Explaning income inequality in Italy: results from a regression-based decomposition approach

Regoli, Andrea
University of Naples "Parthenope", Department of Statistics and Mathematics for Economic Research
via Medina, 40
Naples 80133, Italy
andrea.regoli@uniparthenope.it

Manna, Rosalba
University of Naples "Parthenope", Department of Statistics and Mathematics for Economic Research
via Medina, 40
Naples 80133, Italy
rosalba.manna@uniparthenope.it

1.Introduction

This paper focuses upon the determinants of the observed income differentials. More precisely, the aim of this study is to decompose the income inequality among individuals while measuring the contribution of different individual and household factors through a regression-based decomposition strategy.

Heterogeneity across individuals and across time is accounted for by using the longitudinal information of the sample, thereby fully exploiting the potential of panel data.

A wide literature exists on the decomposition of inequality measures. The traditional methods include the decomposition by income sources (Shorrocks, 1982) and by population subgroups (Shorrocks, 1984).

The former method estimates the contribution of individual income components to the observed inequality, whereas the latter allows to measure inequality both within and between subgroups of the population. Both of them are typically descriptive methods that tell us what sources of incomes or subgroups account for inequality but they fail to detect and measure the contributions of individual determinants to income inequality. For this reason, the information provided by those methods is of limited usefulness for policy-makers seeking to address income inequality problems.

Unlike the traditional methods, the regression-based approach followed in this work has the advantage of going beyond decomposing inequality simply in terms of income components or discrete population categories. Indeed, it enables to include any factor that may drive the observed inequality, such as economic, social, demographic and policy variables, both discrete and continuous. Moreover the regression-based method can manage problems of endogeneity due to reverse causality.

The regression-based decomposition methodology was proposed in the early 1970s (Blinder, 1973; Oaxaca, 1973), but failed to arouse much interest until Morduch and Sicular (2002) and Fields (2003) devised a regression-based decomposition by income determinants through the extension of the decomposition by income sources. Regression-based decompositions start with the estimation of an income-generating function, and then use the estimated coefficients to derive the inequality weight of every explanatory variable.

In the context of regression-based decomposition, many recent studies proposed the application of the Shapley value approach, a concept taken from cooperative game theory (Sastre and Trannoy, 2002; Wan, 2004; Israeli, 2007; Guanatilaka and Chotikapanich, 2009; Devicienti, 2010).

In the wake of these methodological contributions, the present paper applies the Shapley approach to Italian panel data by using the Gini index as inequality measure with the aim of measuring the effect of individual and household factors on the income inequality.

This paper is organized as follows. In Section 2 the theoretical background on the subject is presented

with reference to the Shapley value approach. Section 3 deals with the model selection and specification whereas the empirical data from the Historical Database of the Bank of Italy's survey are presented in Section 4. Section 5 summarizes the estimation results and some conclusions are drawn.

2.The methodology of the Shapley value approach

The Shapley value approach, as introduced by Shorrocks (1999), yields an exact additive decomposition of any inequality measure into its contributory factors. The inequality measure calculated on the predicted income values $I(Y \mid X_1, X_2,..., X_k)$ is expressed as the sum of the contributory factors:

$$I(Y \mid X_1, X_2,..., X_k) = \Phi(X_1, I) + \Phi(X_2, I) + ... + \Phi(X_k, I) \quad (2.1)$$

The Shapley decomposition calculates the marginal impact of each factor $\Phi(X_i, I)$ $i = 1,2,...,k$ through the estimation of a sequence of regression models starting from the specification which includes all the regressors and then successively eliminating each of them. The overall marginal contribution of each variable is then obtained as the average of its marginal effects: since the contribution of any factor depends on the order in which the factors appear in the elimination sequence, this average is calculated over all the possible elimination sequences.

The contribution $\Phi(X_i, I)$ of the factor X_i to the explanation of the inequality measure I is given by the following formula:

$$\Phi(X_i, I) = \frac{1}{k!} \sum_{\pi \in \Pi_k} \left[I\left(Y \mid B(\pi, X_i) \cup \{X_i\}\right) - I\left(Y \mid B(\pi, X_i)\right) \right] \quad (2.2)$$

where $I(Y \mid \mathbf{X})$ is the inequality indicator calculated on the predicted income values from the regression on the vector of explanatory variables $\mathbf{X}$;
Π_k is the set of all the possible orderings (permutations) of the k variables;
$B(\pi, X_i)$ is the set of the variables preceding X_i in the given ordering π.

The calculation of each factor's contribution requires the estimation of $2^k - 1$ income generating models, and then the derivation of the inequality indicator I using the income predicted values for every model.

Finally, the proportion of unexplained inequality $I_R(Y)$ is obtained as the difference between the inequality measure calculated on the observed income values $I(Y)$ and the same measure calculated on the predicted income values, as follows:

$$I_R(Y) = I(Y) - I(Y \mid X_1, X_2,..., X_K) \quad (2.3)$$

3.The income generating function for panel data

The first step in the regression-based decomposition of income inequality requires the specification and the estimation of an income generating function, that is a model where income is regressed on some explanatory variables accounting for individual and household characteristics.

We specified a panel data regression model with time-invariant unobserved effects (Wooldridge, 2002), which can be written as:

$$\ln y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \quad t = 1,2,...,T \quad i = 1,2,...,N \quad (3.1)$$

where $\mathbf{x}_{it}$ is a $1 \times K$ vector of regressors, c_i is the time-constant, individual-specific effect and u_{it} is the disturbance term for which the strict exogeneity condition is assumed to hold, that is

$$E(u_{it} \mid \mathbf{x}_{i1},...,\mathbf{x}_{iT}, c_i) = 0 \quad t = 1,2,...,T \quad (3.2)$$

This assumption implies that each error term u_i is uncorrelated with the regressors at all time periods, namely

$$E(\mathbf{x}'_{is} u_{it}) = 0 \quad s,t = 1,2,...,T \quad (3.3)$$

The two core specifications of such models are known as Random Effects (RE) and Fixed Effects (FE) models. In particular, we have specified a RE model, where the individual effect c_i is treated as a random variable that adds to the error term u_{it}. This choice is justified primarily by the RE model using both the "between variation" (the variability across individuals) and the "within variation" (the variability over time). For this reason, unlike the FE model, it allows both to estimate the coefficients of the regressors that do not vary at all over time (with null within variation) and to measure with no efficiency loss the effects of regressors that display a small within variation. In this study, the dependent variable is represented by the (log of) individual net disposable income, whereas the regressors include, among others, gender (that is invariant over time) and the years of completed studies (that exhibit a little variation over time).

Our preference for the RE model is also explained by the fact that we are not interested in estimating the values of the unobserved term for some specific individuals, but instead we concentrate on the influence of individual and household factors on the disposable income of hypothetical individuals with given characteristics. In the situation where the individuals are drawn randomly from a large population, as is usually the case for household panel studies, the RE model is an appropriate specification (Baltagi, 2008).

The RE estimator is derived under the further assumption of uncorrelation (orthogonality) between the individual effect c_i and the observed explanatory variables $\mathbf{x}_{it}$:

$$E(\mathbf{x}'_{it} c_i) = 0 \quad t = 1,2,...T \quad (3.4)$$

This means that all the regressors $\mathbf{x}_{it}$ are considered to be exogenous.

The choice of the semi-log functional form along with the selection of the explanatory variables were informed by the human capital theory suggesting that the ability to earn income is influenced by the educational level, age and other individual characteristics such as gender and work status, plus the geographical area of residence and a measure of household wealth. Along with age, squared age was included in order to capture the concavity of the income-age profile. The effects of age and squared age on inequality may be then easily added up in order to get a single contribution.

4.Data source, variables and summary statistics

The data used in this work are drawn from the Survey of Household Income and Wealth (SHIW) conducted every two years by the Bank of Italy on a sample of about 8,000 Italian households.

For every survey, the sample is composed of both households that have been already interviewed in previous years (panel households) and fresh households.

We have referred to the Historical Database of the survey (Banca d'Italia, 2010) from which we have selected information on the income earners who have been successfully interviewed from 1998 to 2008. Such information took the form of a balanced micro panel where a large number of individuals N (N=1712) have been observed over a short time period T (T=6 years covering on the whole a time span of 10 years).

Descriptive statistics for the variables introduced in the model are presented in Table 1.

Table 1: Descriptive statistics

Variable	Definition	Obs	Mean	Std. dev.
logY	(Log of) Net disposable income	10272	9.65	0.75
Gender	=1 for male; =0 for female	10272	0.60	0.49
Education	Years of completed study	10272	9.25	4.16
Age	Age (in years)	10272	56.18	14.13
Age squared		10272	3355.22	1594.78
Work status	=1 for employed;=0 for not employed	10272	0.51	0.50
Area	=1 for North and Centre;=0 for South and Islands	10272	0.71	0.45
Wealth	Real and financial wealth (in thousands of euro)	10272	265.18	368.86

The net disposable income is defined as the sum of individual income from wages, self-employment, pensions and other transfers, and property income, from both real and financial assets. Every income item is reported after tax and social security contributions. Negative or null income values were given null log (income) values.

5. Results and discussions

The regression coefficients in Table 2 come from the estimation of the Random Effects saturated model, that is the model including all the explanatory variables.

The signs of the estimated coefficients are in line with the theoretical expectations. The concavity of the income-age profile is confirmed by the positive coefficient for age and by the negative coefficient for squared age. Larger income flows are associated with larger stocks of wealth. Significant income gaps are due to gender, level of education, work status and area of residence: ceteris paribus, on average the males, the more educated, the employed and those who live in northern or central regions enjoy higher income levels. An overall R^2 equal to 0.35 indicates a satisfactory fit of the income regression model, when compared with other studies on the same phenomenon. We might have improved the fit by including interaction terms, but this would have created some problems in correctly assigning the resulting effect to the variables included in the interaction term.

Table. 2 Random effects model estimation

Explanatory variable	Coefficient	Standard error
Gender	0.5045***	0.0218
Education	0.0536***	0.0026
Age	0.0466***	0.0039
Age squared	-0.0002***	0.0000
Work status	0.3633***	0.0192
Geographical area	0.1852***	0.0237
Household wealth	0.0004***	0.0000
Constant	6.5951***	0.1164
R^2 overall = 0.35		
N=1712; T=6		
Wald chi-squared(7)=2473.69;		
p-value=0.00		
***: significant at the 1% level		

The results for the Shapley decomposition of Gini index are presented in Table 3.

The contributions of individual and household factors altogether account for nearly 75% of the observed inequality. The largest part of income inequality (more than 19%) is explained by the gender. In Italy women usually find some difficulties in combining work and family and for this reason they are likely to choose not to participate in the labour force. On the other hand, women who have a job earn on average

lower salaries and have usually fewer opportunities to reach positions of leadership compared to men.

The educational level is the factor that shows the second largest contribution to the Gini index (18.3%). Differences in the years of education and therefore in the returns to education are indeed crucial for the observed income differences.

The contribution of educational level added up to the contribution of age shows that the variables related to human capital (education and age) together explain little more than one third of the sample income inequality.

A non-negligible weight is associated with the wealth stock of the family of origin. While apparently different, individual and household factors are intertwined. Indeed the human capital endowments are quite strongly correlated with the financial wealth of the family of origin.

The remaining variables, occupational status (4.9%) and geographical area (3.4%), are much less essential as determinants of inequality.

Table 3: Gini inequality decomposition

Variables	Absolute Contribution	Percentage Contribution
GENDER	6.6	19.4
EDUCATION	6.3	18.3
AGE	5.7	16.5
WORK STATUS	1.7	4.9
GEOGRAPHICAL AREA	1.2	3.4
HOUSEHOLD WEALTH	4.2	12.3
Total Explained Inequality	25.6	74.8
Unexplained Inequality	8.6	25.2
Observed Inequality	34.3	100.0

This is tantamount to saying that, once human capital, gender and wealth are taken into account, whether an individual is unemployed or not, and whether he or she lives in the North or in the South, have only a minor impact on income gaps.

These results seem to run counter to the ingrained belief that the North-South divide is the key driver for the economic inequality in Italy.

REFERENCES

Baltagi B.H. (2008), *Econometric analysis of panel data*, Third edition, John Wiley & Sons Ltd, Chichester, England.

Banca d'Italia (2010), Historical Database of the Survey of Italian Household Budgets, 1977-2008, SHIW-HD, version 6.0, February 2010, On line at: http://www.bancaditalia.it/statistiche/indcamp/bilfait/docum/Shiw-Historical-Database.pdf

Blinder A.S. (1973), "Wage Discrimination: Reduced Form and Structural Estimates", *Journal of Human Resources*, 8, pp. 436-455.

Devicienti F. (2010), "Shapley-Value Decomposition of Changes in Wage Distribution: A note", *Journal of Economic Inequality*, 8 (1), pp.199-212.

Fields G. (2003), "Accounting for Income Inequality and Its Changes: A New Method with Application to the Distribution of Earnings in the United States", *Research in Labor Economics*, 22, pp. 1-38.

Guanatilaka R., Chotikapanich D. (2009), "Accounting for Sri Lanka's Expenditure Inequality 1980-2002: Regression-Based Decomposition Approaches", *Review of Income and Wealth*, 55 (4), pp. 882-906.

Israeli O. (2007), "A Shapley Based Decomposition of R-Squared of Linear Regression", *Journal of Economic Inequality*, 5 (2), pp.199-212.

Morduch J., Sicular T. (2002), "Rethinking Inequality Decomposition, with Evidence from Rural China", *The Economic Journal*, 112, pp.93-106.

Oaxaca R. (1973), "Male-Female Wage Differences in Urban Labour Markets", *International Economic Review*, 14, pp.693-709.

Sastre M., Trannoy A. (2002), "Shapley Inequality Decomposition by Factor Components: Some Methodological Issues", *Journal of Economics*, 9, pp. 51-89.

Shorrocks A. F. (1982), "Inequality Decomposition by Factor Components", *Econometrica*, 50, pp. 193-211.

Shorrocks A. F. (1984), "Inequality Decomposition by Population Subgroups", *Econometrica*, 52, pp. 1369-85.

Shorrocks A. F. (1999), "*Decomposition Procedures for Distributional analysis: A Unified Framework Based on the Shapley Value*", mimeo, University of Essex.

Wan G.H. (2004), "Accounting for Income Inequality in Rural China: a Regression-Based Approach", *Journal of Comparative Economics*, 32, pp.348-363.

Wooldridge J.M. (2002), *Econometric Analysis of cross section and panel data*, The MIT Press, Cambridge, Massachusetts, London, England.

ABSTRACT

The regression-based decomposition method combined with the Shapley value approach gives the opportunity of quantifying the contribution to the inequality of a set of factors, while taking the correlations among them into account.

The aim of this paper is to measure the relative contributions of individual as well as household factors to the explanation of the inequality in individual disposable incomes. The factors are introduced as explanatory variables in an income generating model that is estimated through a time-invariant unobserved random effects model on panel data from the Italian Survey of Household Income and Wealth (SHIW). The factors that play a dominant role in explaining the observed inequality are gender and educational level.

Less importance is accorded to the age and the household wealth whereas the work status and the area of residence affect the income differentials only in a marginal way.