Società italiana di economia pubblica

WORKING PAPER
No 613
October 2008

INEQUALITY DECOMPOSITIONS - A RECONCILIATION

FRANK A. COWELL AND CARLO V. FIORIO

London School of Economics and STICERD. Address: Houghton Street, London WC2A 2AE, UK. email: f.cowell@lse.ac.uk and
University of Milan and Econpubblica. Address: DEAS, via Conservatorio, 7. 20133 Milan, Italy. email: carlo..orio@unimi.it

JEL Classification: D63

Keywords: inequality, decomposition

Abstract

We show how classic source-decomposition and subgroup-decomposition methods can be reconciled with regression methodology used in recent literature. We also highlight some pitfalls that arise from uncritical use of the regression approach. Examples are provided using the LIS database.

- Keywords: inequality, decomposition
- JEL Classification: D63
- Correspondence to: F. A. Cowell, STICERD, LSE, Houghton St, London WC2A 2AE. (f.cowell@lse.ac.uk)

1 Introduction

What is the point of decomposing income inequality and how should we do it? For some researchers the questions resolve essentially to a series of formal propositions that characterise a particular class of inequality measures. For others the issues are essentially pragmatic: in the same way as one attempts to understand the factors underlying, say, wage discrimination (Blinder 1973) one is also interested in the factors underlying income inequality and it might seem reasonable to use the same sort of applied econometric method of investigation. Clearly, although theorists and pragmatists are both talking about the components of inequality, they could be talking about very different things. We might wonder whether they are even on speaking terms.

In this paper we show how the two main strands of decomposition analysis that are often treated as entirely separate can be approached within a common analytical framework. We employ regression-based methods which are commonly used in empirical applications in various field of economics.

The paper is organised as follows. Section 2 offers an overview of the decomposition literature. Our basic model is developed in section 3 and this is developed into a treatment of factor-source decomposition and subgroup decomposition in sections 4 and 5 respectively. Section 6 provides an empirical application, Section 7 discusses related literature and Section 8 concludes.

2 Approaches to decomposition

The two main strands of inequality-decomposition analysis that we mentioned in the introduction could be broadly labelled as "*a priori*" approaches and "explanatory models."

2.1 *A priori* approaches

Underlying this approach is the essential question "what is meant by inequality decomposition?" The answer to this question is established through an appropriate axiomatisation.

This way of characterising the problem is perhaps most familiar in terms of decomposition by subgroups. A coherent approach to subgroup decomposition essentially requires (1) the specification of a collection of admissible partitions – ways of dividing up the population into mutually exclusive and exhaustive subsets – and (2) a concept of representative income for each group. Requirement (1) usually involves taking as a valid partition any arbitrary grouping of population members, although other specifications also make sense (Ebert 1988); requirement (2) is usually met by taking subgroup-mean income as being representative of the group, although other representative income concepts have been considered (Blackorby et al. 1981; Foster and Shneyerov 1999, 2000). A minimal requirement for an inequality measure to be used for decomposition analysis is that it must satisfy a subgroup consistency or aggregability condition – inequality in a component subgroup increases implies, ceteris paribus, that inequality overall goes up (Shorrocks 1984, 1988); the "ceteris paribus" clause involves a condition that the subgroup-representative incomes remain unchanged. This minimal property therefore allows one to rule out certain measures that do not satisfy the axioms from which the meaning is derived (Cowell 1988), but one can go further. By imposing more structure – i.e. further conditions – on the decomposition method one can derive particular inequality indices with convenient properties (Bourguignon 1979, Cowell 1980, Shorrocks 1980), a consistent procedure for accounting for inequality trends (Jenkins 1995) and an exact decomposition method that can be applied for example to regions Yu et al. (2007) or to the world income distribution Sala-i-Martin (2006). By using progressively finer partitions it is possible to apply the subgroup-decomposition approach to a method of "explaining" the contributory factors to inequality (Cowell and Jenkins 1995, Elbers et al. 2008).

The *a priori* approach is also applicable to the other principal type of decomposability – the break-down by factor-source (Paul 2004, Shorrocks 1982, 1983, Theil 1979). As we will see the formal requirements for factor-source decomposition are straightforward and the decomposition method in practice has a certain amount in common with decomposition by population subgroups. Furthermore the linear structure of the decomposition (given that income components sum to total income) means that the formal factor-source problem has elements in common with the regression-analysis approach that we review in Section 2.2.

Relatively few attempts have been made to construct a single framework for both principle types of decomposition - by subgroup and by factor source. A notable exception is the Shapley value decomposition (Chantreuil and Trannoy 1999, Shorrocks 1999), which defines an inequality measure as an aggregation (ideally a sum) of a set of contributory factors, whose marginal effects are accounted eliminating each of them in sequence and computing the average of the marginal contributions in all possible elimination sequences. However, despite its internal consistency and attractive interpretation, the Shapley value decomposition in empirical applications raises some dilemmas that cannot be solved on purely theoretical grounds. As argued by Sastre and Trannoy 2002, provided all ambiguities about different possible marginalistic interpretations of the Shapley rule are cleared, this decomposition is dependent on the aggregation level of remaining income components and is highly nonrobust. Some refinements have been proposed to improve the Shapley inequality decomposition, including the Nested Shapley (Chantreuil and Trannoy 1999) and the Owen decomposition (Shorrocks 1999), based on defining a hierarchical structure of incomes. However, these solutions might face troubles finding a sensible economic interpretation and some empirical solutions can only circumvent the problem without solving it (Sastre and Trannoy 2000, 2002).

2.2 Explanatory models

The second analytical strand of analysis that concerns us here derives from a mainstream econometric tradition in applied economics. Perhaps richest method within this strand is the development of a structural model for inequality decomposition exemplified by Bourguignon et al. (2001), in the tradition of the DiNardo et al. (1996) approach to analysing the distribution of wages. This method is particularly attractive as an "explanatory model" in that it carefully specifies a counterfactual in order to examine the influence of each supposedly causal factor. However, its attractiveness comes at a price: a common criticism is that it is data hungry and, as such, it may be unsuitable in many empirical applications. Furthermore, the modelling procedure can be cumbersome and is likely to be sensitive to model specification.

A less ambitious version of the explanatory-model approach is the use of a simple regression model as in Fields (2003), Fields and Yoo (2000) and Morduch and Sicular (2002). As with the structural models just mentioned, regression models enjoy one special advantage over the methods reviewed in Section 2.1. Potential influences on inequality that might require separate modelling as decomposition by groups or by income components can usually be easily and uniformly incorporated within an econometric model by appropriate specification of the explanatory variables.

2.3 An integrated approach?

It is evident that, with some care in modelling and interpretation, the *a priori* method can be developed from an exercise in logic to an economic tool that can be used to address important questions that are relevant to policy making. One can use the subgroup-decomposition method to assign importance to personal, social or other characteristics that may be considered to affect overall inequality. The essential step involves the way that between-group inequality is treated which, in turn, focuses on the types of partition that are considered relevant. One has to be careful: the fact that there is a higher between-group component for decomposition using partition A rather than partition B does not necessarily mean that A has more significance for policy rather than B (Kanbur 2006). However, despite this caveat, it is clear that there should be some connection between the between-group/within-group breakdown in the Section 2.1 approach and the explained/unexplained variation in the Section 2.2 approach.

We want to examine this connection using a fairly basic model.

3 Basic model

To make progress it is necessary focus on the bridge between formal analysis and the appropriate treatment of data. Hence we introduce the idea of data generating process (DGP), i.e. the joint probability distribution that is supposed to characterize the entire population from which the data set has been drawn.

Consider a set of random variables $\mathbf{H}$ with a given joint distribution $F(\mathbf{H})$, where $\mathbf{H}$ is partitioned into $[Y, \mathbf{X}]$, where $\mathbf{X} = \{X_1, X_2, ..., X_k\}$. Assume that we aim to explain Y as a function of explanatory variables $\mathbf{X}$ and a purely random disturbance variable U and that we can write the relation in an explicit form with Y as function of $(\mathbf{X}, U)$

$$Y = f(\mathbf{X}, U|\boldsymbol{\beta}) \tag{1}$$

where $\boldsymbol{\beta} := (\beta_1, ..., \beta_K)'$ is a vector of parameters. For example, we could think of Y as individual income, of $\mathbf{X}$ as a set of observable individual characteristics, such as age, sex, education, and of U as an unobservable random variable such as ability or luck.

Provided the functional form of f is known, and it is additively separable in $\mathbf{X}$ and U, we can write

$$Y = g(\mathbf{X}|\boldsymbol{\beta}) + U = E(Y|\mathbf{X}) + U \tag{2}$$

where $E(Y|\mathbf{X})$ is the regression function of Y on $\mathbf{X}$, which is used to estimate $\boldsymbol{\beta}$. For simplicity let us assume that the DGP represented by g takes a linear form:

$$Y = \beta_0 + \sum_{k=1}^{K} \beta_k X_k + U \tag{3}$$

Typically one observes a random sample of size n from $F(\mathbf{H})$,

$$\{(y_i, \mathbf{x}_i) = (y_i, x_{1i}, ...x_{ki}), i = 1, ...n\},$$

where the observations are independent over i. One then generates predictions of income for assigned values of individual characteristics using regression methods to compute a vector $\mathbf{b}$, as an estimate of $\boldsymbol{\beta}$. The true marginal distribution function of each random variable, which might be either continuous or discrete, is often unknown in economic applications, as data do not come from laboratory experiments, and one only knows the empirical distribution functions (EDF). The sample analogue of model (3) can be written as:

$$y = \beta_0 + \sum_{k=1}^{K} \beta_k x_k + \upsilon.$$

Provided that the functional form for g in (2) is correctly specified, and that standard assumptions such as exogenous covariates and spherical error variance hold, one could use OLS methods to estimate the income model obtaining

$$y = b_0 + \sum_{k=1}^{K} b_k x_k + u, \tag{4}$$

where b_k is the OLS estimate of β_k, $k = 0, ...k$, $u = y - E(y|x)$ is the OLS residual.

Using the upper case letter for denoting a random variable (whose distribution function is not known in typical survey settings) and the lower case letter for denoting a size-n random sample from the same distribution function, the mean and inequality function of Y are denoted with $\mu(Y)$ and $I(Y)$, the mean and the inequality statistics (i.e. functions of the data) with $\mu(y) = \mu(y_1, ..., y_n)$ and $I(y) = I(y_1, ..., y_n)$.

We can analyse the structure of the inequality of y (or of Y) in two different ways

- *Subgroup decomposition.* Suppose that a subset $T \subseteq \{1, ..., K\}$ of the observables consists of discrete variables such that x_k (X_k) can take the values ξ_{kj}, $j = 1, .., t_k$ where $k \in T$ and t_k is the number of values (categories) that can logically be taken by the kth discrete observable. Then in this case we could perform a decomposition by population subgroups, where the subgroups are determined by the t categories, where $t := \prod_{k \in T} t_k$. This decomposition could be informative – what you get from the within-group component is an aggregate of the amount of inequality that is attributable to the dispersion of the unobservable υ (U) and the remaining continuous observables x_k, $k \notin T$ $(X_k, k \notin T)$. If all the observables were discrete the within-group component would be an aggregation of $I_{y|x}$ $(I_{Y|X})$ and the between-group component would give the amount of inequality that would arise if there were no variation in υ (U).

- *Factor source decomposition.* We can also interpret (3) as the basis for inequality by factor source expressing $I(Y)$ in terms of component incomes $C_1, ..., C_{K+1}$, where

$$C_k := \beta_k X_k, k = 1, ..., K \tag{5}$$

$$C_{K+1} := U \tag{6}$$

– see section 4 below. In this case the term β_0 is irrelevant.

The application of these decomposition methods has been criticised on a number of grounds. Subgroup decomposition is criticised because it requires partitioning the population into discrete categories although some factors (for example, age) are clearly continuous variables. Moreover, handling more than very few subgroups at the same time can be cumbersome. The factor-source decomposition presented in the Shorrocks (1982) form presents the useful property of being invariant to the inequality measure adopted,[1] however it can be criticised as being limited to a natural decomposition rule where total income is the sum of different types of income (for example pension, employment income and capital income). The subgroup and factor source decomposition methods are sometimes criticised as being purely descriptive rather than analytical and as being irreconcilable one with another. Moreover they are tools which are often not well known in some fields of economics where the main focus is on the determinants of income or the market price of personal characteristics, which are estimated as the OLS coefficient in a Mincer-type wage regression.

The two decomposition methods – by population subgroup and by factor source – can be shown to be related to each other. This can be conveniently done using the model that we have just introduced.

4 Decomposition by factor source

Equation (3) is analogous to the case analysed by Shorrocks (1982) where income is the sum of income components (such as labour income, transfers

[1] Actually in some situations this might be regarded as a shortcoming, especially when the the change of inequality has different sign depending on the inequality measure adopted.

and so on). The inequality of total income, $I(Y)$, can be written using a natural decomposition rule such as:

$$I(Y) = \sum_{k=1}^{K+1} \Theta_k \tag{7}$$

where Θ_k depends on C_k and can be regarded as the contribution of factor k to overall income inequality. Define also the proportional contribution of factor k to inequality

$$\theta_k := \frac{\Theta_k}{I(Y)}.$$

Using (5) and (6) Shorrocks (1982)'s results yield:

$$\theta_k = \frac{\sigma\left(C_k, Y\right)}{\sigma^2\left(Y\right)} = \frac{\sigma^2\left(C_k\right)}{\sigma^2\left(Y\right)} + \sum_{j \neq k}^{K+1} \rho(C_k, C_j) \frac{\sigma\left(C_k\right)\sigma(C_j)}{\sigma^2\left(Y\right)}, k = 1, ..., K+1$$

where $\sigma(X) := \sqrt{\text{var}(X)}$, $\sigma(X, Y) := \text{cov}(X, Y)$ and $\rho(C_i, C_j) := \text{corr}(C_i, C_j)$.

Since $\sigma\left(\beta_k X_k, Y\right) = \beta_k \sigma\left(X_k, Y\right)$ we have:

$$\theta_k = \beta_k^2 \frac{\sigma^2\left(X_k\right)}{\sigma^2\left(Y\right)} + \sum_{j \neq k}^{K+1} \beta_k \beta_j \frac{\sigma\left(X_k, X_j\right)}{\sigma^2\left(Y\right)} + \beta_k \frac{\sigma\left(X_k, U\right)}{\sigma^2\left(Y\right)} \tag{8}$$

from which we obtain

$$\theta_k = \beta_k^2 \frac{\sigma^2\left(X_k\right)}{\sigma^2\left(Y\right)} + \sum_{j \neq k}^{K+1} \beta_k \beta_j \rho(X_k, X_j) \frac{\sigma(X_j)\sigma(X_k)}{\sigma^2\left(Y\right)} + \beta_k \rho(X_k, U) \frac{\sigma\left(X_k\right)\sigma(U)}{\sigma^2\left(Y\right)}, \tag{9}$$

for $k = 1, ..., K$ and

$$\theta_{K+1} = \frac{\sigma^2\left(U\right)}{\sigma^2\left(Y\right)} + \sum_{k=1}^{K} \beta_k \rho(X_k, U) \frac{\sigma\left(X_k, U\right)}{\sigma^2\left(Y\right)}. \tag{10}$$

Replacing β_k by its OLS estimate (b_k), and variances, covariances and correlation by their unbiased sample analogues, the estimate of θ_k, (z_k), can be obtained. A similar approach was followed by Fields (2003). Equations (9)-(10) provides a simple and intuitive interpretation and allows one to discuss the contribution of the value of characteristic $k, c_{k,}$ to inequality $I(y)$. Under the assumption that $\text{corr}(C_k, C_r) = 0, r \neq k$, i.e. there is no multicollinearity among regressors and all regressors are non-endogenous, (8) can be simplified to

$$\theta_k = \begin{cases} \beta_k^2 \frac{\sigma^2(X_k)}{\sigma^2(Y)}, k = 1, ..., K \\ \frac{\sigma^2(U)}{\sigma^2(Y)}, k = K+1 \end{cases} \tag{11}$$

and it can be estimated as

$$z_k = \begin{cases} b_k^2 \frac{\sigma^2(x_k)}{\sigma^2(y)}, k = 1, ..., K \\ \frac{\sigma^2(u)}{\sigma^2(y)}, k = K+1 \end{cases} \tag{12}$$

where $\sigma^2(x_k), \sigma^2(y), \sigma^2(u)$ stand for the unbiased sample variance of x_k, y, u, respectively. The sample analogue of the inequality decomposition as in (7) can be written as:

$$I(y) = \sum_{k=1}^{K+1} Z_k = \sum_{k=1}^{K+1} I(y) z_k = \sum_{k=1}^{K} I(y) b_k^2 \frac{\sigma^2(x_k)}{\sigma^2(y)} + I(y) \frac{\sigma^2(\upsilon)}{\sigma^2(y)}. \tag{13}$$

With some simplifications, the right-hand-side of equation (13) might be interpreted as the sum of the effects of the K characteristics and of the error term, although one should consider it as the sum of the *total value* of the K characteristics, i.e. the product of its "price" of each component as estimated in the income regression ($b_k, k = 1, ..., K$) and its quantity ($\mathbf{x}_k, k = 1, ..., K$). One should also notice that the standard errors of (13) are not trivial to compute as they involve the ratio of variances of random variables coming from a joint distribution and the variance of inequality indices can be rather cumbersome to derive analytically (see for instance Cowell 1989). Bootstrap methods are suggested for derivation of standard errors of (13), although they are not presented for the empirical analysis to follow.

Equation (8) shows that θ_k $(k = 1, ..., K)$ can only be negative if

$$\beta_k (\textstyle\sum_{j \neq k} \beta_j \sigma(X_k, X_j) + \sigma(X_k, U)) < -\beta_k^2 \sigma^2(X_k), k = 1, ...K$$

for which a necessary condition is that there be either a nonzero correlation among RHS variables or at least one endogenous RHS variable.

It should be noted here that the decomposition (7) applies for natural decompositions only, i.e. if the LHS variable can be represented as a sum of factors. In the labour-economics literature it is customary to estimate a log-linear relation, such as

$$\log(y) = b_0 + \sum_{k=1}^{K} b_k x_k + u$$

based on arguments of better regression fit and error properties. In this case, the decomposition (7) can only be undertaken with $I(\log(y))$ on the LHS.

5 Decomposition by population subgroups

Assume now that X_1 is a discrete random variable that can take only the values $\{X_{1,j} : j = 1, ..., t_1\}$. If $\text{corr}(X_{1,j}, X_{k,j}) = 0$ and there is no endogeneity of regressors, equation (3) can be represented for each sub-group j as:

$$Y_j = \beta_0 + \beta_{1,j} X_{1,j} + \sum_{k=2}^{K} \beta_k X_{k,j} + U_j. \tag{14}$$

Define $P_j = \Pr(X_1 = X_{1,j})$, the proportion of the population for which $X_1 = X_{1,j}$. Then within-group inequality can be written as

$$I_{\text{w}}(Y) = \sum_{j=1}^{t_1} W_j I(Y_j), \tag{15}$$

where t_1 is the number of groups considered, W_j is a weight that is a function of the P_j, $\mu(Y_j)$ is mean income for particular subgroup j, and Y_j is given by (14). The decomposition by population subgroups allows one to write:

$$I(Y) = I_{\text{b}}(Y) + I_{\text{w}}(Y), \tag{16}$$

where I_{b} is between-group inequality, implicitly defined by (15) and (16) as

$$I_{\text{b}}(Y) := I(Y) - \sum_{j=1}^{t_1} W_j I(Y_j).$$

In the case of the Generalised Entropy (GE) indices we have, for any $\alpha \in (-\infty, \infty)$,

$$W_j = P_j \left[\frac{\mu(Y_j)}{\mu(Y)} \right]^{\alpha} = R_j^{\alpha} P_j^{1-\alpha}, \tag{17}$$

where $R_j := P_j \mu(Y_j)/\mu(Y)$ is the income share of group j; we also have

$$I(Y) = \frac{1}{\alpha^2 - \alpha} \left[\int \left[\frac{Y}{\mu(Y)} \right]^{\alpha} dF(Y) - 1 \right], \tag{18}$$

from which we obtain

$$I_{\text{w}}(Y) = \frac{1}{\alpha^2 - \alpha} \left[\sum_{j=1}^{t_1} P_j \left[\frac{\mu(Y_j)}{\mu(Y)} \right]^{\alpha} \int \left[\frac{Y_j}{\mu(Y_j)} \right]^{\alpha} dF(Y_j) - 1 \right] \tag{19}$$

and

$$I_{\text{b}}(Y) = \frac{1}{\alpha^2 - \alpha} \left[\sum_{j=1}^{t_1} P_j \left[\frac{\mu(Y_j)}{\mu(Y)} \right]^{\alpha} - 1 \right]. \tag{20}$$

Let us now see how decomposition by population subgroups could be adapted to an approach which uses the estimated DGP. Assuming that all standard OLS conditions are fulfilled, and using a $n-$size random sample $\mathbf{y}, \mathbf{x}_1, .., \mathbf{x}_k$ from the joint distribution function $F(Y, X_1, .., X_k)$ one can estimate equation (14) by using dummy variables for identifying different groups obtaining:

$$y_j = b_{0,j} + \sum_{k=2}^{K} b_k x_{k,j} + u_j \tag{21}$$

where $b_{0,j}$ are OLS estimates of $\beta_0 + \beta_{1,j}\mu(x_{1,j})$ in subsample j and u_j are the OLS residuals of each group.

Given the OLS assumptions, the unbiasedness property of OLS estimates allows one to write the mean of y_j in (21), $\mu(y_j) = b_{0,j} + \sum_{k=2}^{k} b_{k,j}\mu(x_{k,j})$.

The estimated between-group inequality I_b can then be written as:

$$I_{\mathrm{b}}(y) = \frac{1}{\alpha^2 - \alpha}\left[\sum_{j=1}^{t_1} p_j \left[\frac{b_{0,j} + \sum_{k=2}^{K} b_k \mu(x_{k,j})}{b_0 + \sum_{k=1}^{K} b_k \mu(x_k)}\right]^{\alpha} - 1\right] \tag{22}$$

where $p_j := n_j/n$ is the population share and n_j is the size of group j. The estimated within-group inequality is written as:

$$I_{\mathrm{w}}(y) = \sum_{j=1}^{t_1} w_j I(y_j)\left(\sum_{k=2}^{K} \frac{b_{k,j}\sigma^2(x_{k,j}) + \sigma^2(u_j)}{\sigma^2(y_j)}\right)$$

where $w_j = (q_j)^{\alpha}(p_j)^{1-\alpha}$ and $q_j := p_j\mu(y_j)/\mu(y)$ is the income share of group j.

In the general case allowing for the possibility that $\mathrm{corr}(X_{1,j}, X_{k,j}) \neq 0$ and that $\mathrm{corr}(X_{1,j}, U) \neq 0$, equation (3) can be represented as:

$$Y_j = \beta_{0,j} + \beta_{1,j}X_{1,j} + \sum_{k=2}^{K} \beta_{k,j}X_{k,j} + U_j \tag{23}$$

and each equation has to be estimated separately. Subgroups decomposition becomes:

$$I_{\mathrm{b}}(y) = \frac{1}{\alpha^2 - \alpha}\left[\sum_{j=1}^{t_1} p_j \left[\frac{b_{0,j} + \sum_{k=2}^{K} b_{k,j}\mu\left(x_{k,j}\right)}{b_0 + \sum_{k=1}^{K} b_k \mu\left(x_k\right)}\right]^{\alpha} - 1\right] \tag{24}$$

where $b_{0,j}$ is now the OLS estimate of $\beta_{0,j} + \beta_{1,j}\mu(x_{1,j})$ and

$$I_{\mathrm{w}}(y) = \sum_{j=1}^{t_1} w_j I(y_j)\left[\left(\sum_{k=2}^{K} b_{k,j}^2 \frac{\sigma^2(x_{k,j})}{\sigma^2(y_j)} + b_{k,j}\sum_{r\neq k} b_{r,j}\rho(x_{r,j}, x_k)\frac{\sigma(x_{r,j})\sigma(x_k)}{\sigma(y)} + b_{k,j}\rho(x_{k,j}, u_j)\frac{\sigma(x_{k,j})\sigma(u)}{\sigma(y)} + b_{k,j}\frac{\sigma\left(x_{k,j}, u\right)}{\sigma^2(y)}\right) + \frac{\sigma^2(u)}{\sigma^2(y)}\right]. \tag{25}$$

It should be noticed that although equations (22) and (24) look the same, they yield different results as the first uses the whole sample while the second only subgroup samples.

6 Empirical application

The method outlined above is applied to real data using the Luxembourg Income Study (LIS) data set.[2] We look at net disposable income for the United States and Finland in mid 1980s and in 2004. We chose United States and Finland as they are two relevant examples of countries belonging to the group of Anglo-Saxon and Scandinavian countries, the first being characterised by higher inequality of after-tax income and a light welfare state, the second being characterised by relatively lower inequality and heavy welfare state – see for example Brandolini and Smeeding (2008a, 2008b). We focus on equivalent income inequality, i.e. on inequality computed over equivalent income, where the square root equivalence scale is conventionally adopted, meaning that each individual is given his family's income normalised by the square root of the family size.

We use these data also because they allow us to compare the distribution of an uniformly defined income variable at approximately the same periods. In fact, four data sets are considered: United States in 1987 and 2004 and Finland in 1987 and 2004. As Table 1 shows that equivalent income inequality in mid 1980s Finland was between 42% and 86% smaller than that in the US, using overall inequality measures such as some GE and the Gini indices, and between 29% and 59% smaller, using some common quantile ratios. Nearly twenty years later, inequality of equivalent income increased in both countries, especially for higher incomes, as GE(2) shows. Although equivalised-income inequality increased relatively more in Finland, it remained consistently lower in Finland with respect to the US.

[2] Data are available from http://www.lisproject.org/. All empirical results can be replicated downloading relevant files as discussed in Appendix B. For a description of the Luxembourg Income Study, see Gornick and Smeeding (2008).

Equivalent disposable income inequality								
	United States			Finland			Finland/US	
	1986	2004	change	1987	2004	change	1986-87	2004
p90/p10	5.778	5.380	-7%	2.375	2.775	17%	-59%	-48%
p90/p50	2.076	2.080	0%	1.482	1.636	10%	-29%	-21%
p50/p10	2.786	2.584	-7%	1.603	1.698	6%	-42%	-34%
p75/p25	2.406	2.402	0%	1.557	1.687	8%	-35%	-30%
GE(0)	0.212	0.256	21%	0.066	0.101	54%	-69%	-60%
GE(1)	0.183	0.244	33%	0.063	0.124	96%	-65%	-49%
GE(2)	0.199	0.350	76%	0.070	0.315	347%	-65%	-10%
Gini	0.335	0.365	9%	0.193	0.240	24%	-42%	-34%

Table 1: Inequality statistics

Although data limitation would not allow very detailed investigation of the difference in inequality between these two countries in two periods, one might start by looking at two important subgroups, i.e. those by sex and by education of the household head, where education is coded into four categories (less than high school, high school, college and Master/PhD). A way to investigate these issues is by using a decomposition by population subgroups of GE indices, which can be decomposed with no residual. Table 2 presents results by education and by sex subgroups. It first presents the measures of inequality computed in each subgroup considered and than shows the within and within decomposition of inequality for the three GE indices, for United States and then Finland in both periods considered. By the exact decomposability property of GE indices, the sum of the within and between components is equal to total inequality. Looking at this table one might conclude that by decomposing it by educational subgroups both the inequality within groups and the inequality between groups increased in both countries. In particular, between group inequality nearly doubled in both countries, while the trend of within-group inequality was more pronounced in Finland. A decomposition by sex of the household head shows a roughly reversed trend of the within and between components: while the former clearly increased in both countries the latter was roughly stable in absolute value in Finland and clearly decreasing in the United States.[3]

From this analysis one cannot disentangle the changed contribution of a demographic characteristic of the population (e.g. education) while controlling for the other (e.g. sex). A possible solution would be to create a finer partition of the sample by interacting education and sex, as proposed in Cow-

[3] A careful analysis of these inequality statistics should also assess the magnitude of the sampling error (Cowell (1989)), however in this paper we use the empirical application as an illustration of the methodologies presented in the previous sections. Further discussions about confidence intervals estimation of inequality measures and its decompositions will be presented in Section 7.

ell and Jenkins (1995). However, this method could become cumbersome if one wanted to control for some additional characteristics (e.g. ethnicity, area of residence), would need a discretisation of variables which might reasonably considered as continuous (e.g. age) and would reduce the sample size in each subgroup, hence the precision of the estimate.

6.1 Implementation of basic model

What additional insights might a regression-based approach yield? To answer this question we estimated a model of equivalent disposable income as (3) where Y is the household equivalent income and as covariates we used, for both countries in both periods, family variables (number of earners, number of children under age 18, whether the family rents or owns its own dwelling) and variables referring to the household head only (age, age squared, sex and four categories dummies for education).[4]

In Table 3 we present results first for the United States and then for Finland. The sample sizes are quite different: in the US there were 32,452 observations in 1986 and 210,648 in 2004, in Finland the sample size decreased from 33,771 in 1987 to 29,112 in 2004, although according to the LIS documentation all four samples are representative of their respective population and this does not seem to have any relevant effect on statistical significance of each regressors included. The first two columns under each year and country presents the OLS coefficient estimates of an equivalent income regression with their p-values, as in equation (4). While number of earners in the family, age and high education of the household head are always positively associated with equivalent household income, number of children younger than 18, a rented dwelling and a female household head are consistently associated with lower equivalent household income in all the four samples considered. These controls are all individually and jointly statistically significant. Their contribution to total variability of the dependent variable in the specified model ranges from over 40% in the case of 1986 US to less than 11% in the case of 2004 Finland.

Clearly this is not a structural model and its specification is unsuitable for a causal interpretation, however it is still informative about the correla-

[4]This is a clearly simplified model of equivalent income generation, however available data would not allow the develpment of a more complex structural model of household income. From a econometric point of view this is an unsatisfactory model to explain the GDP of equivalent disposable income as can also be assessed by the relatively low R-squared statistics of the OLS regressions. For further discussion of this issue, see Section 7.

Subgroups by education

United States

education	1986 GE(0)	1986 GE(1)	1986 GE(2)	2004 GE(0)	2004 GE(1)	2004 GE(2)
< high school	0.222	0.203	0.230	0.223	0.210	0.308
high school	0.177	0.150	0.156	0.210	0.192	0.262
college	0.135	0.127	0.144	0.185	0.182	0.248
Master/PhD	0.144	0.122	0.124	0.217	0.222	0.306
Within	0.179	0.150	0.165	0.206	0.195	0.298
Between	0.033	0.033	0.034	0.050	0.050	0.052

Finland

education	1987 GE(0)	1987 GE(1)	1987 GE(2)	2004 GE(0)	2004 GE(1)	2004 GE(2)
< high school	0.062	0.059	0.061	0.092	0.099	0.131
high school	0.058	0.055	0.061	0.075	0.082	0.193
college	0.051	0.051	0.063	0.102	0.144	0.424
Master/PhD	0.045	0.046	0.048	0.085	0.094	0.121
Within	0.059	0.056	0.062	0.088	0.110	0.300
Between	0.007	0.007	0.008	0.013	0.014	0.014

Subgroups by sex

United States

sex	1986 GE(0)	1986 GE(1)	1986 GE(2)	2004 GE(0)	2004 GE(1)	2004 GE(2)
male	0.183	0.162	0.176	0.226	0.225	0.323
female	0.270	0.246	0.290	0.283	0.263	0.377
Within	0.197	0.170	0.187	0.252	0.241	0.346
Between	0.015	0.013	0.012	0.004	0.003	0.003

Finland

sex	1987 GE(0)	1987 GE(1)	1987 GE(2)	2004 GE(0)	2004 GE(1)	2004 GE(2)
male	0.062	0.060	0.066	0.095	0.116	0.294
female	0.078	0.079	0.093	0.112	0.141	0.369
Within	0.063	0.061	0.068	0.100	0.122	0.313
Between	0.003	0.003	0.002	0.002	0.002	0.002

Table 2: Subgroup inequality decomposition by educational attainment and by sex of the householder.

tion of some key variables on equivalent household income. It should also be noted that the dependent variable (equivalent household income, y) was normalised to its mean in each sample to ensure scale consistency between different samples and coefficient should be interpreted carefully. The constant captures the difference of the welfare state in the US and in Finland. equiv. income, an average twenty-year-old, uneducated, unemployed woman, living alone with no kids, in a rented house would have an income to live on equal to 27% of the mean in US 1986 and even negative (-9% of the mean) in US 2004. The same person would have an income equal to 0.39% and 0.37% of the average income in Finland 1987 and 2004, respectively. In all data sets considered, educational variables are highly relevant and their impact on income is important. Also the gender variable coefficient is relatively large and statistically significant in all samples.

The third column in Table 3 shows the results of the decomposition proposed in Section 4, presenting estimates (z_k) as in (12). Controlling for all the covariates jointly, it emerges that in the US the number of earners in the household, number of children aged less than 18 and a rented dwelling explained about 22% of inequality in 1986 but less than 11% in 2004. Higher education (namely, college and Master/PhD degrees) accounted for roughly 15% of inequality in both years considered, with Master/PhD consistently explaining nearly 10%. In Finland in 1987, number of earners in the household, number of children aged less than 18 and a rented dwelling explained about 14% of total equivalent income inequality and in 2004 about 5%. Higher levels of education are also very important in explaining inequality in Finland, accounting for over 11% and 4% in 1987 and 2004 respectively, although college education is between 3 and 10 times more important than a Master/PhD degree. High school education always has an equalising effect. Female-headed households are associated to higher inequality, although it emerges that for the US the contribution decreased across time roughly by 90% and by 75% in the US and in Finland, respectively.

The inequality decomposition proposed is exact only if the contribution of the residual is not ignored. Indeed, Table 3 shows that after controlling for a set of individual and family characteristics, the residual still accounts for nearly 60% of inequality in the 1986 US and nearly 90% of inequality in the 2004 Finland. It is also worth recalling that this inequality decomposition enjoys the same properties as the factor source decomposition suggested in Shorrocks (1982), namely the fact that it is invariant to the inequality measure used.

It would now be interesting to assess the contribution of (the total value of) each the right-hand-side variable to inequality applying a regression-based factor source decomposition as discussed in Section 4. In other words, our subgroup decomposition would allow us to assess whether one variable contributes uniformly to explain inequality in each subgroup or it has a larger effect only on some of them. As we cannot accept the zero correlation hypothesis for all covariates in each subgroup (see Appendix A), we estimate separate regression for each subgroup as in (23) and present within inequality decomposition estimates as in (25) for education subgroups in tables 4 and 5, and for gender subgroups in tables 6 and 7. The decompositions by education subgroups show that while the number of young children at home contributes similarly to inequality in all education subgroups in the US and the more highly-educated in Finland, the number of earners accounts for a large proportion of inequality in low educated households in both countries. It also shows that the female penalty uniformly decreased across time in both countries, being particularly low for higher levels of education. Looking at the gender subgroups, the highest level of educational attainment contributes to much more of the inequality among males than among females in the US, while college education accounts for roughly the same proportion of inequality in Finland. In all these subgroup decompositions, the largest part of inequality accounting is left in the unobserved characteristics of households.

6.2 Robustness checks

Ideally one would like to provide an analysis of reliability of estimated by producing standard error of the calculations produced. This is however not a trivial task in this context and it would involve the use of the bootstrap. We intend to discuss the correct bootstrap specification of our methodology in a separate paper. However, it is worth providing here a robustness analysis of our results by testing whether they would change if different variables were included. In Table 8 the decomposition estimated provided before are accompanied by a decomposition which also controls for age of the youngest child, number of people aged 65-74 and 75 or over, marital status, ethnicity (black and white for the US and Finnish or Swedish speaking for Finland) and a great number of regional dummies (when available in the data) and area dummies. It shows that although also these variable play a role to account for inequality (especially ethnicity in the US and age of youngest child in Finland), they do not greatly modify the conclusions outlined above. As our methodology is based on regression methods, it also allows us to interpret change of contributions of different individual and household characteristics

	United States					
	1986			2004		
	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.
number of earners	0.117	0.000	7.164	0.148	0.000	4.400
num. children < 18	-0.118	0.000	11.166	-0.083	0.000	2.755
housing rented	-0.154	0.000	3.440	-0.240	0.000	3.201
age	0.020	0.000	7.405	0.024	0.000	4.199
age squared	0.000	0.000	-4.924	0.000	0.000	-2.723
female	-0.200	0.000	2.805	-0.049	0.000	0.290
high school	0.206	0.000	-1.006	0.196	0.000	-1.614
college	0.443	0.000	4.581	0.497	0.000	4.451
master/PhD	0.685	0.000	9.619	0.964	0.000	9.254
cons	0.291	0.000		0.086	0.000	
residual			59.752			75.787
obs	32452			210648		
F	2428.100			7477.260		
Prob > F	0.000			0.000		
R-squared	0.403			0.242		
Adj R-squared	0.402			0.242		
Root MSE	0.484			0.730		

	Finland					
	1987			2004		
	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.
number of earners	0.094	0.000	9.446	0.110	0.000	2.814
num. children < 18	-0.058	0.000	3.677	-0.072	0.000	1.525
housing rented	-0.077	0.000	1.681	-0.122	0.000	1.060
age	0.017	0.000	1.188	0.018	0.000	1.069
age squared	0.000	0.000	1.246	0.000	0.000	-0.264
female	-0.133	0.000	2.001	-0.109	0.000	0.508
high school	0.076	0.000	-0.373	0.030	0.010	-0.212
college	0.359	0.000	10.214	0.279	0.000	3.255
master/PhD	0.458	0.000	1.342	0.676	0.000	0.972
cons	0.394	0.000		0.365	0.000	
residual			69.578			89.273
obs	33771			29112		
F	1640.180			388.540		
Prob > F	0.000			0.000		
R-squared	0.304			0.107		
Adj R-squared	0.304			0.107		
Root MSE	0.293			0.697		

Notes: LHS is equivalent household income. Omitted variables are: housing owned, male, less than high school.

Table 3: OLS equivalent income regression and equivalent income decomposition by factor source as in eq. (12).

							United States 1986					
Education	**Less than high school**			**High school**				**College**		**Master/PhD**		
Popn. share	0.247			0.512			0.139			0.102		
Income share	0.169			0.492			0.181			0.158		
	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.
num. Of earners	0.123	0.000	14.189	0.126	0.000	8.844	0.044	0.000	1.201	0.091	0.000	3.539
num. < 18	-0.074	0.000	11.053	-0.122	0.000	13.804	-0.205	0.000	14.748	-0.202	0.000	14.908
housing rented	-0.122	0.000	4.277	-0.158	0.000	4.127	-0.180	0.000	1.530	-0.324	0.000	3.717
age	0.009	0.000	7.364	0.019	0.000	12.356	0.045	0.000	15.369	0.044	0.000	15.852
age squared	0.000	0.000	-4.461	0.000	0.000	-8.718	0.000	0.000	-13.267	0.000	0.000	-12.578
female	-0.136	0.000	3.319	-0.224	0.000	3.611	-0.277	0.000	1.339	-0.420	0.000	2.212
constant	0.411	0.000		0.497	0.000		0.486	0.000		0.599	0.002	
residual			64.259			65.977			79.080			72.350
obs.	7970			16511			4728			3243		
F	738.160			1418.490			208.150			206.120		
Prob > F	0.000			0.000			0.000			0.000		
R-squared	0.357			0.340			0.209			0.277		
Adj R-squared	0.357			0.340			0.208			0.275		
Root MSE	0.368			0.434			0.623			0.653		
							United States 2004					
Education	**Less than high school**			**High school**				**College**		**Master/PhD**		
Popn. share	0.168			0.478			0.263			0.091		
Income share	0.096			0.420			0.325			0.159		
	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.
num. Of earners	0.153	0.000	17.912	0.161	0.000	8.294	0.116	0.000	2.132	0.095	0.000	0.770
num. < 18	-0.051	0.000	3.709	-0.077	0.000	3.752	-0.112	0.000	2.914	-0.151	0.000	1.595
housing rented	-0.122	0.000	2.712	-0.229	0.000	3.915	-0.344	0.000	2.550	-0.533	0.000	1.498
age	0.004	0.000	1.443	0.021	0.000	6.051	0.040	0.000	5.668	0.085	0.000	3.315
age squared	0.000	0.000	-0.791	0.000	0.000	-4.191	0.000	0.000	-4.299	-0.001	0.000	-1.727
female	-0.032	0.000	0.526	-0.040	0.000	0.299	-0.058	0.000	0.207	-0.137	0.000	0.269
constant	0.361	0.000		0.282	0.000		0.336	0.000		-0.077	0.575	
residual			74.490			81.878			90.828			94.280
obs.	31544			101919			57426			19759		
F	1800.040			3759.270			966.340			199.740		
Prob > F	0.000			0.000			0.000			0.000		
R-squared	0.255			0.181			0.092			0.057		
Adj R-squared	0.255			0.181			0.092			0.057		
Root MSE	0.388			0.578			0.835			1.325		

Notes: LHS variable is equivalent household income

Table 4: Analysis of inequality decomposition by education subgroups.

Finland 1987

Education	**Less than high school**			**High school**			**College**			**Master/PhD**		
Popn. share	0.439			0.431			0.122			0.009		
Income share	0.404			0.424			0.159			0.013		
	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.
num. Of earners	0.101	0.000	17.992	0.092	0.000	8.557	0.039	0.000	0.957	0.096	0.000	4.909
num. < 18	-0.050	0.000	4.354	-0.063	0.000	5.873	-0.084	0.000	5.871	-0.127	0.000	17.572
housing rented	-0.036	0.000	0.713	-0.102	0.000	3.008	-0.158	0.000	2.132	-0.096	0.169	0.446
age	0.011	0.000	-4.236	0.018	0.000	10.671	0.031	0.000	14.348	0.010	0.459	8.858
age squared	0.000	0.000	7.354	0.000	0.000	-7.746	0.000	0.000	-10.601	0.000	0.862	-1.908
female	-0.113	0.000	2.420	-0.146	0.000	2.267	-0.231	0.000	1.511	-0.197	0.237	0.251
constant	0.523	0.000		0.450	0.000		0.502	0.000		0.971	0.005	
residual			71.403			77.369			85.783			69.873
obs.	15672			12927			4803			369		
F	1045.660			629.860			132.480			26.010		
Prob > F	0.000			0.000			0.000			0.000		
R-squared	0.286			0.226			0.142			0.301		
Adj R-squared	0.286			0.226			0.141			0.290		
Root MSE	0.254			0.284			0.403			0.349		

Finland 2004

Education	**Less than high school**			**High school**			**College**			**Master/PhD**		
Popn. share	0.213			0.432			0.342			0.013		
Income share	0.179			0.391			0.410			0.021		
	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.
num. Of earners	0.118	0.000	13.753	0.097	0.000	4.382	0.118	0.000	0.844	0.156	0.004	0.919
num. < 18	-0.058	0.000	1.538	-0.055	0.000	1.979	-0.110	0.000	2.364	-0.105	0.001	5.114
housing rented	-0.117	0.000	3.194	-0.119	0.000	1.907	-0.148	0.000	0.393	-0.414	0.011	1.534
age	0.013	0.000	-4.186	0.019	0.000	4.725	0.011	0.102	0.889	-0.031	0.224	-12.044
age squared	0.000	0.000	7.737	0.000	0.000	-3.403	0.000	0.186	-0.667	0.000	0.057	19.319
female	-0.079	0.000	1.595	-0.115	0.000	1.206	-0.123	0.000	0.286	0.151	0.066	0.254
constant	0.468	0.000		0.352	0.000		0.830	0.000		1.741	0.005	
residual			76.369			89.205			95.892			84.904
obs.	6937			12074			9685			416		
F	357.390			243.390			69.110			12.120		
Prob > F	0.000			0.000			0.000			0.000		
R-squared	0.236			0.108			0.041			0.151		
Adj R-squared	0.236			0.108			0.041			0.139		
Root MSE	0.348			0.493			1.005			0.707		

Notes: LHS variable is equivalent household income

Table 5: Analysis of inequality decomposition by education subgroups.

				United States 1986		
sex		**Male**			**Female**	
Popn. share	0.848			0.152		
Income share	0.903			0.097		
	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.
num. of earners	0.118	0.000	5.834	0.124	0.000	11.725
num. < 18	-0.129	0.000	12.273	-0.068	0.000	10.337
housing rented	-0.166	0.000	2.971	-0.108	0.000	3.405
age	0.023	0.000	8.937	0.004	0.023	2.800
age squared	0.000	0.000	-6.231	0.000	0.491	-0.762
high school	0.227	0.000	-1.426	0.143	0.000	0.610
college	0.457	0.000	4.372	0.415	0.000	5.951
master/PhD	0.711	0.000	10.263	0.499	0.000	4.797
constant	0.208	0.001		0.364	0.000	
residual			63.008			61.138
obs.	26064			6388		
F	1912.100			506.850		
Prob > F	0.000			0.000		
R-squared	0.370			0.389		
Adj R-squared	0.370			0.388		
Root MSE	0.499			0.375		

				United States 2004		
sex		**Male**			**Female**	
Popn. share	0.538			0.462		
Income share	0.579			0.421		
	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.
num. of earners	0.137	0.000	3.136	0.162	0.000	6.019
num. < 18	-0.092	0.000	2.739	-0.074	0.000	2.671
housing rented	-0.231	0.000	2.248	-0.252	0.000	4.313
age	0.026	0.000	4.160	0.022	0.000	3.814
age squared	0.000	0.000	-2.810	0.000	0.000	-2.250
high school	0.215	0.000	-1.883	0.174	0.000	-1.288
college	0.528	0.000	3.960	0.459	0.000	5.104
master/PhD	1.017	0.000	10.659	0.879	0.000	7.045
constant	0.042	0.075		0.067		
residual			77.791			74.572
obs.	112447			98201		
F	4012.560			4185.280		
Prob > F	0.000			0.000		
R-squared	0.222			0.254		
Adj R-squared	0.222			0.254		
Root MSE	0.766			0.685		

Notes: LHS variable is equivalent household income

Table 6: Analysis of inequality decomposition by sex subgroups.

Finland 1987

sex	**Male**			**Female**		
Popn. share	0.912			0.088		
Income share	0.932			0.068		
	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.
num. of earners	0.093	0.000	8.005	0.107	0.000	14.199
num. < 18	-0.059	0.000	5.050	-0.028	0.000	-0.111
housing rented	-0.079	0.000	1.368	-0.068	0.000	2.602
age	0.016	0.000	2.996	0.020	0.000	-7.151
age squared	0.000	0.000	-1.183	0.000	0.000	13.315
high school	0.078	0.000	-0.461	0.056	0.000	-0.043
college	0.361	0.000	10.458	0.345	0.000	9.694
master/PhD	0.460	0.000	1.418	0.439	0.002	0.253
constant	0.413	0.000		0.132	0.004	
residual			72.349			67.243
obs.	31093			2678		
F	1485.020			162.520		
Prob > F	0.000			0.000		
R-squared	0.277			0.328		
Adj R-squared	0.276			0.326		
Root MSE	0.296			0.257		

Finland 2004

sex	**Male**			**Female**		
Popn. share	0.709			0.291		
Income share	0.737			0.263		
	Coef.	$P > \|t\|$	Decomp.	Coef.	$P > \|t\|$	Decomp.
num. of earners	0.109	0.000	2.372	0.111	0.000	3.358
num. < 18	-0.074	0.000	2.108	-0.066	0.000	0.729
housing rented	-0.154	0.000	1.195	-0.069	0.000	0.622
age	0.023	0.000	2.466	0.008	0.006	-0.345
age squared	0.000	0.000	-1.439	0.000	0.002	0.845
high school	0.040	0.003	-0.283	0.001	0.950	-0.011
college	0.283	0.000	3.141	0.272	0.000	3.809
master/PhD	0.626	0.000	0.875	0.822	0.000	1.247
constant	0.239	0.000		0.476	0.000	
residual			89.566			89.745
obs.	20573			8539		
F	299.450			121.840		
Prob > F	0.000			0.000		
R-squared	0.104			0.103		
Adj R-squared	0.104			0.102		
Root MSE	0.700			0.686		

Notes: LHS variable is equivalent household income

Table 7: Analysis of inequality decomposition by sex subgroups.

	United States				Finland			
	1986		2004		1987		2004	
	y	y		y		y		y
number of earners	7.164	7.538	4.400	4.343	9.446	9.447	2.814	2.755
num. children $<$ 18	11.166	9.848	2.755	2.659	3.677	2.874	1.525	1.333
age of youngest child	no	-0.090	no	-0.028	no	0.696	no	-0.101
number aged 65-74	no	0.071	no	0.000	no	-0.110	no	0.084
number aged 75+	no	0.000	no	0.023	no	-0.166	no	-0.020
regional dummies	no	yes	no	yes	no	yes	no	yes
area dummies	no	yes	no	yes	no	no	no	yes
housing rented	3.440	3.811	3.201	3.227	1.681	2.082	1.060	1.374
age	7.405	9.727	4.199	4.164	1.188	1.267	1.069	1.272
age squared	-4.924	-6.948	-2.723	-2.701	1.246	1.419	-0.264	-0.299
female	2.805	2.545	0.290	0.294	2.001	1.429	0.508	0.476
married	no	0.054	no	0.453	no	0.855	no	0.163
ethnicity	no	3.020	no	0.818	no	0.011	no	0.281
high school	-1.006	-0.814	-1.614	-1.283	-0.373	-0.364	-0.212	-0.253
college	4.581	3.872	4.451	3.932	10.214	9.426	3.255	2.945
master/PhD	9.619	8.283	9.254	8.593	1.342	1.241	0.972	0.891
residual	59.752	55.105	75.787	73.772	69.578	66.361	89.273	87.845
Total	100.000	100.000	100.000	100.000	100.000	100.000	100.000	100.000

Table 8: Equivalent income inequality decomposition including additional controls.

as an effect of omitted variable bias, which is relevant only in case omitted and included explanatory variables are strongly correlated.

7 Discussion

Clearly any empirical methodology should come with a set of warnings about implementation: so too with the techniques illustrated in Section 6.

First, it is an exact decomposition only if the residual is not ignored. In other words, it is important to be clear whether inequality of income or inequality of predicted income is being considered. To illustrate how important this may be Table 9 gives the decomposition of equivalent household income inequality ($I(y)$) and the *predicted* equivalent household income inequality ($I(\hat{y})$) for the same data sets considered in Section 6. For instance, on taking a casual glance at inequality decomposition in Finland, one might conclude that college education contribution to inequality did not change substantially between 1987 and 2004, as its contribution to the decomposition of $I(\hat{y})$ decreased only from 33% to 30%. However, this is true only if the focus of the analysis is *predicted* income. Looking at the break-down of inequality of total income, in Finland one may conclude that the contribution of total value of college to equivalent income inequality decreased by over a third, from 10% to 3%, and most of the contribution now lies in the residual.

Second, although the computation of standard errors is sometimes treated

	United States				Finland			
	1986		2004		1987		2004	
Decomposition of:	$I(y)$	$I(\hat{y})$	$I(y)$	$I(\hat{y})$	$I(y)$	$I(\hat{y})$	$I(y)$	$I(\hat{y})$
number of earners	7.164	17.798	4.400	18.173	9.446	31.049	2.814	26.230
num. children < 18	11.166	27.742	2.755	11.376	3.677	12.086	1.525	14.220
housing rented	3.440	8.546	3.201	13.218	1.681	5.527	1.060	9.881
age	7.405	18.399	4.199	17.340	1.188	3.905	1.069	9.967
age squared	-4.924	-12.234	-2.723	-11.245	1.246	4.094	-0.264	-2.462
female	2.805	6.969	0.290	1.199	2.001	6.579	0.508	4.735
high school	-1.006	-2.500	-1.614	-6.665	-0.373	-1.225	-0.212	-1.972
college	4.581	11.381	4.451	18.384	10.214	33.574	3.255	30.343
master/PhD	9.619	23.899	9.254	38.218	1.342	4.410	0.972	9.058
residual	59.752		75.787		69.578		89.273	
Total	100.000	40.249	100.000	24.213	100.000	30.422	100.000	10.727

Notes: LHS is equivalent household income. Omitted variables are: housing owned, male, less than high school. $y = b_0 + \sum_{k=1}^{k} b_k x_k + u$ and $\hat{y} = b_0 + \sum_{k=1}^{k} b_k x_k$.

Table 9: Equivalent income inequality decomposition of total and explained equivalent household incomes.

as a trivial problem (see for instance Morduch and Sicular 2002), this is far from being the case; the main reason for the complexity is that the inequality index computed from a random sample is itself a random variable and cannot be treated as deterministic in the calculation of standard errors (see Section 4); moreover, $I(y)$ often appears at the denominator of these decompositions making theoretical computation of standard errors cumbersome. A viable way to assess the reliability of calculation is by providing different specifications of the regression models used assessing the robustness of results to the inclusion or exclusions of some explanatory variables, as in Section 6.2, or even better by computing standard errors using the bootstrap.

Third, a single-equation model, such as that developed above, should only be interpreted as a descriptive model, showing correlations rather than causal relationships. Could we have done better by opting for a richer model such as the Bourguignon et al. (2001) simultaneous-equation extension of the Blinder-Oaxaca decomposition? Their interest is in the change across time of the full distribution of income and related statistics. The components of their model are an earnings equation for each household member (linking individual characteristics to their remuneration), a labour supply equation (explaining the decision of entering the labour force depending on individual and other household's members decisions) and a household income equation (aggregating the individuals' contributions to household income formation). The estimation of such an econometric model at two different dates allows one to disentangle: (i) a "price effect" (people with given characteristics and the same occupation get a different income because the remuneration structure has changed) (ii) a "participation" or "occupation effect" (individuals with given characteristics do not make the same choices as for entering the labour force because their household may have changed) and (iii) a "population effect" (individual and household incomes change because socio-demographic characteristics of population of households and individuals change). The main merit of such an approach is that it builds a comprehensive model of how decisions regarding income formation are taken, including the individual decision of entering the labour force and wage formation mechanism, into a household-based decision process, extracting part of the information left in the residuals of single-equation linear models as the one used in this paper. Bourguignon et al. (2001) used this methodology to argue persuasively that the apparent stability of Taiwan's income inequality was just due to the offsetting of different forces. However, the rich structural model comes at the expense of increasing the complication of the estimation process and of introducing additional and perhaps questionable assumptions. Among the most important limitations of the Bourguignon et al. (2001) approach are: the robustness of the estimates of some coefficients, the problem of simultaneity between household members' labour-supply decisions, the issue of understanding what is left in the residuals of the labour supply equations and the counterfactual wage equations, the path-dependence problem (i.e. which counterfactual is computed first) is also a problem.[5] In sum, the full structural model approach for inequality analysis can be cumbersome and is likely to be sensitive to model specification.

8 Concluding comments

Our approach to reconciling the various strands of inequality-decomposition analysis is based on a single-equation regression, builds on the Shorrocks (1982) methodology and is aimed at providing a tool for understanding inequality especially when data are not sufficiently detailed to allow a structural model specification. It shares some features with the approach suggested by Fields (2003),[6] but improves on it by including in the analysis the decomposition by subgroups and in showing how this might also be useful to identify differences in determinants of inequality.

It is fairly robust, providing an improvement on other methods, but it

[5] To get some idea of the magnitude of the path-dependence problem the authors computed all possible evaluations of price, participation and population effects, although the complex problem of computing proper confidence intervals for the structural model is not tackled. The problem has something in common with that of the Shapley-value method discussed in section 2.1.

[6] See also Fields and Yoo (2000), Morduch and Sicular (2002).

provides results consistent with other decomposition methods. The simple specification enables one to distinguish clearly between "explanations" of inequality that rely solely on a breakdown of the factors that underlie predicted income and the breakdown of inequality of observed income.

References

Blackorby, C., D. Donaldson, and M. Auersperg (1981). A new procedure for the measurement of inequality within and among population subgroups. *Canadian Journal of Economics 14*, 665–685.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources 8*, 436–455.

Bourguignon, F. (1979). Decomposable income inequality measures. *Econometrica 47*, 901–920.

Bourguignon, F., M. Fournier, and M. Gurgand (2001). Fast development with a stable income distribution: Taiwan, 1979-94. *Review of Income and Wealth 47*, 139–163.

Brandolini, A. and T. M. Smeeding (2008a). Income inequality in richer and OECD countries. In W. Salverda, N. Nolan, and T. M. Smeeding (Eds.), *Oxford Handbook on Economic Inequality*, Chapter 4. Oxford: Oxford University Press.

Brandolini, A. and T. M. Smeeding (2008b). Inequality patterns in western-type democracies: Cross-country differences and time changes. In I. Democracy, P. B. Representation, and R. S. F. C. J. Anderson (eds), New York (Eds.), *Democracy, Inequality and Representation.* New York: Russell Sage Foundation.

Chantreuil, F. and A. Trannoy (1999). Inequality decomposition values: The trade-off between marginality and consistency. Working Papers 99-24, THEMA, Université de Cergy-Pontoise.

Cowell, F. A. (1980). On the structure of additive inequality measures. *Review of Economic Studies 47*, 521–531.

Cowell, F. A. (1988). Inequality decomposition - three bad measures. *Bulletin of Economic Research 40*, 309–312.

Cowell, F. A. (1989). Sampling variance and decomposable inequality measures. *Journal of Econometrics 42*, 27–41.

Cowell, F. A. and S. P. Jenkins (1995). How much inequality can we explain? A methodology and an application to the USA. *The Economic Journal 105*, 421–430.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica 64*, 1001–1044.

Ebert, U. (1988). On the decomposition of inequality: Partitions into nonoverlapping sub-groups. In W. Eichhorn (Ed.), *Measurement in Economics*. Heidelberg: Physica Verlag.

Elbers, C., P. Lanjouw, J. A. Mistiaen, and B. Özler (2008). Reinterpreting between-group inequality. *Journal of Economic Inequality 6*, 231–245.

Fields, G. S. (2003). Accounting for income inequality and its change: a new method with application to distribution of earnings in the United States. *Research in Labor Economics 22*, 1–38.

Fields, G. S. and G. Yoo (2000). Falling labor income inequality in Korea's economic growth: patterns and underlying causes. *Review of Income and Wealth 46*, 139–159.

Fiorio, C. and S. P. Jenkins (2007). Regression-based inequality decomposition, following fields (2003). *UK Stata User Group meeting, 10 September*.

Foster, J. E. and A. A. Shneyerov (1999). A general class of additively decomposable inequality measures. *Economic Theory 14*, 89–111.

Foster, J. E. and A. A. Shneyerov (2000). Path independent inequality measures. *Journal of Economic Theory 91*(2), 199–222. 33.

Gornick, J. C. and T. M. Smeeding (2008). The Luxembourg Income Study. In W. A. J. Darity (Ed.), *International Encyclopedia of the Social Sciences* (2nd ed.)., pp. 419–422. Detroit: Macmillan.

Jenkins, S. P. (1995). Accounting for inequality trends: Decomposition analyses for the UK. *Economica 62*, 29–64.

Kanbur, S. M. N. (2006). The policy significance of decompositions. *Journal of Economic Inequality 4*, 367–374.

Morduch, J. and T. Sicular (2002). Rethinking inequality decomposition, with evidence from rural China. *The Economic Journal 112*, 93–106.

Paul, S. (2004). Income sources effects on inequality. *Journal of Development Economics 73*, 435–451.

Sala-i-Martin, X. (2006). The world distribution of income: Falling poverty and ... convergence, period. *Quarterly Journal of Economics 121*, 351–397.

Sastre, M. and A. Trannoy (2000, December). Changing income inequality in advanced countries: a nested marginalist decomposition analysis. *mimeo*.

Sastre, M. and A. Trannoy (2002). Shapley inequality decomposition by factor components: some methodological issues. *Journal of Economics Supplement 9*, 51Ű90.

Shorrocks, A. (1999, June). Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value. mimeo, Department of Economics, University of Essex.

Shorrocks, A. F. (1980). The class of additively decomposable inequality measures. *Econometrica 48*, 613–625.

Shorrocks, A. F. (1982). Inequality decomposition by factor components. *Econometrica 50*(1), 193–211.

Shorrocks, A. F. (1983). The impact of income components on the distribution of family income. *Quarterly Journal of Economics 98*, 311–326.

Shorrocks, A. F. (1984). Inequality decomposition by population subgroups. *Econometrica 52*, 1369–1385.

Shorrocks, A. F. (1988). Aggregation issues in inequality measurement. In W. Eichhorn (Ed.), *Measurement in Economics.* Physica Verlag Heidelberg.

Theil, H. (1979). The measurement of inequality by components of income. *Economics Letters 2*, 197–199.

Yu, L., R. Luo, and L. Zhan (2007). Decomposing income inequality and policy implications in rural China. *China and World Economy 15*, 44 – 58.

A Appendix A: ancillary empirical results

In Table 10-13 the correlation matrix between RHS variables are presented.

United States 1986

	equiv. income	num. earners	num. <18	rented	age	age squared	female	high school	college	master/ PhD	residual
equiv. income	1.000										
num. earners	0.322	1.000									
	0.000										
num. <18	-0.363	0.022	1.000								
	0.000	*0.000*									
rented	-0.307	-0.182	0.132	1.000							
	0.000	*0.000*	*0.000*								
age	0.176	-0.006	-0.323	-0.292	1.000						
	0.000	*0.269*	*0.000*	*0.000*							
age squared	0.141	-0.069	-0.346	-0.250	0.985	1.000					
	0.000	*0.000*	*0.000*	*0.000*	*0.000*						
female	-0.244	-0.233	0.006	0.240	0.023	0.045	1.000				
	0.000	*0.000*	*0.299*	*0.000*	*0.000*	*0.000*					
high school	-0.062	0.012	-0.007	0.011	-0.170	-0.164	-0.003	1.000			
	0.000	*0.029*	*0.221*	*0.039*	*0.000*	*0.000*	*0.623*				
college	0.187	0.061	-0.070	-0.089	-0.025	-0.035	-0.068	-0.412	1.000		
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*			
master/PhD	0.291	0.061	-0.029	-0.092	0.012	-0.010	-0.077	-0.345	-0.136	1.000	
	0.000	*0.000*	*0.000*	*0.000*	*0.028*	*0.073*	*0.000*	*0.000*	*0.000*		
residual	0.772	0.000	0.000	0.001	0.000	0.000	0.000	0.000	-0.001	0.000	1.000
	0.000	*0.938*	*0.953*	*0.931*	*0.998*	*0.987*	*0.971*	*0.982*	*0.929*	*0.993*	

Table 10: Pairwise correlation matrix of dependent, independent variables and residual . P-value in italics.

United States 2004

	equiv. income	num. earn.	num. <18	rented	age	age sq.	female	high sch.	college	mast/PhD	residual
equiv. income	1.000										
num. earn.	0.232	1.000									
	0.000										
num. <18	-0.182	0.028	1.000								
	0.000	*0.000*									
rented	-0.255	-0.119	0.070	1.000							
	0.000	*0.000*	*0.000*								
age	0.108	-0.132	-0.335	-0.284	1.000						
	0.000	*0.000*	*0.000*	*0.000*							
age sq.	0.074	-0.177	-0.360	-0.243	0.982	1.000					
	0.000	*0.000*	*0.000*	*0.000*	*0.000*						
female	-0.100	-0.090	0.060	0.122	-0.083	-0.070	1.000				
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*					
high sch.	-0.138	0.000	-0.022	0.038	-0.038	-0.030	0.023	1.000			
	0.000	*0.953*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*				
college	0.171	0.045	-0.034	-0.140	-0.011	-0.032	-0.008	-0.571	1.000		
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*			
mast/PhD	0.280	0.010	-0.046	-0.119	0.075	0.057	-0.071	-0.303	-0.188	1.000	
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*		
residual	0.871	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
	0.000	*0.999*	*1.000*	*1.000*	*1.000*	*1.000*	*0.999*	*1.000*	*1.000*	*1.000*	

Table 11: Pairwise correlation matrix of dependent, independent variables and residual . P-value in italics.

Finland 1987

	equiv. income	num. earn.	num. <18	rented	age	age sq.	female	high sch.	college	mast/PhD	residual
equiv. income	1.000										
num. earn.	0.338	1.000									
	0.000										
num. <18	-0.143	0.158	1.000								
	0.000	*0.000*									
rented	-0.185	-0.169	-0.052	1.000							
	0.000	*0.000*	*0.000*								
age	0.020	-0.114	-0.279	-0.183	1.000						
	0.000	*0.000*	*0.000*	*0.000*							
age sq.	-0.021	-0.186	-0.314	-0.141	0.984	1.000					
	0.000	*0.000*	*0.000*	*0.000*	*0.000*						
female	-0.186	-0.257	-0.158	0.165	0.107	0.138	1.000				
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*					
high sch.	-0.035	0.004	0.042	0.036	-0.330	-0.312	-0.018	1.000			
	0.000	*0.457*	*0.000*	*0.000*	*0.000*	*0.000*	*0.001*				
college	0.305	0.036	0.066	-0.052	-0.019	-0.034	-0.031	-0.324	1.000		
	0.000	*0.000*	*0.000*	*0.000*	*0.001*	*0.000*	*0.000*	*0.000*			
mast/PhD	0.109	0.017	0.031	-0.030	0.010	0.002	-0.025	-0.083	-0.035	1.000	
	0.000	*0.002*	*0.000*	*0.000*	*0.082*	*0.673*	*0.000*	*0.000*	*0.000*		
residual	0.834	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
	0.000	*0.998*	*0.986*	*0.996*	*0.993*	*0.993*	*1.000*	*1.000*	*0.981*	*0.988*	

Table 12: Pairwise correlation matrix of dependent, independent variables and residual . P-value in italics.

Finland 2004

	equiv. income	num. earn.	num. <18	rented	age	age sq.	female	high sch.	college	mast/PhD	residual
equiv. income	1.000										
num. earn.	0.170	1.000									
	0.000										
num. <18	-0.086	0.359	1.000								
	0.000	*0.000*									
rented	-0.151	-0.235	-0.052	1.000							
	0.000	*0.000*	*0.000*								
age	0.034	-0.157	-0.290	-0.245	1.000						
	0.000	*0.000*	*0.000*	*0.000*							
age sq.	0.009	-0.228	-0.325	-0.197	0.983	1.000					
	0.147	*0.000*	*0.000*	*0.000*	*0.000*						
female	-0.076	-0.138	-0.118	0.142	0.036	0.045	1.000				
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*					
high sch.	-0.106	0.068	0.064	0.070	-0.196	-0.193	-0.049	1.000			
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*				
college	0.181	0.095	0.044	-0.123	-0.057	-0.085	0.045	-0.629	1.000		
	0.000	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*			
mast/PhD	0.095	0.013	0.001	-0.049	0.022	0.014	-0.011	-0.099	-0.082	1.000	
	0.000	*0.023*	*0.912*	*0.000*	*0.000*	*0.020*	*0.054*	*0.000*	*0.000*		
residual	0.945	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
	0.000	*0.999*	*0.996*	*0.993*	*0.995*	*0.996*	*0.997*	*0.996*	*0.993*	*0.998*	

Table 13: Pairwise correlation matrix of dependent, independent variables and residual . P-value in italics.

B Appendix B: Files to replicate empirical results

All empirical results are computed using Stata (www.stata.com) on a remote machine, resident at LIS, and can be replicated using the relevant files from: http://www.economia.unimi.it/users/fiorio/ftp/projects/CowelFiorio08_IneqDec/CowellFiorio08_IneqDec.zip. The main results are obtained using a modification of the Stata routine ineqrbd (Fiorio and Jenkins 2007), which can also be downloaded from Stata typing "ssc install ineqrbd, replace" in the Stata command line.