Decomposition of Rank-Dependent Measures of Inequality by Subgroups

Abstract:

The purpose of additive subgroup decomposition is to study the relationship between overall inequality and inequality within and between population subgroups defined by variables like gender, age, education and region of residence. As opposed to the inequality measures that are additively decomposable, the so-called generalized entropy family of inequality measures, the Gini coefficient does not admit decomposition into within- and between-group components but does also require an interaction (overlapping) term. The purpose of this paper is to introduce an alternative decomposition method that can be considered to be a parallel to Lerman and Yitzhaki’s (1985) elasticity approach for decomposing the Gini coefficient by income sources, which means that the elasticity of the Gini coefficient with respect to various income components is treated as the basic quantities of the decomposition method. Thus, rather than decomposing the Gini coefficient or any other inequality measure into a within-inequality term, a between-inequality term and eventually an interaction term, the basic quantities of the introduced method are the effects of marginal changes in variables that are used to specify the population subgroups.

Keywords: The Gini coefficient; the Bonferroni coefficient; rank-dependent measures of inequality; decomposition by subgroups.

JEL classification: [Klikk for å skrive inn tekst]

Acknowledgement: We would like to thank Giovanni Maria Giorgi for helpful comments. Steinar Bjerve gratefully acknowledges the support of The Wessmann Society during the course of this work. Kjell Doksum’s work was supported in part by NSF grants DMS-9971301 and DMS-0505651.

Address: Rolf Aaberge, Statistics Norway, Research Department. E-mail: rolf.aaberge@ssb.no

Steinar Bjerve, Department of Mathematics, University of Oslo. E-mail: steinar@math.uio.no

Kjell Doksum, Department of Statistics, University of Wisconsin-Madison, USA. E-mail: doksum@stat.wisc.edu
Discussion Papers

comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers in PDF are available on the Internet:
http://www.ssb.no
http://ideas.repec.org/s/ssb/dispap.html

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
NO-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no
1. Introduction

The most widely used measure of income inequality is the Gini coefficient, which is defined equal to twice the area between the Lorenz curve and its equality reference\(^1\). The simple and direct relationship between the Gini coefficient and the Lorenz curve appears to be the major reason for its popularity in applied work. However, since empirical analyses of income inequality normally deals with issues that require use of decomposition methods numerous proposals on how to decompose the Gini coefficient by income sources as well as by subgroups has occurred in the literature\(^2\). The purpose of subgroup decomposition is to study the relationship between overall inequality and inequality within and between population subgroups defined by variables like gender, age, education and region of residence\(^3\). As opposed to the inequality measures that are additively decomposable, the so-called generalized entropy family of inequality measures, the Gini coefficient does not admit decomposition into within- and between-group components. However, by adding an extra term that captures the overlap between the marginal income distributions of subgroups it can be demonstrated that the Gini coefficient can be decomposed into three terms, the within-group term, the between-group term and an interaction term\(^4\). Note that the interaction term vanishes when there is no overlapping of income ranks between income units belonging to different subgroups; i.e. when the income distributions of subgroups do not overlap. However, a number of alternative approaches for decomposing the Gini coefficient and other measures of inequality by subgroups could be defined, see Shorrocks (1984). The purpose of this paper is to introduce a new method that can be considered to be a parallel to Lerman and Yitzhaki’s (1985) elasticity approach for decomposing the Gini coefficient by income sources, which means that the elasticity of the Gini coefficient with respect to various income components is treated as the basic quantities of the decomposition method. Thus, we turn the focus from decomposing the Gini coefficient or any other inequality measure into a within-inequality term, a between-inequality term and eventually an interaction (overlapping) term to the effects of marginal changes in the variables that are used to specify the population subgroups.

---

\(^1\) See Giorgi (1990) for a bibliographical portrait of the Gini coefficient.


\(^3\) See e.g Shorrocks (1984).

2. Decomposition of Lorenz curves and rank-dependent measures of inequality

Let $Y$ be a positive, continuous random variable representing wage or income, $X$ a random covariate vector. Leaving out the influence of $X$, the overall Lorenz curve is

$$L(u) = \mu \int_0^1 \left[ I \left[ t \leq u \right] F^{-1}(t) \right] dt = \mu \int_0^1 \left[ I \left[ y \leq F^{-1}(u) \right] \right] y dF(y)$$

where $\mu = \int_0^\infty y dF(y)$ is the mean of $Y$, $F^{-1}$ denotes the left inverse of the distribution function $F$ of $Y$, and $I$ is the indicator function. $L(u)$ gives the proportion of the total amount of income that is owned by the 100u poorest percent of the population. We extend this definition to include the influence of covariates by considering the proportion of the total amount of income that is owned by the subpopulation with covariate values $x$ and with income below the $u$'th quantile in the entire population. To this end we define the pseudo-Lorenz regression curve as

$$\Lambda(u|x) = \frac{1}{\mu} E \left[ Y I \left[ Y \leq F^{-1}(u) \right] | X = x \right] = \frac{1}{\mu} \int_0^1 \int y I \left[ y \leq F^{-1}(u) \right] dF(y|x)$$

(2.1)

$$= \int y \left\{ I \left[ y \leq F^{-1}(u) \right] \right\} \left[ \int_0^1 F^{-1}(u) du \right] dF(y|x), \ 0 \leq u \leq 1.$$

where $F(y|x)$ denotes the distribution function of $Y$ given $X = x$. Although this curve differs from the standard Lorenz curve it has the nice property that it is a decomposition of the Lorenz curve in the sense that its expected value equals the Lorenz curve for the total population, i.e. by using the iterated expectation theorem, see Bickel and Doksum (2001), we find

(2.2)

$$E \left[ \Lambda(u|X) \right] = L(u).$$

As (2.1) shows, this definition of Lorenz regression aggregates incomes from the subgroup with covariate vector $x$, but uses $F^{-1}(u)$ as a common reference when computing proportions\(^5\). This reference quantile $F^{-1}(u)$ is the $u$'th quantile of the overall income distribution $F(y)$ which is obtained by averaging out $x$, that is, $F(y) = E \left[ F(y|X) \right]$. If $X$ is used to partition the sample space into distinct categories $C_1, ..., C_s$ with probabilities $P(C_j) = P(X \in C_j)$, $j = 1, ..., s$, then (2.2) becomes

\(^5\) See Aaberge, Bjerve and Doksum (2005) who have used conditional Lorenz curves for deriving a regression framework for the Lorenz curve and the Gini coefficient.
\[ L(u) = \sum_{j=1}^{n} P(C_j) \Lambda(u|C_j) \]

where

\[ \Lambda(u|C_j) = \frac{1}{\mu} E \{ Y I \left[ Y \leq F^{-1}(u) \right] \mid x \in C_j \} = \frac{\mu(C_j)}{\mu} L(F(F^{-1}(u)) \mid C_j), \]

and \( L(C_j) \) is the Lorenz curve for sub-population \( C_j \). Note that \( \sum P(C_j) \Lambda(u|C_j) = 1 \), but \( \Lambda(I|C_j) = \frac{\mu(C_j)}{\mu} \neq 1 \) except when \( \mu(C_j) = \mu \). Thus, the above decomposition of the Lorenz curve gives a method for identifying the contribution to overall inequality from each subgroup, where the subgroup contributions can be expressed as the product of three components; the proportion of the population that belong to the subgroup, the ratio between the subgroup mean income and the overall mean income and an interaction component that depends on income inequality within the subgroup as well as the relative location of the subgroup distribution.

Similar to (2.3) for the discrete case we get the following expression for the continuous case,

\[ \Lambda(u|x) = \frac{\mu(x)}{\mu} L(g(u)|x) \]

where \( g(u) = F(F^{-1}(u)|x) \) and \( L(|x) \) is the (conditional) Lorenz curve for \( F(y|x) \).

To summarize the information provided by the pseudo Lorenz curve \( \Lambda(u|x) \) we may use the pseudo-Gini coefficient \( 6 \) defined by

\[ \Gamma(x) = \frac{1}{2} \int_0^1 \left[ u - \Lambda(u|x) \right] du = \frac{1}{\mu} E \left[ Y \left[ 2F(Y) - 1 \right] \mid x \right], \]

or alternatively any member of the following family of pseudo inequality measures

\[ \Psi_p(x) = 1 + \int_0^1 P^p(u) \Lambda(|x) du = E \left( Y \left[ 1 - P^p(F(Y)) \right] \mid x \right), \]

---

6 Kakwani (1980) introduced a similar definition in cases where \( x \) is a vector of discrete variable. See also Mahalanobis (1960).
where the weight-function $P'$ is the derivative of a concave function $P$ defined on the unit interval that satisfies the conditions $P(0) = 0$, $P(1) = 1$ and $P'(1) = 0$. Note that the unconditional counterpart of (2.6) is the family of rank-dependent measures of inequality introduced by Mehran (1976). By inserting for $P(u) = 2u - u^2$ in (2.5) we find that $\Psi_P(x) = \Gamma(x)$. As for the pseudo-Lorenz curve we find the following convenient aggregation property for the pseudo inequality measures $\Psi_P(x)$

\begin{equation}
J_P = E[\Psi_P(X)] = EY[1 - P'(F(Y))] = 1 - \int_0^1 P'(u)dL(u) = 1 - \int_0^1 \frac{1}{\mu}P'(u)F^{-1}(u)du.
\end{equation}

As demonstrated by Aaberge (2000) the Gini coefficient attaches an equal weight to a given transfer of income irrespective of where it takes place in the income distribution, as long as the income transfer occurs between individuals with the same difference in ranks. Thus, in general the Gini coefficient favors neither the lower nor the upper part of the Lorenz curve. To supplement the information provided by the Gini coefficient it might be relevant to use the Bonferroni coefficient defined by

\begin{equation}
B = \int_0^1 \left[1 - u^{-1}L(u)\right]du = 1 + \frac{1}{\mu}E[Y \log F(Y)]
\end{equation}

and the pseudo-Bonferroni coefficient defined by

\begin{equation}
B(x) = \int_0^1 \left[1 - u^{-1}A(u|x)\right]du = 1 + \frac{1}{\mu}E\left[Y \log F(Y)\right|x
\end{equation}

Note that $B$ and $B(x)$ corresponds to $J_P$ and $\Psi_P$ for $P(u) = u(1 - \log u)$. As demonstrated by Aaberge (2000) the Bonferroni coefficient $B$ satisfies Mehran's principle of positional transfer sensitivity for any distribution function $F$ and Kolm's principle of diminishing transfers for all $F$ for which $\log F(x)$ is strictly concave. Thus, $B$ is particular sensitive to transfers that occur in the lower part of the income distribution for logconcave distribution functions.

As suggested in Section 1 the main purpose of this paper is not to focus attention on the various components defined by the covariable vector $x$ in cases where $x$ is a vector of discrete variables, but to treat $x$ as a vector of continuous variables and develop a framework that can be

---

\(^7\) Mehran (1976) introduced the $J_P$-family by relying on descriptive arguments, whereas alternative normative motivations of the $J_P$-family and various subfamilies of the $J_P$-family have been provided by Donaldson and Weymark (1980, 1983), Weymark (1981), Yaari (1987,1988), Ben Porath and Gilboa (1994) and Aaberge (2001).


\(^9\) See also Nygård and Sandstrøm (1981) and Giorgi (1998).
considered to provide similar information as the decomposition method in a situation with discrete variables. To this end we introduce the regression coefficients of the regression functions (2.1), (2.5) and (2.6) as quantities that provide information on the influence of covariates on overall inequality.

3. Measuring the effect of covariates on rank-dependent measures of inequality

By exploiting the parallel with the quantile regression approach, Aaberge, Bjerve and Doksum (2005) developed a regression framework for the conditional Lorenz curve, the conditional Gini coefficient and conditional rank-dependent measures of inequality, which can be used to examine the influence of covariates $x$ on income inequality in the conditional distribution ($F(y|x)$) of $Y$ given given $X = x$.

However, since the overall Lorenz curve and the overall Gini coefficient will not be attained by averaging out the covariates in the conditional Lorenz curve and the conditional Gini coefficient, the effects of covariates on the conditional Lorenz curve and the conditional Gini coefficient do not immediately carry over to the overall Lorenz curve and the overall Gini coefficient. Thus, the (conditional) Lorenz and Gini regression coefficients are not the appropriate quantities when focus is turned to the effects of covariates on overall inequality. To this end it appears more relevant to consider the regression coefficients of the pseudo-Lorenz curve and the pseudo-Gini coefficient introduced in Section 2. The pseudo-Lorenz regression coefficient curves are defined by

$$
(3.1) \quad \lambda_j(u|x) = \frac{\partial \Lambda(u|x)}{\partial x_j}, \quad 0 \leq u \leq 1, \quad j = 1, 2, \ldots, s,
$$

and can be considered as measures of the relative importance of the covariate $x_j$ on income inequality$^{10}$. They show how much a small perturbation of $x_j$ for $j=1, 2, \ldots, s$ changes the pseudo-Lorenz curves and allows the effects of the covariates to depend on whether the response is located in the lower, the central or the upper segment of the income distribution. Similarly as for the quantile regression coefficients curves it may be useful to summarize the pseudo-Lorenz regression coefficient curves across the covariates by

$$
(3.2) \quad \lambda_j(u) = E\lambda_j(u; X), \quad 0 \leq u \leq 1, \quad j = 1, 2, \ldots, s.
$$

---

$^{10}$ A similar approach for quantile regression was introduced by Chaudhury et al. (1997).
Note that $\lambda_j(u)$ gives the average change of the pseudo-Lorenz curves due to small change in the j’th covariate when the remaining covariates are first kept fixed, then averaged out. We call $\lambda_j(\cdot)$ the j’th marginal pseudo-Lorenz curve.

To complete the summarization of the pseudo-Lorenz regression coefficients provided by $\lambda_j(u)$ a summary measure that captures the variation across quantiles will be introduced. To this end we may use the pseudo-Gini coefficient as a summary measure of the information content of the pseudo-Lorenz curve. The pseudo-Gini regression coefficients that correspond to (3.1) are defined by

$$(3.3) \quad \gamma_j(x) = \frac{\partial \Gamma(x)}{\partial x_j} = -2 \int_0^1 \lambda_j(u, x) du, \quad j = 1, 2, ..., s.$$ 

Moreover, by summarizing over $x$ we get

$$(3.4) \quad \gamma_j = E \gamma_j(X) = -2 \int_0^1 \lambda_j(u) du, \quad j = 1, 2, ..., s.$$ 

The corresponding pseudo-Bonferroni summary measures are given by

$$(3.5) \quad b_j(x) = \frac{\partial B(x)}{\partial x_j}, \quad b_j = E b_j(X).$$ 

Since alternative methods for summarizing the pseudo-Lorenz regression coefficients may be called for, we introduce the $\Psi_P$-regression coefficients derived from the pseudo-inequality measures defined by (2.5),

$$(3.6) \quad \xi_{Pj}(x) = \frac{\partial \Psi_P(x)}{\partial x_j} = \int_0^1 P''(u) \lambda_j(u, x) du, \quad j = 1, 2, ..., s,$$ 

where $P''$ is the second derivative of the weight-function $P$. By summarizing over $x$ we get

$$(3.7) \quad \xi_{Pj} = E \xi_{Pj}(X) = \int_0^1 P''(u) \lambda_j(u) du, \quad j = 1, 2, ..., s.$$ 

Note that $P(u) = 2u - u^2$ is the $P$-function that corresponds to the Gini coefficient, whilst $P(u) = u(1 - \log u)$ corresponds to the Bonferroni coefficient ($P''(u) = -1/u$).
4. Estimation

We have considered a variety of maps $m: \mathbb{R}^s \to \mathbb{R}$ that measure inequality in income $Y$ as a function of covariates $x \in \mathbb{R}^s$. These surfaces $m(\cdot)$, which are referred to as “curves” in the literature and this paper, can not be displayed effectively, nor estimated efficiently unless the sample sizes are enormous. For this reason we turn to summary measures: The average derivative nonparametric parameter is the gradient vector

$$\nabla m(x) = \left[ \frac{\partial}{\partial x_j} m(x) \right]_{j=1,\ldots,s}^T$$

In the case of single index models, $\nabla m(x)$ is proportional to the single index parameter vector.

Average Derivative Estimates (ADE’s) have been proposed and analysed by Stoker (1986), Härdle and Stoker (1989), Härdle et al. (1993), Chaudhury et al. (1997), and Hristache et al. (2001), among others. Related work on projection pursuit regression appears in Friedman and Stuetzle (1981) and Hall (1989). One basic idea is to estimate the gradient $\nabla m$ locally near a sample point $x_i$ by using locally weighted least squares. That is, use $\hat{\nabla} m_j$, where

$$\hat{\nabla} m_j(x_i) = \arg \min_{a \in \mathbb{R}, \beta \in \mathbb{R}^s} \sum_{j=1}^n \left[ V_j - \left[ a + \beta^T (X_j - X_i) \right] \right]^2 K \left( \frac{|X_j - X_i|^2}{h^2} \right)$$

$$= \sum_{j=1}^n \left( \frac{1}{X_{ij}} \right) \left( \frac{X_j - X_i}{h^2} \right) \left( \frac{X_j - X_i}{h^2} \right)^{-1} \sum_{j=1}^n V_j \left( \frac{1}{X_{ij}} \right) K \left( \frac{|X_j|^2}{h^2} \right).$$

Here $h$ is a tuning parameter selected using the data, $X_{ij} = X_j - X_i$, is Euclidean distance, and the basic data $\{(X_i, V_i)_{i=1}^n \}$ is assumed to be i.i.d. The proceeding references give various modifications of this basic formula in order to deal with regions with sparse data.

To use these methods we need further specification. We have considered the following three m’s:

$$m_1(u;x) = \Lambda(u|x) = \mu^{-1} E[Y I[F(Y) \leq u]|x],$$

$$m_2(x) = \Gamma(x) = 1 - 2\mu^{-1} \left[ E(Y|x) - E(Y F(Y)|x) \right],$$

$$m_3(x) = B(x) = 1 + \mu^{-1} E[Y \log F(Y)|x].$$
Thus, we need ADE’s for the four cases where \( V = Y I[F(Y) \leq u] \), \( V = Y \), \( V = Y F(Y) \) and \( V = Y \log F(Y) \).

Because \( F \) is unknown, we need to replace \( F(Y_i) \) by its empirical version \( \hat{F}(Y_i) = i/n \), where the incomes \( \{Y_i\} \) have been arranged in increasing order and \( X_i \) now denotes the covariate vector that belongs with the \( i \)’th ordered \( Y \). For ease of interpretation and display the ADE algorithms require that each \( X_{ij} \) in the sample have the sample mean \( \bar{X}_j \) subtracted and be divided by the sample standard deviation \( s_j, j = 1,\ldots,s \). Our curves require an estimate of \( \mu = E(Y) \), which we take as \( \hat{\mu} = \bar{Y} \).

We label the outputs from the ADE algorithms as

\[ \hat{\lambda}(u) = n^{-1} \sum_{i=1}^{n} \hat{V}(u; X_i) \] (Lorenz curve in direction \( X_i \))

\[ \hat{\gamma}_j = n^{-1} \sum_{i=1}^{n} \hat{V}(X_j) \] (Gini coefficient in direction \( X_j \))

\[ \hat{b}_j = n^{-1} \sum_{i=1}^{n} \hat{V}(X_j) \] (Bonferroni coefficient in direction \( X_j \))

When there is only one covariate \( X \), estimation is more straightforward. In the case of \( \Lambda(u|X) \) we can apply any nonparametric regression estimator to the data \( (X_i, V_i(u),\ldots,X_n, V_n(u)) \) where

\[ V_i(u) = I(i \leq \lfloor un \rfloor) Y_i \]

where \( \lfloor . \rfloor \) is the greatest integer function. One simple such estimator would be

\[ \hat{\Lambda}(u|x) = \frac{\sum_{i=1}^{n} V_i(u) K_h(X_i - x)}{\bar{Y} \sum_{i=1}^{n} K_h(X_i - x)} \]

where \( K_h(u) = h^{-1} K(u/h) \), \( K(u) \) is a kernel on \( R \) with \( \int K(u) du = 1 \) and \( h > 0 \) a tuning parameter.

The Gini regression index can be estimated as

\[ \hat{\Gamma}(x) = 1 - 2(\bar{Y})^{-1} \hat{\mu}_a(x) \]
where

\[
\hat{\mu}_G(x) = \frac{\sum_{i=1}^{n} \left(1 - \frac{i}{n+1}\right) Y_i K_b(x_i - x)}{\sum_{i=1}^{n} K_b(x_i - x)}
\]

Here the \( \{Y_i\} \) are in increasing order and \( x_i \) is the covariate value the case with ordered response \( Y_i \).

Similarly, the Bonferroni regression index can be estimated as

\[
\hat{B}(x) = 1 + \left( \overline{Y} \right)^{-1} \hat{\mu}_B(x)
\]

where

\[
\hat{\mu}_B(x) = \frac{\sum \log \left( \frac{i}{n+1} \right) Y_i K_L(x_i - x)}{\sum K_L(x_i - x)}
\]
References


Recent publications in the series Discussion Papers

346 B.M. Larsen and R. Nesbakk (2003): How to quantify household electricity end-use consumption
347 B. Halvorsen, B. M. Larsen and R. Nesbakk (2003): Possibility for hedging from price increases in residential energy demand
349 B. Holtsmark (2003): The Kyoto Protocol without USA and Australia - with the Russian Federation as a strategic permit seller
350 J. Larsson (2003): Testing the Multiproduct Hypothesis on Norwegian Aluminium Industry Plants
352 E. Holmøy (2003): Aggregate Industry Behaviour in a Monopolistic Competition Model with Heterogeneous Firms
363 E. Røed Larsen and Dag Einar Sommervoll (2003): Technology, Trade and Inequality
369 T. Skjerpen (2004): The dynamic factor model revisited: the identification problem remains
371 M. Greaker (2004): Industrial Competitiveness and Diffusion of New Pollution Abatement Technology - a new look at the Porter-hypothesis
374 K. Telle and J. Larsson (2004): Do environmental regulations hamper productivity growth? How accounting for improvements of firms' environmental performance can change the conclusion
379 E. Lund Sagen and F. R. Aune (2004): The Future European Natural Gas Market - are lower gas prices attainable?
380 A. Langangen and D. Romøningen (2004): Local government preferences, individual needs, and the allocation of social assistance
381 M. Rege and K. Telle (2004): Effects of inspections on plants' regulatory and environmental performance - evidence from Norwegian manufacturing industries
382 T. A. Galloway (2004): To What Extent Is a Transition into Employment Associated with an Exit from Poverty
383 J. F. Bjørnstad and E. Ytstredt (2004): Two-Stage Sampling from a Prediction Point of View
384 A. Bruvoll and T. Fæhn (2004): Transboundary environmental policy effects: Markets and emission leakages
386 N. Keilman and D. Q. Pham (2004): Empirical errors and predicted errors in fertility, mortality and migration forecasts in the European Economic Area
387 G. H. Bjørtnæs and T. Fæhn (2004): Energy Taxation in a Small, Open Economy: Efficiency Gains under Political Restraints
389 B. Halvorsen (2004): Effects of norms, warm-glow and time use on household recycling