure 13: Mean μ and quantile x_p as functions of *p*

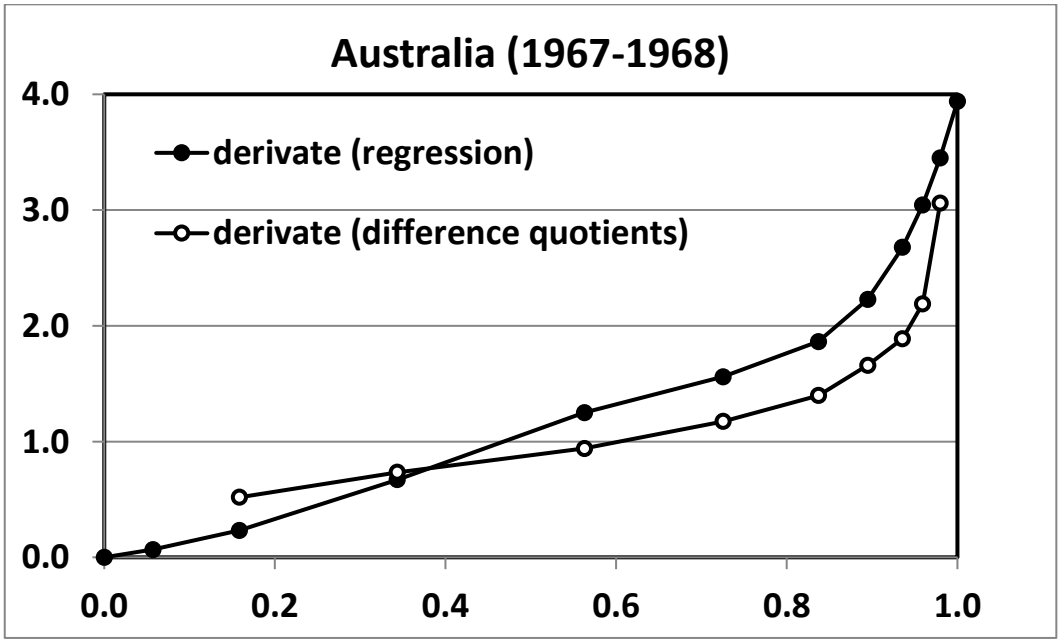

Figure 14: Derivates of the Lorenz curve estimated by the difference quotients and the regression models are included. The correspondence between the findings is poor. The derivatives estimated by the difference quotients and the regression model show marked discrepancies mainly caused by limited data

5. Discussion

Various attempts have been made to produce mostly exact Gini index estimates. Gastwirth [7] introduced interval estimates of the Gini coefficient in order to measure the accuracy of the estimates. Needleman [8] starts from the trapezium estimate of the Gini coefficient G_L. He then introduces an improved upper estimate G_U His final estimate follows the "two-thirds rule", that is, $G = \frac{G_L}{3} + \frac{2G_U}{3}$.

He considered the Γ density and applied Monte Carlo methods.

An obviously better alternative is to approximate the Lorenz curve with Lagrange´s interpolation [9]. Lagrange polynomials of the second degree can be considered as a generalization of Simpson´s rule and do not demand subintervals of equal length, but the number of subintervals should still be even. The polynomials obtained have to be integrated in order to yield approximate areas and Gini coefficients. If the subintervals are of the same length, the Lagrange polynomial method is identical to Simpson´s rule.

Golden [10] showed how a quick approximation of the Gini coefficient can be calculated empirically using numerical data in cumulative income quintiles. Fellman [2] compared different methods. He applied Simpson´s rule and considered Lorenz curves with deciles. In addition, Fellman used Lagrange polynomials and generalizations of Golden´s method.

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 when more detailed studies have already resulted in very accurate estimates, the comparisons are possible. Some authors ([7], [11]-[14]) have introduced interval estimates, but these are often rather broad, and it is still difficult to identify the best method. Gastwirth [7] presents interval estimations of the Gini coefficient. The exact Gini estimate on Current Population Surveys (CPS) income data for 1968 was computed by Tepping, his result being 0.4014. Gastwirth´s Table 2 shows Tepping´s data grouped into a 10-subgroup Lorenz curve. He compares his Gini interval estimates with Tepping´s finding. Gastwirth [7] considers a minimum of restrictive conditions, obtaining the interval $0.3883 < G < 0.4083$. Mehran [11] suggests an alternative estimation method, obtaining the interval estimate $0.3883 < G < 0.4087$. The grouping limits are not equidistant and one cannot apply Simpson´s rule. Applying the trapezium rule yields *0.3883*, and the negative bias is apparent. The Lagrange rule yields *0.4033*, and the modification of the Golden rule yields the rather inaccurate estimate *0.3740*. Such comparison problems are eliminated if the numerical estimations are applied to theoretical distributions with known theoretical indices [5].

Needleman [8] stated that, as the Lorenz curve is convex, the trapezium approximation is always greater than the actual area under the curve, so that the estimate of *G* based on this approximation is always less than the actual value of the coefficient. Furthermore, he noted that most authors using the trapezium approximation indicate that they are aware of the bias involved, but either assume

the error so small as to be insignificant or else use a large number of intervals in the belief, usually justified, that the bias will then be negligible.

In order to perform comparisons between the estimated and theoretical Gini coefficients, Fellman [2] analysed classes of theoretical Lorenz curves with varying Gini coefficients. As an alternative to income distributions, some scientists have built models for the Lorenz curve. Among these, we list the following studies: [5], [15]-[24].

The theoretical step from Lorenz curve to distribution function is more difficult than that from distribution function to Lorenz curve. Fellman [5] noted that with respect to the numbers of the parameters there is a difference between advanced and simple Lorenz models. Advanced Lorenz models yield a better fit to data, but are difficult to exactly connect to 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.

Kakwani and Podder [16] applied their Lorenz model to the Australian data, comparing four alternatives, all of which resulted in accurate estimates. The estimates varied between 0.3195 and 0.3208, when the actual value was 0.3196. Rao and Tam [20] applied the Kakwani-Podder, the generalized Pareto, the Rao-Tam model, the Gupta and the simplified Rao-Tam models to the same data. Their comparison of the models indicates that the Kakwani-Podder, the generalized Pareto and the Rao-Tam model yielded the best estimates. The Gupta and the simplified Rao-Tam model resulted in estimates with the largest errors. For the Gupta model, the estimate was too high (0.3691), and for the simplified Rao-Tam model it was too low (0.2508). These findings support the criticism of the estimation based on simple one-parameter Lorenz models. In this study, the trapezium model yields 0.3134 and our regression model yields 0.3188, which is close to the correct value of 0.3196. Dedduwakumara and Prendergast [6] estimated the Gini index for the Australian data. They used linear interpolation and obtained 0.319 and using Generalized Lambda distribution they obtained the estimate 0.329. Consequently, the agreement between the results presented in different studies is adequate.

Fellman [1] applied the Pareto model on the high incomes of the Finnish data. He assumed that the Pareto model may start from $Y = 25000$ €. For values equal to or greater than that, he obtained the estimate $\hat{\alpha} = 2.637$, and in addition, the coefficient of determination was $R^2 = 0.99241$. For the income distribution for incomes greater than 25000 € the Gini coefficient was $G = 0.234$ t goes here.

6. Conclusions

The composition of the empirical data may vary from study to study. The proposed studies may vary and no general optimal rules are available. Consequently, the attempt proposed in this study cannot be suggested to be a universally applicable strategy. Hence, our study must be considered as an alternative for ongoing discussions, and new alternatives and suggestions are appreciated.

References

[1] Fellman, J, (2015). Mathematical analysis of distribution and redistribution of income. Science Publishing Group 166 pp ISBN: 978-1-940366-25-8, (2015). http://www.sciencepublishinggroup.com/book/B-978-1-940366-25-8.aspx

[2] Fellman, J. (2012a). Estimation of Gini coefficients using Lorenz curves. Journal of Statistical and Econometric Methods. 1(2):2012:31-38.

[3] Fellman, J.(2018).
Regression analyses of income inequality indices. Theoretical Economics Letters, 8:1793-1802. https://doi.org/10.4236/tel.2018.810117

[4] Lee, W.-C. (1999). Probabilistic analysis of global performances of diagnostic tests: Interpreting the Lorenz curve based summary measures. Statistics in Medicine 18:455-471.

[5] Fellman. J. (2012b). Modelling Lorenz curves. Journal of Statistical and Econometric Methods, 1(3):53-62.

[6] Dedduwakumara, D. S. & Prendergast, L. (2019). Interval estimators for inequality measures using grouped data. Preprint 18 pp.

[7] Gastwirth, J. L. (1972). The estimation of the Lorenz curve and Gini coefficient. Rev. Economics and Statistics 54:306-316.

[8] Needleman, L. (1978). On the approximation of the Gini coefficient of concentration. The Manchester School 46:105-122.

[9] Berrut, J.-P. & Trefethen, L. N. (2004). Barycentric Lagrange interpolation. SIAM Review 46 (3):501-517.

[10] Golden, J. (2008). A simple geometric approach to approximating the Gini coefficient. J. Economic Education 39(1):68-77.

[11] Mehran, F. (1975). Bounds on the Gini index based on observed points of the Lorenz curve. J. Amer. Statist. Assoc. JASA 70:64-66.

[12] McDonald, J. B & Ransom, M. R. (1981). An analysis of the bounds for the Gini coefficient. Journal of Econometrics 17:177–188

[13] Rigo, P. (1985). Lower and upper distribution free bounds for Gini´s concentration ratio. Proceedings International Statistical Institute, 45th Session, Amsterdam, Contributed Papers, Book 2:629-630.

[14] Giorgi, G. M. & Pallini, A. (1987). About a general method for the lower and upper distribution-free bounds on Gini´s concentration ratio from grouped data. Statistica 47:171-184.

[15] Kakwani N. C. & Podder N. (1973). On the Estimation of Lorenz Curves from Grouped Observations. International Economic Review 14: 278-292

[16] Kakwani N. C. & Podder N. (1976). Efficient estimation of the Lorenz curve and the associated inequality measures from grouped observations. Econometrica 44:137-148.

[17] Kakwani, N. (1980). On a Class of Poverty Measures Econometrica 4: 437-446

[18] Rasche, R. H., Gaffney, J., Koo A. Y. C. & Obst, N. (1980). Functional Forms for Estimating the Lorenz Curve. Econometrica 48:1061-1062

[19] Gupta, M. R. (1984). Functional form for estimating the Lorenz curve. Econometrica 52:1313-1314.
[20] Rao, U. L. G. & Tam, A. Y.-P. (1987). An empirical study of selection and estimation of alternative models of the Lorenz curve. J. of Applied Statistics 14:275-280.
[21] Chotikapanich, D. (1993). A comparison of alternative functional forms for the Lorenz curve. Economics Letters 41:129-138.
[22] Ogwang, T. & Rao, U. L. G, (2000). Hybrid models of the Lorenz curve, Economics Letters, 69:39-44.
[23] Cheong, K. S. (2002). An empirical comparison of alternative functional forms for the Lorenz curve. Applied Economics Letters 9:171-176
[24] Rohde, N. (2009). An alternative functional form for estimating the Lorenz curve. Economics Letters 105:61-63.