Microsimulation and analysis of income distribution: an application to Italy

by

Carlo V. Fiorio

A thesis submitted for the degree of Doctor of Philosophy of the University of London

Department of Economics

London School of Economics and Political Science

University of London

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

The first chapters of the thesis put special emphasis on tax-benefit microsimulation models. The state of the art in the economic literature of tax-benefit microsimulation models is reviewed and discussed. Particular attention is paid to issues such as the reliability of estimation and the grossing-up of the sample. In order to analyze tax-benefit microsimulation, a new model is developed focusing on the case of Italy: it shares many features with other country-specific tax-benefit microsimulation models. The model, appropriately calibrated to population totals, is also used for an estimation of tax evasion via comparison with a number of different data sources. Non-parametric density estimation is used to improve the understanding of policy simulations and to analyze the effect of fiscal reform: an application to the 1998 Italian personal income taxation reform is provided. The first part concludes with an analysis of the reliability of microsimulation models, which has been addressed by few authors before. The analysis is undertaken using the bootstrap, which tends to show a better performance in finite sample than asymptotic approximations. The main result is that static microsimulation does not by itself make confidence intervals larger: on the contrary they can also make it smaller. To improve the reliability of microsimulation models the best way to proceed is to reduce the sampling error of the available data sets.

In the remaining chapters the thesis analyzes how microsimulation models can be useful in understanding the causes of inequality trends. As a preliminary step, the review and discussion of the literature about the main methods for inequality decomposition is provided. Based on this, a combination of two recent microsimulation methods is proposed to analyze the trend of inequality in Italy in 1977-2000. It is found that analysis using traditional methods of inequality decomposition can be seriously misleading if the sample is not representative of the whole population in some of its dimensions, such as female labor force participation. Microsimulation techniques can overcome this problem and can account for the major factors that driving inequality. Finally, the thesis discusses the issue of inference with thick-tailed distributions, such as the Pareto distribution with infinite second moment, that is of special relevance to empirical analysis of income distribution. It is shown that inference based on the standard t-ratio statistic can induce a non negligible error in rejection probability. Some solutions are suggested with an application to Italian household income data.

ACKNOWLEDGMENTS

It has been a privilege for me to spend these past years at the London School of Economics. I have greatly benefited from the courses I took, the seminars I attended, the researchers I interacted with. There are many people I would like to thank for their advice, feedback and support, without whom this thesis may not have come about.

I am truly indebted to Frank Cowell for his careful supervision, support, and great patience over the past five years. It was through him that I received access to the office facilities in STICERD, a resource that proved invaluable to my research.

Vassilis Hajivassiliou also provided thoughtful supervision, his insights and help with Chapter 7 are particularly appreciated.

I acknowledge financial support from ESRC, STICERD, ONAOSI, Università di Milano, Università di Padova, Università di Pavia, for which I am thankful.

I would also like to thank the Bank of Italy for providing me with the Survey of Household of Income and Wealth data set.

Further thanks go to Roberto Artoni for his continued support. Sanghamitra Bandyopadhyay, Andrea Brandolini, Conchita D'Ambrosio, Joanna Gomulka, Fabrizio Iacone, Marco Manacorda, Daniela Mantovani, Ceema Namazie for their generous gifts of time and useful comments.

Emmanuel Flachaire patiently introduced me to the bootstrap which plays a relevant role in these thesis. How many times since have I had to let you win at squash? Frédéric Robert-Nicoud helped me put the world to rights over numerous frozen pizzas as we shared the LSE experience. Marco Piersimoni for his polite humor and great cooking. Tim Walker for his immediate friendship, the scars from frisbee matches remain. The LSE would have not been such a pleasant working environment - and London not such a pleasant place to live in - without the company of Heski Bar-Isaac, Ralph Bayer, Sabine Bernabè and Anders, Virginie Blanchard, Thomas Buettner, Chiara Candelise, David Chilosi, Paolo Cossu, Guillermo Cruces, Juan De Laiglesia, Marta Foresti, Sinead Lester, Pete Longyear, Jez Longyear, Rocco Macchiavello, Maryla and Wojtek Maliszewski, Maria Luisa Mancusi, Felix Muennich, Silvia Pezzini, Michele Pellizzari, Amedeo and Annarita Poli, Riccardo Puglisi, Annamaria Reforgiato-Recupero, Toril Skjetne, Cecilia Testa, among many others.

Above all thanks to my family, whose support has been tremendous throughout. And Mara, whose love kept me going even in the most difficult times. This thesis is dedicated to them and to a good friend of mine who is fighting against cancer.

Table of Contents

1 Microsimulation: a tool for economic analysis **1**

1.1 Microsimulation in economics 2

1.2 Static microsimulation models 7

1.2.1 Static microsimulation models with behavioral responses . . . 9

1.3 Dynamic microsimulation models 10

1.3.1 Dynamic cross-section microsimulation models 11

1.3.2 Dynamic longitudinal microsimulation models 12

1.3.3 Limitations of dynamic microsimulation models 13

1.4 Issues in microsimulation modelling 14

1.4.1 Grossing-up 14

1.4.2 Validation 17

1.4.3 Reliability 18

1.5 Conclusion 20

2 A static tax-benefit microsimulation model for Italy **22**

2.1 The data set for Italian MSMs 23

2.1.1 Some limitations of the database 25

2.2 Structure of the model 26

2.3 Building the tax base 29

2.4 Grossing-up 31

2.5 Estimation of tax evasion and validation 39

2.6 Summary: why another MSM? 43

3 Microsimulation and non-parametric estimation **46**

3.1 The data set, the microsimulation model and the IRPEF reform . . . 48

3.2 The density estimation technique 50

3.3 The non-parametric density estimation and the MSM combined using counterfactuals 53

3.3.1 Decomposing the sample 56

3.3.2 Losers and gainers 59

3.4 A revenue-neutral reform simulation 62

3.5 Decomposing the fiscal reform 66

3.6 Conclusions 71

3.7 Appendix A: 1998 IRPEF vs. counterfactual 1991 IRPEF 72

3.8 Appendix B: inequality analysis of BT, actual and counterfactual AT incomes 74

4 Assessing the reliability of MSMs using the bootstrap **76**

4.1 Methodology for assessing the sampling error 78

4.2 The data set and the microsimulation model 84

4.3 Non-linear transformation and confidence intervals: results of the analysis 86

4.4 Conclusions 101

5 Review of the literature on income inequality decomposition **106**

5.1 The "traditional" approach 107

5.2 The regression-based approach 112

5.3 Microsimulation approaches to inequality decomposition 119

5.4 Conclusions 129

6 Understanding inequality trends in Italy **131**

6.1 Analysis of Italian household income distribution: available evidence . 132

6.2 Data, hypothesis and aims 136

6.2.1 The data set: pros and cons 136

6.2.2 Description of demographic trends 138

6.2.3 Preliminary hypothesis for inequality analysis 145
6.2.4 Analysis of inequality estimates and data contamination issues 148
6.3 Description of the methodology 156
6.3.1 Effects of individual and household characteristics on household inequality 156
6.3.2 Effects of changing dispersion of individual incomes on household incomes 161
6.3.3 Testing the change of inequality 163
6.4 Results of the analysis 164
6.5 Conclusions 179
6.6 Appendix C: The Generalized Entropy class of inequality indices . . . 181

7 Inference issues with thick tail distributions 183
7.1 The t-ratio distributions of some infinite variance distributions 186
7.2 Simulation results 191
7.3 Testing with TT distribution 197
7.4 Solutions 201
7.5 An application to income data 205
7.6 Conclusions 207
7.7 Appendix D: Moments of the symmetric Pareto distribution 209
7.8 Appendix E: TT in the economics literature 211

8 Conclusions 214
8.1 The value of microsimulation 214
8.2 Practical methods for analyzing the causes of inequality 217
8.3 Special methods required for analysis of income distribution data . . 218

List of Tables

2.1 Structure of 1998 IRPEF tax brackets in Lit '000 (Lit 1,936.27 = €1). 30

2.2 Tax allowances for dependent spouse and for other dependent relatives in Lit '000 (Lit 1,936.27 = €1) 31

2.3 Tax allowances depending on amount and type of income received in Lit '000 (Lit 1,936.27 = €1) 32

2.4 Grossed-up variables using Banca d'Italia (2000) grossing-up weights compared with population totals. *External source: ISTAT (2004). **External source: CNEL (2004). 35

2.5 Grossed-up variables using alternative grossing-up weights compared with population totals. *External source: ISTAT (2004). **External source: CNEL (2004). 37

2.6 Grossed-up variables using alternative grossing-up weights compared with population totals. *External source: ISTAT (2004). **External source: CNEL (2004). 38

2.7 Summary statistics for initial SHIW and final grossing-up weights. . . 39

2.8 Tax evasion estimation for TABEITA98 41

2.9 Validation of the TABEITA98 output, at the national and area level. Ext. Sources: Ministero delle Finanze (2002); CNEL (2004) 44

3.1 Abbreviations used in tables and figures 56

3.2 Actual 1998 and counterfactual structure of IRPEF tax brackets (Lit 1936.27=€1) 72

3.3 Actual 1998 and counterfactual tax allowances for dependent spouse (Lit 1936.27=€1) 72
3.4 Actual 1998 and counterfactual tax allowances for dependent children and other relatives (Lit 1936.27=€1) 73
3.5 Actual 1998 and counterfactual tax allowances for employment income. * is an average. The actual tax allowance (in Lit '000) was computed as 851-[$(y_B - 12,400) \times 0.78$], where y_B is BT income. 73
3.6 Self-employment tax allowances. * is an average. The actual tax allowance (in Lit '000) was computed as 168-[$(y_B - 6,800) \times 0.78$], where y_B is BT income. 74
3.7 Lorenz curve for different type income 75
3.8 Some inequality indices for different type of income 75

4.1 Proportion of households by occupation of the household head 86
4.2 Occupation of household head if not working 86
4.3 Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the sample mean; $\epsilon = 0$ 89
4.4 Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the sample mean; $\epsilon = 0.5$ 90
4.5 Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the sample mean; $\epsilon = 1$ 90
4.6 Asymptotic and bootstrap 90% confidence intervals for the 20th percentile; $\epsilon = 0$ 91
4.7 Asymptotic and bootstrap 90% confidence intervals for the 20th percentile; $\epsilon = 0.5$ 91
4.8 Asymptotic and bootstrap 90% confidence intervals for the 20th percentile; $\epsilon = 1$ 91
4.9 Asymptotic and bootstrap 90% confidence intervals for the 40th percentile; $\epsilon = 0$ 92

4.10 Asymptotic and bootstrap 90% confidence intervals for the 40th percentile; $\epsilon = 0.5$ 92
4.11 Asymptotic and bootstrap 90% confidence intervals for the 40th percentile; $\epsilon = 1$ 92
4.12 Asymptotic and bootstrap 90% confidence intervals for the 50th percentile; $\epsilon = 0$ 93
4.13 Asymptotic and bootstrap 90% confidence intervals for the 50th percentile; $\epsilon = 0.5$ 93
4.14 Asymptotic and bootstrap 90% confidence intervals for the 50th percentile; $\epsilon = 1$ 93
4.15 Asymptotic and bootstrap 90% confidence intervals for the 60th percentile; $\epsilon = 0$ 94
4.16 Asymptotic and bootstrap 90% confidence intervals for the 60th percentile; $\epsilon = 0.5$ 94
4.17 Asymptotic and bootstrap 90% confidence intervals for the 60th percentile; $\epsilon = 1$ 94
4.18 Asymptotic and bootstrap 90% confidence intervals for the 80th percentile; $\epsilon = 0$ 95
4.19 Asymptotic and bootstrap 90% confidence intervals for the 80th percentile; $\epsilon = 0.5$ 95
4.20 Asymptotic and bootstrap 90% confidence intervals for the 80th percentile; $\epsilon = 1$ 95
4.21 Asymptotic and bootstrap 90% confidence intervals for the GE(0); $\epsilon = 0$ 96
4.22 Asymptotic and bootstrap 90% confidence intervals for the GE(0); $\epsilon =$ 0.5 97
4.23 Asymptotic and bootstrap 90% confidence intervals for the GE(0); $\epsilon = 1$ 97
4.24 Asymptotic and bootstrap 90% confidence intervals for the GE(1); $\epsilon = 0$ 97
4.25 Asymptotic and bootstrap 90% confidence intervals for the GE(1); $\epsilon =$ 0.5 98
4.26 Asymptotic and bootstrap 90% confidence intervals for the GE(1); $\epsilon = 1$ 98

4.27 Asymptotic and bootstrap 90% confidence intervals for the GE(2); $\epsilon = 0$ 98
4.28 Asymptotic and bootstrap 90% confidence intervals for the GE(2); $\epsilon =$ 0.5 99
4.29 Asymptotic and bootstrap 90% confidence intervals for the GE(2); $\epsilon = 1$ 99
4.30 MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the sample mean; $\epsilon = 0.5$ 101
4.31 MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the 20th percentile; $\epsilon = 0.5$ 101
4.32 MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the 40th percentile; $\epsilon = 0.5$ 102
4.33 MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the 50th percentile; $\epsilon = 0.5$ 102
4.34 MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the 60th percentile; $\epsilon = 0.5$ 103
4.35 MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the 80th percentile; $\epsilon = 0.5$ 103
4.36 MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the GE(0); $\epsilon = 0.5$ 104
4.37 MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the GE(1); $\epsilon = 0.5$ 104
4.38 MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the GE(2); $\epsilon = 0.5$ 105

6.1 Decomposition of the population by age groups 139
6.2 Decomposition of the population by household type 140
6.3 Decomposition of the population by number of components 141
6.4 Labor force participation: Total, by sex, by age 143
6.5 Inequality by different type of incomes 155
6.6 Inequality of equivalent household income 156
6.7 Abbreviations used in tables and figures 167

6.8 Counterfactuals using DFL methodology - Base year is 1991 171
6.9 Counterfactuals using DFL + Burtelss methodology - Base year is 1991 177
6.10 Counterfactuals using Burtless methodology - Base year is 1991 . . . 178

7.1 Summary statistics for BT income over Lit 30 millions 207

List of Figures

3-1 Analysis of density estimates using different bandwidths on the whole sample; in Lit '000 (Lit 1936.27=€1) 57

3-2 AT income densities by occupation of the householder, with 90% confidence bands; in Lit '000 (Lit 1936.27=€1) 58

3-3 Difference between counterfactual AT and actual 1998 AT income densities, by occupation of the householder and with different bandwidths; in Lit '000 (Lit 1936.27=€1) 59

3-4 Losers and gainers, whole sample; different bandwidhts, 90% confidence bands; in Lit '000 (Lit 1936.27=€1) 61

3-5 Losers and gainers, by occupation of the householder; Silverman's bandwidth, 90% confidence bands; in Lit '000 (Lit 1936.27=€1) . . . 63

3-6 Revenue-neutral simulation: lump-sum redistribution; with 90% confidence bands; in Lit '000 (Lit 1936.27=€1) 64

3-7 Revenue-neutral simulation: lump-sum redistribution; with 90% confidence bands; in Lit '000 (Lit 1936.27=€1) 65

3-8 Scenario 1 vs. actual AT; with 90% confidence bands; in Lit '000 (Lit 1936.27=€1) 67

3-9 Scenario 2 vs. actual AT; with 90% confidence bands; in Lit '000 (Lit 1936.27=€1) 68

3-10 Losers and gainers: simulation D1, with different bandwidths and 90% confidence bands; in Lit '000 (Lit 1936.27=€1) 69

3-11 Losers and gainers: simulation D2, with different bandwidths and 90% confidence bands; in Lit '000 (Lit 1936.27=€1) 70

6-1 Decomposition of the population by age groups 138
6-2 Decomposition of the population by HH type 141
6-3 Decomposition of the population by HH type 142
6-4 Labor force participation: Total, by sex, by age 144
6-5 Frequency of income earners in av. HH or population 146
6-6 Inequality indices, individual (monthly) incomes 151
6-7 Inequality indices, individual (monthly) incomes 152
6-8 Inequality indices, individual (monthly) incomes 153
6-9 Frequency of income earners in av. HH or population 154
6-10 DFL methodology: number of income receivers and number of components in HH 166
6-11 DFL methodology: number of income receivers in HH and probability of being pensioners 168
6-12 DFL methodology: number of income receivers and probability of female in the labor force 169
6-13 Decomposition of the population by sex 170
6-14 Counterfactuals using Burtless methodology 173
6-15 Counterfactuals using Burtless methodology 174
6-16 Counterfactuals using Burtless methodology 175
6-17 Counterfactuals using DFL+Burtless methodology 176

7-1 t-ratio of Cauchy and infinite-first-moment symmetric Pareto distributions 192
7-2 t-ratio of symmetric Pareto distributions with $1 < \alpha \leq 2$ 193
7-3 t_1 Pareto distributions with $0 < \alpha \leq 1$ 193
7-4 t_2 of Pareto distributions with $1 < \alpha \leq 2$. 194
7-5 Distribution of the variance of some distributions with infinite mean. 195
7-6 Distribution of the variance of some distributions with infinite mean. 195
7-7 ERP for two-tail test with a sample from a symmetric or a positive definite Pareto with $1 \leq \alpha$ 2. 200

7-8 ERP for one-tail test with a sample from a symmetric or a positive definite Pareto with $1 \leq \alpha$ 2. 200
7-9 Test of difference in mean of Pareto distributions with $\beta = 3$. 202
7-10 t_3 with infinite first moment distributions 204
7-11 t_3 with Pareto distribution with $1 < \alpha \leq 2$ 204
7-12 ERP with t_3 from a Pareto distribution with $1 < \alpha \leq 2$ 205

Abbreviations

MSM	microsimulation model
FES	Family Expenditure Study (U.K.)
BT	before-tax
AT	after-tax
SHIW	Survey of Household Income and Wealth (Italy)
IRPEF	Imposta sui Redditi delle PErsone Fisiche (personal income tax)
NW	North-West
NE	North-East
C	Center
S	South
MF	Ministero delle Finanze (Italian Ministry of Finance)
ISTAT	Italian Institute of Statistics
CPI	Consumer Price Index
GE	Generalized Entropy inequality indices
LIS	Luxembourg Income Study
DGP	data generating process
EDF	empirical distribution function
CDF	cumulative distribution function
LFP	labor force participation
DF	density function
TT	thick tail distribution

Chapter 1

Microsimulation: a tool for economic analysis

Simulation can be described as a process of imitating the behavior of complex systems, such as economic or biological systems, a set of tax rules or the computer network of a large firm: given a set of available information, simulation allows one to build a system that imitates the "reality". Simulation, as a method of solving problems, becomes of relevance when conventional analytical, numerical, physical or experimental methods are too expensive, complicated or time demanding.

Simulation models in economics were originally developed by Orcutt (1957) and Orcutt et al. (1961) to investigate socioeconomic systems using a microanalytic approach. Economic modelling using simulation was seen as an alternative to aggregate-type national income models originated by the work of Tinbergen (1939), and input-output models that followed the work of Leontief (1951). Tinbergen's approach uses major sectors, such as the household and business sector, as basic components, es-

timating and testing macroeconomic relationships on the basis of annual or quarterly time series data of such variables as aggregate consumption and income of the household sector. Leontief's approach uses industries as basic components and places emphasis on the cross-sectional structure of the economy rather than on its dynamic features. Orcutt's approach instead develops the most general model in terms of its statistical structure, since it is developed to include both Tinbergen's and Leontief's models features and to increase details of the micro unit involved (Orcutt et al., 1976).

Since the pioneering work of Orcutt, simulation has played an increasing role in economic research and has greatly evolved with respect to its initial aims. In Section 1.1 the role of simulation in economics will be discussed, focusing mainly on microsimulaton models, i.e. on models that use microeconomic units as the basis of their simulations. The following sections will focus mainly on tax-benefit microsimulation models, broadly divided in static and dynamic models: Section 1.2 deals with the former, Section 1.3 with the latter. A distinction is often made between cross-section and longitudinal dynamic microsimulation models: they will be discussed in Sections 1.3.1 and 1.3.2. In Section 1.4 some issues of particular relevance to microsimulation modelling will be discussed, namely the grossing-up procedure, the validation and reliability of microsimulation models. Section 1.5 concludes this chapter.

1.1 Microsimulation in economics

In recent years, there has been an extensive development of simulation models for quantitative research in economics. Simulation has been used more and more fre-

quently as a research tool in economics also thanks to the rapid development of computers and their easy access to users: fast computers allowed improved accuracy and the development of more complex simulation systems.

Simulation models are used as conditional forecasting tools to forecast the effect of shocks or policies on individual units or larger systems. Forecasting can be ex ante or ex post. In the ex ante case, forecasts are computed on the basis of distinct conditions described by a given scenario: they are a conditional look into future developments. In the ex post case, the given real world situation is compared with alternative interventions (Merz, 1991).

Simulation can be performed at macro or micro level. Macrosimulation models analyze relationships between national economic sectoral and aggregate variables. They are developed using aggregates or sub-aggregates of the totals. They have been used mainly in government offices for tax forecasting but in the last decades have been replaced when possible by micro models, which obtain more reliable results the larger is the complexity of each tax, the diversity of taxpayers and the variation in the tax base (Eason, 1996, 2000). Macrosimulation models are still used in particular contexts, such as forecasting of North Sea taxation (Blow et al., 2002). Macrosimulation models are developed by central banks to describe long-run equilibrium in the economy. An example of such models is the Macro Model (MM) at the Bank of England, which is mainly used by the Monetary Policy Committee to set interest rates and attain the inflation target. Macromodels are also used in applied general equilibrium models, which can include detailed scenarios on macroeconomic developments in both domestic and foreign economy, changes in world prices of internationally traded

goods, changing in trade policies. An example of these complex models is MONASH for the Australian economy (Meagher, 1996).

Microsimulation models focus directly on micro units such as individuals, households and firms. If these micro units are firms they can be identified by their organization, their occupation structure, the set of products, etc.; if they are individuals they can be identified by demographic characteristics such as age, sex, occupation, residence.

Microeconomic models of firms are relatively less common, the main limitation being in the availability of data for firms. Microsimulation of firms within a model of the economy was first developed by Eliasson (1978) in a model for Sweden. This model is a complete micro- to macro-model that contains a selection of the most important corporate firms, which are simulated in conjunction with residual firms to replicate the Swedish economy. These models often incorporate production aspects as well as financial aspects, using a firm in an uncertain and changing environment that is assumed to behave with bounded rationality. These models are very demanding in terms of data and this is one of the main reasons why not all developed economies have such a model. Among the few exceptions, see van Tongeren (1995) for the Dutch economy.

Microsimulation models on individuals or households have developed rapidly in recent decades. Simulation is used in economics mainly to develop counterfactual analysis to forecast possible effects of different economic policies, but it also serves in particular situations for generating data that are missing. The structure of a simulation model is principally expressed in terms of logical mathematical relations:

a simulation model is a set of algebraic equations and decision structures, which can be characterized as a complex set of "if... then" relations. Relations, procedures and data can be either completely determined or incorporate random errors. If relations are all completely specified the simulation is known as deterministic; if randomness is included the simulation is known as stochastic.

The main aim of microsimulation models based on individual or household data is to analyze the impact of policy changes on the distribution of some target variables rather than on their mean, as it happens using regression techniques. The development of microsimulation models on households goes together with increasing availability and reliability of micro data sets and improving computer capacity. Static models generally are based on sample surveys, which provide detailed information about individual and family characteristics, labor force status, housing status, earnings. They typically contain the receipt of social security benefits and income tax liabilities, or incorporate enough information for their calculation. With a microsimulation model the immediate distributional impact of fiscal policies, such as an increase in child benefits, in income tax rates or in the minimum wage, can be modelled, and estimates of the characteristics of winners and losers and total cost can be computed. Microsimulation models can also be used to project into the future and to assess the socio-economic consequences of an ageing population, or of changes in educational structure and in marriage patterns. In recent years various microsimulation modelling methodologies have been developed: some of them will be discussed in Section 5.3 in the context of inequality decomposition and Chapter 6 will provide an application to Italy combining two microsimulation methodologies.

A field of economics where microsimulation has been widely exploited is the analysis of the effects of tax-benefit reforms on income, welfare and behavior of individuals. Until the early 1980s tax-benefit microsimulation models were not widespread and it was rather common to analyze tax-benefit effects using a range of representative households. Government agencies and academic departments decided to invest effort and monetary resources in developing microsimulation models since it was clear that representative household analysis is unable to give a broad picture of the effect of the policy on the whole population. For instance, Atkinson and Sutherland (1983) compared the family composition and circumstances recorded in the UK Family Expenditure Survey 1980 with the hypothetical family types used in the Department of Health and Social Security (DHSS) tax-benefit model. They found that some 4% of actual families were covered by the assumption of the complex DHSS hypothetical family model. This concern is even more relevant for some of the theoretical simulation models used to investigate the effects of government policy in a complex intertemporal setting.

Nowadays virtually all developed countries have at least one model to simulate changes of taxes and benefits on individual incomes (see among others Mitton et al., 2000; Gupta and Kapur, 2000; Sutherland, 1994) and after the transition from centralized planning to a market economy and the consequent need of financing government policy through taxation, also Eastern and Central European countries showed an increased interest in tax-benefit microsimulation models (see for instance Coulter et al., 1998; Juhász, 1998). In the European Union the EUROMOD project developed a 15-country Europe-wide microsimulation model to provide estimates of the distributional

impact of changes to personal tax and transfer policy each taking place at either the national or the European level, to assess the consequences of consolidated social policies and to understand how different policies in different countries may contribute to common objectives (Sutherland, 2001). The rest of this chapter will mainly focus on tax-benefit microsimulation models (MSMs).

1.2 Static microsimulation models

Static MSMs are based on instantaneous pictures of characteristics of a sample of a population in a given period. They are appropriate models for the analysis of the impact of policy changes, where these effects can be deduced, completely or in large part, from knowledge of the current circumstances of individuals in the sample.

In static MSMs behavioral relations and institutional conditions are varied exogenously. Micro-data bases are comprised of a cross-section of micro units in a given period. These micro units are generally assigned a sampling weight, which allows one to infer about the population of origin.

Static microsimulation is first developed for the specific period to which the data relate. A static MSM can then be applied to different time periods using static "aging procedures". Such a procedure consists in re-weighting the available information using given aggregate of another time period. After re-weighting a sample, a new weight will be assigned to each micro unit, i.e. each micro unit will represent a different number of units in the whole population. However, the number of observations in the data set will remain unchanged after performing any simulation or aging of the sample.

A static ageing procedure is the easiest ageing procedure and can be reasonably applied for short- or medium-run forecasts, where it can reasonably be assumed that demographic characteristics of the underlying population do not change significantly.

A static aging procedure involves updating income and wealth data. It is common to update income and wealth variables using for all units growth in money income or a price index such as the retail price index (RPI) or the consumer price index. This solution can lead to important bias. Sutherland (1989) showed that while in the period 1982-1988 RPI increased by 32%, the earnings of full-time adult men in the bottom decile grew by 42% and those in the top decile grew by an average of 64% in the same period. Notwithstanding this limitation, straightforward indexing by published data is often the only possible way to go, mainly for reasons of data limitation. Updating bias can be a very limited problem if considered over a short period of time.

The term "static" may seem limiting when compared to "dynamic". However, in some contexts dynamic models are not very useful. In fact, static models allow one to hold constant a large number of variables so that it becomes possible to isolate some elements of particular interest. In the setting of fiscal policies, for instance, they can separate direct effects on income of changing the structure of tax and benefits from other possible effects (see for instance, Chapter 3). Static models, which are sometimes also called "arithmetic", are the building block of more complex models, such as behavioral or dynamic ones (see among many others, Atkinson and Sutherland (1988a,b); Dilnot et al. (1988); Bourguignon et al. (1988)).

1.2.1 Static microsimulation models with behavioral responses

One of the shortcomings of static MSMs is the assumption that individual behavior is exogenous. However, many tax and benefit policies are designed specifically to have behavioral effects. For instance many policy reform are designed to encourage more labor force participation; other transfer policies may have unintended negative incentive consequences; government revenues and expenditures calculations may be misleading if potential behavioral responses are not properly taken into account and estimated.

The type of static MSMs discussed in this subsection are intended as an improvement through the introduction of explicit modelling of behavioral responses to policy reforms. They hold certain characteristics fixed (such as family composition) but allow other characteristics to change, like labor force participation and, consequently, earnings. This type of modelling presents many computational and analytical problems and is nowadays one of the most dynamic field in the microsimulation literature. To include behavioral response it is necessary to handle complex budget constraints that allow each individual's constraint to be unique, along with the desire to model heterogeneity. All these issues impose demanding modelling and computer programming requirements (Duncan, 2003). In fact, all components of a tax-benefit MSM have to be closely integrated and, given the large number of sample units involved, it is important to invest in developing efficient computer routines. Researchers have to decide about the underlying economic model and its econometric specification. They also have to decide whether to use discrete or continuous methods when modelling

observed outcome, and simulating the effects of policies on individual behavior.

MSMs with behavioral responses are developed from a static MSM without behavioral responses. Since they introduce additional complication to modelling issues, they also increase their limitations in terms of grossing-up, validation and reliability of results issues. Moreover they are more demanding in terms of data quality. Most data sets include fewer information about non-workers than about workers: the researcher has then to introduce "reasonable" assumptions, which will affect the reliability of the estimates. Although MSMs with behavioral response have been criticized for being rather unreliable, particularly in estimating figures of losers and gainers, tax revenue and overall inequality after-tax reforms (Pudney and Sutherland, 1996), their use has proved very successful in the labor economics literature, namely in estimating the labor supply responses to changes in net wages (see, among others Blundell et al., 1992, 1998; Duncan and Giles, 1996).

1.3 Dynamic microsimulation models

Dynamic MSMs differ from static ones mainly in terms of the ageing procedure. In dynamic MSMs each micro unit is aged individually using survivor probabilities estimated empirically. Dynamic models not only include the possibility of death but introduce events such as marriage, household composition evolution, such as births, inclusion of other relatives in the household, divorce. Hence, dynamic demographic ageing will create a new data set whose dimension will typically be different from the original one.

Demographic dynamics will also rises the issue of the inclusion of behavioral response to demographic changes: for instance, if a child is born it is possible that the mother's decision to participate in the labor market will change. However, the magnitude of behavioral change is difficult to assess due to the widely divergent estimates of relevant elasticities, and simulations are generally presented for a number of different estimates (Hagenaars, 1990, p. 31).

It is also debatable whether the elasticities obtained from cross-section data can be assumed to closely reflect lifetime behavioral response. For instance, there is panel data evidence that labor force participation decisions are made with a very long time horizon in mind (Heckman and MaCurdy, 1980, p. 67) and that future expected values of variables determined current labor supply decisions (MaCurdy, 1981, 1983). Therefore it could be that a higher real wage increases labor force participation in the short term but life-cycle income will be partly offset by a decision to retire earlier (Harding, 1990, p. 15). Among dynamic MSMs, the main distinction is between dynamic cross-section and dynamic longitudinal models.

1.3.1 Dynamic cross-section microsimulation models

Dynamic cross-section MSMs (or dynamic population MSMs) attempt to project micro units forward through time simulating demographic events such as death, birth, marriage, divorce, etc. Recently also immigration has been introduced (see for instance Walker, 2000). After a main demographic event has been modelled, other characteristics can be imputed, such as education, labor force status, housing. The

data base for dynamic cross-section MSM is the same as in the static model, i.e. a random sample of the population.

Dynamic cross-section models are aimed at depicting the future structure of the population and typically map only a few decades of the lives of individuals from many age cohorts. They are useful to forecast the future characteristics of the population and to model the effects of policy changes during the next years or decades.

These models have been mainly applied for analyzing the distributional impacts of social security system (see, for instance Favreault and Caldwell, 2000; Nelissen, 1996), the evolution of the pension systems (see Galler, 1996; Eklind et al., 1996; Andreassen et al., 1996) or lifetime analysis of poverty alleviation programs (see Falkingham and Harding, 1996).

1.3.2 Dynamic longitudinal microsimulation models

In a dynamic longitudinal MSMs (or dynamic cohort MSMs) the aging process is as in the cross-section models but only one cohort is aged rather than the entire population. In general, one cohort is aged from birth to death so that a whole life cycle of one cohort is simulated. The same life cycle profiles can be generated with dynamic cross-section models, however it would be inefficient if lifetime circumstances of one or two cohorts are of interest.

Dynamic longitudinal models are generally used for analyzing lifetime earnings and income distributions, to assess the lifetime incidence of taxes and government spending programs (Harding, 1990). Dynamic longitudinal MSMs are much quicker to

process than cross-section MSMs and are easier to model since they restrict attention to the demographics and socio-economic dynamics of a single cohort rather to all cohorts in a population.

1.3.3 Limitations of dynamic microsimulation models

Dynamic models, as extensions of static MSMs, present additional limitations.

Dynamic MSMs are more data hungry: they need additional information to estimate demographic changes to be included in dynamic models. Ideally these data sets include death rates by age, sex and socio-economic status, marriage rates by age, sex, education level and previous marital status; divorce rates by age, sex, duration of marriage, and number of age of children; attendance rates at primary, secondary and tertiary levels by age, sex, parental socio-economic status and previous education; labor force participation rate by age, sex, education, marital status, age of children, duration of current employment and of unemployment spells, etc. Moreover, cross-section data are not usually adequate for setting the parameters in dynamic models. For instance, the probability of transition between states can only be obtained from longitudinal data (Harding, 1990).

Since it is rare for a data set to be suitable for every kind of analysis, dynamic models generally rely on whatever piece of data is available, using matching techniques to put information together. This procedure is done at the expense of reducing the accuracy of the models and of requiring frequent updates.

The reliability issue is even stronger here than in static models because sampling

error is likely to happen in estimating survivor and transition probabilities and demographic changes. An additional source of error arises from the simulation of transitions to different states and survivor probabilities: since these events are obtained using Monte Carlo simulations, different simulations introduce different state-transition in the single micro unit. The model sensitivity to different simulation can be assessed by using a large number of replications. However, this is often difficult to perform since dynamic models (and especially dynamic cross-section models) require huge computing resources to run: the characteristics of the micro units in the initial year have to be stored and the final analysis is thus frequently based upon a very large number of observations.

1.4 Issues in microsimulation modelling

Among the various challenges that microsimulation modelling poses, the issues of grossing-up, validation and reliability are worth additional attention. They have been studied mainly in the context of static MSM, though they are of relevance for all types of microsimulation models.

1.4.1 Grossing-up

The procedure of grossing-up is concerned with generating figures to cover the population being modelled from the data set under use. The procedure should adjust for differences between the sample data and the characteristics of the population to be modelled at the date of sampling.

The grossing-up procedure is basically aimed at adjusting the data set to reflect differential non-response between different groups in the sample. It involves stratifying the sample, by some relevant characteristics, after the data have been collected and applying known proportions. This procedure is also sometimes referred to as post-stratification (see for instance Atkinson and Micklewright (1983)).

The grossing-up procedure consists in assigning to each unit in a sample of dimension N a weight p_j with $j = 1, ..., N$, such that some chosen statistics of interest calculated on the weighted sample coincide with the population statistics. The procedure is trivial if we want to reconcile the sample with the population using only one discrete statistic, s_k with $k = 1, ...K$, such as family types or income ranges. In this case, we compute the probability of having the characteristic s_k in the sample, say $P(s_k)$, and make it equal to the probability of having the same characteristic in the population, say $p(s_k)$. If the dimension of the sample and of the population are N and n respectively, then the grossing-up weight is $p_j = np(s_k)/NP(s_k)$, i.e. the size of the cell with characteristic s_k in the population divided by the size of the cell with characteristic s_k in the sample. If more variables are considered for the grossing-up procedure it should be necessary to consider the interactions between the different variables, i.e. consider the joint distribution of the control variables considered. However, this conflicts with available information from external sources, that in general, do not report the joint distribution of population variables but only the totals for each variable. For instance, it is possible to know the total number of single-parent families and the total number of self-employed in the population but not how many single-parent families have self-employment income. Hence, the conditions imposed

on the weights p_j are far less stringent than in the "full information" case we would have if the joint distribution were known, and in general there are many possible sets of weights p_j achieving the desired adjustment. To choose among them Atkinson et al. (1988) suggest the requirement that given a data set of dimension N, with original sampling weights q_j, $j = 1, 2, ..., N$, the set of grossing-up weights p_j have the least deviation from original weights, q_j. The original weights could reflect the sampling procedure or be uniform. Both grossing-up and initial weights have to sum up to the population size: $\sum q_j = \sum p_j = n$. If original and sample weights sum up to the sample dimension, they first have to be multiplied by n/N. It is then common practice to impose the condition that the new weights minimize the distance from initial weights. Hollenbeck (1976) proposed to use as a measure of distance the half of the squared sum of the difference between final and initial weights. However, in order to avoid negative weights, Atkinson et al. (1988) suggest minimizing a measure of distance derived from information theory (Theil, 1967; Cowell, 1980):

$$d(p, q) = \sum p_j log\left(\frac{p_j}{q_j}\right) \tag{1.1}$$

As for the optimal number of control totals to be included, no result is currently available. Although it is more common to face the problem of not having enough external sources than to have too many, Sutherland (1989, p. 15) warns on the risk of increasing the variance of weights since the larger the number of control totals becomes, the smaller the number of observations in each "cell' (i.e. with each combination of characteristics being controlled for).

Atkinson et al. (1988) applied their methodology[1] to TAXMOD, a MSM for the UK, and compared their results with what could be obtained with uniform weights, i.e. multiplying the sampling weights by n/N. The grossed-up results were significantly more plausible. The conclusion from their analysis is that the use of uniform weights can be seriously misleading.

1.4.2 Validation

Model validation is a task consisting of two distinct but related aspects. First it consists in comparing the primary data set to external data from a number of sources. This procedure is complementary to grossing-up in that detailed information not used as control totals are used as control data after grossing-up has been performed. Secondly, validation involves looking at the results of the model's output and analyzing them in relation to estimates published elsewhere. If the external source for validation is using the same sample this validation ends up in a comparison of different models. If validation is with estimates using different data sets, the comparison analyzes the robustness of results of both estimates. Hence, model validation is an exercise with a broader scope than grossing-up since it deals with model construction as well as its output (Hope, 1988).

However, model validation is also less easy to perform than grossing-up, at least as for the comparison of different estimates: it presumes that there is at least another model with a comparable level of accuracy to compare results with.

[1] As control totals they used the variables (i) family composition, (ii) employment status, (iii) income range and (iv) housing tenure.

1.4.3 Reliability

Although MSMs are widely used nowadays few authors working in this field have paid explicit attention to the statistical reliability of MSM output. MSMs typically produce summary statistics such as average income, income percentile, inequality indices, number of gainers and losers after a policy simulation. This output is an estimate of effects of changes in some policy instruments and, as well as any other statistical estimate, it should be accompanied by a standard error or confidence intervals. MSMs, as all survey-based models, can be affected by measurement and misreporting errors. They can also introduce peculiar errors such as errors in updating data to a later year, bad or no specification of behavioral response and market adjustments, stochastic simulation error. However, simulated figures are often quoted without standard error or confidence intervals and it is often hard to distinguish reliable results from those badly affected by sampling or other source of error. The main reason for not providing confidence intervals or standard error to MSM output is probably due to the technical problems involved in their calculation.

Pudney and Sutherland (1994) provide the first contribution on the reliability of MSMs. They investigate the asymptotic sampling properties of a set of typical simulation results, focusing on the sampling error assuming that finite population corrections can be avoided. They used equivalent income to account for economies of scale in the family or household. Although equivalence scales are non linear transformations, Pudney and Sutherland simply assume that errors are normally distributed and derive the $(1-\alpha)$ asymptotic confidence intervals as $c_{\alpha/2}$ deviations from the

standard error, where $c_{\alpha/2}$ is the $\alpha/2$ critical value for the normal distribution. Their application on 1998 Family Expenditure Study (FES, U.K.) data using a static MSM without behavioral response shows that the basic process of simulation seems reasonably reliable. However, they find that while some statistics are estimated with acceptable precision, others, like number of gainers and losers, poverty and inequality indices, can have a wide margin of sampling error and suggest that there should be some doubt about the reliability of some widely-quoted simulation results.

Even more pessimistic conclusions are reached in Pudney and Sutherland (1996) where asymptotic confidence intervals are estimated in a static MSM that incorporates a multinomial logit model of female labor supply. The resulting confidence intervals allow errors associated with sampling variability, parameter estimation and stochastic simulation. Pudney and Sutherland found that the sampling error is the main source of variability for most summary statistics, but that the measures describing the impact of policy on female participation are very uncertain and may be of no practical use to economists, mainly because of the variability of parameter estimates.

Chapter 4 contributes to the analysis of static MSM reliability using the bootstrap to compute confidence intervals. The bootstrap is considered for two main reasons: (a) it allows one to remove the hypothesis of infinite population to compute confidence intervals using a simulation-based methodology rather than finite sample corrections, and (b) there is a growing body of literature showing that bootstrap often performs better, or not worse, than asymptotic approximation in small samples. It is found that bootstrap confidence intervals are in general less conservative than asymptotic confidence intervals, especially the smaller are sub-samples used. However, because

of the complex non-linear setting of MSMs, to derive the exact or Monte Carlo finite population confidence intervals is a formidable task and it is not possible to evaluate the improved precision of bootstrap confidence intervals compared with asymptotic confidence intervals. However, the generally good performance of the bootstrap in finite samples should increase the concern about the reliability of estimates on particular sub-samples of widely used surveys. It was also found that MSMs itself do not necessarily make confidence intervals larger. In some cases, summary statistics on simulated incomes have narrower confidence intervals, as percentage of the computed statistic, than before the tax-benefit simulation. These results show that concerns in sampling error with MSMs are sometimes misplaced: it is not microsimulation that necessarily makes the estimation less reliable. A poor coverage of the population of some current surveys is often the main cause of error and improvement in data collection should be pursued.

1.5 Conclusion

Microsimulation models and, in particular tax-benefit microsimulation models are powerful tools for analyzing effects of demographic trends or to assess the effects on living standards of various public policies. However they cannot provide an answer for every question and must be handled with care. A great deal of attention should also be devoted to the presentation and analysis of data. In Chapter 3 non-parametric density estimation is proposed to increase the understanding of MSM output.

Microsimulation modelling requires a large effort in programming but its aims

should never be just the production of numbers: grossing-up, validation procedures and confidence interval estimation should be carefully addressed. A reliable MSM will also increase the credibility of a modelling technique that has only recently started being extensively exploited by academics.

A great deal of attention should also be devoted to improve data collection, since quality of the data sets is the key ingredient for a reliable MSM.

Chapter 2

A static tax-benefit microsimulation model for Italy

MSMs are powerful research tools with high fixed costs. In recent years the development of personal computers allowed single researchers to build their own MSM, however, building a MSM to simulate the complexity of economic systems still often requires team work. In the case of tax-benefit MSMs good programmers, who manage to develop fast computer codes, have to work together with experts of the tax and benefit legislation and of its implementation problems, and with econometricians who are able to analyze and treat data with rigor. For this thesis it was decided to look at Italian models for personal interest and also because my knowledge of the Italian tax-benefit system is deeper than that of any other national system, although the results of this and the following two chapters should also be of interest for non-Italian MSMs. The starting point for any MSM is the choice of the data set: the more it is representative of the population of interest and the wider the information it provides,

the more reliable the model becomes. Section 2.1 describes and discusses the main limitations of the 1998 SHIW data set, which is used for the MSM presented in this chapter.

To gain full access to a MSM, a new model had to be developed. When this project started no model using 1998 SHIW data set had been completed. The MSM developed for this thesis is TABEITA98: it was constructed using 1998 SHIW data set and STATA software, partly following the structure of Dirimod95[1]. The main features of TABEITA98 are briefly described in Section 2.2. Section 2.3 describes how the tax base was built from the available data set and Section 2.4 deals with the issue of grossing-up the model. Section 2.5 shows how the model was employed to estimate tax evasion and how these findings were used to validate the model. Section 2.6 concludes.

2.1 The data set for Italian MSMs

The data set used in this chapter is the Survey of Household Income and Wealth (SHIW) published by the Bank of Italy and based on interviews run in 1998. This data set will also be used in Chapters 3, 4 and 7. The SHIW is a long standing survey: it was started in the mid 1960s, was run about annually up to 1987, henceforth about every two years. The Bank of Italy paid particular attention to improve the quality of the data. For instance since 1995 an increasing number of interviews were performed using a computer to check consistency of answers and particular attention was paid

[1]Dirimod95 is a MSM developed at Prometeia, Bologna, using 1995 SHIW data and SAS software. It was kindly provided by Daniela Mantovani.

in formulating questions as clearly as possible with several trial interviews (Banca d'Italia, 2000, p.29). At present the SHIWs are the main, if not the only, data set for Italian household MSMs and among the most frequently used for any kind of household income analysis at the national level in Italy (for a review of other data sets, see Brandolini, 1999).

The 1998 data set collects detailed micro data for about 7,147 households and 20,901 individuals on disposable income, consumption, labor market, monetary and financial variables. The sample was drawn in two stages (municipalities and households) with the stratification of the primary sampling units (municipalities) by units and size, to make it representative of the national population. Within each stratum, all municipalities with population of more than 40,000 were selected, while smaller towns were randomly included. Households were then selected randomly and a sampling weight, defined as the inverse of the probability of inclusion of each household in the sample, was attached to each observation. Since 1989 a number of households who had been interviewed previously have been interviewed again, to start producing a panel data set. Although in the present and following chapters the panel will not be considered, it has an effect on the probability of a household being included in the sample. These issues have been addressed and resolved by the Bank of Italy, which provides a set of appropriate sampling weights. Data are checked before release: the strategy is either to drop the interview for the whole household if missing data cannot be reasonably inferred from other characteristics of the individual/household or to impute the missing data, often using regression models to forecast missing variables based on the personal characteristics of the individual/household involved. Data

imputation is less 0.1% for most variables (Banca d'Italia, 2000, p.35).

2.1.1 Some limitations of the database

Among the limitations of the SHIW data set some affect any analysis of Italian household income, some others are specifically of interest for the reliability of MSMs.

A first limitation of the data set is the low rate of response. Participation in the survey is voluntary and not paid. Although all households were granted total anonymity, in 1998 only 52.6% of contacted households agreed to being interviewed. The low rate of response can cause a selectivity bias as some households seem to be more likely to refuse an interview. In fact, the likelihood of accepting an interview decreases with increases in income, wealth and education of the household head, and the size of the town of residence (Banca d'Italia, 2000, p.31). In order to mitigate the selectivity bias some measures are adopted, such as the replacement of refusing households with others from the same town. Some estimations of the selectivity bias on incomes recorded in SHIW show that the underestimation of household income is on average rather limited (Cannari and D'Alessio (1992) estimate it at about 5%). Other limitations of this data set include the fact that the household is interviewed rather than the family. This leads to an overestimation of the average number of components, which cannot be corrected at all since the relevant information is missing. The interviews include only recall questions, i.e. questions referring to the previous year, reducing the precision of the reporting. An alternative approach would be to ask households to record all their incomes and expenditures of the coming week or month

but it was discarded to keep a reasonable rate of response and to avoid approximations that come from extending the week or month to cover the whole year. Finally, data do not include information about people who do not have a registered dwelling or are in a hospital or other kind of institution.

As for the limitations which are more relevant for MSMs, the main one refers to the type of income recorded: it refers to disposable income, excluding taxes and social contributions paid and benefit received. Hence, the first role of a MSM is to simulate the before-tax income before introducing any other policy simulation. This feature implies that, in contrast to other MSMs, no simulation error[2] can be properly assessed (for the U.K. see Pudney and Sutherland (1994)).

2.2 Structure of the model

The MSM developed for this thesis is TABEITA98, a TAx-BEnefit microsimulation model on ITAlian 1998 SHIW data. TABEITA98 refers to 1998 personal income taxation (IRPEF and "imposte sostitutive")[3] net of social contributions. TABEITA98 is a static model without behavioral response. It can be described as a deterministic transformation of a given sample into a new one. Let $\mathbf{y^A}$ and $\mathbf{y^B}$ be the vectors of after-tax (AT) and before-tax (BT) income, respectively: the former vector is obtained from the latter through a tax transformation, say τ_i, $i = 1, 2..., N$, where N is the number of individuals in the sample. Since the data are net of taxes and social

[2]The simulation error is defined as the difference between the simulated after-tax income, obtained applying the MSM to the declared before-tax income, and the declared after-tax income.

[3]IRPEF and "imposte sostitutive" on interest and capital gains accounted for 75,8% of 1998 total Italian revenues from direct taxation (Banca d'Italia, 1999).

contributions, the first role of the model is to recover individual BT income:

$$y_i^B = \tau_i^{-1}(y_i^A) \tag{2.1}$$

for all $i = 1, ...N$. There are two major complications here. First, the tax transformation τ_i is not the same for all individuals. Personal income taxation in Italy is on individual base; the amount of tax each individual has to pay depends on the type of incomes she receives and her family characteristics. For instance, arrears do not enter the personal income tax (IRPEF) base: they are taxed with a proportional tax rate while work and pension income is taxed with progressive tax rates; there are several tax allowances which depend on a set of individual and family characteristics, such as the number of dependent children, whether the spouse is dependant, whether income comes from self-employment, employment or pension, etc. Secondly, the tax transformation in (2.1) is highly non linear. This implies that y_B has to be obtained numerically, by recursive approximations. The tax transformation, τ_i, used to recover y^B is obtained from the 1998 tax code. Various assumptions about take-up rates of tax allowances could be introduced, however no uncertainty is considered here. Although the analysis of benefits take-up is a relevant issue in countries where welfare programs are widespread[4] in Italy there are no generalized unemployment benefit, income maintenance or house benefit schemes. The issue of non take-up is limited to tax allowances and tax deductions which do not involve issues of stigma or psycho-

[4]For instance, see among others Fry and Stark (1993), Duclos (1995), Bollinger and David (1997), Pudney (2001) and Pudney et al. (2002).

logical dependency and is then less relevant. Moreover, this choice, together with the idea of not considering behavioral responses, allowed the model to be as simple and robust as possible[5].

In this model the main assumptions are that (a) the sample is representative of the population and contains enough details for simulation, (b) the tax and benefit legislation, τ_i, is perfectly known by the individual and applied without error. Although the first assumption is granted by the Bank of Italy who produced the data, the second is meant to keep variability to a minimum. An alternative solution would be to randomly include errors in the model assessing the relevance of such changes in the MSM. Here, instead, only systematic errors leading to under-reporting are considered and treated as tax evasion or tax avoidance in Section 2.5; involuntary under-reporting is assumed to be off-setting with involuntary over-reporting errors. The probability of programming mistakes is kept to a minimum with a number of checks and a validation procedure (see Section 2.5).

TABEITA98 is developed in four different modules using Stata 7 on a personal computer. The first module involves the preparation of the data base, and in particular the data consistency checks. The second is the grossing-up procedure. The third involves an estimation of tax evasion to correct the underreporting of certain population groups and validation of the output. The last deals with the simulation of alternative scenarios. The first three stages will be briefly presented in the following sections. Chapter 3 will provide an illustration of simulations that can be performed with TABEITA98.

[5]For a discussion of the low reliability of MSMs with behavioral responses, recall Section 1.4.3

The model allows one to compute the amount of tax allowances, of tax base deductions, the amount of personal income tax paid, the BT income[6], transfers and pensions not liable of personal income tax. All incomes can be computed at the individual and household level, allowing for different equivalence scales.

2.3 Building the tax base

As a first step the model has to put together the different types of incomes to build the tax base. TABEITA98 was initially built and developed following Dirimod95, using 1995 SHIW data (Mantovani, 1998)[7]. Analysis of data is performed starting from employment, pension, self-employment incomes and rents. They all form the IRPEF tax base. During the reconstruction of the tax base, the data imputation of the Bank of Italy is analyzed: it appears to be particularly relevant especially for self-employment and rental income. Although the Bank of Italy does not explain in detail how the data imputation is performed and what kind of additional information are known and used, imputed data are used here.

Once all the components of the IRPEF tax base have been assembled, household relations are recovered identifying those who are required to present the tax form and those who are not, the family relations and the right to use tax allowances depending on family composition according to the 1998 tax code. This analysis is performed

[6]BT income is divided into five components: (a) employment, (b) self-employment, (c) pension, (d) rental and estate income, (e) capital, interests and participation.

[7]In particular TABEITA98 follows Dirimod95 quite strictly for reconstructing the BT income while the algorithm for obtaining AT income was developed ad hoc. However, so many were the changes made on Dirimod95, especially as for estate and rental incomes, grossing-up procedure and tax evasion estimation, that the TABEITA98 is quite different from Dirimod95, the latter bearing no responsibility for possible mistakes in the former.

income brackets	tax rates
0-15,000	19%
15,000-30,000	27%
30,000-60,000	34%
60,000-135,000	40%
over 135,000	46%

Table 2.1: Structure of 1998 IRPEF tax brackets in Lit '000 (Lit 1,936.27 = €1).

assuming a coincidence of household and family since the data do not allow one to disentangle the presence of more than one family under the same dwelling.

Up to this point, income is only AT. The program then recovers the BT income excluding those components of AT income which are not liable to IRPEF (e.g. invalidity pensions), taking into account those tax allowances which do not depend on income, then applying an iterative procedure to recover numerically the BT income. For the Italian personal income tax law, the AT income of individual i can be derived as:

$$y_i^A = y_i^B - t_i(\underbrace{y_i^B - yex_i}_{yc_i} - d_i) + D_i + Dy_i \tag{2.2}$$

where $t_i(\dot{})$ is the five-bracket tax function applied to taxable income; yc_i is the IRPEF gross income, which is equal to BT income minus IRPEF-exempt incomes, yex_i; d_i is a set of deductions to be subtracted from yc_i; D_i and Dy_i are a the set of tax allowances that do and do not depend on BT income, respectively, and that reduce the tax to be paid. Table 2.1 shows the structure of the five-bracket tax schedule: the lower tax rate is 19% the higher 46%, showing the strong progressivity of IRPEF. Table 2.2 shows the main tax allowances for family burdens: some are depending on the BT income, some other are constant regardless of the BT income. Table 2.3

Dependent spouse		Other dependent relatives	
income brackets	amount	type	amount
0-30,000	1,057.55	each child	336
30,000-60,000	951.55	other relative	336
60,000-100,000	889.55		
over 100,000	817.55		

Table 2.2: Tax allowances for dependent spouse and for other dependent relatives in Lit '000 (Lit 1,936.27 = €1)

presents the other main tax allowances aimed at reducing the tax due by individuals whose income comes mainly from work income. Traditionally tax allowances are lower for the self-employed, mainly because, in contrast to the employed, their tax base can be reduced by the costs of producing income.

The algorithm developed in ITAXMOD98 then recovers the BT income as:

$$y_i^B = y_i^A - D_i - Dy_i + t_i(y_i^B - yex_i - d_i) \tag{2.3}$$

Clearly, in the first step of (2.3) y_i^B and Dy_i on the RHS are unknown, then they are set to zero. From the second step onward, y_i^B on the RHS is replaced with y_i^B obtained in the previous iteration, as well as Dy_i is computed accordingly to y_i^B of the previous iteration. The process goes on until y_i^B does not change for two successive iterations for all individuals in the sample.

2.4 Grossing-up

As discussed in Chapter 1, MSMs are mainly used for forecasting and analyzing the impact of a change in the structure of the tax and benefit system on (a) the

Employment		Self-employment	
income brackets	amount	income brackets	amount
0-9,100	1,680	0-9,100	700
9,100-9,300	1,600	9,100-9,300	600
9,300-15,000	1,500	9,300-9,600	500
15,000-15,300	1,350	9,600-9,900	400
15,300-15,600	1,250	9,900-15,000	300
15,600-15,900	1,150	15,000-30,000	200
15,900-30,000	1,050	30,000-60,000	100
30,000-40,000	950		
40-000-50,000	850		
50,000-60,000	750		
60,000-60,300	650		
60,300-70,000	550		
70,000-80,000	450		
80,000-90,000	350		
90,000-90,400	250		
90,400-100,000	150		
over 100,000	100		

Table 2.3: Tax allowances depending on amount and type of income received in Lit '000 (Lit 1,936.27 = €1)

distribution of income and (b) the public accounts.

While the analysis of the redistributive effects of tax-benefit policies could be conducted in relative terms, the forecasting of the aggregate effects on public accounts requires the projection of the sample to country totals. This projection can be obtained using a simple proportion between the dimension of the sample and that of the national population, generally weighted using the sampling weights provided in the data set. More often, data sets come with weights to be used for national projections, which are obtained from a process of post-stratification of the sample to known population totals. Post-stratification is an issue that have been extensively analyzed in survey statistics (see for instance Sarndal et al. (1992)) and consists in calibrating some sub-samples (post-strata) of a data set to given totals. In the MSM literature

post-stratification is more commonly referred to as grossing-up, since the problem consists in grossing the sample up to the population under study. The aim of the grossing-up procedure is to make the sample as close as possible to the true population, although it depends on the variable used for performing the grossing-up as well as on the procedure implemented. For instance, the grossing-up procedure proposed by Atkinson et al. (1988) aims at minimizing the distance between the initial weights, provided with the data set, and the grossed-up weights using a measure of distance derived from information theory (recall Section 1.4.1). However, a particular set of grossing-up weights can be able to closely reflect the characteristics of the population as for some variables but not for others.

The SHIW data set is post-stratified using the variables sex, age class, area and dimension of the town of residence (Banca d'Italia, 2000, p. 40). However, it is not clearly stated what methodology was used and, for instance, which age classes were considered. Table 2.4 shows the population totals taken for ISTAT (Italian Institute of Statistics) as for a set of variables. Using the weights provided in the SHIW data set, the differences between the grossed-up and actual figures are less than 0.2% for sex and area of residence (North-West (NW), North-East (NE), Center (C) and South (S)). As for the age classes, the difference between grossed-up and population figures is however more relevant: for instance, the grossed-up totals are under-estimated by 3.7% and over-estimate by 6.4% for the 18-30 and the over-65 years groups, respectively. Since the Bank of Italy does not make public the age groups considered, it could be possible that this difference is due to the different age groups used here. None the less, this shows a problem with grossed-up simulation: a

redistributive policy in favor of the old age would imply an overstatement of its cost due to the over-sampling of this age group. These distortions could be even worse for other subsamples. For instance, using the same age groups divided in the 4 macro areas considered before, Table 2.4 shows that the SHIW grossed-up weights would over-represent the elderly living in the South (S) by about 25%. They would also induce an over-representation of the self-employed in the Center (C) and an under-representation the self-employed in the S (Table 2.4). Moreover, while the difference between the actual and the grossed-up total number of employed and self-employed is smaller than 1%, these figures hide a 29% over-representation of the self-employed in the C and an under-representation of self-employed in the NE and in the S by 13% and 21%, respectively. All these issues are of relevance whenever an analysis of income by population sub-groups is performed.

For these reasons a set of alternative grossing-up weights were estimated using the same methodology as Atkinson et al. (1988) using control totals found in ISTAT (2004); CNEL (2004). Tables 2.5 and 2.6 show the results for six different weights: "weight 1" uses total population, by area, by sex, by age groups (below 18, between 19 and 30, between 31 and 65, over-65). Although the grossing-up methodology performs well for the variables considered, it does also have an effect on other variables. For instance, it has a positive effect on the over-65 living in the S, reducing its over-representation but it has a negative effect on the over-65 living in the C, increasing its under-representation. The "weight 2" improves on "weight 1" for grossing-up also on the age groups by area of residence. Although for some variables its effect is positive, reducing the discrepancy from actual totals, it is still rather unsatisfactory

Variables	Actual	SHIW	diff.
Tot Population*	57,612,615	57,612,568	0.00%
Males*	27,967,670	27,951,136	-0.06%
Females*	29,644,945	29,661,432	0.06%
Pop NW*	15,069,493	15,099,744	0.20%
Pop NE*	10,560,820	10,547,936	-0.12%
Pop C*	11,071,715	11,064,505	-0.07%
Pop S*	20,910,587	20,900,383	-0.05%
age≤18*	10,845,419	11,032,994	1.73%
18<age≤30*	9,987,651	9,619,324	-3.69%
30<age≤65*	27,218,646	26,787,452	-1.58%
age>65*	9,560,899	10,172,798	6.40%
age≤18NW*	2,409,663	2,497,552	3.65%
age≤18NE*	1,687,699	1,853,786	9.84%
age≤18C*	1,873,809	2,073,762	10.67%
age≤18S*	4,874,248	4,607,894	-5.46%
18<age≤30NW*	2,498,184	2,411,373	-3.47%
18<age≤30 NE*	1,766,221	1,630,855	-7.66%
18<age≤30 C*	1,824,075	1,988,102	8.99%
18<age≤30 S*	3,899,171	3,588,994	-7.95%
30<age≤65 NW*	7,509,728	7,523,230	0.18%
30<age≤65 NE*	5,174,474	5,070,504	-2.01%
30<age≤65 C*	5,368,887	5,191,858	-3.30%
30<age≤65 S*	9,165,557	9,001,860	-1.79%
age>65 NW*	2,651,918	2,667,589	0.59%
age>65 NE*	1,932,426	1,992,791	3.12%
age>65 C*	2,004,944	1,810,783	-9.68%
age>65 S*	2,971,611	3,701,635	24.57%
employed**	14,549,000	14,530,169	-0.13%
self-employed**	5,886,000	5,852,953	-0.56%
employed NO**	4,470,000	4,345,113	-2.79%
employed NE**	3,104,000	3,199,310	3.07%
employed C**	2,911,000	2,821,364	-3.08%
employed S**	4,086,000	4,164,382	1.92%
self-empl NO**	1,643,000	1,793,760	9.18%
self-empl NE**	1,330,000	1,156,244	-13.06%
self-empl C**	1,184,000	1,532,649	29.45%
self-empl S**	1,730,000	1,370,300	-20.79%
Elementary schooling*	16,104,000	15,625,930	-2.97%
Compulsory schooling*	16,118,000	13,975,447	-13.29%
High School degree*	13,365,000	15,402,757	15.25%
Laurea*	3,066,000	3,641,053	18.76%
Agriculture**	1,201,000	1,038,245	-13.55%
Industry**	6,730,000	6,548,547	-2.70%
Services**	12,504,000	12,796,330	2.34%
Single*	4,982,000	4,380,481	-12.07%
Single parent*	1,655,000	1,666,809	0.71%
Couple no kids*	3,828,000	4,373,753	14.26%
Couple w/ kids*	9,410,000	9,978,556	6.04%
Others*	1,440,000	773,930	-46.25%
All families*	21,315,000	21,173,529	-0.66%

Table 2.4: Grossed-up variables using Banca d'Italia (2000) grossing-up weights compared with population totals. *External source: ISTAT (2004). **External source: CNEL (2004).

as for the number of employed and self-employed in different parts of the countries. Since this is relevant for detecting the possibility of tax evasion/avoidance, some more grossing-up weights are estimated. Eventually, "weight 6" is chosen to replace the SHIW initial weights. It allows one to gross-up the sample to a population that is very close to the true population for a large number of relevant variables, including occupation by area of residence, education and sector of activity. However, as the last column of Table 2.6 shows, this weight is still unable to represent correctly the distribution of family by type: single households tend to be under-represented while couple with kids tend to be over-represented. It is chosen not to perform the grossing-up procedure also for type of families mainly because external data refer to families and SHIW refers uniquely to households. In chosing "weight 6" as final weight it was also considered the risk of increasing the variance of weights with respect to initial weights. The increased variance could come from two type of factors. First, in contrast to SHIW initial weights "weight 6" is not uniform within the household since the grossing-up procedure is performed at the individual level. Second, the larger the number of control totals the smaller the number of observations with each combination of characteristics being controlled (Sutherland, 1989). However, as Table 2.7 shows, the variance of "weight 6" is not larger than original SHIW weights.

In comparable Italian MSMs the issue of estimation of grossing-up weights alternative to those provided in the data set is often overlooked. Neither MASTRICT (Proto, 2000), nor Dirimod95 (Mantovani, 1998) and its updated version Mapp98 (Baldini, 1998), nor the Italian module in EUROMOD (Atella et al., 2001) address the problem and the weights provided in the SHIW data set are used instead.

Variables	Actual	weight 1	diff.	weight 2	diff.	weight 3	diff.
Tot Population*	57,612,615	57,612,734	0.00%	57,612,769	0.00%	57,612,689	0.00%
Males*	27,967,670	27,967,738	0.00%	27,967,745	0.00%	27,967,768	0.00%
Females*	29,644,945	29,644,996	0.00%	29,645,024	0.00%	29,644,921	0.00%
Pop NW*	15,069,493	15,069,523	0.00%	15,069,538	0.00%	15,069,537	0.00%
Pop NE*	10,560,820	10,560,844	0.00%	10,560,808	0.00%	10,560,852	0.00%
Pop C*	11,071,715	11,071,708	0.00%	11,071,783	0.00%	11,071,708	0.00%
Pop S*	20,910,587	20,910,659	0.00%	20,910,640	0.00%	20,910,592	0.00%
age≤18*	10,845,419	10,845,442	0.00%	10,845,466	0.00%	10,845,422	0.00%
18<age≤30*	9,987,651	9,987,683	0.00%	9,987,658	0.00%	9,987,698	0.00%
30<age≤65*	27,218,646	27,218,712	0.00%	27,218,727	0.00%	27,218,684	0.00%
age>65*	9,560,899	9,560,897	0.00%	9,560,918	0.00%	9,560,885	0.00%
age≤18NW*	2,409,663	2,448,122	1.60%	2,409,678	0.00%	2,452,333	1.77%
age≤18NE*	1,687,699	1,825,501	8.17%	1,687,702	0.00%	1,829,320	8.39%
age≤18C*	1,873,809	2,036,606	8.69%	1,873,838	0.00%	2,040,402	8.89%
age≤18S*	4,874,248	4,535,213	-6.96%	4,874,248	0.00%	4,523,367	-7.20%
18<age≤30NW*	2,498,184	2,496,925	-0.05%	2,498,198	0.00%	2,485,397	-0.51%
18<age≤30 NE*	1,766,221	1,696,552	-3.94%	1,766,211	0.00%	1,685,816	-4.55%
18<age≤30 C*	1,824,075	2,062,713	13.08%	1,824,076	0.00%	2,064,014	13.15%
18<age≤30 S*	3,899,171	3,731,493	-4.30%	3,899,173	0.00%	3,752,471	-3.76%
30<age≤65 NW*	7,509,728	7,624,243	1.52%	7,509,742	0.00%	7,628,840	1.59%
30<age≤65 NE*	5,174,474	5,162,376	-0.23%	5,174,474	0.00%	5,164,466	-0.19%
30<age≤65 C*	5,368,887	5,271,826	-1.81%	5,368,918	0.00%	5,265,679	-1.92%
30<age≤65 S*	9,165,557	9,160,267	-0.06%	9,165,593	0.00%	9,159,699	-0.06%
age>65 NW*	2,651,918	2,500,233	-5.72%	2,651,920	0.00%	2,502,967	-5.62%
age>65 NE*	1,932,426	1,876,415	-2.90%	1,932,421	0.00%	1,881,250	-2.65%
age>65 C*	2,004,944	1,700,563	-15.18%	2,004,951	0.00%	1,701,613	-15.13%
age>65 S*	2,971,611	3,483,686	17.23%	2,971,626	0.00%	3,475,055	16.94%
employed**	14,549,000	14,832,997	1.95%	14,830,481	1.93%	14,549,021	0.00%
self-employed**	5,886,000	5,951,129	1.11%	5,942,476	0.96%	5,886,031	0.00%
employed NO**	4,470,000	4,428,138	-0.94%	4,381,285	-1.98%	4,346,883	-2.75%
employed NE**	3,104,000	3,276,169	5.55%	3,317,881	6.89%	3,218,263	3.68%
employed C**	2,911,000	2,877,905	-1.14%	2,838,694	-2.48%	2,824,913	-2.96%
employed S**	4,086,000	4,250,785	4.03%	4,292,621	5.06%	4,158,962	1.79%
self-empl NO**	1,643,000	1,817,118	10.60%	1,802,766	9.72%	1,798,636	9.47%
self-empl NE**	1,330,000	1,178,261	-11.41%	1,190,253	-10.51%	1,167,382	-12.23%
self-empl C**	1,184,000	1,560,119	31.77%	1,548,420	30.78%	1,543,659	30.38%
self-empl S**	1,730,000	1,395,631	-19.33%	1,401,037	-19.02%	1,376,354	-20.44%
Elementary schooling*	16,104,000	15,303,360	-4.97%	15,235,698	-5.39%	15,381,891	-4.48%
Compulsory schooling*	16,118,000	14,110,613	-12.45%	14,156,361	-12.17%	14,100,825	-12.52%
High School degree*	13,365,000	15,692,853	17.42%	15,725,012	17.66%	15,646,311	17.07%
Laurea*	3,066,000	3,691,111	20.39%	3,687,272	20.26%	3,663,811	19.50%
Agriculture**	1,201,000	1,058,337	-11.88%	1,064,699	-11.35%	1,039,454	-13.45%
Industry**	6,730,000	6,686,566	-0.65%	6,684,699	-0.67%	6,579,136	-2.24%
Services**	12,504,000	13,039,223	4.28%	13,023,559	4.16%	12,816,462	2.50%
Single*	4,982,000	4,245,869	-14.78%	4,251,030	-14.67%	4,240,640	-14.88%
Single parent*	1,655,000	1,676,363	1.29%	1,664,537	0.58%	1,674,495	1.18%
Couple no kids*	3,828,000	4,289,382	12.05%	4,296,591	12.24%	4,297,968	12.28%
Couple w/ kids*	9,410,000	10,061,245	6.92%	10,052,744	6.83%	10,059,334	6.90%
Others*	1,440,000	760,855	-47.16%	753,073	-47.70%	760,403	-47.19%
All families*	21,315,000	21,033,714	-1.32%	21,017,975	-1.39%	21,032,840	-1.32%

Table 2.5: Grossed-up variables using alternative grossing-up weights compared with population totals. *External source: ISTAT (2004). **External source: CNEL (2004).

Variables	Actual	weight 4	diff.	weight 5	diff.	weight 6	diff.
Tot Population*	57,612,615	57,612,738	0.00%	57,612,693	0.00%	57,612,819	0.00%
Males*	27,967,670	27,967,691	0.00%	27,967,666	0.00%	27,967,812	0.00%
Females*	29,644,945	29,645,047	0.00%	29,645,027	0.00%	29,645,007	0.00%
Pop NW*	15,069,493	15,069,535	0.00%	15,069,524	0.00%	15,069,552	0.00%
Pop NE*	10,560,820	10,560,850	0.00%	10,560,811	0.00%	10,560,825	0.00%
Pop C*	11,071,715	11,071,748	0.00%	11,071,755	0.00%	11,071,777	0.00%
Pop S*	20,910,587	20,910,605	0.00%	20,910,603	0.00%	20,910,665	0.00%
age≤18*	10,845,419	10,845,437	0.00%	10,845,455	0.00%	10,845,487	0.00%
18<age≤30*	9,987,651	9,987,679	0.00%	9,987,642	0.00%	9,987,667	0.00%
30<age≤65*	27,218,646	27,218,670	0.00%	27,218,696	0.00%	27,218,729	0.00%
age>65*	9,560,899	9,560,952	0.00%	9,560,900	0.00%	9,560,936	0.00%
age≤18NW*	2,409,663	2,462,053	2.17%	2,464,446	2.27%	2,409,685	0.00%
age≤18NE*	1,687,699	1,813,836	7.47%	1,816,658	7.64%	1,687,714	0.00%
age≤18C*	1,873,809	2,122,875	13.29%	2,146,916	14.57%	1,873,829	0.00%
age≤18S*	4,874,248	4,446,673	-8.77%	4,417,435	-9.37%	4,874,259	0.00%
18<age≤30NW*	2,498,184	2,498,413	0.01%	2,508,562	0.42%	2,498,199	0.00%
18<age≤30 NE*	1,766,221	1,680,691	-4.84%	1,673,535	-5.25%	1,766,212	0.00%
18<age≤30 C*	1,824,075	2,052,503	12.52%	2,036,362	11.64%	1,824,073	0.00%
18<age≤30 S*	3,899,171	3,756,072	-3.67%	3,769,183	-3.33%	3,899,183	0.00%
30<age≤65 NW*	7,509,728	7,605,731	1.28%	7,607,615	1.30%	7,509,730	0.00%
30<age≤65 NE*	5,174,474	5,194,626	0.39%	5,205,307	0.60%	5,174,472	0.00%
30<age≤65 C*	5,368,887	5,136,223	-4.33%	5,117,690	-4.68%	5,368,921	0.00%
30<age≤65 S*	9,165,557	9,282,090	1.27%	9,288,084	1.34%	9,165,606	0.00%
age>65 NW*	2,651,918	2,503,338	-5.60%	2,488,901	-6.15%	2,651,938	0.00%
age>65 NE*	1,932,426	1,871,697	-3.14%	1,865,311	-3.47%	1,932,427	0.00%
age>65 C*	2,004,944	1,760,147	-12.21%	1,770,787	-11.68%	2,004,954	0.00%
age>65 S*	2,971,611	3,425,770	15.28%	3,435,901	15.62%	2,971,617	0.00%
employed**	14,549,000	14,549,021	0.00%	14,549,022	0.00%	14,549,025	0.00%
self-employed**	5,886,000	5,885,996	0.00%	5,885,999	0.00%	5,886,023	0.00%
employed NO**	4,470,000	4,470,026	0.00%	4,470,001	0.00%	4,470,008	0.00%
employed NE**	3,104,000	3,103,998	0.00%	3,103,985	0.00%	3,103,999	0.00%
employed C**	2,911,000	2,911,003	0.00%	2,911,012	0.00%	2,911,011	0.00%
employed S**	4,086,000	4,063,994	-0.54%	4,064,024	-0.54%	4,064,007	-0.54%
self-empl NO**	1,643,000	1,642,995	0.00%	1,642,994	0.00%	1,643,006	0.00%
self-empl NE**	1,330,000	1,330,003	0.00%	1,330,004	0.00%	1,329,990	0.00%
self-empl C**	1,184,000	1,184,001	0.00%	1,184,002	0.00%	1,184,014	0.00%
self-empl S**	1,730,000	1,728,997	-0.06%	1,728,999	-0.06%	1,729,013	-0.06%
Elementary schooling*	16,104,000	15,387,753	-4.45%	16,104,016	0.00%	16,104,072	0.00%
Compulsory schooling*	16,118,000	14,121,594	-12.39%	16,118,021	0.00%	16,118,041	0.00%
High School degree*	13,365,000	15,619,639	16.87%	13,365,006	0.00%	13,365,007	0.00%
Laurea*	3,066,000	3,661,898	19.44%	3,065,999	0.00%	3,066,023	0.00%
Agriculture**	1,201,000	1,055,137	-12.15%	1,200,997	0.00%	1,201,005	0.00%
Industry**	6,730,000	6,574,907	-2.30%	6,730,011	0.00%	6,730,028	0.00%
Services**	12,504,000	12,804,973	2.41%	12,504,013	0.00%	12,504,015	0.00%
Single*	4,982,000	4,226,747	-15.16%	4,230,030	-15.09%	4,236,991	-14.95%
Single parent*	1,655,000	1,669,811	0.89%	1,685,693	1.85%	1,672,088	1.03%
Couple no kids*	3,828,000	4,290,829	12.09%	4,278,283	11.76%	4,289,241	12.05%
Couple w/ kids*	9,410,000	10,069,620	7.01%	10,056,183	6.87%	10,047,093	6.77%
Others*	1,440,000	762,475	-47.05%	762,409	-47.05%	754,947	-47.57%
All families*	21,315,000	21,019,482	-1.39%	21,012,598	-1.42%	21,000,360	-1.48%

Table 2.6: Grossed-up variables using alternative grossing-up weights compared with population totals. *External source: ISTAT (2004). **External source: CNEL (2004).

weight	Obs	Mean	Std. Dev.	Min	Max
SHIW	20901	2756.453	2688.614	228.1136	28395.24
weight 1	20901	2756.453	2672.33	232.074	28755
weight 2	20901	2756.453	2669.014	232.642	28326
weight 3	20901	2756.453	2671.421	228.981	28569.5
weight 4	20901	2756.453	2669.172	220.594	27490.8
weight 5	20901	2756.453	2760.695	186.943	33526.2
weight 6	20901	2756.453	2759.995	183.083	33711.1

Table 2.7: Summary statistics for initial SHIW and final grossing-up weights.

2.5 Estimation of tax evasion and validation

A common finding from the SHIW data set is that total income in the survey is on average higher than what declared to fiscal authorities and that the difference is larger for some incomes (e.g. self-employment) and smaller for others (e.g. employment) (see among others Marenzi, 1996). Disregarding this fact would than imply a simulation of a larger tax yield than that actually obtained, hence an incorrect forecasting of redistributive and revenue effects of different fiscal policies. Any MSM for Italy that uses SHIW data, needs to consider this discrepancy. The common practice is to assume that the difference comes from the fact that taxpayers are more honest with an interviewer that grants anonymity than with the fiscal authorities. The difference between the total amount of income grossed-up from individual incomes declared in the SHIW and the total amount of income declared to the fiscal authorities is therefore attributed to tax evasion or tax avoidance. This approach has also been used in recent year to provide an estimate of tax evasion to be compared with other methods of tax evasion estimation. Alternative methodologies include the direct

approach[8], the indirect approach[9] or the latent variable approach[10] (for a review of results from different methods to estimate underground economy in Italy, see Zizza (2002)).

In Dirimod95 an estimate of evasion/avoidance is obtained making use of the analysis of 1991 tax forms, relative to 1990 incomes (Ministero delle Finanze, 1995). Tax evasion and tax avoidance is estimated in two stages. At first the model is run assuming zero evasion, data are grossed-up and total BT income is compared to BT income as found from aggregated tax forms. In the second stage increasing level of tax evasion and tax avoidance is estimated and the resulting income is compared with data from aggregate tax forms. It is then found that grossed-up employment income is about the same as what was declared to fiscal authorities while self-employment income is about 15% lower (Mantovani, 1998). In Dirimod95, however, tax form data refer to 1990 incomes and their updating to 1995 using a constant consumer price index (CPI) adds a bias in the estimation procedure. The module of EUROMOD for Italy deals with tax evasion and tax avoidance in a similar way, using results from MASTRICT, a MSM developed at ISTAT (Proto, 2000). In particular, in the EUROMOD module for Italy employment income tax evasion and tax avoidance is

[8]The direct approach is based on tax audits or on census and labor force data and on analysis of expenditures. The assumptions are that the labor force participation obtained from population census and labor force surveys include unregistered employment, and that consumption is more truthfully declared than income, clearly an analogous assumption to what used here.

[9]The indirect approach is based on the currency demand approach, which assumes that hidden transactions use cash. A demand for currency is than estimated with regression methodologies (for a review of the approach and updated results, see Schneider (2000))

[10]The latent variable approach considers the underground economy as a non observable variable and estimates the links with a set of determinants, including the production and the labor market activities. It employs latent variable econometric tools combined with factorial analysis (see, for instance Frey and Weck-Hanneman, 1984)

Type of income	total of tax forms (Lit '000)	total of SHIW (Lit '000)	Difference
rents and estates	40,713	53,100	30%
employment	636,553	667,800	5%
self-employment	115,213	166,300	44%
capital and participation	65,022	53,100	-18%

Table 2.8: Tax evasion estimation for TABEITA98

estimated to 0%, self-employment income to 50% (Atella et al., 2001).

The methodology to estimate tax evasion in TABEITA98 is analogous to the one in Dirimod95. Results differ as for the estimation obtained for two main reasons. First, the data set used is 1998, not 1995 SHIW; second, the external source is the analysis of 1999 tax forms, relative to 1998 incomes (Ministero delle Finanze, 2002). The time coincidence of the external source and the data set allows one to avoid the adjustments to different periods and, possibly, different conditions of the economy. In order to compare the SHIW and the Ministry of Finance (MF) data a MSM is needed since MF data presents information only about BT. In the first stage the model is run exactly as described in Section 2.3. The weighted sum of individual yc_i is compared with MF external data for the population of taxpayers.

Results show that tax evasion is particularly relevant for self-employment income (44%) and for rental and estate income (30%) and is small, although positive, for for employment income (5%) (Table 2.8). Capital income are underestimated with SHIW compared to tax forms data by 18%. Although this results may seem odd, it was also noted in other MSMs using SHIW data (for instance, see Atella et al., 2001).

Equation (2.3) is modified to include tax evasion/avoidance in TABEITA98 ,

subtracting evaded/avoided individual income, yev_i, from the tax base.

$$y_i^B = y_i^A - yev_i - D_i - Dy_i + t_i(y_i^B - yex_i - d_i - yev_i) \tag{2.4}$$

y_i^B is then BT income as it appears to the fiscal authorities, but it is not the actual BT. To obtain the actual BT income, ya_i^B, the concealed income has to be added back:

$$ya_i^B = y_i^B + yev_i \tag{2.5}$$

The amount of evaded income is derived calibrating yev_i at the aggregate level starting from tax evasion results reported in Table 2.8. The individual evaded income is obtained multiplying individual employment, self-employment and rental and estate income by 6%, 27% and 31%, respectively. No correction is applied to capital and participation income.

Using this procedure it is then possible to provide a first validation of the model's output. At the national level employment, self-employment and rental and estate incomes are about the same as found from tax returns. Capital income is still underestimated because no correction was introduced. At the national level actual BT income is lower than the declared BT income by 13%, a result in line with other findings (for instance, see Calzaroni, 2000). Comparing the model's results with MF data by geographical areas, the differences with MF data are generally larger than at the national level.

These results are consistent with other Italian MSMs and in some cases TABEITA98

performs better than other models. For instance, the Italian module of EUROMOD overestimates family tax allowances by 44%, TABEITA98 only by 3% (Table 2.9) . Some variables reported in Table 2.9 are not easily comparable because they are not reported in other MSMs documentation. For instance, validation results are reported only at the national level and no results about work tax allowances are produced.

2.6 Summary: why another MSM?

In this chapter TABEITA98, a static MSM for Italy developed using 1998 SHIW data, was presented. TABEITA98 allows one to simulate personal income taxation, which accounted for about 3/4 of Italian direct taxation in 1998. It does not introduce behavioral responses and avoids simulating different take-up rates to keep its structure as simple as possible. This chapter discussed the importance of addressing the grossing-up of the sample to population totals, both for redistributive analysis and the forecasting of effects of fiscal reforms on public finances. The grossing-up weights estimated here will also be used in the following chapters. The use of TABEITA98 provided an updated estimation of tax evasion and tax avoidance in Italy: it seems to be low for employment income, over 40% for self-employment income and about 30% for rental and estate income.

TABEITA98 produces results that are comparable to other Italian MSMs, improving on the issue of grossing-up and validation and on the estimation of tax evasion/tax avoidance. However, in contrast to some of them, it is limited to personal income taxation. Notwithstanding this limitation, TABEITA98 is suitable to study tax-benefit

variable	SHIW (Lit '000,000)	Ext. Source (Lit '000,000)	Difference (%)
Italy			
declared BT income	844,700,000	866,004,870	-2
employment inc.	635,000,000	636,553,037	0
self-empl inc.	116,100,000	115,212,773	1
capital inc.	52,910,000	65,022,227	-19
rental& est. inc.	40,730,000	40,712,541	0
actual BT inc.	980,200,000	866,004,870	13
NW area			
declared BT income	279,400,000	290,741,450	-4
employment inc.	207,800,000	208,730,788	0
self-empl inc.	39,050,000	41,435,160	-6
capital inc.	20,590,000	24,881,868	-17
rental & est. inc.	11,980,000	12,886,631	-7
actual BT inc.	323,400,000	290,741,450	11
NE area			
declared BT income	188,200,000	202,179,990	-7
employment inc.	137,100,000	143,368,984	-4
self-empl inc.	24,810,000	28,569,365	-13
capital inc.	16,020,000	19,050,314	-16
rental & est. inc.	10,250,000	9,291,672	10
actual BT inc.	216,800,000	202,179,990	-7
C area			
declared BT income	174,900,000	180,395,573	-3
employment inc.	132,700,000	133,010,136	0
self-empl inc.	23,810,000	23,398,024	2
capital inc.	9,055,000	12,486,490	-27
rental & est. inc.	9,374,000	9,531,839	-2
actual BT inc.	202,800,000	180,395,573	12
S area			
declared BT income	202,100,000	192,687,857	5
employment inc.	157,300,000	151,443,129	4
self-empl inc.	28,390,000	21,810,224	30
capital inc.	7,248,000	8,603,555	-16
rental & est. inc.	9,117,000	9,002,399	1
actual BT inc.	237,200,000	192,687,857	23
family tax. allowances	8,137,920	9,186,184	3
work tax. allowances	31,392,390	23,454,708	34

Table 2.9: Validation of the TABEITA98 output, at the national and area level. Ext. Sources: Ministero delle Finanze (2002); CNEL (2004)

microsimulation modelling. TABEITA98 was validated with external sources, especially with those coming from the analysis of 1998 tax forms: its results are at least as good as those of comparable MSMs.

The main limitation of this MSM lies in the input data: only AT incomes are asked for in the survey and BT incomes have to be simulated. Moreover, there are problems in identifying the appropriate variable for the model since the SHIW survey was not primarily collected for MSMs. Sometimes these problems are resolved using other information in the data set (e.g. tax allowances depending on family burdens are estimated using information about household relations), in others dropping the variable (e.g. capital and participation incomes is often disregarded from the analysis). However, these limitations are common to all Italian MSMs since, unfortunately, no alternative data set for tax-benefit simulation is currently available.

Chapter 3 proposes the use of non-parametric density estimation to assess the effect of fiscal reforms and Chapter 4 addresses the issue of reliability of MSMs. Both chapters make use of TABEITA98.

Chapter 3

Microsimulation and non-parametric estimation

Nonparametric density estimation can be regarded as a development of the more intuitive histogram technique for density estimation. In contrast to histograms, nonparametric density estimation does not suffer major limitations such as choice of origin, limited robustness of estimates, ragged picture, absence of derivative and low flexibility for multivariate density analysis. Although the theory of nonparametric density estimation is well established and has developed since the 1950s, the use of nonparametric density estimation in empirical research has spread widely in more recent years mainly because of the rapid growth in computing power. The interpolation of pointwise estimation of density provides a smooth picture, useful in detecting unusual behavior of the distribution such as bimodality.

Nonparametric density estimation has proven to be an effective research tool in economics. For instance, in income inequality literature kernel estimation has been

used to show the evolution of income distributions from a unimodal to a bimodal shape. Cowell et al. (1996); Jenkins (1994) among others called this phenomenon the "shrinking of the middle class". They showed that the widening gap between least and most well-off households in the UK was mainly due to stagnating income of non-working and relatively large households opposed to dynamic earnings of working households. Pudney (1993) used nonparametric methods to analyze income and wealth inequality in the life-cycle using Chinese data: he found that only a small part of observed inequality can be explained by life-cycle factors in contrast to what other authors had found using dummy-variable regression methods. In the empirical growth literature Quah (1997) used nonparametric estimation to point out the emergence of a "twin peaks effect" - the clustering of a large number of countries at lower per capita incomes and the increasing gap between poor and rich countries. In labor economics DiNardo et al. (1996) used a semi-parametric analysis based on kernels to show the importance of institutional factors, such as unionization and minimum wage, for the evolution of wage distribution in the US labor market.

Kernel density estimation has been used also on Italian data, mainly to assess household income inequality between the late 1980s and early 1990s. Bimodality of disposable equivalent income density has been found using polarization indices (D'Ambrosio, 2001) and bimodality tests (Pittau and Zelli, 2001).

This chapter suggests a combination of MSMs and the descriptive power of nonparametric density estimation to analyze tax policies. Using a static MSM developed for personal income taxation in Italy, it focuses on the 1998 personal income taxation reform and shows that nonparametric density estimation is a useful tool for tax-benefit

policy analysis.

Section 3.1 deals with the data set and the microsimulation model used and briefly describes the main novelties of the 1998 IRPEF reform. Section 3.2 presents the non-parametric methodology for density estimation. In Section 3.3 the methodology proposed is developed and the results obtained are discussed. In Sections 3.4 and 3.5 some other simulations are performed and results commented. Section 3.6 concludes.

3.1 The data set, the microsimulation model and the IRPEF reform

The data set used is the 1998 Survey of Household Income and Wealth (SHIW) described in Section 2.1. The SHIW database is the most frequently used data set for Italian household microeconomic analysis, including tax policy analysis, since it comprises information about all members of the interviewed household and their relationships. However, since all data are net of taxes and social contributions, TABEITA98 model is used to recover gross income prior to any simulation of change in taxes and benefits, as described in Section 2.2.

Because of the financial and currency crisis that hit Italy in 1992, several tax policies were introduced during the 1990s, affecting both direct and indirect taxation. In particular, two clearly different periods may be distinguished: the first up to 1996 and the second starting from 1996. The first of these periods was characterized by constant political instability and frequent changes of the Minister of Finance, with

several temporary taxes without a clear overall design, while during the second higher political stability favored the design of a comprehensive tax reform. Here, the focus will be on some aspects of the 1997-98 tax reform, and in particular on the effects caused on the distribution of income and inequality on Italian households caused by the personal income tax (IRPEF) reform.

The two main novelties of the 1998 IRPEF reform with respect to previous years IRPEF concern the modification of the tax brackets and of the tax allowances structure, while no relevant change in income base definition was introduced. As shown in Appendix A, the number of fiscal brackets were reduced, from seven to five, with the reduction of the highest tax rate (from 50% of 1991, increased to 51% from 1992 onwards, to 45.5% of 1998), the increase of the first tax rate (from 10% of 1991 to 18.5% of 1998) and a substantial change of the others. Tax allowances for employment and self employment were increased in amount and in number, tax allowances for "family burdens" were increased, a new tax allowance for pension recipient where introduced depending on income and a few other attributes. The center-left government which passed the reform claimed that the increase in tax allowances would have offset the effect of the first bracket tax rate increase.

These topics have been analyzed by other authors. Among others, Bosi et al. (1999), CER (1998b,a) and Birindelli et al. (1998) analyzed in detail the 1998 reform compared with the previous year legislation, while Giannini and Guerra (1999) compared 1999 taxation system with the 1990 one. They all conclude that the reform caused an overall increase of IRPEF liability on Italian households but there is less agreement in detecting the most and least affected group dividing the sample by area

of residence and the occupation of the household head. Moreover the results are at times numerically quite different and an abundance of numbers tends to obscure the main picture of the distributional effects of the reforms.

3.2 The density estimation technique

The nonparametric density estimation method used here is derived from a generalization of the adaptive kernel density estimator to take into account sampling weights. The adaptive kernel is obtained in a two-stage procedure. In the first stage, a pilot estimate with fixed-bandwidth is performed to get a rough idea of the density along the data range. In the second stage, the fixed bandwidth parameter is replaced by a function of the fixed-bandwidth and of the pilot density estimation such that the bandwidth is larger where the pilot density estimate is smaller (i.e. the data are more sparse), and is smaller where density estimation is larger (i.e. the data are more concentrated). Such an estimation technique is particularly suitable for thick-tailed distributions such as income densities, since a variable bandwidth tends to dampen fluctuations in the tails and increase precision in the bulk of the distribution. In detail the procedure is as follows (Abramson, 1982; Silverman, 1986). Let $\mathbf{X} = \{X_i, i = 1, ..., N\}$ be a univariate random sample from an unknown distribution f. The pilot estimate $\widetilde{f}_N(x)$ to be estimated for the first stage is:

$$\widetilde{f}_N(x) = \frac{1}{Nh_N} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h_N}\right) \tag{3.1}$$

where h_N is a fixed-bandwidth, K is the kernel function and $\widetilde{f}_N(X_i) > 0$ for all i.

The second stage begins with the estimation of a local bandwidth factor λ_i :

$$\lambda_i = \left(\frac{\widetilde{f}_N(X_i)}{g} \right)^{\alpha} \tag{3.2}$$

where $0 \leq \alpha \leq 1$, and g is the geometric mean of $\widetilde{f}_N(X_i)$,:

$$g = \Pi_{i=1}^{N} \left(\widetilde{f}_N(X_i) \right)^{1/N} \tag{3.3}$$

The final estimation $\widehat{f}_N$ is given by:

$$\widehat{f}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h_N \lambda_i} K \left(\frac{x - X_i}{h_N \lambda_i} \right) \tag{3.4}$$

The adaptive kernel can then be modified to take into account the sampling weights, θ_i, normalized to sum to N. Every observation is then weighted by $\frac{\theta_i}{N}$ and not by $\frac{1}{N}$ implying that (3.1) becomes:

$$\widehat{f}(x) = \frac{1}{N h_N} \sum_{i=1}^{N} \theta_i K \left(\frac{x - X_i}{h_N} \right) \tag{3.5}$$

then (3.4) becomes:

$$\widehat{f}_a(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{\theta_i}{h_N \lambda_i} K \left(\frac{x - X_i}{h_N \lambda_i} \right) \tag{3.6}$$

The choice of the bandwidth parameter h_N is a delicate issue: a larger than optimal h_N will oversmooth the density increasing the bias, a smaller than optimal

h_N will increase the variance of the estimate reducing the bias. Moreover, it was proved that the optimal bandwidth, defined as the parameter that minimizes the mean integrated square error, depends on the unknown density being estimated (Parzen, 1962). Among the various optimal bandwidth parameter proposed in the literature[1], I chose the Silverman's rule-of-thumb bandwidth:

$$h_N = 0.9A(n)^{(1/5)} \tag{3.7}$$

where $A = min\{\text{standard deviation, interquantile range}/1.34\}$ is an adaptive estimate of spread. This bandwidth parameter was proposed by Silverman (1986, p. 48) as a parameter that copes well with a wide range of densities and is trivial to compute. However, other bandwidth have also been used and some results will be presented for bandwidths that are equal to a given proportion of the optimal bandwidth in (3.7).

Making the local bandwidth factor dependent on a power of the pilot density gives flexibility in the design of the method: the larger the power α, the more sensitive the method will be to variations in the pilot density. Following Silverman (1986, p.48)'s suggestion, it was set at $\alpha = 1/2$.

As for the choice of the kernel function, K, the Epanechnikov kernel was used since it maximizes efficiency (see among others Silverman, 1986, Section 3.3.2).

Finally, in some relevant cases, to assess the reliability of density estimates the 90% confidence bands are computed as 1.645 standard errors around $\widehat{f_\alpha}(x)$. Standard

[1]For an introductory description of the choice of the smoothing parameter, see Silverman (1986, Section 3.4) or Bowman and Azzalini (1997, Section 2.4)

errors come from the following expression for the variance of $\widehat{f_\alpha}(x)$ given in Burkhauser et al. (1999, p. 261):

$$V(\widehat{f_\alpha}(x)) = \left(\sum_{i=1}^{N} \frac{\theta_i^2}{N^2}\right) \frac{f_\alpha(x)}{h_N \lambda_i} \int \{K(z)\}^2 dz \tag{3.8}$$

3.3 The non-parametric density estimation and the MSM combined using counterfactuals

The combination of microsimulation and nonparametric density estimation here is undertaken comparing the existing 1998 IRPEF system with actual 1991 IRPEF system, weighted by CPI.

In the first stage, using the disposable income data contained in the N-dimensional 1998 SHIW sample (Y_{A98j}, $j = 1, ..., N$) and the 1998 IRPEF legislation (β_{98j}), the MSM is used to obtain the BT income (Y_{B98j}). In the second stage, the counterfactual estimation is performed starting from BT income (Y_{B98j}, for $j = 1, ..., N$), simulating 1991 IRPEF system (β_{91j}). The density estimation of disposable income under actual 1998 legislation ($f(Y_{A98})$) and counterfactual 1991 legislation ($f^c(Y_{A91})$) are then estimated and compared. The counterfactual distribution can be described as the "distribution of income that would have prevailed if 1998 IRPEF had been replaced by the 1991 IRPEF"[2]. It would be more precise to say that $f^c(Y_{A91})$ is the density that would have prevailed in 1998 if personal taxation had been replaced by

[2]Of course, given the assumptions of the MSM, if β_{98j} was applied in the second stage it would give Y_{A98}.

1991 IRPEF and each income recipient had obtained exactly the same income, before personal taxation"[3].

Given the focus on household welfare an equivalence scale was adopted for individual incomes, assuming equal distribution among members of household income. Due to the impossibility of obtaining a unique equivalence scale (see Cowell and Mercader-Prats (1997) and Blundell and Lewbel (1991)), the normal practice is to present results with different equivalence scales to show the sensitivity of results to different hypothesis, or to use a conventional equivalence scale. In this chapter, the Italian Poverty Commission equivalence scale, derived from the Engel methodology, was conventionally adopted. With this approach the issue of zero expenditures for the estimation of Engel curves (Pudney, 1990) is overlooked: the elasticity of total consumption on family magnitudes is estimated by a weighted regression with the proportion of food expenditure on household expenditure (c_f) as dependent variable, and the log of total household expenditure (c) and the log of the number of the members of the household (m) as regressors (De Santis, 1998):

$$c_f = \gamma_0 + \gamma_1 \ln c + \gamma_2 \ln m + u \tag{3.9}$$

The elasticity estimate is obtained as $\varepsilon = (-\gamma_2/\gamma_1)$ and the equivalent income of

[3]It is not claimed that this simulation is fully suitable to compare 1998 and 1991 Italian personal income taxation because I do not estimate behavioral response to taxation that possibly played a role in changing distribution of income in the 7-year-period considered. I would rather say that 1998 personal taxation is compared with a simulated one, which is equal to the one in place in Italy in 1991. The year 1991 was considered as the comparison year since 1992 is regarded as the year before the "turning point" of Italian public finance management. The year 1991 is the last year before the financial and currency crisis and that prompted the recovery process.

each member of household h can be estimated as:

$$X_h = \frac{Y_h}{m^{\varepsilon}} \tag{3.10}$$

where Y_h is the sum of all incomes in household h. From the regression performed on 1998 SHIW data, an elasticity equal to 0.757 was obtained.

Figure 3-1 presents density estimations on the whole sample for BT, actual and counterfactual AT income distributions with three different bandwidths: h equal to the Silverman's optimal bandwidth (3.7), $h_2 = 0.75h$ and $h_3 = 2h$. This figure produces a clear picture of the concentration effect induced by personal income taxation: the 1998 BT income density presents a lower maximum and a higher mode than AT income. The AT density presents a thinner upper tail than the BT income density, showing that the 1998 IRPEF system is very effective in reducing the overall density at medium-high incomes. AT income density is clearly bimodal if h_N is equal to or smaller than the Silverman's bandwidth (3.7). This result add something to findings of Pittau and Zelli (2001) and D'Ambrosio (2001): the bimodality of equivalent AT income is due to or at least magnified by personal income taxation.

The comparisons between the actual and the counterfactual AT income distributions show that the counterfactual AT density reaches lower maxima than the actual 1998 AT income density, although the location of the modes do not widely change.

The last panel of Figure 3-1 depicts the difference between counterfactual and actual AT distributions for the three bandwidths considered. It shows that density of incomes at income levels below about Lit 18 millions (approximately €9,000) was

higher in 1998 than with the counterfactual 1991 tax system. Given this pattern we can reasonably expect that inequality indices will show a decreasing trend from BT income to counterfactual AT income and to actual AT income[4]. Another relevant issue is clearly evident from the kernel density estimation: the personal income tax-benefit system is not effective in tackling poverty. In fact the BT income density at zero equivalent income is different from zero, showing that there are households with zero income. However, BT and AT incomes do not differ at zero equivalent income level. This is mainly due to the fact that IRPEF does not allow tax credits in case of negative tax or of tax allowances larger than BT income[5]

act AT	actual 1998 AT income density, $f(y_{A98})$
ctf AT	countefactual AT income density, $f^c(y_{A91})$
act BT	actual 1998 BT income density, $f(y_{B98})$

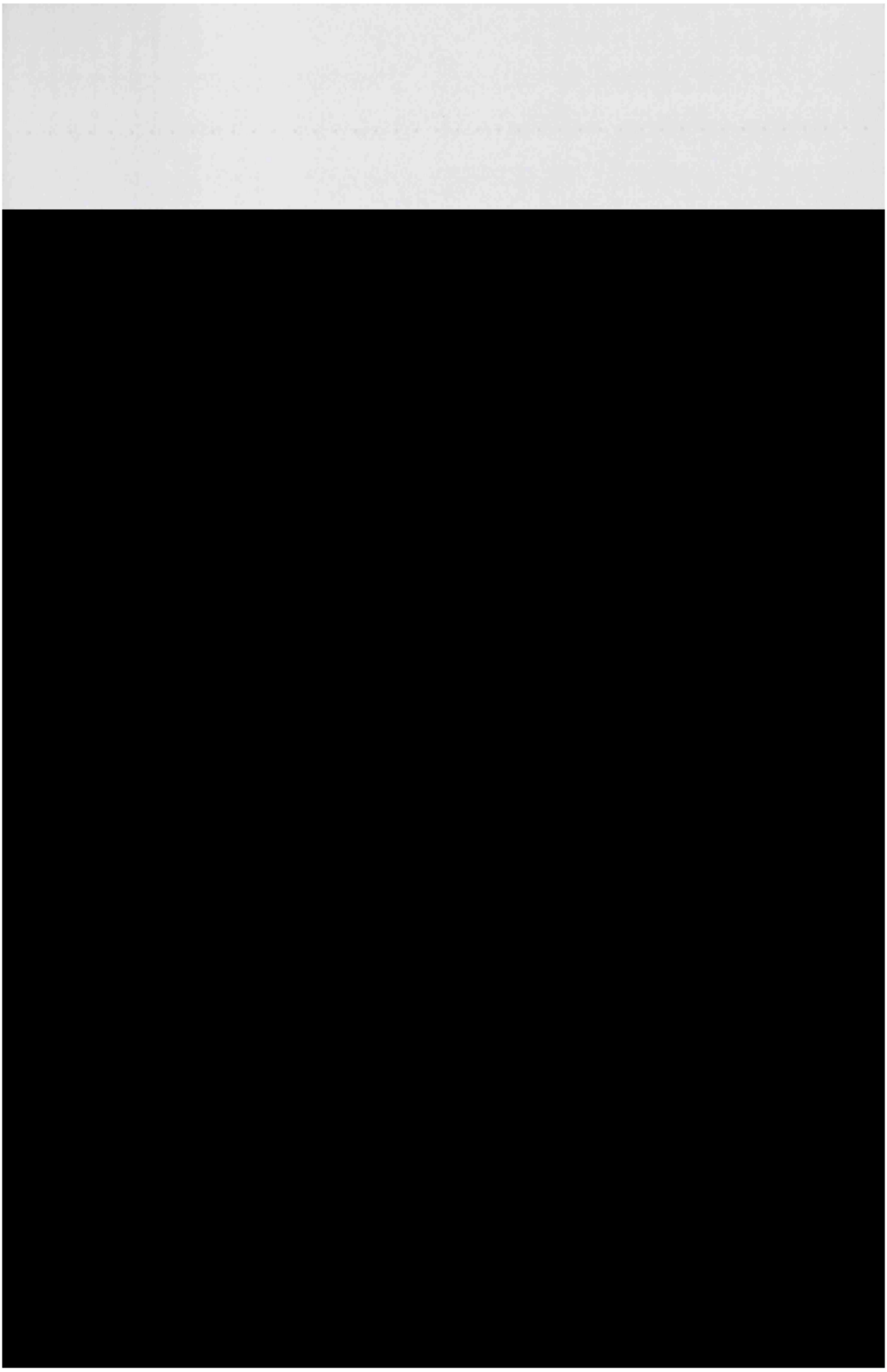
Table 3.1: Abbreviations used in tables and figures

3.3.1 Decomposing the sample

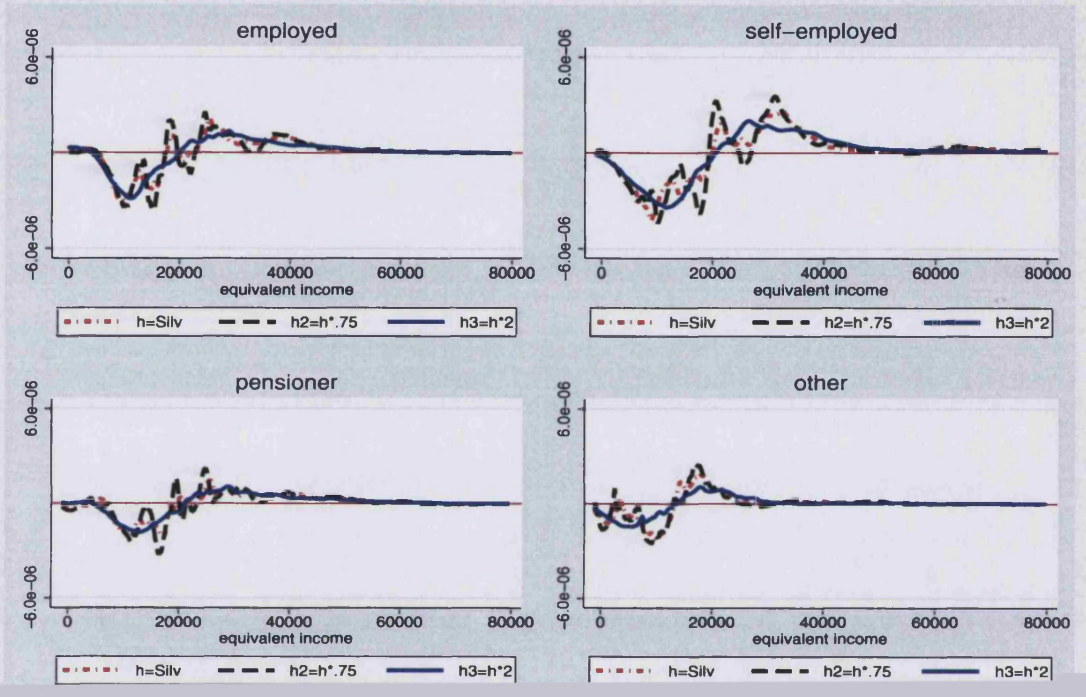
Other interesting observations can be provided breaking down the sample by occupation of the householder. Figure 3-2 depicts the actual and counterfactual AT density estimates with a bandwidth as in (3.7) using continuous and dashed lines, respectively. Thinner lines are the 90% confidence bands of the two density estimates. Figure 3-2 shows that bimodality of AT incomes is very clear in employed household and is

[4]Indeed such a guess is correct as shown in Appendix B using Lorenz curves.

[5]The analysis of Figure 3-1 is performed without confidence bands mainly for clarity reasons. It should however be noted that the confidence regions greatly overlaps for the two AT densities. BT density is instead significantly different from AT incomes densities. Finally, confidence bands also show that the density is significantly different from zero at zero income. Figure 3-1 inclusive of confidence bands can be obtained on request from the author.



the difference is at first negative and becomes positive at higher levels of income for all sub-samples, although it obviously presents more fluctuations the smaller is the bandwidth.



An analysis of losers and gainers is performed analyzing the joint distribution of X and Y, where X is the actual 1998 BT income and Y is the difference between counterfactual and actual 1998 AT income relative to 1998 BT income, i.e. the relative loss of income caused by the 1998 personal taxation system. Given the N pairs $\{x_i, y_i, i = 1, ..., N\}$, the relationship between Y and X can be estimated as:

$$y_i = m(x_i) + \epsilon_i \tag{3.11}$$

$$E(Y|X = x) = m(x) \tag{3.12}$$

where ϵ_i is a random error. The estimation of $m(x)$ is then performed nonparametrically using the Nadaraya-Watson estimator. As in the kernel density estimation the bandwidth h_N determines the degree of smoothing of $\widehat{m}$. As the h_N goes to zero, $\widehat{m}(X_i)$ converges to Y_i, i.e. we obtain an interpolation of the data. On the other hand, if h_N goes to infinity the estimator is a constant function that assigns the sample mean of Y to each x (see, among others Härdle et al., 2004). Choosing the smoothing parameter for the covariate vector is again a crucial problem: as before I started by using Silverman's rule-of-thumb bandwidth (3.7) and analyzed the sensitivity of the estimates to different bandwidths.

Figure 3-5 shows the results on the whole sample of a nonparametric regression of the relative loss caused by 1998 tax system on 1998 BT income for different values of the bandwidth, i.e. a nonparametric regression of Y over X as defined above. It shows that those in the lowest part of the income range suffered the highest relative loss: in some cases they had to pay up 6-8% more with 1998 personal income reform than

than others. In Figure 3-5 the nonparametric regression by occupation of the householder are reported with a bandwidth chosen as in (3.7) and 90% confidence bands. Relative loss was increasing from zero level for employed, but it was higher than average for poor pensioner households and the residual ("other") group and for some of the self-employed households below 20 millions Lit.

Estimates are reasonably reliable with the only exception of higher incomes in the residual group due to the small sample size. These results show that the probability of experiencing poverty is higher after the reform than before it especially for households with a non-employed head. This result is likely to come from the fact that the "entry" marginal tax rate in 1998 (18.5%) was higher than that in 1991 (10%), and tax-allowances were increased more for working than for non-working tax-payers.

3.4 A revenue-neutral reform simulation

Since the 1998 IRPEF reform induced a relevant increase of revenue compared to the 1991 system, two different revenue-neutral simulations were performed. In the first of these the 1998 excess revenue is distributed equally to each individual in the population, in the second the excess revenue is distributed to each individual in proportion to her BT income.

It is of course difficult to assess how the increased tax revenue was employed, partly because other tax and welfare reforms were introduced in the same period, and partly because the increased tax revenue was not constrained to be used for any particular policy. However we can equally think of redistribution as happening in

cash or in kind or in reductions of other taxes liabilities.

A lump-sum redistribution of excess revenue, if compared to 1998 actual AT income, would induce a reduction of the density at a lower levels of income and an increase in the mode of the distribution basically due to a shif of the former distribution to the left (Figure 3-6). Since the transfer is per capita, it would benefit more larger households.

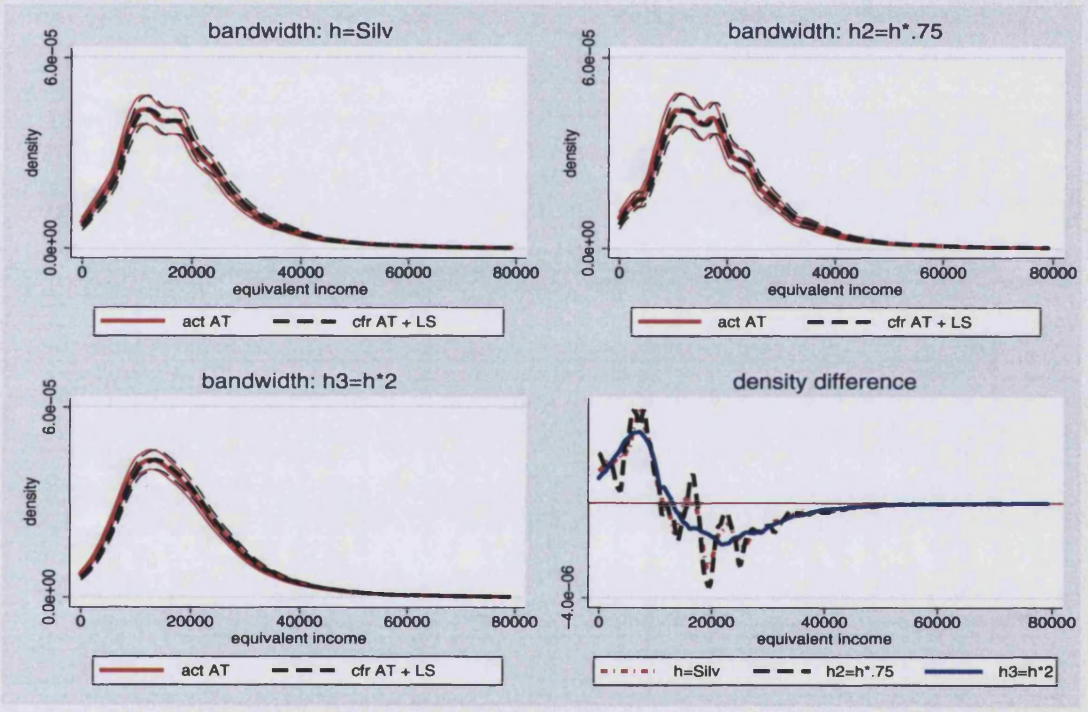


Figure 3-6: Revenue-neutral simulation: lump-sum redistribution; with 90% confidence bands; in Lit '000 (Lit 1936.27=€1)

A proportional redistribution have a slightly smaller effect on equivalent income distribution (Figure 3-7), showing that the excess revenue of 1998 IRPEF compared with 1991 IRPEF revenue was mainly due to a proportional (to BT income) increase of tax liability rather than to an equal increase of tax liability for all.

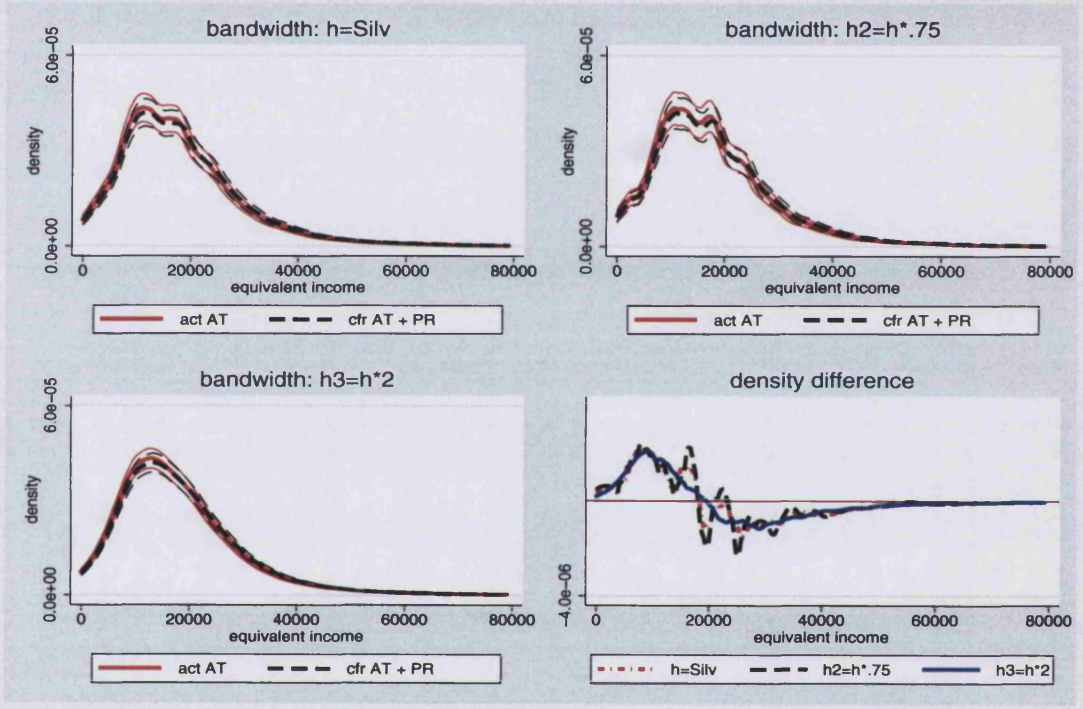


Figure 3-7: Revenue-neutral simulation: lump-sum redistribution: with 90% confi-

3.5 Decomposing the fiscal reform

The difference between the counterfactual and actual densities was decomposed[7] to investigate the overall importance of tax allowances for altering the shape of equivalent income distribution and inequality, as opposed to that of the bracket structure.

Using $\beta_d = 98$ to denote the set of tax allowances from gross tax liability in 1998 and $\beta_b = 98$ the tax bracket structure of 1998 and similarly for 1991, the difference between the two densities mentioned earlier can be represented in equation (3.9a) and decomposed as in (3.9b)-(3.9c) and (3.9d)-(3.9e).

$$f(Y_A; \beta_d = 91, \beta_b = 91) - f(Y_A; \beta_d = 98, \beta_b = 98) = \tag{3.9a}$$

$$[f(Y_A; \beta_d = 91, \beta_b = 91) - f(Y_A; \beta_d = 98, \beta_b = 91)] + \tag{3.9b}$$

$$[f(Y_A; \beta_d = 98, \beta_b = 91) - f(Y_A; \beta_d = 98, \beta_b = 98)] = \tag{3.9c}$$

$$[f(Y_A; \beta_d = 91, \beta_b = 91) - f(Y_A; \beta_d = 91, \beta_b = 98)] + \tag{3.9d}$$

$$[f(Y_A; \beta_d = 91, \beta_b = 98) - f(Y_A; \beta_d = 98, \beta_b = 98)] \tag{3.9e}$$

Line (3.9c) presents the difference between a counterfactual AT income density, which was obtained using the 1998 tax allowance system and the 1991 IRPEF brackets, and the actual 1998 AT income density. In the last part of this chapter two alternative scenarios are simulated, both on 1998 data. Scenario 1 is characterized by a tax allowance system equal to the one actually in use in 1991 but with an income

[7] A brief description of the 1998 IRPEF tax reform is provided in Section 3.7

bracket structure and tax rates like that in 1998 (i.e. the first part of (3.9e)). On the other hand, Scenario 2 considers the case in which the tax allowance system is like that in 1998 and the income brackets structure and tax rates as in 1991 (the first part of (3.9c)). Results are shown in Figures 3-8 and 3-9 together with actual 1998 AT density.

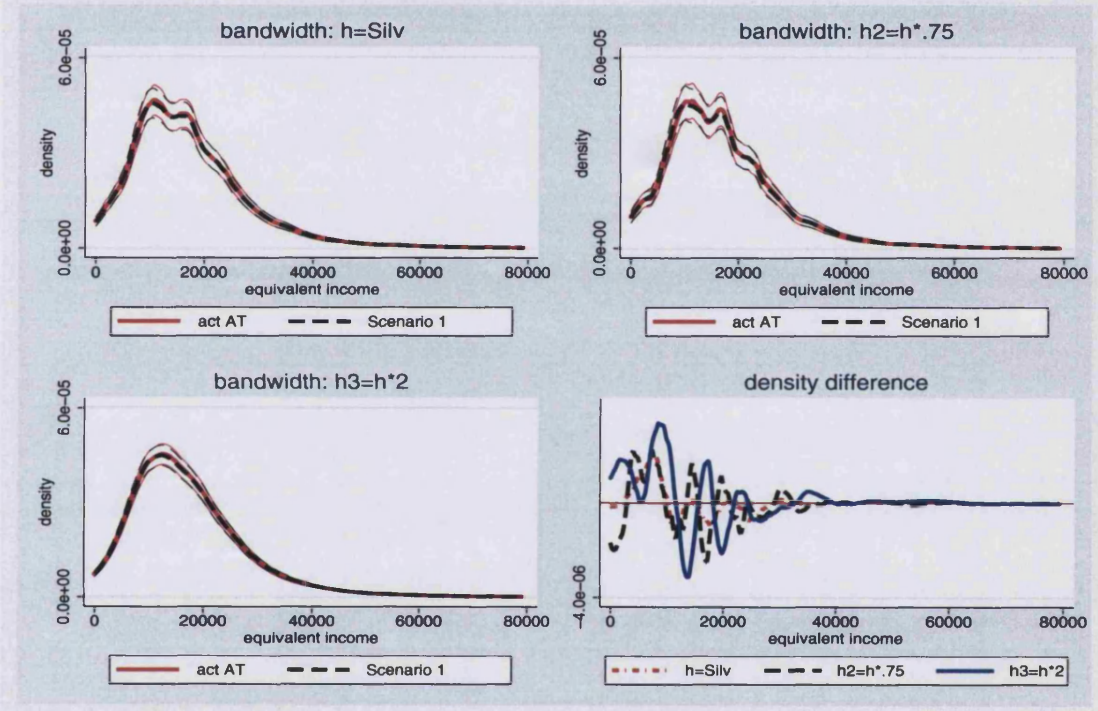


Figure 3-8: Scenario 1 vs. actual AT; with 90% confidence bands; in Lit '000 (Lit 1936.27=€1)

From the comparison of densities it is evident that the Scenario 1 density approaches the actual 1998 AT density more closely than Scenario 2 one. These results show that the new tax rates and brackets were more effective than the new tax allowances in modifying the distribution of income

These simulations have been analyzed also by using nonparametric regression. In Figure 3-10 the relative loss variable, Y is defined as the difference between Scenario

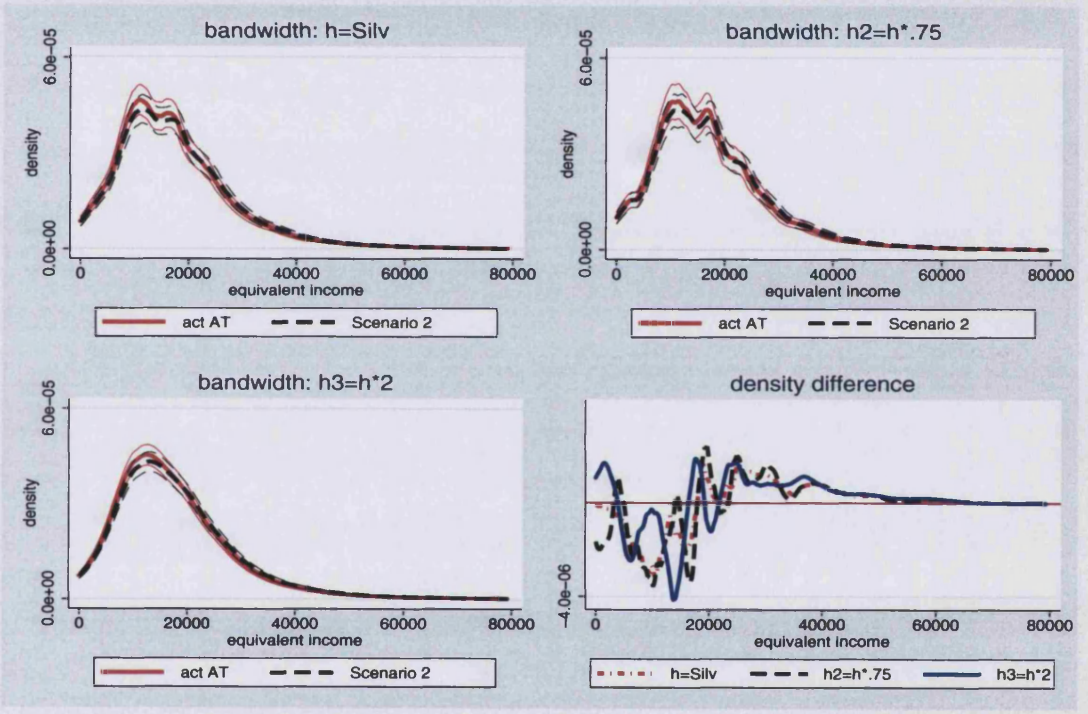


Figure 3.9: Scenario 2 vs. actual AT: with 90%

1 and actual AT income (i.e. the difference in line 3.9e) relative to actual BT income. Figure 3-10 shows that had tax allowances been set as in 1991 the average loss of households with equivalent incomes higher than 20 millions Lit would have been about zero, however, poorest income would have suffered significant losses because of tax allowances higher than in 1998.

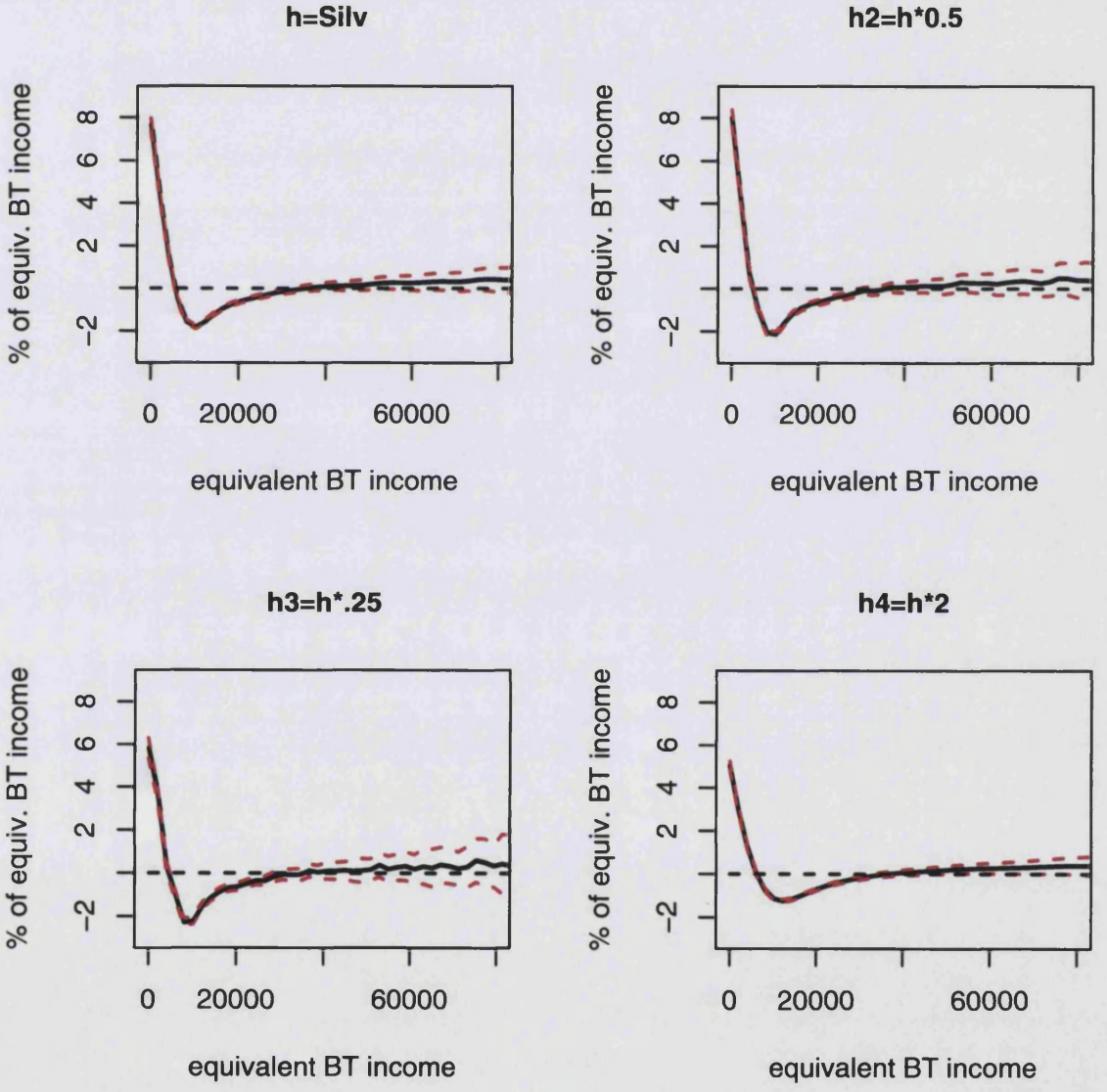


Figure 3-10: Losers and gainers: simulation D1, with different bandwidths and 90% confidence bands; in Lit '000 (Lit 1936.27=€1)

In Figure 3-11 the relative loss variable, Y is defined as the difference between

Scenario 2 and actual AT income (i.e. the difference in line 3.9c) relative to actual BT income. Figure 3-11 shows that had tax brackets and tax rates been set at the 1991 level, losses would have been more evenly spread across different levels of income and would approximately average 3%. The lower than average losses of equivalent incomes below 10 millions Lit are likely to be due to the effect of 1998 tax allowances.

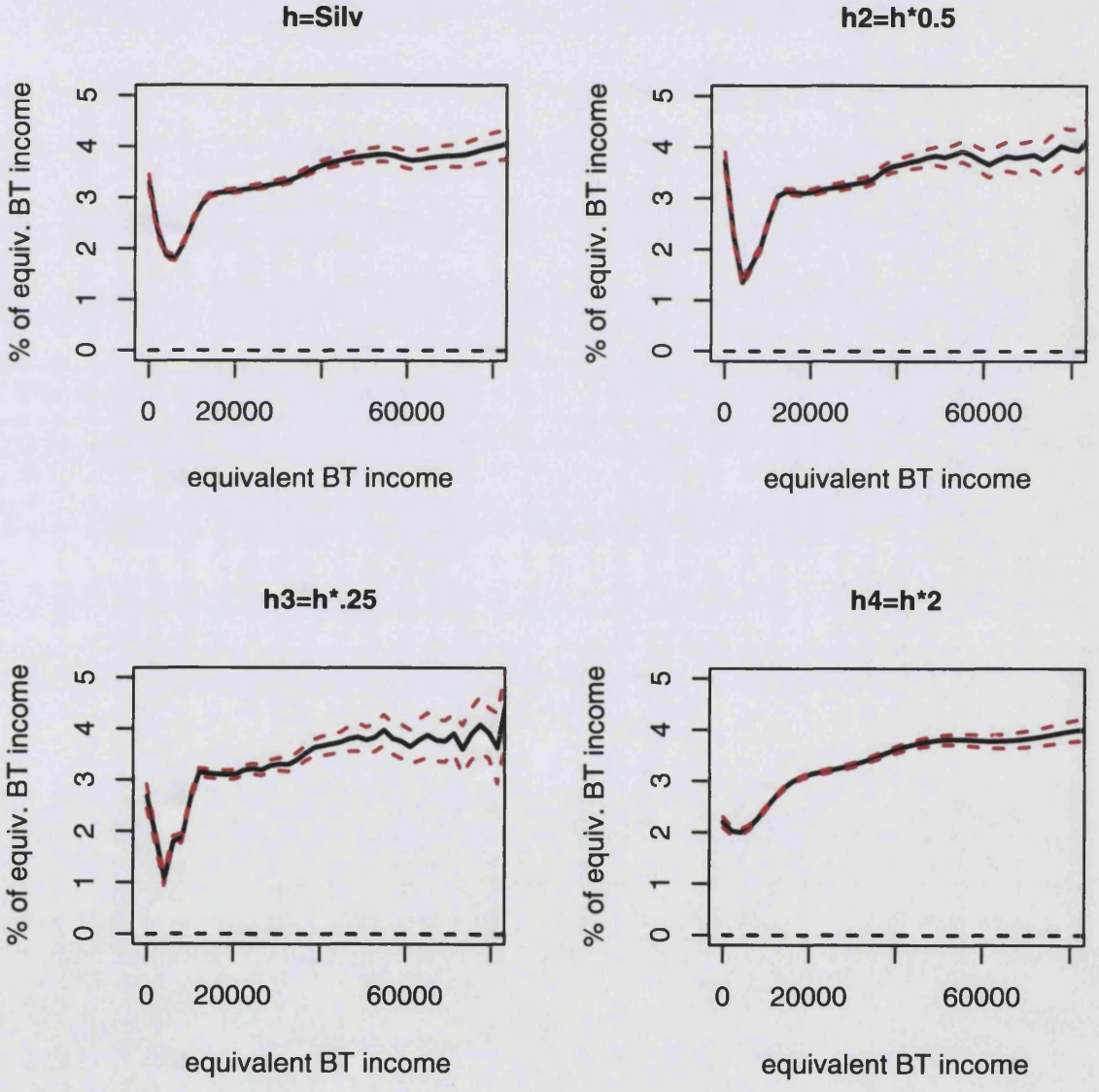


Figure 3-11: Losers and gainers: simulation D2, with different bandwidths and 90% confidence bands; in Lit '000 (Lit 1936.27=€1)

3.6 Conclusions

This chapter proposed a methodology consisting in a combination of microsimulation and non-parametric estimation for tax-benefit policy analysis. It was shown that this combination can increase the understanding and access of microsimulation results and provide useful insights regarding income distribution profiles.

Simulating AT income for 1998, the higher concentration induced by the taxation system was illustrated. The combination of microsimulation and nonparametric density estimation allowed us to show that the 1998 IRPEF system increased the concentration around the mode with respect to the updated 1991 IRPEF system. The higher concentration was obtained with a movement of part of the density mass from upper to lower levels of equivalent income, resulting in an overall decrease of inequality. Nonparametric density estimation also showed that personal income taxation has an important role for the emergence of bimodality in the Italian AT equivalent income density.

Decomposing the sample into different subgroups, it was shown that some households have been affected differently from others. While poor employed households had been basically unaffected by the reform, other households, namely those headed by a non-working head, suffered major losses increasing their probability of experiencing poverty. In fact, for these groups of households the increased tax allowances were more than offset by the increased tax rate of the first bracket.

Finally, it was shown that the increased tax liability was roughly proportionally spread across tax payers and that changes in income brackets were more effective in

changing the overall distribution of equivalent incomes than changes in tax allowances.

3.7 Appendix A: 1998 IRPEF vs. counterfactual 1991 IRPEF

The following provides detail about the difference between the 1991 and the 1998 legislations. The counterfactual density has been simulated updating 1991 IRPEF tax brackets and tax allowances to 1998 price. Income brackets were reduced from seven to five (Table 3.2).

1991 IRPEF (Lit '000)	**Counterfactual** (Lit '000)	**tax rate** (%)	**1998 IRPEF** (Lit '000)	**tax rate** (%)
0-6,800	0-8,786	10	15,000	18.5
6,800-13,500	8,876-17,442	22	15,000-30,000	26.5
13,500-33,700	17,442-43,540	26	30,000-60,000	33.5
33,700-67,600	43,540-87,339	33	60,000-135,000	39.5
67,600-168,800	87,339-218,090	40	over 135,000	45.5
168,800-337,700	218,090-436,308	45		
over 337,700	over 436,308	50		

Table 3.2: Actual 1998 and counterfactual structure of IRPEF tax brackets (Lit 1936.27=€1)

For a fiscally dependent spouse, tax allowance depends on BT income (Table 3.3).

1991 IRPEF	**tax all.** (Lit '000)	**Counter. tax all.** (Lit '000)	**1998 IRPEF** (Lit '000)	**tax all.** (Lit '000)
any income	675	872.1	0-30,000	1,057.552
			30,000-60,000	951.552
			60,000-100,000	889.552
			over 100,000	817.552

Table 3.3: Actual 1998 and counterfactual tax allowances for dependent spouse (Lit 1936.27=€1)

Tax allowances for dependent children and other dependent relatives do not depend on BT income (Table 3.4).

	1991 tax all. (Lit '000)	**counterf tax all.** (Lit '000)	**1998 tax all.** (Lit '000)
depend. child	78	101	336
other dep. relative	108	139.5	336

Table 3.4: Actual 1998 and counterfactual tax allowances for dependent children and other relatives (Lit 1936.27=€1)

Tax allowances for employment and self-employment income are different and depend on income (Table 3.5).

1991 IRPEF (Lit '000)	**Fisc.all.** (Lit '000)	**count. brack.** (Lit '000)	**t.a.** (Lit '000)	**1998 brack.** (Lit '000)	**t.a.** (Lit '000)
12,400	851	16,021	1,099	0-9,100	1,680
12,400-12,659	750**	16,021-16,355	969	9,100-9,300	1,600
over 12,659	648	16,355	837	9,300-15,000	1,500
				15,000-15,300	1,350
				15,300-15,600	1,250
				15,600-15,900	1,150
				15,900-30,000	1,050
				30,000-40,000	950
				40,000-50,000	850
				50,000-60,000	750
				60,000-60,300	650
				60,300-70,000	550
				70,000-80,000	450
				80,000-90,000	350
				90,000-90,400	250
				90,400-100,000.	150
				over 100,000	100

Table 3.5: Actual 1998 and counterfactual tax allowances for employment income. * is an average. The actual tax allowance (in Lit '000) was computed as 851-$[(y_B - 12,400) \times 0.78]$, where y_B is BT income.

If the individual receives only a pension income which is less than Lit 18 million, and he is not owner of other building apart from the principal dwelling, there is an

1991 inc. brack (Lit '000)	**tax.all.** (Lit '000)	**Counterf.** (Lit '000)	**t.a.** (Lit '000)	**New IRPEF** (Lit '000)	**t. a.** (Lit '000)
0-6,800	168	0-8,786	217	0-9,100	700
6,800-7,000	90*	8,876-9,044	116	9,100-9,300	600
				9,300-9,600	500
				9,600-9,900	400
				9,900-15,000	300
				15,000-30,000	200
				30,000-60,000	100

Table 3.6: Self-employment tax allowances. * is an average. The actual tax allowance (in Lit '000) was computed as 168-$[(y_B - 6,800) \times 0.78]$, where y_B is BT income.

additional tax allowance equal to Lit 70,000 in 1998, while he did not receive anything in 1991.

3.8 Appendix B: inequality analysis of BT, actual and counterfactual AT incomes

A traditional analysis of inequality of BT, actual and counterfactual AT incomes can be performed using Lorenz curves.

Provided that the density function is non zero throughout the range $[Y_1, Y_N]$, where N is the number of observation, and $Y_1 < Y_N$, then for each $p \in (0,1)$, there is just one income level y, which satisfies $p = F(y)$, the income of the first $100p$ percent of income recipients is $N\int_0^y yf(y)dy$ and the total income is $N\int_0^\infty yf(y)dy = N\mu$, where μ is the mean income (Lambert, 1993). Hence, using $\widehat{f}(y)$ for the density estimation of income, a finite-sample Lorenz curve $L(p)$ is defined by

$$p = F(y) \Rightarrow L(p) = \frac{1}{N}\sum_{j=1}^{J}\frac{X_i\widehat{f}(X_i)}{N}, 0 < p < 1$$

In Table 3.7 the Lorenz curve for BT income, counterfactual AT income and actual AT income is provided. Lorenz curves do not cross, i.e. 1998 AT income Lorenz-dominate (i.e. is more equally distributed of) the counterfactual AT income, which Lorenz-dominates the 1998 BT income. This results will be confirmed by a large class of inequality indices satisfying axioms of anonymity, mean independence and the transfer principle (Cowell, 1995). In Table 3.8 some inequality indices are reported: they show, coherently with what stated before, that equality is increased by 1991 IRPEF and even more by 1998 IRPEF.

Pop. share	BT income.	count. AT y	AT y 1998
1/10	0.0117	0.0161	0.0176
2/10	0.0396	0.0560	0.0611
3/10	0.0811	0.1146	0.1247
4/10	0.1363	0.1920	0.2093
5/10	0.2066	0.2880	0.3123
6/10	0.2936	0.4015	0.4315
7/10	0.3997	0.5282	0.5615
8/10	0.5308	0.6631	0.6939
9/10	0.6986	0.8086	0.8313
10/10	1	1	1

Table 3.7: Lorenz curve for different type income

	AT income	count AT y	AT income
rel. mean deviation	0.306	0.288	0.285
coeff. of variation	0.923	0.826	0.818
Gini index	0.429	0.404	0.401
Theil entropy meas.	0.325	0.283	0.279
Theil mean log dev.	0.359	0.315	0.312

Table 3.8: Some inequality indices for different type of income

Chapter 4

Assessing the reliability of MSMs using the bootstrap

Static MSMs are widespread techniques for analyzing the "morning after" effects of government policy on welfare and distribution of income. In recent years research on MSMs has mainly focused on finding the most efficient way to introduce behavioral relations, to improve computer routines, to calibrate the sample to known totals or develop a general equilibrium model of the economy starting from microeconomic relationships (see Chapter 1).

However, little attention has been paid to assessing the reliability of the output of frequently quoted tax-benefit MSMs, even in the simplest case of MSMs without behavioral response. One of the few exceptions remains Pudney and Sutherland (1994), which investigates the asymptotic sampling properties of a set of typical simulation results using a UK tax-benefit model on FES data. They start with the consideration that even the simplest static deterministic MSM without behavioral response pro-

duces an output that is affected by stochastic error, as any survey-based model. This error could come from a variety of sources such as sampling error, misreporting or underreporting behavior, programming mistakes, and it might increase variability in the model estimation. The importance of the stochastic error has to be assessed, and estimates from an MSM should be accompanied with standard errors or confidence intervals. This remark, which might seem obvious to many, is indeed well placed since it is still common to see MSM results quoted without any standard error or confidence interval.

> Although there are formidable technical problems involved in the derivation of confidence intervals, there is no reason why the sampling properties of simulation results should not receive as much attention from econometricians as do the properties of estimators of behavioural relationships (Pudney and Sutherland, 1994, p. 328).

The method used by Pudney and Sutherland (1994) is to assume that the underlying population is infinite so that finite population correction can be avoided and asymptotic theory can be applied. The income vector used is that of net equivalent income, which is a transformation of the sum of all net incomes of all members of a family that takes into account economies of scale within the family. Net incomes are obtained by deducting income tax and national insurance contributions. Then some common statistics for the analysis of MSM output are produced, properly taking into consideration the sampling weights contained in the FES data set. The weighted statistics and their corresponding variance are then used to compute confidence in-

tervals.

Section 4.1 presents the approach used here. This approach differs from Pudney and Sutherland (1994) since no infinite population assumption is introduced. Instead of introducing finite sample corrections a simulation based method, namely the bootstrap, is used to approximate the true distribution and estimate confidence intervals. The choice of the bootstrap is motivated by the finding that in small samples it performs significantly better and in other cases no worse than asymptotic theory. The consequences for asymptotic confidence interval estimation of using a nonlinear transformation to equivalize household incomes and to perform simulations will be discussed. Section 4.2 will briefly present the data, MSM and hypotheses adopted in this chapter. Section 4.3 will discuss some results obtained using a tax-benefit model on Italian household data. Section 4.4 concludes.

4.1 Methodology for assessing the sampling error

Supposing we have a family of tests, with the same level, α, for testing a set of hypotheses about a scalar parameter $\theta \in \mathbb{R}$, it is possible to use them for constructing a confidence interval for the parameter of interest. A confidence interval is defined as the interval on the real line which encompasses all values of θ for which the hypothesis that $\theta = \theta_0$ is not rejected by the appropriate test in the family, where θ_0 is a certain parameter value. The coverage probability of the interval is defined as the probability that the true parameter will lie in the computed interval. Confidence intervals can be either exact or approximate. The confidence intervals are exact if

the finite-sample distribution of the test statistic is known, while they can only be asymptotic confidence intervals if the distribution of the test statistic is known only asymptotically. In the latter case, the smaller is the sample the more inaccurate the coverage of the confidence interval tends to be, i.e. the interval will not "cover" the true parameter value with the specified probability.

Let $\tau(\mathbf{y}, \theta_0)$ be the test statistic for testing the null hypothesis that $\theta = \theta_0$, where $\mathbf{y}$ is the vector of realizations of a random variable Y from an unknown distribution that is used to compute the particular realization of the statistic, $\widehat{\theta}$. Let $c^U_{\alpha/2}$ denote the value on the real line such that the probability under the null that the test statistic is larger than $c^U_{\alpha/2}$ is equal $\alpha/2$; correspondingly $c^L_{\alpha/2}$ denotes the value on the real line such that the probability under the null that the test statistic is smaller than $c^L_{\alpha/2}$ is equal $\alpha/2$. Hence, by definition

$$\Pr_{\theta_0}(c^L_{\alpha/2} \leq \tau(\mathbf{y}, \theta_0) \leq c^U_{\alpha/2}) = 1 - \alpha \tag{4.1}$$

It is then possible to obtain the limit of the confidence interval of level α, by inverting the test statistic $\tau(\mathbf{y}, \theta_0)$. If the distribution of $\tau(\mathbf{y}, \theta_0)$ is symmetric and centered around zero then $c^U_{\alpha/2} = c^L_{\alpha/2} \equiv c_{\alpha/2}$, and the confidence interval in (4.1) can be rewritten as:

$$\Pr_{\theta_0}(|\tau(\mathbf{y}, \theta_0)| \leq c_{\alpha/2}) = 1 - \alpha \tag{4.2}$$

If the test statistic is the classical t-ratio, $\tau = (\widehat{\theta} - \theta_0)/s_{\widehat{\theta}}$, the confidence interval

can also be written as:

$$\theta = \widehat{\theta} \pm c_{\alpha/2} \times s_{\widehat{\theta}} \tag{4.3}$$

where $s_{\widehat{\theta}}$ is the standard error of $\widehat{\theta}$. This is the confidence interval computed by Pudney and Sutherland (1994), where all statistics are weighted using the sampling weights provided. Assuming that the underlying population is infinite and that the the random variable Y is extracted from a finite variance distribution, the classical t-ratio will be distributed as a standard normal and $c_{\alpha/2}$ can be obtained by standard normal statistical tables (for the analysis of the t-ratio statistic of a random sample from an infinite variance distribution see Chapter 7).

Such a confidence interval is rather conservative as a measure for evaluating the reliability of estimates. It only assesses the magnitude of the sampling error and avoids considering the simulation, misreporting and underreporting errors and, where relevant, the approximation errors due to the updating of the sample to different years or external data sources. The asymptotic confidence intervals can also be more conservative than the true ones if the asymptotic approximation tends to over-reject in small samples.

An alternative way to construct confidence intervals is to use simulation-based methods such as the bootstrap. For simplicity, let us start by considering the classical t-ratio statistic $\widehat{t}(\theta_0) \equiv \tau(\mathbf{y}, \theta_0) = (\widehat{\theta} - \theta)/s_{\widehat{\theta}}$, where $s_{\widehat{\theta}}$ is the standard error computed using the data $\mathbf{y}$ and τ is an asymptotically pivotal statistic[1]. First $\widehat{\theta}$ and $s_{\widehat{\theta}}$ are

[1]A test statistic τ is pivotal for a given model if, for each sample size, its distribution is independent of the DGP that generates the data from which τ is calculated. The statistic τ is asymptotically pivotal for a given model if its asymptotic distribution exists for all possible DGP and it is independent of the DGP that generate the data (Davidson and MacKinnon, 2001). The

computed using the original data $\mathbf{y}$, then, through random sampling with replacement from the original data set, B bootstrap samples, $\mathbf{y}_j^*, j = 1, ..., B$ of the same size as $\mathbf{y}$, need to be generated. For each of these samples, an estimate θ_j^* and its standard error s_j^*, are computed in the same way $\widehat{\theta}$ and $s_{\widehat{\theta}}$ were computed in the original sample. Then the bootstrap test statistic is computed

$$t_j^* \equiv \tau(\mathbf{y}_j^*, \widehat{\theta}) = \frac{\theta_j^* - \widehat{\theta}}{s_j^*} \tag{4.4}$$

This statistic tests the null hypothesis that $\theta = \widehat{\theta}$, because $\widehat{\theta}$ is the true value of θ for the bootstrap data generating process (DGP). However, if τ is an asymptotic pivot, the difference between $\widehat{\theta}$ and θ_0 is negligible (see for instance, Hall (1992b)).

The limits of the bootstrap confidence intervals will then depend on the quantiles of $\widehat{F}^*$, the empirical distribution function (EDF) of the t_j^*.

If $\widehat{F}^*$ were the cumulative distribution function (CDF) of a continuous distribution, the confidence interval could be expressed in terms of the quantiles of this distribution but since the distribution of t_j^* is always discrete in practice, a little more attention has to be paid. Assuming for simplicity that $\widehat{t}(\theta_0)$ is on the left side of the distribution, the bootstrap P value is

$$2\widehat{F}^*(\widehat{t}(\theta_0)) = \frac{2}{B}\sum_{j=1}^{B} I(t_j^* \leq \widehat{t}(\theta_0)) = \frac{2r(\theta_0)}{B}$$

method of computing confidence intervals based on an asymptotically pivotal statistic as $\widehat{t}(\theta_0)$ is called bootstrap-t or percentile-t. Other methods are also available such as the percentile and the BCa (Efron and Tibshirani, 1986), although asymptotically pivotal statistics has generally a better performance (Beran, 1988).

where $r(\theta_0)$ is the number of bootstrap t statistics that are less than or equal to $\widehat{t}_{\theta_0}$. Thus θ_0 belongs to the $1-\alpha$ confidence interval if and only if $2r(\theta_0)/B \geq \alpha$, that is, if $r(\theta_0) \geq \alpha B/2$. Since $r(\theta_0)$ is an integer, while $\alpha B/2$ in general is not, this inequality is equivalent to $r(\theta_0) \geq r_{\alpha/2}$, where $r_{\alpha/2}$ is the smallest integer not less than $\alpha B/2$.

After sorting t_j^* from smallest to largest, let $c_{\alpha/2}^*$ denote the entry in the sorted list indexed by $r_{\alpha/2}$, where $r_{\alpha/2}$, is the smallest integer not less than $\alpha B/2$. So, if $\widehat{t}(\theta_0) = c_{\alpha/2}^*$, the number of the t_j^* less than or equal to $\widehat{t}(\theta_0)$ is precisely $r_{\alpha/2}$, if $\widehat{t}(\theta_0)$ is any smaller than $c_{\alpha/2}^*$, then this number is strictly less than $r_{\alpha/2}$. Thus θ_u, the upper limit of the confidence interval of the estimate $\widehat{\theta}$, is defined implicitly by $\widehat{t}(\theta_u) = c_{\alpha/2}^*$, i.e.:

$$\theta_u = \widehat{\theta} - s_\theta \times c_{\alpha/2}^*$$

To obtain the lower limit of the confidence interval, let us start by assuming that $\widehat{t}(\theta_0)$ is on the upper part of the distribution and following the same reasoning as before it can be found that

$$\theta_l = \widehat{\theta} - s_\theta \times c_{1-(\alpha/2)}^*$$

where $c_{1-(\alpha/2)}^*$ is the entry indexed by $r_{1-(\alpha/2)}$ when t_j^* are ordered from smallest to largest.

The asymmetric equal-tail bootstrap confidence interval can then be written as

$$[\theta_l, \theta_u] = [\widehat{\theta} - s_\theta c_{1-(\alpha/2)}^*,\ \widehat{\theta} - s_\theta c_{\alpha/2}^*]$$

If the statistic is an exact pivot, then the probability that the true value of θ is greater than θ_u is exactly equal to $\alpha/2$ only if $\alpha(B+1)/2$ is an integer (Dufour and Kiviet, 1998; Davidson and MacKinnon, 2000).

The bootstrap as a method for deriving confidence intervals is used to overcome some of the limitations of the asymptotic confidence intervals. If c_α is the critical value for the asymptotic distribution of $\tau(\mathbf{y}, \theta_0)$, but not for the exact finite-sample distribution, then (4.1) is only approximately true. Although the bootstrap t-statistic (4.4), as most test statistics used in econometrics is asymptotically pivotal, it is likely to be non-pivotal in finite samples. This means that the distribution of the t-statistic depends on unknown parameters or other unknown features of the DGP. This leads to bootstrap p-values that are inaccurate in some way because of the difference between the true and the bootstrap DGP. However, theory suggests that the bootstrap test generally performs better in finite samples than tests based on asymptotic theory, in the sense that they commit errors that are of lower order in the sample size (see, among others, Hall (1992a), Davidson and MacKinnon (1999b)). There is also a growing body of literature that shows that bootstrap tests indeed performs better than asymptotic tests (see, for instance, Horowitz (1994), Davidson and MacKinnon (1999a), Godfrey (1998)) or at least as well as asymptotic tests (see, for inequality indices, Cowell and Flachaire (2002)). Of course, a smaller error in rejection probability (ERP) of a test is mirrored by a smaller error in coverage probability of the confidence interval.

Moreover the bootstrap allows one to compute asymmetric confidence intervals, i.e. it does not impose the distribution of $\tau(\mathbf{y}, \theta_0)$ to be symmetric. In finite samples

it is likely that $(\mathbf{y}, \theta_0)$ is not symmetric.

There is at least one more reason for using the bootstrap. It is used when the distribution of a statistic is unknown or is too difficult to derive analytically. For instance, if all the elements of the vector of random realizations $\mathbf{y}$ of the random variable Y is multiplied by the same scalar, $k \in \mathbb{R}_0$, the confidence interval in (4.3) does not change: because of the mean and variance properties with linear transformations, all elements in the left and right hand side of (4.3) are multiplied by k. However, if different elements of the vector $\mathbf{y}$ are multiplied by different scalars the confidence interval does change. In particular we cannot say a priori if the confidence interval as percentage of the estimated parameter is any different from (4.3) and, in general, there is no way to define the distribution of the t-ratio statistic in finite samples. We cannot say if the finite sample properties of the distribution of the t-ratio changes at all, hence we cannot evaluate how accurate is its asymptotic approximation. This point is of relevance for equivalent income data that are generally used in MSM literature. The equivalent income transformation is in fact highly non linear.

The main drawback of using the bootstrap is that it takes a long time to process. However, thanks to the rapid development in computing power this limitation is becoming less relevant.

4.2 The data set and the microsimulation model

The data set used in the present chapter is the 1998 Survey of Household Income and Wealth (SHIW) published by the Bank of Italy and described in Section 2.1. The

MSM used is TABEITA98, which assumes that household and family coincide. This assumption is due to the fact that questions are asked to households and there is no information to infer if there are more family units living under the same dwelling (see Chapter 2). In contrast to other countries, the SHIW one, which is the only suitable data set for Italian MSM, does not record the amount of taxes paid and benefits received. Hence, the first role of a MSM is to simulate the gross income before performing any other policy simulation. This feature implies that, differently to what performed in Pudney and Sutherland (1994), no simulation error (difference between the simulated tax and the actual one) can be properly assessed.

Before analyzing the MSM output, the income of each individual j belonging to household h, z_{jh}, has been equivalized using an equivalence scale. As mentioned in Section 3.3 there is no unique equivalence scale. The equivalence scale used here consists in scaling the household income z_h by b_h^ϵ, where ϵ is a constant $0 \leq \epsilon \leq 1$ and b_h is number of components in the household, i.e.:

$$y_{jh} = \frac{z_h}{b_h^\epsilon} \tag{4.5}$$

The Luxemburg Income Study's approach is to use $\epsilon = 0.5$. The Italian poverty commission's approach is to fix ϵ equal to the elasticity of total consumption on family size (De Santis, 1998)[2].

The sample was divided by main occupation of the household head, distinguishing between employed, self-employed, pensioners and "others", the residual group. The

[2]Applying the Italian Poverty Commission approach to 1998 SHIW we would have $\epsilon = 0.757$ (see Chapter 3).

household head is considered to be the one declared in the interview, regardless of contribution to household income. The largest sub-group is that of pensioners, followed by the employed (see Table 4.1). Of the "other" households most declare that they have a household head who is unemployed and actively looking for a job or who is a housewife (see Table 4.2).

Sample	Freq.	Percent
0. all	7112	100.00
1. employment	2715	38.17
2. self-empl	1049	14.75
3. pensioner	2760	38.41
4. other	588	8.27

Table 4.1: Proportion of households by occupation of the household head

Status	Freq.	Percent
Looking for first job	27	4.59
Unemployed	278	47.28
Housewife	241	40.99
Well-off	18	3.06
Student	24	4.08
Total	588	100.00

Table 4.2: Occupation of household head if not working

4.3 Non-linear transformation and confidence intervals: results of the analysis

Applying the model to net incomes to recover gross incomes means imposing a strongly non-linear transformation on the original data set. A non-linear transformation is also applied to the data set whenever attention is focused on the vector of equivalized incomes. This section shows the effect of these non-linear transformations

using the 1998 SHIW data set and the TABEITA98 model. The data are analyzed using some common summary statistics such as the mean, the 20th, 40th, 50th, 60th and 80th quantiles, the GE indices with parameter $\alpha = 0, 1, 2$ (for a definition of quantiles and GE indices used see Section 6.2.3 and 6.6, respectively). Since GE indices are only defined for positive incomes, households with non positive equivalent incomes are dropped from the analysis. Both the asymptotic and the bootstrap confidence intervals are computed as described in Section 4.1. The confidence intervals are set at 90% in both cases. Grossing-up and validation issues are not considered since they would significantly complicate the analysis without additional value for the purpose of this chapter.

In bootstrap re-sampling I have used the sampling procedure used by the Bank of Italy. Because of its stratified sampling procedure the SHIW data set includes sampling weights, which record the inverse of the probability of the households to be included in the sample. These sampling weights are defined for households, i.e. all members of the same household have the same weight. Sampling weights are normalized to sum to the number of households in the sample, provided each household is counted only once. No additional information about the sampling stages and strata is provided. Call these weights "original", p_o.

All asymptotic confidence intervals are weighted as in Pudney and Sutherland (1994). As for the bootstrap confidence intervals a different procedure is undertaken. An analysis of sampling weights in the 1998 SHIW data set shows that the number of household-units with positive household income is $N = 7112$ and the sampling-weight distribution is highly skewed to the right: it has mean 1, median at 0.69, a

minimum value of 0.08 and a maximum value of 10.11. In the first stage the original sampling weights, p_o, have been normalized dividing by their minimum and rounded to the nearest integer, yielding weight p_n. The smallest value of p_n is 1. Then the data set has been expanded using p_n. For instance, a household with p_n equal to 5 before expansion will be replicated 5 times after the expansion. This expansion makes the sample a close image of the underlying population, provided sampling weights have first been correctly estimated. From the expanded data set, a number B of bootstrap samples of dimension $N = 7112$ is built using random sampling with replacement. By construction, the new data set will present sampling weights uniformly equal to 1[3]. In this procedure households rather than individuals are re-sampled, as is the case in the SHIW data set. Whenever a household unit is sampled more than once it is renamed so that it is considered a completely different household from its clone.

For each bootstrap sample a t-ratio statistic is produced as in (4.4) for each summary statistic. They are then used to compute the EDF of the t_j^* and the bootstrap confidence intervals as described in Section 4.1. For the computation of the sampling variance of sub-sample means and quantiles I followed Pudney and Sutherland (1994); for the GE indices, I followed Cowell (1989).

In the experiments presented the number of bootstrap replications were fixed at $B = 999$.

The first point analyzed is the role of an equivalence scale on assessing the importance of sampling weights. For this procedure, no MSM is used. Hence, all incomes

[3]A similar algorithm to deal with sampling weights and the bootstrap was also taken in Fiorio (2004).

are net of taxes and social contributions. Since the effect of the equivalence scale on the confidence interval cannot be assessed analytically, both asymptotic and bootstrap confidence intervals are computed. Ideally we would like to find that there is no large difference between asymptotic and bootstrap confidence intervals so that the former can be used also in smaller samples, as they are quicker to compute. Secondly we would hope to find that confidence intervals as percentage of the estimated statistic do not change with the equivalence scale adopted or, otherwise, see how different equivalence scales affect reliability (i.e. modify confidence intervals). The equivalence scales considered are as in (4.5), with $\epsilon = 0, 0.5, 1$.

As for the mean, there is no significant difference between the confidence intervals as different economies of scale are considered. There is instead quite a large difference between asymptotic and bootstrap confidence intervals. Bootstrap confidence intervals tend to be larger than asymptotic ones and the difference is larger the smaller is the sample size, e.g. for the self-employed and the "other" households (Tables 4.3,4.4 and 4.5).

Sample	mean	AS lb	AS ub	BS lb	BS ub
0. all	53416	-0.90	0.90	-1.56	1.72
1. employment	54483	-1.04	1.04	-2.01	2.15
2. s-e. wrk	73545	-2.55	2.55	-4.44	5.22
3. pens.	47884	-1.48	1.48	-2.39	2.55
4. other	30908	-3.73	3.73	-7.01	8.69

Table 4.3: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the sample mean; $\epsilon = 0$

Almost the same conclusions can be reached using the percentiles as summary

Sample	mean	AS lb	AS ub	BS lb	BS ub
0. all	29854	-0.92	0.92	-1.47	1.71
1. employment	29322	-1.05	1.05	-1.87	1.97
2. s-e. wrk	39269	-2.80	2.80	-4.46	5.32
3. pens.	29418	-1.39	1.39	-2.03	2.38
4. other	16575	-3.82	3.82	-5.58	7.59

Table 4.4: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the sample mean; $\epsilon = 0.5$

Sample	mean	AS lb	AS ub	BS lb	BS ub
0. all	17486	-1.06	1.06	-1.52	1.76
1. employment	16338	-1.24	1.24	-1.92	1.86
2. s-e. wrk	21742	-3.28	3.28	-4.52	5.93
3. pens.	19078	-1.48	1.48	-1.93	2.33
4. other	9427	-4.66	4.66	-5.70	8.19

Table 4.5: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the sample mean; $\epsilon = 1$

statistics: the bootstrap confidence intervals tend to be wider than the asymptotic ones and there is no clear trend as for the role of the equivalence scales on the reliability of the estimates. However, in some cases the difference between the asymptotic and bootstrap confidence intervals is very small or even reversed (for instance, see Table 4.7 for pensioners). Moreover, the distribution of the bootstrap t-ratio statistic seems highly non symmetric, hence the assumption of symmetric confidence intervals employed with the asymptotic approach seems very strong (see Tables 4.6 to 4.20).

Although confidence intervals for sample means and percentiles can be worryingly wide at least for certain sub-samples, they become even wider for the inequality

Sample	20th pct	AS lb	AS ub	BS lb	BS ub
0. all	25600	-1.78	1.78	-2.16	2.22
1. employment	31135	-1.94	1.94	-1.92	2.04
2. s-e. wrk	32900	-3.99	3.99	-5.77	7.92
3. pens.	22002	-3.19	3.19	-2.32	1.90
4. other	9757	-9.79	9.79	-9.02	13.50

Table 4.6: Asymptotic and bootstrap 90% confidence intervals for the 20th percentile; $\epsilon = 0$

Sample	20th pct	AS lb	AS ub	BS lb	BS ub
0. all	15127	-1.61	1.61	-1.84	1.80
1. employment	16766	-2.04	2.04	-3.09	3.03
2. s-e. wrk	17489	-4.04	4.04	-8.35	8.49
3. pens.	15600	-2.51	2.51	-2.32	2.15
4. other	5403	-9.19	9.19	-22.16	16.26

Table 4.7: Asymptotic and bootstrap 90% confidence intervals for the 20th percentile; $\epsilon = 0.5$

Sample	20th pct	AS lb	AS ub	BS lb	BS ub
0. all	8370	-1.70	1.70	-2.29	2.71
1. employment	8435	-2.22	2.22	-3.52	2.01
2. s-e. wrk	8980	-3.97	3.97	-9.26	9.87
3. pens.	10495	-2.64	2.64	-2.97	2.69
4. other	2900	-8.82	8.82	-18.49	12.33

Table 4.8: Asymptotic and bootstrap 90% confidence intervals for the 20th percentile; $\epsilon = 1$

Sample	40th pct	AS lb	AS ub	BS lb	BS ub
0. all	37920	-1.38	1.38	-2.03	1.24
1. employment	43440	-1.77	1.77	-2.45	2.15
2. s-e. wrk	48975	-2.87	2.87	-4.11	5.03
3. pens.	33150	-2.59	2.59	-2.36	3.54
4. other	19542	-7.37	7.37	-12.95	9.33

Table 4.9: Asymptotic and bootstrap 90% confidence intervals for the 40th percentile; $\epsilon = 0$

Sample	40th pct	AS lb	AS ub	BS lb	BS ub
0. all	22066	-1.27	1.27	-1.83	1.97
1. employment	23074	-1.75	1.75	-2.56	2.35
2. s-e. wrk	26176	-2.77	2.77	-2.06	2.97
3. pens.	22167	-2.13	2.13	-2.04	2.76
4. other	10441	-6.90	6.90	-7.38	15.42

Table 4.10: Asymptotic and bootstrap 90% confidence intervals for the 40th percentile; $\epsilon = 0.5$

Sample	40th pct	AS lb	AS ub	BS lb	BS ub
0. all	12505	-1.28	1.28	-1.79	2.13
1. employment	12048	-1.89	1.89	-1.97	2.40
2. s-e. wrk	13505	-2.91	2.91	-3.40	2.76
3. pens.	14421	-1.94	1.94	-2.15	2.59
4. other	5221	-7.16	7.16	-13.36	13.19

Table 4.11: Asymptotic and bootstrap 90% confidence intervals for the 40th percentile; $\epsilon = 1$

Sample	50th pct	AS lb	AS ub	BS lb	BS ub
0. all	45241	-1.32	1.32	-1.82	2.07
1. employment	50052	-1.55	1.55	-1.22	3.24
2. s-e. wrk	55364	-2.80	2.80	-3.58	3.09
3. pens.	39166	-2.52	2.52	-2.27	4.32
4. other	24405	-6.02	6.02	-13.55	6.54

Table 4.12: Asymptotic and bootstrap 90% confidence intervals for the 50th percentile; $\epsilon = 0$

Sample	50th pct	AS lb	AS ub	BS lb	BS ub
0. all	25643	-1.11	1.11	-1.61	1.88
1. employment	26832	-1.61	1.61	-1.79	2.76
2. s-e. wrk	29958	-2.61	2.61	-3.29	5.01
3. pens.	25086	-1.98	1.98	-3.02	2.05
4. other	13386	-6.39	6.39	-9.05	6.58

Table 4.13: Asymptotic and bootstrap 90% confidence intervals for the 50th percentile; $\epsilon = 0.5$

Sample	50th pct	AS lb	AS ub	BS lb	BS ub
0. all	14532	-1.14	1.14	-1.67	1.62
1. employment	14154	-1.72	1.72	-2.54	2.73
2. s-e. wrk	15587	-2.87	2.87	-2.65	3.13
3. pens.	16576	-1.74	1.74	-1.68	2.88
4. other	6984	-7.40	7.40	-17.42	5.71

Table 4.14: Asymptotic and bootstrap 90% confidence intervals for the 50th percentile; $\epsilon = 1$

Sample	60th pct	AS lb	AS ub	BS lb	BS ub
0. all	53089	-1.16	1.16	-1.33	1.60
1. employment	56684	-1.42	1.42	-1.72	2.66
2. s-e. wrk	65294	-3.03	3.03	-4.79	4.75
3. pens.	46666	-2.58	2.58	-4.25	3.77
4. other	29831	-4.99	4.99	-12.86	6.36

Table 4.15: Asymptotic and bootstrap 90% confidence intervals for the 60th percentile; $\epsilon = 0$

Sample	60th pct	AS lb	AS ub	BS lb	BS ub
0. all	29418	-1.07	1.07	-1.43	1.85
1. employment	30541	-1.40	1.40	-1.83	2.27
2. s-e. wrk	34832	-2.77	2.77	-1.55	5.49
3. pens.	28854	-2.01	2.01	-2.49	3.22
4. other	16900	-4.58	4.58	-6.98	8.63

Table 4.16: Asymptotic and bootstrap 90% confidence intervals for the 60th percentile; $\epsilon = 0.5$

Sample	60th pct	AS lb	AS ub	BS lb	BS ub
0. all	16735	-1.07	1.07	-1.31	2.11
1. employment	16161	-1.56	1.56	-1.68	1.71
2. s-e. wrk	18317	-3.22	3.22	-4.30	5.03
3. pens.	18474	-1.65	1.65	-1.62	2.45
4. other	9333	-5.16	5.16	-6.50	6.45

Table 4.17: Asymptotic and bootstrap 90% confidence intervals for the 60th percentile; $\epsilon = 1$

Sample	80th pct	AS lb	AS ub	BS lb	BS ub
0. all	73880	-1.18	1.18	-2.14	2.45
1. employment	73738	-1.44	1.44	-2.48	2.91
2. s-e. wrk	94043	-3.20	3.20	-3.87	6.70
3. pens.	68758	-2.41	2.41	-4.75	3.93
4. other	42519	-3.47	3.47	-10.73	4.93

Table 4.18: Asymptotic and bootstrap 90% confidence intervals for the 80th percentile; $\epsilon = 0$

Sample	80th pct	AS lb	AS ub	BS lb	BS ub
0. all	39424	-1.03	1.03	-1.91	1.39
1. employment	39451	-1.39	1.39	-2.40	1.49
2. s-e. wrk	48498	-3.25	3.25	-5.31	4.46
3. pens.	38656	-1.73	1.73	-2.32	2.38
4. other	24071	-4.03	4.03	-8.80	11.25

Table 4.19: Asymptotic and bootstrap 90% confidence intervals for the 80th percentile; $\epsilon = 0.5$

Sample	80th pct	AS lb	AS ub	BS lb	BS ub
0. all	23254	-1.31	1.31	-1.37	1.78
1. employment	22208	-1.76	1.76	-2.53	2.94
2. s-e. wrk	27565	-3.90	3.90	-5.30	6.42
3. pens.	24872	-2.22	2.22	-2.90	2.66
4. other	13597	-4.56	4.56	-6.80	4.02

Table 4.20: Asymptotic and bootstrap 90% confidence intervals for the 80th percentile; $\epsilon = 1$

Sample	GE(0)	AS lb	AS ub	BS lb	BS ub
0. all	0.234	-3.38	3.38	-5.56	6.26
1. employment	0.139	-4.10	4.10	-7.58	7.84
2. s-e. wrk	0.268	-7.80	7.80	-12.79	15.76
3. pens.	0.212	-4.63	4.63	-6.59	9.24
4. other	0.446	-9.16	9.16	-16.63	19.00

Table 4.21: Asymptotic and bootstrap 90% confidence intervals for the GE(0); $\epsilon = 0$

indices considered. According to my calculations, all bootstrap confidence intervals for GE indices are asymmetric: the t-ratio statistic is highly skewed to the right, i.e. there is more uncertainty on the upper bound. Again, bootstrap confidence intervals are significantly wider than asymptotic ones, the difference being larger the smaller the size of the sub-sample considered. In various cases, the bootstrap upper bound is more than 100% larger than the estimated statistic (for instance, see Table 4.29). Finally, there is a clear effect of the equivalent income transformation, especially in smaller sub-samples. In particular, the larger the economies of scale assumed in the household (i.e. the smaller ϵ in (4.5)) the narrower the confidence intervals (see Tables 4.21-4.29). This result applies both using asymptotic and bootstrap confidence intervals.

The final part of this chapter focuses on the effect of MSM transformation on sampling error. To derive the t-ratio statistic analytically starting from the original data set and the knowledge of the tax rules used in the microsimulation procedure requires a formidable effort even for trivial tax-benefit schemes. Alternatively, we can proceed along two alternative paths: assume that the underlying population is

Sample	GE(0)	AS lb	AS ub	BS lb	BS ub
0. all	0.219	-3.79	3.79	-5.72	7.20
1. employment	0.147	-4.50	4.50	-7.89	8.99
2. s-e. wrk	0.269	-8.66	8.66	-11.68	17.39
3. pens.	0.166	-6.09	6.09	-7.34	12.45
4. other	0.436	-10.53	10.53	-14.81	22.63

Table 4.22: Asymptotic and bootstrap 90% confidence intervals for the GE(0); $\epsilon = 0.5$

Sample	GE(0)	AS lb	AS ub	BS lb	BS ub
0. all	0.251	-3.83	3.83	-5.56	6.19
1. employment	0.189	-4.92	4.92	-7.31	8.39
2. s-e. wrk	0.306	-8.82	8.82	-11.61	17.12
3. pens.	0.173	-6.81	6.81	-8.82	11.49
4. other	0.485	-13.72	13.72	-16.41	25.01

Table 4.23: Asymptotic and bootstrap 90% confidence intervals for the GE(0); $\epsilon = 1$

Sample	GE(1)	AS lb	AS ub	BS lb	BS ub
0. all	0.224	-4.62	4.62	-6.88	9.20
1. employment	0.130	-3.60	3.60	-6.20	7.31
2. s-e. wrk	0.290	-9.58	9.58	-13.87	20.84
3. pens.	0.210	-6.85	6.85	-8.16	19.30
4. other	0.360	-10.19	10.19	-17.43	23.76

Table 4.24: Asymptotic and bootstrap 90% confidence intervals for the GE(1); $\epsilon = 0$

Sample	GE(1)	AS lb	AS ub	BS lb	BS ub
0. all	0.215	-5.67	5.67	-6.96	10.66
1. employment	0.139	-4.61	4.61	-6.79	8.73
2. s-e. wrk	0.303	-11.47	11.47	-13.53	27.27
3. pens.	0.175	-10.06	10.06	-10.74	29.52
4. other	0.360	-18.36	18.36	-20.61	43.98

Table 4.25: Asymptotic and bootstrap 90% confidence intervals for the GE(1); $\epsilon = 0.5$

Sample	GE(1)	AS lb	AS ub	BS lb	BS ub
0. all	0.253	-6.14	6.14	-7.05	10.28
1. employment	0.187	-6.83	6.83	-7.33	10.37
2. s-e. wrk	0.356	-12.20	12.20	-14.36	27.67
3. pens.	0.188	-11.32	11.32	-12.96	24.94
4. other	0.440	-31.87	31.87	-28.52	115.59

Table 4.26: Asymptotic and bootstrap 90% confidence intervals for the GE(1); $\epsilon = 1$

Sample	GE(2)	AS lb	AS ub	BS lb	BS ub
0. all	0.312	-9.54	9.54	-11.73	21.95
1. employment	0.144	-4.30	4.30	-6.72	10.89
2. s-e. wrk	0.448	-14.99	14.99	-17.47	40.79
3. pens.	0.273	-17.49	17.49	-19.11	98.32
4. other	0.500	-17.60	17.60	-27.09	42.45

Table 4.27: Asymptotic and bootstrap 90% confidence intervals for the GE(2); $\epsilon = 0$

Sample	GE(2)	AS lb	AS ub	BS lb	BS ub
0. all	0.328	-12.99	12.99	-14.40	32.65
1. employment	0.159	-7.92	7.92	-9.22	17.30
2. s-e. wrk	0.524	-20.44	20.44	-21.66	73.21
3. pens.	0.253	-25.75	25.75	-26.06	147.90
4. other	0.578	-48.51	48.51	-41.41	241.20

Table 4.28: Asymptotic and bootstrap 90% confidence intervals for the GE(2); $\epsilon = 0.5$

Sample	GE(2)	AS lb	AS ub	BS lb	BS ub
0. all	0.428	-14.88	14.88	-15.34	33.02
1. employment	0.248	-16.81	16.81	-14.89	33.97
2. s-e. wrk	0.693	-24.08	24.08	-25.21	86.25
3. pens.	0.299	-26.95	26.95	-28.95	102.31
4. other	1.051	-85.38	85.38	-73.99	822.57

Table 4.29: Asymptotic and bootstrap 90% confidence intervals for the GE(2); $\epsilon = 1$

infinite, hence use the normal critical values for confidence interval estimation, or use simulation-based inference methods such as the bootstrap. Since our interest is to see whether an MSM in itself makes sampling error more serious, simulated BT income was used for the asymptotic confidence intervals: the simulation was from AT income as recorded in the original data set using the TABEITA98 model. For the bootstrap confidence intervals the procedure is more time demanding. In fact the MSM model needs to be applied to each of the B bootstrap samples produced from the AT income as recorded in SHIW data set[4]. The income vector is equivalized using (4.5) with $\epsilon = 0.5$.

Estimates on the whole sample are quite reliable. Confidence intervals are never larger than $\pm 3.3\%$ of estimated statistics for the mean and the quantiles; they are also relatively narrow for the GE indices. The most striking result is that there is no clear evidence that microsimulation transformations make summary statistic estimation less reliable. Comparing simulated BT income with original AT income with $\epsilon = 0.5$ in various cases the confidence interval, as a percentage of the estimated statistic, is smaller for simulated income than for the original income. For instance comparing Tables 4.25 and 4.37 confidence intervals are smaller as percentage of the estimated statistic for simulated income in the sub-sample of employed, pensioners, and other households, using both asymptotic and bootstrap confidence intervals. The confidence intervals are even narrower if the $GE(2)$ index is considered (compare Tables 4.28 and 4.38).

[4]It took us 16 hours to compute these confidence intervals with $B = 999$ and Intel Pentium M, 1.6GHz processor, 512Mb of RAM.

Sample	mean	AS lb	AS ub	BS lb	BS ub
0. all	28855	-1.00	1.00	-1.57	1.74
1. employment	30965	-1.10	1.10	-1.96	1.93
2. s-e. wrk	36316	-3.24	3.24	-4.77	6.30
3. pens.	26137	-1.38	1.38	-2.13	2.20
4. other	13996	-4.62	4.62	-8.39	11.22

Table 4.30: MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the sample mean; $\epsilon = 0.5$

Sample	20th pct	AS lb	AS ub	BS lb	BS ub
0. all	13015	-2.19	2.19	-3.29	2.80
1. employment	16629	-2.31	2.31	-2.95	3.32
2. s-e. wrk	13880	-5.14	5.14	-5.54	9.21
3. pens.	12435	-3.73	3.73	-3.06	5.65
4. other	3221	-13.34	13.34	-15.99	32.99

Table 4.31: MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the 20th percentile; $\epsilon = 0.5$

The non-linear microsimulation transformation that simulates BT income has a different distribution (and in particular a different first and second moment) from the original AT income. The MSM transformation also modifies the distribution of the t-ratio statistic. Hence, the confidence intervals do change, but not necessarily for the worse.

4.4 Conclusions

This chapter contributes to the analysis of static MSMs reliability using the bootstrap to compute confidence intervals. The bootstrap is considered mainly for two reasons:

Sample	40th pct	AS lb	AS ub	BS lb	BS ub
0. all	20703	-1.44	1.44	-2.12	1.95
1. employment	24150	-1.88	1.88	-2.06	3.13
2. s-e. wrk	23147	-3.66	3.66	-3.90	4.89
3. pens.	19178	-2.67	2.67	-3.96	3.43
4. other	7025	-9.72	9.72	-14.39	11.79

Table 4.32: MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the 40th percentile; $\epsilon = 0.5$

Sample	50th pct	AS lb	AS ub	BS lb	BS ub
0. all	24421	-1.31	1.31	-2.28	1.14
1. employment	27752	-1.68	1.68	-2.41	2.08
2. s-e. wrk	27275	-3.13	3.13	-4.68	4.14
3. pens.	22519	-2.35	2.35	-2.86	2.95
4. other	10341	-8.31	8.31	-14.24	15.37

Table 4.33: MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the 50th percentile; $\epsilon = 0.5$

(a) It allows one to drop the assumption of an infinite population to compute confidence intervals. (b) There is a growing body of literature showing that bootstrap often performs better, or not worse, than asymptotic approximation.

At first, using the original sample before performing any tax-benefit microsimulation, the bootstrap confidence intervals for some summary statistics (mean, percentiles and *GE* indices) are computed and compared with asymptotic ones, as suggested in Pudney and Sutherland (1994). It is found that bootstrap confidence intervals are less conservative than asymptotic confidence intervals, especially for the smaller sub-samples used. Although there is no clear theoretical result to compare

Sample	60th pct	AS lb	AS ub	BS lb	BS ub
0. all	28753	-1.23	1.23	-1.86	1.70
1. employment	32034	-1.55	1.55	-2.42	2.21
2. s-e. wrk	32124	-3.09	3.09	-3.15	4.18
3. pens.	26651	-2.32	2.32	-2.21	4.09
4. other	13912	-7.06	7.06	-11.00	10.34

Table 4.34: MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the 60th percentile; $\epsilon = 0.5$

Sample	80th pct	AS lb	AS ub	BS lb	BS ub
0. all	40531	-1.16	1.16	-1.68	1.95
1. employment	42308	-1.52	1.52	-2.22	1.43
2. s-e. wrk	46938	-3.27	3.27	-6.57	4.59
3. pens.	37834	-2.13	2.13	-3.76	3.03
4. other	21737	-3.94	3.94	-9.12	7.91

Table 4.35: MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the 80th percentile; $\epsilon = 0.5$

asymptotic and bootstrap confidence intervals in the complex non-linear setting considered, the generally better performance of the bootstrap in finite samples should give rise to concern about the reliability of estimates on sub-samples of widely used surveys. It was also shown that the non-linear equivalence scale transformation can have effect on reliability of estimates. It should then become a good practice to provide confidence intervals for different equivalence scales.

Secondly, the chapter analyzed the effect of a typical microsimulation transformation on the reliability of the summary statistics considered. It was found that MSMs do not necessarily make confidence intervals larger. In some cases, summary

Sample	GE(0)	AS lb	AS ub	BS lb	BS ub
0. all	0.298	-3.66	3.66	-5.91	6.80
1. employment	0.173	-4.64	4.64	-8.95	8.78
2. s-e. wrk	0.358	-8.84	8.84	-13.97	17.39
3. pens.	0.262	-6.15	6.15	-7.96	9.04
4. other	0.635	-7.95	7.95	-14.55	18.99

Table 4.36: MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the GE(0); $\epsilon = 0.5$

Sample	GE(1)	AS lb	AS ub	BS lb	BS ub
0. all	0.256	-5.66	5.66	-7.89	11.60
1. employment	0.160	-4.66	4.66	-7.26	7.99
2. s-e. wrk	0.377	-13.27	13.27	-18.24	29.75
3. pens.	0.213	-5.55	5.55	-7.50	10.05
4. other	0.472	-10.36	10.36	-17.93	40.04

Table 4.37: MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the GE(1); $\epsilon = 0.5$

statistics on simulated incomes have narrower confidence intervals, as percentage of the computed statistic, than before any tax-benefit simulation.

These results show that concerns in sampling error with MSM are sometimes misplaced: it is not *microsimulation* that necessarily makes the estimation less reliable. A poor coverage of the population of some current surveys is often the main problem. An improvement of data production should then be a major concern for policy makers interested in understanding in advance the effects of their policies.

Sample	GE(2)	AS lb	AS ub	BS lb	BS ub
0. all	0.378	-15.00	15.00	-17.15	43.70
1. employment	0.185	-7.62	7.62	-9.34	12.32
2. s-e. wrk	0.710	-24.84	24.84	-29.02	83.25
3. pens.	0.246	-9.28	9.28	-10.80	18.56
4. other	0.713	-22.12	22.12	-36.20	148.89

Table 4.38: MSM: Asymptotic and bootstrap 90% confidence intervals as % of the estimate for the GE(2); $\epsilon = 0.5$

Chapter 5

Review of the literature on income inequality decomposition

Decomposition analysis is important in understanding the "sources" of inequality and its main determinants; it may also indicate directions for policy. The first contributions on inequality decomposition started to appear at the end of the 1970s and were based on the analysis of the mathematical properties of inequality indices. This could be referred as the "traditional" approach to inequality decomposition. In the last decade or so, interest in inequality decomposition has been revitalized by the possibility offered by a regression-based approach, which seems to overcome some of the main limitations of the traditional methods and, more recently, by simulation-based approaches. This chapter will focus mainly on (household/family) income inequality. However, inequality indices are, broadly speaking, dispersion measures which can be applied to any random variable.

Section 5.1 deals with the "traditional" approach to inequality decomposition.

Section 5.2 discusses the regression based approach and Section 5.3 the microsimulation methodology. Section 5.4 concludes.

5.1 The "traditional" approach

The "traditional" decomposition analysis has mainly focused attention on levels rather than trends in inequality and has developed two main ways for decomposing inequality levels: (i) the decomposition by population subgroups (the main reference thereby are Bourguignon (1979), Cowell (1980) and Shorrocks (1980)) and (ii) the decomposition by "factor components"[1] (Fei et al. (1978); Pyatt et al. (1980); Shorrocks (1984)).

The decomposition by "population subgroups" requires the division of the sample of interest in non-overlapping sub-samples (e.g. age groups, sex, area of residence, etc.) and the choice of an inequality index, since different indices have different decomposability properties. In some ideal cases (i.e. indices belonging to the Generalized Entropy class), the total inequality can be expressed as a weighted sum of the inequality in each subgroup ("within" inequality) and of the inequality remaining where each person's income is set equal to her sub-group's mean income ("between" inequality). In other cases (e.g. the Gini coefficient), the decomposition is not exact: a residual term has to be added to account for total inequality. However, such a decomposition methodology suffers serious limitations, common to all inequality indices.

[1]The decomposition by "factor components" is often called by "income source", especially in the literature considered here, namely inequality of individual or household/family *income*.

First, the decomposition applies only to discrete categories and while this is reasonable for characteristics like sex it is less so for variables such as age, which could be regarded as continuous. This has often been seen as a minor limitation since analysis by age groups are nonetheless very informative. Second, the relative importance of "between" and "within" inequality is dependent on the number of subgroups considered (see, among others, Cowell and Jenkins, 1995). In the extreme case in which there are as many subgroups as households or individuals, the "between" components account for the whole inequality index. Moreover, the finer is the decomposition, the smaller becomes the number of individuals/households belonging to each cell increasing the variance of the estimates, i.e. their reliability. Finally, there is no control for endogeneity in the sense that the dependent variable (e.g. income) can be a determinant of some characteristics of the unit under observation (e.g. area of residence or education) leaving this decomposition methodology open to the criticism of being simply a descriptive tool, unable to provide insights about the *causes* of inequality.

The decomposition by "factor source" stems from looking at income as the sum of different income sources, which can be either positive (e.g. work income, pension) or negative (e.g taxes) components of the total income. Let $\{y_i, i = 1, ..., N\}$ be the set of individual incomes in a sample of size N. Each individual income can be derived as the sum of incomes coming from M different sources, $y_i = \sum_{m=1}^{M} y_i^m$ and total income from source m can be derived as $y^m = \sum_{i=1}^{N} y_i^m$. Total inequality of income,

$I(\mathbf{y})$, can then be written as a weighted sum of individual incomes:

$$I(\mathbf{y}) = \sum_{i=1}^{N} a_i(\mathbf{y}) y_i \tag{5.1}$$

where $a_i(\mathbf{y})$ is the weighting factor. The proportional contribution of source m to overall inequality is simply

$$s^m = \frac{\sum_{i=1}^{N} a_i(\mathbf{y}) y_i}{I(\mathbf{y})} \tag{5.2}$$

Proposing additional requirements to such decomposition[2] Shorrocks (1982) showed that there exists a unique decomposition rule, invariant to the inequality index used:

$$s^m = \frac{cov(\mathbf{y}^m, \mathbf{y})}{var(\mathbf{y})} \tag{5.3}$$

where $\mathbf{y}^m = \{y_i^m, i = 1, ...N\}$. Factor contributions can be either positive or negative, depending on the factor providing a disequalizing or equalizing contribution. This decomposition rule put a stop to the discussion on which is the best decomposition by "factor components" (see for instance, Fei et al., 1978; Pyatt et al., 1980). However, the invariance to inequality index, besides being one of the strong points of this decomposition rule can also be seen as a drawback when different inequality indices provide different answers regarding direction and amount of the overall change. In particular it can be argued that the relative contribution of a component must depend on the inequality measure chosen, which could depend on how this measure weights

[2]Namely, (i) that a given income source makes no contribution to overall inequality if income receipts from this source are equally distributed and (ii) that if total income is divided into two components whose factor distributions are permutations of each other, then the two components contribute equally to aggregate inequality

a progressive transfer against a regressive one (Chantreuil and Trannoy, 1999).

Both the decompositions by "population subgroup" and by "factor source" are decompositions for the *level*, which are addressing a different question than the decomposition for the *trend* of inequality, namely the change between different moments in time.

The analysis of the trend of inequality with the "traditional" decomposition methodologies has been performed extending the existing decompositions. Mookherjee and Shorrocks (1982) provide the main reference for the extension to the analysis of trend of the decomposition by "population subgroups"[3]. Inequality changes between two different periods can then be decomposed into "pure" inequality changes within groups (weighted by the average of the population in each group in the two periods), changes due to variation in the numbers of individuals in different groups and changes resulting from variations in the relative income of different groups. Jenkins (1995) extends the decomposition by "factor components" using counterfactuals since the direct extension of the Shorrocks (1982) does not provide an intuitive interpretation. This extension allows one to represent the absolute contribution of a factor as the average of two alternative ways of summarizing the statement that factor m makes a contribution $\mathbb{C}_m$ to total income inequality. In particular, such a contribution can be regarded as (a) the inequality which would be observed if income component m were the only source of income differences or (b) the amount by which inequality would fall if factor m income receipts were eliminated. The two counterfactuals can then be

[3]Mookherjee and Shorrocks (1982), provide the extension to the trend only for the Generalized Entropy measure with parameter equal zero, i.e. the mean logarithmic deviation.

estimated and their change compared over time (Jenkins, 1995, p. 40).

Jenkins (1995) is an interesting paper from a methodological point of view, as well as for the evidence it presents on the UK income inequality. In particular, the combination of the two "traditional" decomposition methodologies presented there is not only novel but also instructive for understanding its pros and cons. The pros are clearly that you can get a broader explanation of the causes of a certain trend of inequality, disentangling between "income recipients" and "income source" influences to change of inequality. The cons are that such a combination is performed at the expense of clarity of results, in part due to the need for using different inequality indices in part to the fact that you can only have direction of change rather than exact contribution values of different factors. The variability of the estimates can be very large and, last but not least, the conclusions obtained from such an analysis are highly sensitive to the researcher's choices, as for which years are compared and which types of income are analyzed.

A general limitation of traditional decomposition analysis is that decomposition by "population subgroup" and by "factor source" address different problems and cannot be combined in a single framework. Recently some authors have attempted to put together the two techniques into a unifying framework. Shorrocks (1999) suggests starting from the definition of an inequality index, I, as some function of m factor contributions. He then suggests computing the marginal effect of each of the factors as they are eliminated in succession, and then averaging these marginal effects over all possible elimination sequences. Formally the resulting formula is identical to the Shapley value in cooperative game theory, henceforth it has been referred as the

Shapley decomposition (see also Chantreuil and Trannoy, 1999). However, application of this methodology (which potentially allows studying levels as well as trends of inequality) has been quite disappointing. The main reasons for its unsatisfactory performance lies in the high sensitivity of computation to the level of disaggregation of the factors that account for inequality. Moreover it does not properly handle causality in inequality factors (Sastre and Trannoy, 2000b). To overcome some of its limitations Sastre and Trannoy (2000a) suggested developing a tree of causality and to use Shapley value only when no clear priority of causes can be declared. However, this solution makes the method more cumbersome and less convincing.

5.2 The regression-based approach

Mainly to overcome some of the limitations of the "traditional" approach, recent contributions have looked at regression analysis to decompose inequality.

In the "traditional decomposition" the attention has been focused on expressing total inequality as a function of the inequality in population subgroups or of different sources of income, possibly without residuals, with little attention to the causes of inequality (hence the critique for being "only" a descriptive tool). In the "regression-based approach" instead, the main attention is on the DGP that led to a particular distribution of income, that is on the causal relation of individual and household characteristics on the generation of income. The income-generating function, with $\mathbf{Y}$, an N vector of (log) incomes, $\mathbf{X}$, an $N \times K$ matrix of individual and household characteristics, β, a $K \times 1$ vector of coefficients (or "prices") and ϵ, an $N \times 1$ vector

of residuals, can be expressed as:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{5.4}$$

This model is then estimated for a sample of observations $\{y_i,\ x_i,\ i = 1, 2, ...N\}$. Morduch and Sicular (2002) estimate an income regression as in (5.4) with the vector of per capita household income as dependent variable on a constant and a set of individual characteristics such as age, education, etc. Hence the vector β is interpreted as the effect (or "price") of the independent variables on income flows.

Assuming that there are M different sources of income, such as pension, employment, transfer, etc., per capita household income of household i, y_i can be seen as the sum of y_i^m, $m = 1, 2, ..., M$, such that

$$y_i = \sum_{m=1}^{M} y_i^m \tag{5.5}$$

Per capita income of household i is represented as:

$$y_i = \sum_{k=1}^{K} \widehat{\beta}_k x_i^k + \widehat{\epsilon}_i \tag{5.6}$$

The shares attributable to the characteristic $k = 1, ..., K$, using (5.2) take the form:

$$s^k = \widehat{\beta}_k \left(\frac{\sum_{i=1}^{K} a_i(\mathbf{y}) x_i^k}{I(\mathbf{y})} \right) \tag{5.7}$$

However it should be noted that the sum of all s^k adds up to the predicted value

$\widehat{y_i}$ and not to the actual value y_i.

Since the decomposition (5.7) is linear in the parameters, Morduch and Sicular (2002) claim that the variance of s^k is easily estimated using the variance of $\widehat{\beta_k}$, $var(\widehat{\beta_k})$, which is a standard output in regression analysis:

$$var(s^k) = var(\widehat{\beta_k}) \times \left[\frac{\sum a_i(\mathbf{y})x_i^k}{I(\mathbf{y})}\right]^2 \tag{5.8}$$

Morduch and Sicular (2002) claim that this method "yields an exact allocation of contributions to the identified variables, it is general in that it can be employed with different inequality indices and decomposition rules" (p. 94).

However, in the application of the method to Chinese data, it is also shown that the wide applicability to different inequality measures is counterbalanced by its large variability. For instance, while the net effect[4] of an additional year of education is to reduce per capita household income inequality by about 50% using the Theil index, it is to slightly increase it using the Gini index. They motivate the difference pointing out the fact that the Gini does not fulfill the Corollary of the Uniform Addition Property (CUAP), but even sticking to the Theil index it is difficult to be satisfied with a decomposition method, which leaves over 90% of inequality unexplained[5]. The same critique can be applied to the other regression-based contributions especially when household income is the dependent variable, as is the case, for instance in Fields (2002)[6].

[4]The net effect of education is given by the sum of the effects of the variables *average education of adults* and *education squared.* The same approach is taken also by Fields (2002).

[5]See "Residual regression" line in Table 2, p.103.

[6]In Fields and Mitchell (1999) the dependent variable is *labor income.*

Moreover, their explanation in support of the Theil vis-a-vis the Gini index based on their Corollary to the Uniform Addition Property is not convincing. The CUAP states that if a component of total income (say, income type m) is perfectly equally distributed its contribution to total income inequality, s^m, must be negative. Clearly this property is not satisfied by the decomposition rule (5.3). In fact, using Shorrocks's rule (5.3) the contribution s^k is negative only if there is a negative correlation between total income and income component k, as is often the case for taxes or transfers. The decomposition rule that Morduch and Sicular (2002) are suggesting is then "just" another of a big bunch of possible decomposition rules. However, unless valid reasons are provided for constraining the set of potential decomposition rules, preferably to the point where inequality index can be decomposed in only one way, "the inequality contribution assigned to any income source can vary arbitrarily, depending on the choice of decomposition rule. This turns the calculation of inequality contributions into a meaningless exercise..." (Shorrocks, 1983, p. 315). Besides not fulfilling CUAP, which is crucial in the Morduch and Sicular argument, the decomposition rule (5.3) is the unique one to be invariant to the choice of inequality measure and, to the best of my knowledge, no alternative "unique" rule has been proposed.

Morduch and Sicular (2002) also claim that their method is associated with a simple procedure for deriving standard errors and confidence intervals for the estimated components of inequality, s^k, as described above. However, even assuming that the component $a_i(\mathbf{y})x_i^k$ in the numerator of eq. (5.8) is non-stochastic, the denominator,

$I(\mathbf{y})$, certainly is stochastic and the correct estimated variance should then be:

$$var(s^k) = var\left(\frac{\widehat{\beta}_m}{I(\mathbf{y})}\right) \times \left[\sum a_i(\mathbf{y}) x_i^k\right]^2 \tag{5.9}$$

The variance (5.9) is clearly less easy to compute since the standard error of $\left(\frac{\widehat{\beta}_m}{I(\mathbf{y})}\right)$ implies the use of the bootstrap or of relatively complicated formulae (see for example Cowell, 1989; Cowell and Flachaire, 2002).

Like Morduch and Sicular (2002), Fields (2002) also takes an approach based on estimating a (per capita) income-generating equation without modelling the endogenous decision at the household level, claiming that it is general enough to be applied to household income as well as individual income. The factor share, s^k, proposed by Fields (2002) is based on a decomposition of the Total Sum of Squares into Explained Sum of Squares and in Residual Sum of Squares ($TSS = ESS + RSS$). Given a sample of size N and denoting $r_{jy} = N\sum_{i=1}^{N}(x_{ij} - \overline{x}_j)(y_i - \overline{y})$:

$$\begin{aligned} ESS &= \widehat{\beta}_1 r_{1y} + \widehat{\beta}_2 r_{2y} + ... + \widehat{\beta}_K r_{K,y} \\ &= \widehat{\beta}_1 \frac{cov(x_1, y)}{var(y)} + \widehat{\beta}_2 \frac{cov(x_2, y)}{var(y)} + ... + \widehat{\beta}_K \frac{cov(x_K, y)}{var(y)} \end{aligned}$$

This decomposition of the ESS, and the notation $\hat{y}^k = \hat{\beta}_k x^k$ and $\sigma(x) = \sqrt{var(x)}$ allows one to write down the following:

$$\sum_{k=1}^{K} s^k = \sum_{k=1}^{K} \frac{cov(\hat{y}^k, y)}{var(y)} = \sum_{k=1}^{K} \frac{\hat{\beta}_k \cdot \sigma(x^k) \cdot corr(x^k, y)}{\sigma(y)} = R^2 \tag{5.10}$$

where $corr(x^k, y)$ is the correlation coefficient between x^k and y and R^2 is the centered coefficient of determination. It was claimed that this method allows a "full and exact decomposition of an inequality index" (Fields, 2002, p. 6), however it regards *only* the share of the inequality index that is explained by the variables used in the regression. In general, the decomposition is *not* full and exact, since the R^2 is equal to one only if there is perfect coincidence between the dependent variable predicted by the model and its true value, which happens only in uninteresting cases[7]. Fields (2002) also discusses the problems that regression-based decomposition faces when interaction terms are introduced and found statistically significant in the regression, such as the product between a `sex` dummy and an `education` variable. The proposed solution of splitting the contribution to the interaction of sex and education components evenly is not completely satisfactory because it assumes that they have the same importance in the interaction term. Nor is it satisfactory to split the sample into at least two groups, "males" and "females" and run two different regressions: not only would this solution become very cumbersome if a dummy variable taking more than two values is used or more than one interaction term is introduced, but also a pooled and two (or more) separate regressions imply different estimates for the same coefficients and their standard errors.

The only extension of the regression-based approach to inequality decomposition to the *trend* of inequality is provided by Fields (2002) and Fields and Mitchell (1999).

[7]Namely, when the space spanned by the dependent variable is the same as the space spanned by the vector of columns of the matrix of explanatory variables.

Considering two years, say t and t', the change in inequality can be expressed as:

$$I_t - I_{t'} = \sum_k [s_t^k I_t - s_{t'}^k I_{t'}] \tag{5.11}$$

$$\sum_k \Pi_k \equiv \frac{\sum_k [s_t^k I_t - s_{t'}^k I_{t'}]}{I_t - I_{t'}} = 1 \tag{5.12}$$

and the contribution of the k-th factor to the change in inequality between time t and time t', regardless of the inequality measure I used, is given by:

$$\Pi_k \equiv \frac{[s_t^k I_t - s_{t'}^k I_{t'}]}{I_t - I_{t'}} \tag{5.13}$$

The difference between the s^k in the two years, assuming the change is only marginal, is studied taking the percentage rate of growth (log-differention) of s^k in (5.10) obtaining:

$$[\dot{s^k}] = [\hat{\beta}_k] + [\sigma(\dot{x^k})] + [corr(\dot{x^k}, y)] - [\sigma(\dot{y})] \tag{5.14}$$

the " $\dot{}$ " symbol indicating a percentage rate of growth. However, Fields (2002) and Fields and Mitchell (1999) do not discuss the fact that approximation (5.14) can be very poor[8] and no discussion of confidence intervals for s^k are introduced. Finally, a common limitation of the regression-based approach that applies to both the analysis of level and trend is that only *linear* income-generating models can be considered.

[8]In Fields and Mitchell (1999), Table 5.4 for example, the variable CHJOB ("individual has changed job within the last year") contributes for 163% to inequality change in Taiwan in 1980-1993, but the sum of the element of the RHS of (5.14) is overestimating the actual change, $[\dot{s^k}]$, where $k = CHJOB$, by 39.7%.

In conclusion, the regression-based approach presents more limitations than real advantages, and does not look very appealing for analysis of the trend either. Moreover, the high share of total income that is left unexplained, especially when household income is considered as the dependent variable, suggests that a better model should be considered, possibly estimating the labor participation decisions conditional on the participation decisions of the other members of the household, which are intrinsically non-linear.

5.3 Microsimulation approaches to inequality decomposition

Under the heading of microsimulation there is a broad class of models which basically involve the construction of counterfactuals (i.e. simulated models) and its comparison with the actual model.

Let us consider a general income formation model for individual/household i at time t. Her income can be expressed as $y_{it} = g(x_{it}, w_{it}, \epsilon_{it})$, where $g(\cdot)$ is un unknown function of a set of individual or household characteristics or some policy variables (e.g. taxes), x, a set of sampling weights, w, and some unobservable characteristics or error term, ϵ. In this model the function $g(\cdot)$ is not constrained to be linear nor parametric, and can represent a single income-generating equation or a system of income-generating functions. In most cases however, some constraints are introduced. For instance, the function $g(\cdot)$ is often assumed to be parametric and the parameters,

β_{it}, then represent a set of prices, labor remuneration rates and parameters describing the occupational choice behavior of i at time t. To estimate the parameters, a sample of observations, $\{y_i, x_i, w_i; i = 1, 2, ..., N\}$, is used and the set of parameters constrained to be constant throughout individuals, i.e. $\beta_{it} = \beta_{jt} \equiv \beta_t, j \neq i$.

The income generating model can then be expressed as: $y_{it} = g(x_{it}, \beta_t, w_{it}, \epsilon_{it})$ and can be used to simulate a large number of experiments. Three broad classes of possible simulations can be distinguished:

(i) the observed conditioning variables x_{it} can be replaced with some set of counterfactual explanatory variables, x_{it}^c, to simulate changes in population characteristics;

(ii) the sampling weights, w_{it}, can be changed to modify the relative weight between observations;

(iii) when $g(\cdot)$ is a system of functions generating different (and interdependent) incomes (as work income, pension income, etc.), different counterfactual incomes can be generated and aggregated through the function $g(\cdot)$.

The difference between the simulated and the actual model will finally be measured and interpreted as a policy shock or a demographic or an institutional change. The change can be expressed in terms of:

(a) the change in the expected outcome;

(b) the change in a single random outcome;

(c) the change in the distribution of random outcomes;

(d) the change of some function (e.g. an inequality index) of the distribution of random outcomes.

A common factor in all microsimulation approaches is the need for a detailed and time-consistent micro data set. The functional form of $g(\cdot)$ should be the same for actual and counterfactual data sets but comparison can be either at different times or at the same time.

The famous Blinder-Oaxaca decomposition (Blinder, 1973; Oaxaca, 1973) belongs to the microsimulation models of type (i). It was developed to explain wage differences between different groups of the population and was based on the estimate of an earnings equations. Let $\mathbf{y}$ denote the vector of individual log-earnings and x the matrix of individual characteristics for two non-overlapping groups, say A and B. Assuming there are no major selection problem, it is possible to run a separate regression of y over x for each group: $y^i = x^i\beta^i + \epsilon^i$, with $i = A, B$. Standard econometric techniques allows one to obtain two different sets of parameter estimates, $\widehat{\beta}^A$ and $\widehat{\beta}^B$. These parameters are then used to predict the counterfactual income $\overline{y}_B^A$, which is the average value of income of group A provided it was paid according to prices of group B, and viceversa ($\widehat{\overline{y}}_B^A = \overline{x}^A\widehat{\beta}^B$). The difference $\overline{y}^i - \overline{y}_j^i$, with $i, j = A, B$ and $i \neq j$, then measures the amount of this difference in prices paid to the two groups, where $\overline{y}^i$ is the actual mean wage in each group. Notwithstanding this methodology being very helpful, it presents two main limitations as for household inequality analysis. First it can only be used for analyzing the difference in mean (case (a) above); second, when the dependent variable is household income a single equation model can be a

poor representation of the process that generated household income, neglecting issues such as the combination of employment, self-employment, pension and possibly capital income, household formation decisions, labor participation of single individuals conditional on other household members' decisions.

Bourguignon et al. (2001) propose an extension of the single-wage-equation Blinder-Oaxaca decomposition to a system of simultaneous equations, describing the structural model of income formation. Their interest is in the change across time of the full distribution of income and on some functions of it (e.g. inequality indices) as in cases (c) and (d) above. The components of their model are an earnings equation for each household member (linking individual characteristics to their remuneration), a labor supply equation (explaining the decision of entering the labor 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 labor 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 labor force and wage formation mechanism, into a household-based decision process, extracting part of the information left in the residuals of linear models described in Section 5.2. Bourguignon et al. (2001) have used this methodology to explain that the apparent stability of Taiwan's income inequality was just due to the offsetting of different forces: the increased wage inequality due to increased return of schooling was partly off-set by the increase in relative weight of middle earners due to changes in participation decisions and a change in the socio-demographic structure of the population. This methodology also allows the consideration of the husband-wife sorting that can induce well-educated working women to marry well-educated men, inducing an increase of household income dispersion. However, the other side of the coin is that the structural model is developed at the expense of increasing the complication of the estimation process and of introducing additional and debatable assumptions.

Among what I consider to be the most important limitations of this approach, the robustness of the estimates of some coefficients is an issue. Bourguignon et al. proposed treating the original estimates from the various cross-sections using "time-smoothing", i.e. replacing all the estimated parameter with their predicted value coming from a simple regression on a time polynomial of order two: such a procedure seems a debatable procedure to reduce variability and induces a reduction of the amplitude of price effects. There is the problem of simultaneity between household members' labor-supply decisions, which is taken into account solely by considering the various household members sequentially, starting from the household head. The issue of understanding what is left in the residuals and how to treat them appropriately

in counterfactual analysis still remains. In particular, there are problems with the labor supply equations - where residuals have to be imputed for the inactive since the residuals of the inactive are unknown - and in the counterfactual wage equations - that determine the effect of the change of "price" coefficients. The path-dependence problem (i.e. which counterfactual is computed first) is also a problem. To have an 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. Finally in Bourguignon et al. (2001) the residual is zero only because it is given the name of "population effect" after the sum of price and participation effect are taken out of the actual change in the distribution of individual earnings. This solution does not say anything on the ability of the structural model to reduce the dimension of the regression residual.

The contribution by DiNardo et al. (1996) is aimed at performing counterfactual analysis to evaluate how much distribution of wages has changed due to changes in the minimum wage, unionization and other characteristics of individuals and of the labor market. In their simulations weights are estimated and then applied to re-weight a kernel density estimation (a simulation of type (ii), according to the scheme above). Their method is however disjoint from kernel density estimation and can be applied to other kinds of functional of data, as for example inequality indices. The change of weights induces a variation of the "relative importance" of different groups in the population. For instance, if wage inequality is compared in two years and the frequency of unionized workers increased from year 0 to year 1, one could

assess what wage inequality would have been if unionization in year 1 was as in year 0 changing appropriately the relative weights between unionized and non-unionized workers. A standard probability frequency could be used here, but DiNardo et al. (1996) proposed a method based on propensity scores for conditioning the probability of being unionized for each individual on her own characteristics. The *actual* density of wage, $f(y)$, in two different years, 0 and 1, can be written as the integral of the density of wages conditional on a set of attributes x (which would also include union status) and the date:

$$
\begin{aligned}
f(y|t_y = 0, t_x = 0) &= \frac{f(y, t_y = 0, t_x = 0)}{f(t_y = 0, t_x = 0)} && (5.15)\\
&= \frac{\int f(y, x, t_y = 0)dx}{f(t_y = 0, t_x = 0)} && (5.16)\\
&= \int f(y|x, t_y = 0) f(x|t_y = 0, t_x = 0)dx && (5.17)\\
&= \int f(y|x, t_y = 0) f(x|t_x = 0)dx && (5.18)\\
f(y|t_y = 1, t_x = 1) &= \int f(y|x, t_y = 1) f(x|t_x = 1)dx && (5.19)
\end{aligned}
$$

where t_x and t_y are the time of individual characteristics and incomes, respectively, and $f(\cdot)$ is a density function. (5.17) comes from (5.15) using Bayes' rule and conditioning probability definition. (5.18) comes from the fact that x is independent from t_y. (5.19) is obtained analogously. The counterfactual density, can be read as "what the density of y in $t = 1$ would be if the distribution of x were as in $t = 0$" and can

be computed as:

$$f(y|t_y = 1, t_x = 0) = \int f(y|x, t_y = 1) f(x|t_x = 0) dx \tag{5.20}$$

Clearly, the only element that changes in the RHS of (5.20) with respect to (5.19) is the conditional density $f(x|\cdot)$. It can happen that the integral in (5.20) can be easily computed but more often this is not the case. The idea of DiNardo et al. (1996) was to estimate a weight

$$\psi_i(x) = \frac{f(x_i|t_x = 0)}{f(x_i|t_x = 1)}$$

so that the counterfactual density function can be easily transformed into a re-weighting of the actual density:

$$f(y|t_y = 1, t_x = 0) = \int f(y|x, t_y = 1) \cdot \psi(x) \cdot f(x|t_x = 1) dx$$

DiNardo (2002) proves that this counterfactual is analogous to the Blinder-Oaxaca one and discusses some of its limitations. Among them there is the fact that re-weighting methods are "methods of ignorance" in the sense that it is not made explicit why the treatment effect (i.e. the probability of being unionized, in the previous example) varies across individuals; that very high or low values of the propensity score are a potential problem since they involve having a very small number of either "treatment" or "control" observations; that errors in estimating the denominator

in $\psi(x)$ can result in an imprecise estimation of $\psi(x)$ itself[9]; that there could be a "selection" problem arising in the estimate of the propensity score which is based only on observables; that for the propensity score to be appropriate, the x variables have to be exogenous, ruling out problems of endogeneity or interactions (for a discussion of this issue, see also Bourguignon et al., 2002). However, the DiNardo et al. (1996) approach is still a valuable methodology as far as underlying household inequality is concerned, and it can be extended to analysis of equivalent household income (see Daly and Valletta (2002) and Chapter 6).

The decomposition proposed by Burtless (1999) is a microsimulation technique based on a rank-dependent transformation that does not involve the estimation of a regression model. It does not consider $g(\cdot)$ as the income-generating function but more simply as the function that aggregates different individual incomes (such as self-employment, employment, etc.) into total individual income and, eventually, into household income. Given a vector $\mathbf{y}$, of dimension N, the rank function at time t, R_t, can be defined as that function that assigns to each and every element y_i a number between 0 and N such that if $R_t(y_i) < R_t(y_j)$ than, at time t, $y_i < y_j$, and vice versa[10]. Hence, given the two actual income vectors at time 0 and time 1, $\mathbf{y}^0$ and $\mathbf{y}^1$ respectively, the counterfactual income vector using the Burtless (1999) simulation method is defined using the inverse of the rank function. For instance, a counterfactual income vector, $\mathbf{y}^c$, may be obtained as the vector of income such that the lowest income at time 1 is replaced with the lowest at time 0, the second lowest

[9]This can cause more than a problem since the estimate $\widehat{\psi}_i$ is performed using fitted values from some discrete choice model, setting the prediction error to zero.

[10]The event of ties is often treated assigning the same integer values to each tie.

with the second lowest, etc. and can be expressed as:

$$\mathbf{y}^c = R_0^{-1}(R_1(\mathbf{y}^1)) \tag{5.21}$$

Clearly, the extension of the rank function to a non-integer codomain gives the (discrete) quantile function and in the limit, when y can be regarded as a continuous variable, the counterfactual can be written using the inversion of the cumulative density functions[11].

The Burtless (1999) decomposition can be seen as a step forward to overcome one limitation of the Shorrocks (1982) decomposition, namely the invariance of the decomposition rule to the inequality index. Looking at the evolution of overall distribution as the combination of (α) the evolution of relative share of different sources in total income, (β) the evolution of the marginal distribution of each income source and (γ) the modification of the correlation between the different income sources, the transformation in equation (5.21) corresponds to see what are the changes in overall distribution keeping (β) fixed at a given year, without varying (α) and (γ).

The problems with residuals do not arise here simply because no stochastic income-generation model is ever considered. With this decomposition it is assumed that without caring about the individual personal characteristics, either observable or non-observable, the least paid worker will not change her decision to work in a given year (it is not introduced a change in labor force participation) and she will only be

[11]Since the attention is pointed at the effect of changing dispersion, total incomes is normalized so that they do not change in the different years.

able to get the least paid job on the market. Hence, it is assumed that there is no randomness as to how individuals are assigned which wage out of the ordered set of available wages. This assumptions could seem overly restrictive to some researchers, at least clear and easy to understand to others. This approach does also affect the mobility of the individual income although when focusing on inequality we are adopting the anonymity principle, according to which inequality does not change if two individuals swap their income, hence their position, in the income ranking. Analysis of mobility is also relevant to understand the forces underlying the evolution of inequality, however, in all the microsimulation-based methods considered here, mobility issues are not analyzed. Burtless (1999) applies the rank-dependent transformation to a sample of the US male and female wage distributions in two different years, and extends the simulation to incorporate marriage patterns matching male and female wage distributions accordingly. He finds that the main factors that induced American inequality to increase after 1979 were the growing correlation of husband and wife earned incomes and the increasing percentage of Americans who live in single-adult families, typically characterized by more unequal incomes than husband-wife families. An analogous matching could be performed between different marginal distribution of the same individual (for instance, see Fournier (2001) and Chapter 6).

5.4 Conclusions

This chapter has reexamined the problem of decomposing income inequality in an informative manner. Recent developments in the literature have focused on method-

ologies for overcoming some of the limitations of the "traditional methods" for income decomposition. For instance, although the "traditional approach" has reached a high level of formalization it is still unable to explain the causes of inequality and is subject to the criticism of being merely a descriptive tool, unable to provide convincing policy implications and to be unsuitable to extensions to the analysis of trends of inequality. As discussed in Section 5.2, regression-based inequality decomposition has proved to be unsatisfactory to overcome the limitations of the "traditional decomposition". In fact, the main problems lay in the inadequacy of a single equation model to explain household income, to extract enough information out of the total variability of the inequality index, to extend the methodology to inequality trends. Simulation-based methods seem more promising, although there is so far no single established methodology among the variety of those available. They all rely on building some counterfactual income vectors to be compared with actual outcomes: some are based on developing a comprehensive model to understand the process that generated the data, others are based on re-weighting the sample, others on inverting the rank function. Recently some authors have tried combining different decomposition methods to provide a broader picture of income inequality trends (for instance, see Daly and Valletta, 2002). In Chapter 6 a combination of two microsimulation methods is developed to analyze inequality trends in Italy in the period 1977-2000. It allows one to understand the role of demographic factors, such as the increased female labor force participation, and of the changed dispersion of different income sources, assessing their relative importance for the actual change in inequality.

Chapter 6

Understanding inequality trends in Italy

According to recent comparative studies on OECD countries, the highest income inequality is found in the US, followed by the UK and Italy, the latter two presenting similar figures using standard inequality measures (Atkinson et al., 1995; Smeeding, 2000). However, while the US and the UK present a roughly increasing trend of income inequality since the 1970s, Italian household income distribution exhibits substantial fluctuation but no clear trend (Brandolini and D'Alessio, 2001; D'Alessio and Signorini, 2000). From their decomposition of income distribution by population subgroups Brandolini and D'Alessio (2001) find that demographic characteristics, such as household size, sex of household head, age class of household head and household composition, are able to explain only a limited amount of overall inequality but do not investigate the issue further.

This chapter contributes to the empirical analysis of Italian household inequal-

ity and its determinants by assessing the role of the changed dispersion of different income factors and of the demographic evolution on household income distribution. It also contributes to the literature on household income inequality decomposition by proposing a unified framework for two different microsimulation methods for decomposing inequality and overcoming some limitations of traditional methodologies. Such a decomposition takes into consideration the dispersion of income sources as well as socio-demographic factors, across many years. This methodology maximizes the clarity of results and allows one to study the reliability of the estimates. For robustness of conclusions, more than one inequality index is considered.

Section 6.1 reviews the available evidence about Italian household inequality. Section 6.2 discusses the data, hypotheses and aims of the investigation. Section 6.3 describes the methodology adopted and Section 6.4 presents results. Section 6.5 concludes.

6.1 Analysis of Italian household income distribution: available evidence

In recent years Italy has experienced important demographic and social changes. The population has grown older, the family structure has changed, female labor participation has steadily increased. The impact of some of these demographic changes have been studied in some detail in recent papers, mainly using the Bank of Italy SHIW-HA data set, and their findings are relevant to this chapter.

D'Alessio and Signorini (2000) focused on the role of the household for reducing inequality from work and transfer income and found that, while inequality among income receivers exhibits a clear downward trend since 1977, household inequality presents no trend but substantial fluctuation. Using a decomposition of the Gini index, they explained the decrease of inequality among income receivers in terms of the increase of the number of people receiving income from work, mostly because of an increased female labor force participation, and of the augmented number of people receiving pension income. Using SHIW data Brandolini and Sestito (1993) showed that household inequality tends to have pro-cyclical patterns. These results have been found also using a different data set, that from ISTAT. Brandolini and D'Alessio (2001) focused on household inequality and analogously described the trend of inequality as having "many fluctuations /.../ but no particular medium term tendency" (p. 2). Using the Luxembourg Income Study data set, they also pointed out that elderly Italian households (where the head is over 65) have a higher income than analogous household in other OECD countries. However, their decomposition of the mean logarithmic deviation index trend by population subgroups, such as household size, sex of household head, age class of household head and household type, has not been very satisfactory in understanding the causes of equivalent income inequality. The greatest change is found in the classification by sex of the household head. If the composition of the household heads in 1977 had been as it was in 1995, overall inequality would have been 3.3% higher, mainly due to the greater weight attributed to women, among whom, they say, dispersion was higher. Neither has regional dualism been found to provide useful insights for inequality dynamics.

Baldini (1996) compared the level of household inequality in the period 1987-1993 using the decomposition by "factor components". Using the Gini and $GE(2)$ indices, he claims that the increase of household inequality in the period considered was mainly driven by increased relevance of pension and capital income in household income. However, Baldini (1996) reaches this conclusion comparing the share of inequality explained by different sources in different years, without developing a "factor components" decomposition for inequality trend as Jenkins (1995) did using counterfactuals (see Section 5.1, page 110). Such a methodology is then debatable since it does not properly handle demographic evolution.

As final remark about income inequality during the 1990's it should be noted that during the early 1990s the Italian economy went through one of the deepest recession after the WWII and returned to growth only after 1995. The bad conditions of Italian public finances characterized the 1990s as a period of rising fiscal burden, which in same cases badly affected the poorest households (for an analysis of the 1998 IRPEF reform and the effect on non-employed households, see Chapter 3).

Although not focused on household inequality but on individual inequality, other contributions are relevant for this chapter. Brandolini et al. (2001) focused mainly on primary-job earnings and found a clear downward trend of inequality up to the late 1980s followed by a marked increase in the early nineties. They also found that changes of early 1990s were mainly concentrated among workers at the margin of the labor market; that diffusion of low-paid jobs evolved in parallel with the increase of earnings inequality; and that the probability of being in poverty was more closely correlated with the amount of employment in the household (particularly employment

of members other than the head), rather than with low pay. Moreover, during the period considered, employment fell, exacerbating the cost of the early 1990s recession for low skilled and low experienced workers, who suffered from the deficiencies of the Italian safety net and unemployment benefit scheme (Brandolini et al., 2001).

Erickson and Ichino (1995) and Manacorda (2002) have also studied the distribution of wages in Italy in recent years finding evidence of an effect of the abolition of the automatic indexation of wages, which took place gradually at the end of the 1980s up to 1991. Erickson and Ichino (1995) found that even those workers who kept the same job throughout the period suffered changes in their relative wages, since the automatic indexation of wages was abolished in 1991 and the contribution relief for firms in the South was gradually stopped since 1994, increasing the volatility of wages. Using data up to 1991 partly coming from SHIW and partly from a private source (Federmeccanica-Assolombarda), Erickson and Ichino (1995) concluded that at the end of the 1980s Italy pesented a compressed wage structure which had not experienced the decompression seen elsewhere during the 1980s. Moreover it could be that the spread of part-time and fixed-term employment contracts and the effect of institutional changes had unleashed a decompression of the wage structure, resulting in a larger dispersion of incomes already at work in other countries as shown in Manacorda (2002), using SHIW data. Manacorda also argues with counterfactual analysis that, had the automatic wage indexation been inoperative, earnings inequality in Italy would have increased at a rate similar to that observed in the US.

Finally, Devicienti (2003) analyzed wage inequality using administrative data from the Italian Institute for Social Security (INPS) from 1985 to 1996. These data show

an analogous, though moderate increase of wage inequality in the early nineties. Devicienti (2003) finds a story consistent with earlier papers: earnings have become more dispersed because more senior and skilled workers have been able to obtain greater reward in the labor market after the abolition of the wage indexation mechanism.

The clear evidence about wage inequality trend has not yet been fitted into household inequality analysis. It still is unclear why household inequality is so different from individual inequality and what is the relevance of self-employment and pension income to explain household inequality trend.

6.2 Data, hypothesis and aims

6.2.1 The data set: pros and cons

The SHIW data set, based on interviews run about annually from 1966 to 1987 and about every two years thereafter, collects detailed information on income, wealth, consumption and individual characteristics relative to a sample of resident Italian households. The 1998 data set was described in detail in Section 2.1. Since 1998 the Bank of Italy gathered all SHIWs starting from 1977 and made them consistent in a Historic Archive (SHIW-HA). The last version of the Historic Archive at present covers the period 1977-2000.

Although it is a very attractive data set, it is important to reflect on some of its limitations. As any survey-based data set obtained though direct interview whose participation is voluntary, the SHIWs present problems of non-response or under-

reporting, especially for sensitive data such as income and wealth, problems of low response rates as well as of under representation of household living in very isolated places of the country, whose interview are very costly. Moreover, the SHIW-HA presents the additional limitation of being a collection of data sets: besides recording the same variables and being developed by the same institution, in some cases sample designs and dimensions was not constant through time. For instance, a first important change in the sample selection was introduced in 1984, with units no longer from electoral lists, but rather from registry office records, removing the over-sampling of large households. In 1986 the sample design was completely revised and the number of households interviewed more than doubled. In 1987 there was an over-sampling of high-income households. After 1989 onward, instead, the sampling methodology did not change and the dimension of the sample remained about constant[1]. Some of these shortfalls have been corrected with various sets of sampling weights but the data should still be analyzed with caution (for a comprehensive discussion of the data set, see Brandolini, 1999).

Despite these problems, the SHIW-HA is the only data set that permits measurement of the changes in the whole Italian income distribution through time and relate it to household characteristics and income components.

[1]Actually, in 1989 and 1991 there was a dramatic drop of response rate but this could be due to the fact that interviewers started to be paid also for non-responding unit. The decline could thus be explained either by the under-reporting of non-responding units in previous surveys, or by a tendency to inflate non-responses in those years, or both (Brandolini, 1999). However, inequality indices and quantile ratios, performed also for different kind of income, do not present high variability for all years after 1989.

6.2.2 Description of demographic trends

As for demographic changes[2], the age groups decomposition shows a decrease by about 20% of cohorts younger than 30 and an increase by 45% of the over 65 during the 23 year period considered. The former group was about 43% and the latter about 12% of total population in 1977, at the end of the period they were respectively 34% and 18%. There has been some increase also in the cohort 31-65, mainly due to the sons of the 1960s Italian "baby boom" (Table 6.1 and Figure 6-1).

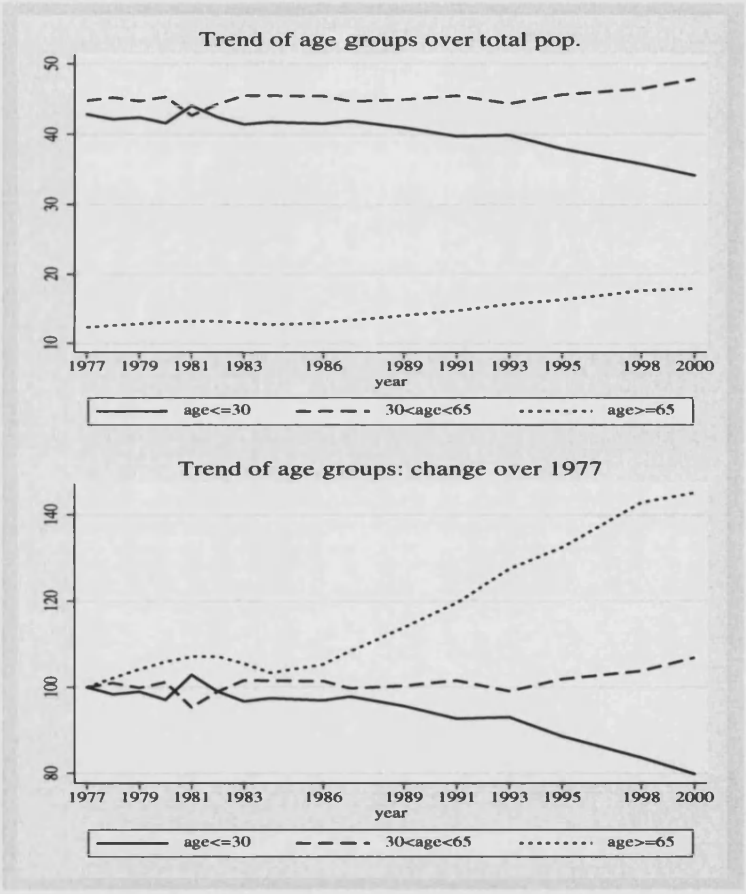


Figure 6-1: Decomposition of the population by age groups

[2]In this chapter SHIW-HA sampling weights are used. The grossing-up procedure described in Section 2.4 was not used because too complicated to be applied to 16 different data sets.

Year	<30 yrs		31<yrs<40		41<yrs<50	
	total	%	total	%	total	%
1977	24,500,000	42.84	7,181,393	12.55	8,082,990	14.13
1978	24,700,000	42.14	7,608,596	12.99	8,000,017	13.65
1979	24,900,000	42.39	7,385,158	12.56	8,462,309	14.39
1980	24,800,000	41.58	7,502,312	12.59	8,193,308	13.75
1981	26,900,000	44.11	7,804,814	12.77	7,910,079	12.95
1982	25,300,000	42.42	7,438,001	12.48	8,242,708	13.83
1983	25,200,000	41.43	7,705,300	12.66	8,633,686	14.19
1984	24,700,000	41.74	8,135,767	13.77	8,223,285	13.92
1986	24,300,000	41.52	8,301,395	14.21	7,853,034	13.44
1987	24,500,000	41.90	7,934,049	13.56	8,037,568	13.73
1989	24,300,000	40.96	7,592,147	12.81	8,504,180	14.34
1991	23,700,000	39.72	7,931,810	13.27	8,276,795	13.85
1993	23,300,000	39.87	8,400,389	14.38	7,827,936	13.40
1995	22,500,000	37.98	8,693,111	14.67	7,821,800	13.20
1998	22,200,000	35.85	9,329,555	15.07	8,507,282	13.74
2000	20,400,000	34.19	9,562,311	15.99	8,243,886	13.78

Year	51<yrs<65		over 65		Total	
	total	%	total	%	total	%
1977	10,400,000	18.11	7,072,426	12.36	57,236,809	100
1978	10,900,000	18.59	7,399,249	12.63	58,607,862	100
1979	10,500,000	17.78	7,576,160	12.88	58,823,627	100
1980	11,300,000	18.99	7,803,431	13.09	59,599,051	100
1981	10,300,000	16.93	8,091,380	13.24	61,006,273	100
1982	10,700,000	18.01	7,895,534	13.25	59,576,243	100
1983	11,400,000	18.67	7,941,044	13.05	60,880,030	100
1984	10,500,000	17.79	7,547,575	12.78	59,106,627	100
1986	10,400,000	17.81	7,612,212	13.03	58,466,641	100
1987	10,200,000	17.41	7,846,751	13.41	58,518,368	100
1989	10,600,000	17.81	8,345,906	14.08	59,342,233	100
1991	11,000,000	18.39	8,832,156	14.78	59,740,761	100
1993	9,695,718	16.60	9,198,317	15.75	58,422,360	100
1995	10,500,000	17.79	9,700,094	16.37	59,215,005	100
1998	11,000,000	17.69	10,900,000	17.66	61,936,837	100
2000	10,800,000	18.11	10,700,000	17.93	59,706,197	100

Source: own calculation on SHIW-HA data

Table 6.1: Decomposition of the population by age groups

The proportion of single-person households more than doubled, so that in 2000 one out of five households had this structure. The proportion of single-parent households with children increased by over 30%, while that of couples with kids decreased by over 20%. Female-headed households became markedly more frequent in the last decades as well as the average dimension of household showed a clear downward trend (Tables 6.2, 6.3 and Figure 6-2). During the period in question, the importance of the male household head income became less relevant, partly because of the increased number of female-headed households and partly because of the increased labor force participation of the other members of the household (Figure 6-3)

Year	cpl. w/ kids	cpl. no kids	sng w/ kids	sng no kids	single only	Male HH head	Female HH head
1977	58.85	21.65	7.16	2.46	9.88	88.15	11.85
1978	59.85	20.96	7.07	2.29	9.83	86.97	13.03
1979	57.54	20.13	6.74	2.17	13.42	86.28	13.72
1980	57.67	20.62	7.06	2.89	11.76	85.68	14.32
1981	57.32	20.23	7.64	2.00	12.80	84.82	15.18
1982	59.20	21.14	6.54	2.45	10.66	87.78	12.22
1983	57.74	20.94	6.37	2.62	12.33	85.17	14.83
1984	56.82	19.54	7.64	1.71	14.29	84.07	15.93
1986	55.65	20.32	7.38	2.14	14.52	81.81	18.19
1987	56.62	18.34	7.66	2.63	14.75	81.82	18.18
1989	53.10	19.58	7.82	2.19	17.32	80.45	19.55
1991	52.82	19.09	8.24	1.64	18.21	78.81	21.19
1993	51.42	19.05	9.55	2.46	17.53	71.94	28.06
1995	50.29	19.52	9.15	2.73	18.31	71.67	28.33
1998	47.18	20.61	8.86	2.67	20.69	71.94	28.06
2000	45.43	21.71	9.51	2.48	20.87	64.68	35.32

Source: own calculation on SHIW-HA data

Table 6.2: Decomposition of the population by household type

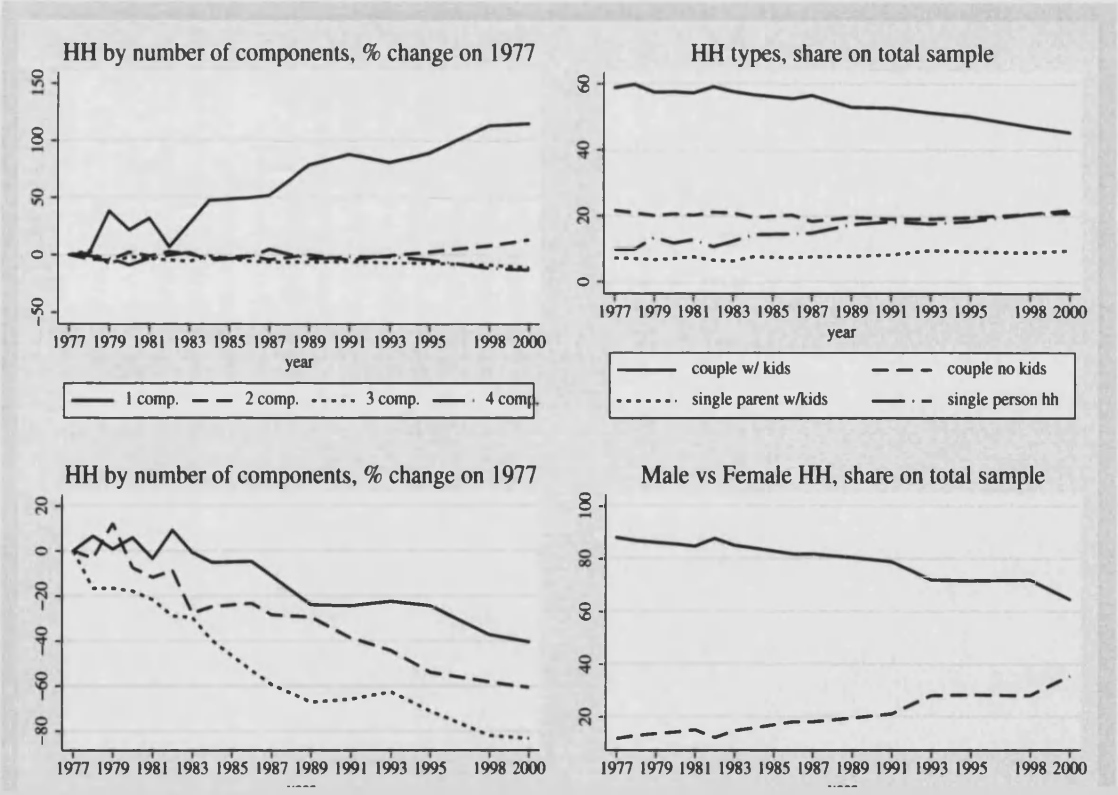
According to the SHIW-HA data, total labor force participation (LFP)[3] had a stable path up to 1986, increased significantly up to 1989 (between 10% and 20% depending on the wage groups considered) and then remained stable at higher values

[3]Total LFP is computed as percentage of working age individuals - i.e. between 15 and 65 years - who declare to be either working or actively looking for an occupation.

Year	1 comp	2 comp	3 comp	4 comp	5 comp	6 comp	+6 comp
1977	9.69	24.79	25.39	24.00	9.79	3.92	2.42
1978	9.83	23.85	26.30	23.76	10.44	3.79	2.02
1979	13.41	23.61	23.39	23.30	9.88	4.40	2.02
1980	11.79	25.43	24.98	21.81	10.38	3.62	1.99
1981	12.79	24.57	24.58	23.26	9.45	3.46	1.90
1982	10.36	25.49	24.24	23.91	10.70	3.58	1.72
1983	12.33	24.96	23.94	24.50	9.72	2.84	1.71
1984	14.29	23.97	25.19	22.85	9.29	2.95	1.46
1986	14.52	24.54	23.95	23.51	9.33	3.01	1.13
1987	14.75	23.76	23.78	25.20	8.71	2.81	0.99
1989	17.32	24.82	23.71	23.14	7.45	2.77	0.80
1991	18.21	23.70	23.86	23.57	7.41	2.41	0.83
1993	17.53	24.64	23.53	23.60	7.61	2.19	0.91
1995	18.31	25.41	23.47	22.89	7.41	1.81	0.70
1998	20.69	26.79	23.12	21.17	6.16	1.64	0.44
2000	20.87	28.05	22.52	20.78	5.84	1.55	0.41

Source: own calculation on SHIW-HA data

Table 6.3: Decomposition of the population by number of components



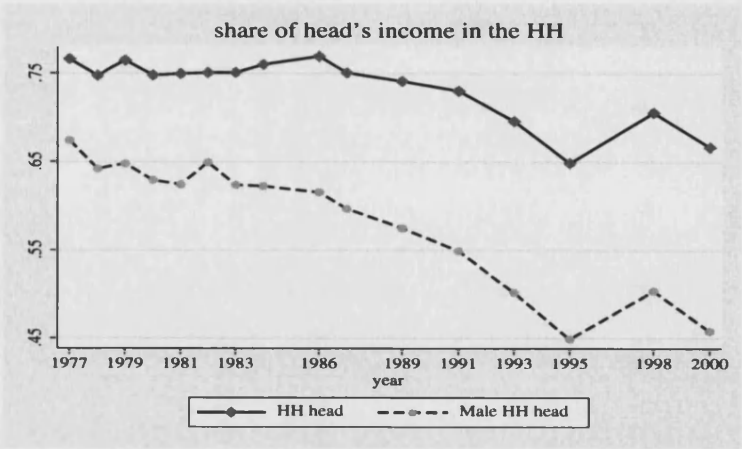


Figure 6-3: Decomposition of the population by HH type

in the following period. However, while male LFP has been fairly stable throughout the period, the increase has been marked for female LFP. This dynamics reduced the differential of male-female LFP by about 10% points. It should also be noted that the very high variability of LFP figures up to mid 1980s was probably due to the small sample size[4] (Table 6.4 and Figure 6-4).

Another feature of the evolution of the population which has attracted the attention of researchers is the change of the ratio between income receivers and number of members of the household (Figure 6-5). Over the period, on average about 35% of the household members received labor income and this percentage remained stable for the whole period. On the other hand, pension income was received on average by an increasing proportion of individuals in the household and, in particular since 1993, the proportion of individuals receiving pension income was higher than the proportion of individuals receiving labor income (regardless of their respective amounts). A similar trend is found also looking at the population instead than at an average

[4]Individual sample size was about 10,000 before 1980, about 13,500 between 1981 and 1984, not less than 20,900 in the rest of the period.

	Total LFP over 15		Total LFP over 24		Male LFP over 15	
Year	total	%	total	%	total	%
1977	19,200,000	48.37	18,100,000	54.67	13,800,000	71.61
1978	19,700,000	49.75	18,500,000	56.14	14,000,000	72.54
1979	20,300,000	51.08	19,100,000	58.00	14,200,000	72.36
1980	20,300,000	50.76	19,300,000	57.52	13,800,000	70.19
1981	20,100,000	51.23	19,000,000	57.63	13,700,000	71.70
1982	20,000,000	49.03	19,000,000	56.94	13,800,000	68.78
1983	20,400,000	49.38	19,400,000	56.67	14,300,000	70.00
1984	20,500,000	51.75	18,800,000	59.94	14,100,000	72.14
1986	19,500,000	49.34	17,800,000	56.78	13,400,000	68.40
1987	20,700,000	52.18	18,800,000	59.52	13,800,000	70.58
1989	23,900,000	59.73	19,800,000	62.71	15,200,000	77.73
1991	23,900,000	59.56	20,600,000	63.98	14,900,000	75.06
1993	23,800,000	60.56	20,500,000	65.22	14,800,000	76.32
1995	23,800,000	60.21	21,000,000	64.63	14,600,000	74.34
1998	24,000,000	61.22	21,500,000	65.34	14,400,000	74.88
2000	23,600,000	60.19	21,600,000	64.72	14,600,000	74.70

	Male LFP over 24		Female LFP over 15		Female LFP over 24	
Year	total	%	total	%	total	%
1977	13,200,000	82.17	5,417,333	26.53	4,914,579	28.80
1978	13,300,000	83.38	5,701,033	28.09	5,156,085	30.44
1979	13,600,000	82.77	6,118,284	30.40	5,565,376	33.55
1980	13,100,000	80.50	6,566,139	32.13	6,113,980	35.64
1981	13,100,000	82.38	6,440,035	31.90	5,911,027	34.64
1982	13,300,000	80.57	6,203,209	29.92	5,768,463	34.01
1983	13,600,000	81.03	6,153,319	29.35	5,786,168	33.21
1984	13,200,000	84.64	6,455,095	32.00	5,683,182	35.78
1986	12,300,000	81.82	6,162,132	30.75	5,410,471	33.44
1987	12,700,000	82.52	6,935,503	34.35	6,089,851	37.61
1989	13,100,000	84.92	8,615,103	42.36	6,726,816	41.57
1991	13,100,000	83.65	8,966,156	44.34	7,522,508	45.41
1993	13,000,000	83.95	8,944,273	45.12	7,481,286	47.00
1995	13,000,000	81.61	9,238,063	46.30	7,952,013	48.19
1998	13,000,000	81.58	9,521,853	47.95	8,446,128	50.00
2000	13,400,000	81.48	9,033,466	45.82	8,129,078	48.30

Source: own calculation on SHIW-HA

Table 6.4: Labor force participation: Total, by sex, by age

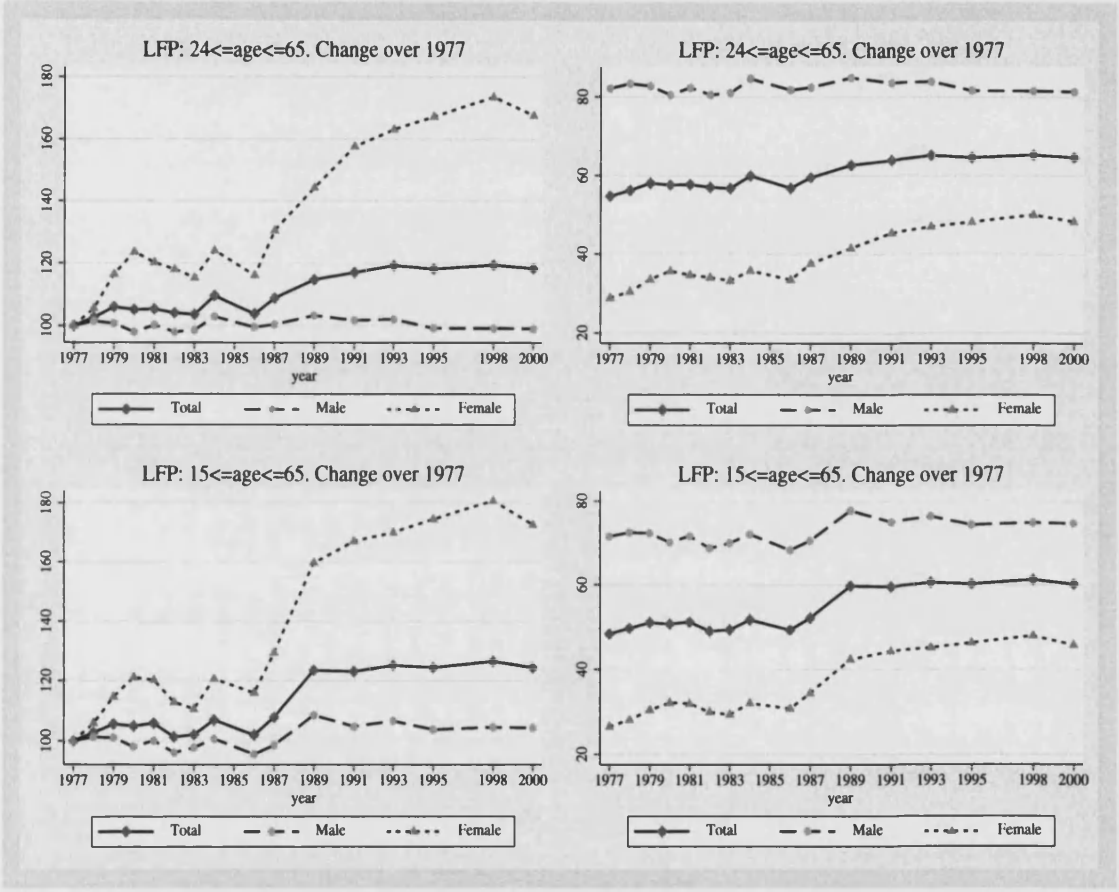


Figure 6-4: Labor force participation: Total, by sex, by age

household.

It should also be noted that the increased proportion of pensioners is only partly due to the fact that life expectancy increased in recent years. As can be seen in Figure 6-5, the proportion of pensioners over 65 did not change significantly, and the increased proportion of pensioners on overall population was induced mainly by the increase of young pensioners (between 45 and 65 years of age). This phenomenon was partly due to the pension system, which allowed very early retirement, some times even before workers were in their forties, without linking the pension to the pensioner's life expectancy. The phenomenon of the so-called "baby pensioners" has been partly reduced with the 1993 and 1995 pension reforms.

6.2.3 Preliminary hypothesis for inequality analysis

This chapter focuses only on disposable income per equivalent adult, using the LIS equivalence scale[5]. It involves assigning to each individual the total income of her household multiplied by $N_h^{-\epsilon}$, where N_h is the number of components of household h ($h = 1, 2, ..., H$) and $\epsilon = 0.5$. Using the LIS equivalence scale it is also assumed that intra-household allocation is egalitarian, i.e. that all members of the household get the same share of income, regardless of their individual income, their role in the household and other characteristics. The individual equivalent income (also referred to as household equivalent income) is considered as the elementary unit of analysis. As

[5]Brandolini and D'Alessio (2001) and D'Alessio and Signorini (2000) also used the LIS equivalence scale. However, it was verified that the conclusion of the present chapter are not strongly dependent of the type of equivalence scale used, changing the value of the parameter ϵ to 0, 0.25, 0.75 and 1. All results from this sensitivity analysis are not presented here for reason of space but they can be obtained from the author.

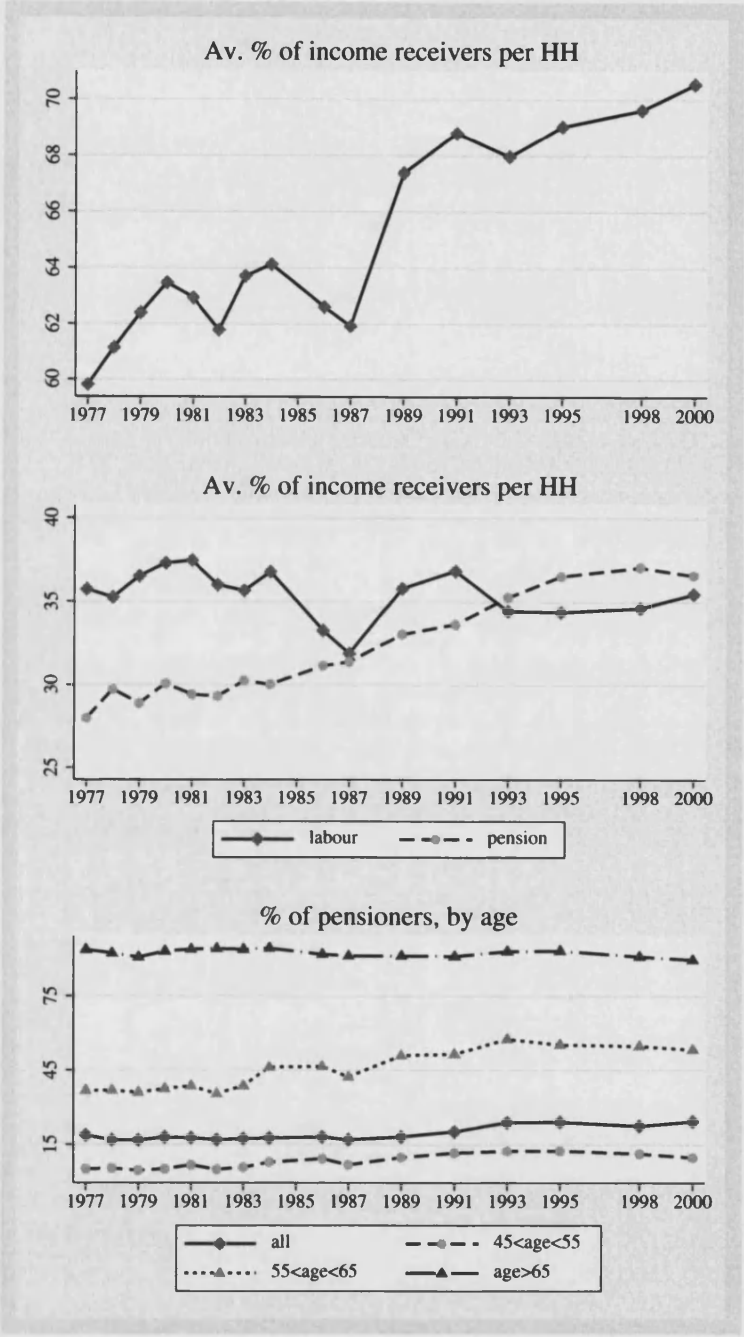


Figure 6-5: Frequency of income earners in av. HH or population

in D'Alessio and Signorini (2000) only income by work and transfers is considered, excluding income from capital since it presents measurement problems and is not uniformly available for all considered years. Finally, the 2.5% of poorest households are attributed the equivalent income of the household at the 2.5th percentile, mainly because some of the inequality indices used here cannot be computed for non positive incomes. This is a peculiar choice since it is generally more frequent to disregard the drop of zero or negative observations. It was however taken to make the comparison across time consistent without losing any data.

Three different inequality indices are considered: the Generalized Entropy (GE) indices, with $a = 0, 1, 2$. These indices are chosen because they should provide a broad picture of the distribution. In fact, these inequality indices differ in their sensitivity to difference in various parts of the distribution: the more positive the parameter a of the GE class is the more $GE(a)$ is sensitive to income differences at the top of the distribution, the smaller a is the more $GE(a)$ is sensitive to differences at the bottom of the distribution (see Appendix C in Section 6.6). Moreover, the use of GE indices allows us to compute their confidence intervals using asymptotic distributions (Cowell, 1989). In some cases, the Gini index and quantile ratios will also be used, since quantiles relative to the median can provide useful insights on where the main changes in the distribution happened[6]. Let y_i be the i-th element of the income vector $\mathbf{y}$ in ascending order, with $i = 1, ..., N$, and w_i be the corresponding sample weight of y_i, with $\sum_{i=1}^{N} w_i = N$. The q-quantile, with $q \in [0, 1]$ can be defined as

[6]Moreover, as proved by Cowell and Victoria-Feser (1996), the quantile function is the only inequality index robust to outliers.

$y_j \in \{y_i, i = 1, ..., N\}$ or its nearest value such that

$$q = \sum_{i=1}^{j} \frac{w_i}{N}$$

6.2.4 Analysis of inequality estimates and data contamination issues

Figures 6-6, 6-7 and 6-8 report inequality indices and quantile ratios for individual monthly incomes[7]. It can be seen that inequality is much higher among self-employed income receivers, and it is generally higher for pensioners than for employed people. As for the trend, it appears decreasing up to the end of the 1980s, increasing between the 1991 and 1993, and finally fairly stable for all types of income, though there are much more fluctuations for self-employment income. The $GE(2)$ index increase after 1980s confirms other researchers' findings that major changes in employment income happened in top incomes[8]. Figure 6-9 shows the share of different types of incomes on household income and some inequality indices and quantile ratios for equivalized household income. As can be seen from the first panel, the share of employment income on household income has been decreasing constantly, pension income has increased at least up to 1993[9], while self-employment has fluctuated[10]. Restricting

[7]Monthly incomes are derived from yearly incomes using information available in the data set.

[8]The $GE(2)$ index was not included in the self-employment and pension income figures since the high spikes presented in the first part of the sample, would make the scale of the figures unclear.

[9]In 1993 and in 1995 two major pension reforms were approved, which among other things aimed at reducing the ratio of paid pension/DGP.

[10]It should be noticed that this type of descriptive statistics yields limited information since, for instance, the increase of the share of pension income does not take into account that households have become older across time.

attention to the period post 1989, all indices show a slight decrease from 1989 to 1991, when the minimum was reached, and a dramatic increase in 1993, followed by a slight decrease only from 1998 to 2000. Comparing their numerical values in 1991 and in 2000, the Gini coefficient increased by 16.2% and the $GE(0)$, $GE(1)$ and $GE(2)$, increased by 46.3%, 41.8% and 55.6% respectively. Since the $GE(0)$ and $GE(2)$ indices are more sensitive to lower and to higher incomes respectively, these changes show that more of the action in the changing income distribution occurred at the bottom and at the top, respectively. This conclusion is confirmed using quantile ratios. In the top quantiles there has been a clear rise of the richer quantiles, the 95/50 ratio increasing comparatively more that the 90/50 ratio, which still showed more dynamics than the relatively stable 75/50 ratio. Changes in the distribution were also quite pronounced in the lower quantiles where the widening of the gap between the lower quantiles and the median is evident: even the 25th percentile saw a decrease with respect to the 50th percentile and only after 1998 signs of improvement appeared.

Based on Figure 6-9 and Tables 6.5 and 6.6 and looking at the whole 1977-2000 period, one could perhaps agree with Brandolini and D'Alessio (2001) and D'Alessio and Signorini (2000) that there is no clear trend, but rather a substantial fluctuation in household inequality. However, looking at the quantile ratios, it should also be pointed out that weird spikes appear in connection with 1980 and 1987. In particular if the year 1987 were removed, the trend would indeed appear to be decreasing up to 1991, then increasing in the following years, similarly to the trend of individual income inequality.

These considerations lead to consider issues of data contamination in years 1980 and 1987 rather than to rule out the presence of a trend. Data may be contaminated as a result of recording errors, measurement errors, data collection problems, and alike. Data contamination reduces the quality of the data introducing a bias that, if not removed, can seriously mislead the analysis. Part of the possible contamination is corrected by the Bank of Italy before publishing the data (see Section 2.1 for 1998 SHIW), but some remains. Although the Historic Archive (SHIW-HA) covers an attractively long period of time and have proved to be useful for the analysis of primary-job income dispersion (Brandolini et al., 2001), it seems to present some limitations for analysis of household income in particular years. Some of these problems may be due to changing sampling procedures (see Section 6.2.1, page 137), some others to difficulties during data collection (see note 1, page 137). Data quality of SHIW-HA is an issue until mid 1980s, before sample size was significantly increased (see note 4, page 142). Data quality might also be a problem in 1987 because of the over-sampling of high income households (see Section 6.2.1, page 137). One possible, though drastic, solution in this case is to remove the contaminated observation. Removing 1987 from the analysis makes the household inequality trend to resemble income-receivers inequality trend quite closely, with a trend that is decreasing up to the end of 1980s and than increasing. Another solution is to try correcting the contamination. Section 6.4 will show how microsimulation can be useful to correct for possible data contamination bias.

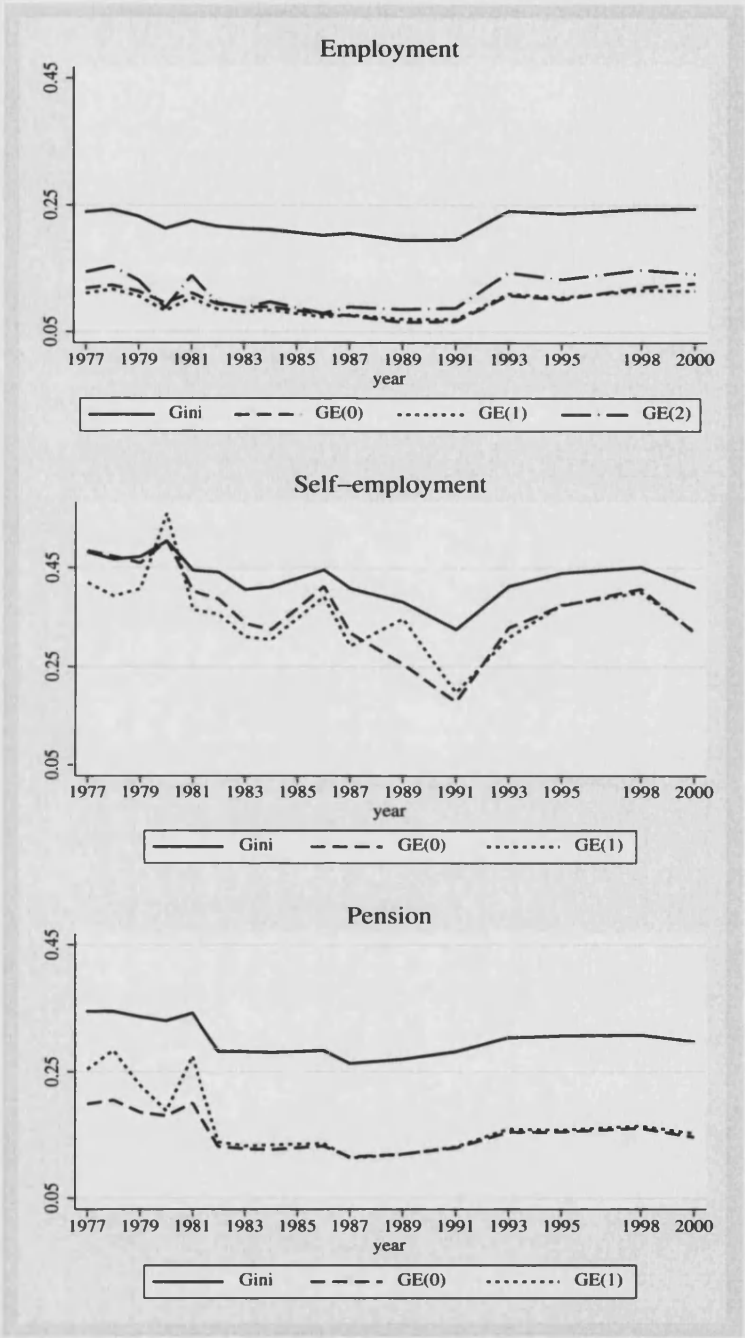


Figure 6-6: Inequality indices, individual (monthly) incomes

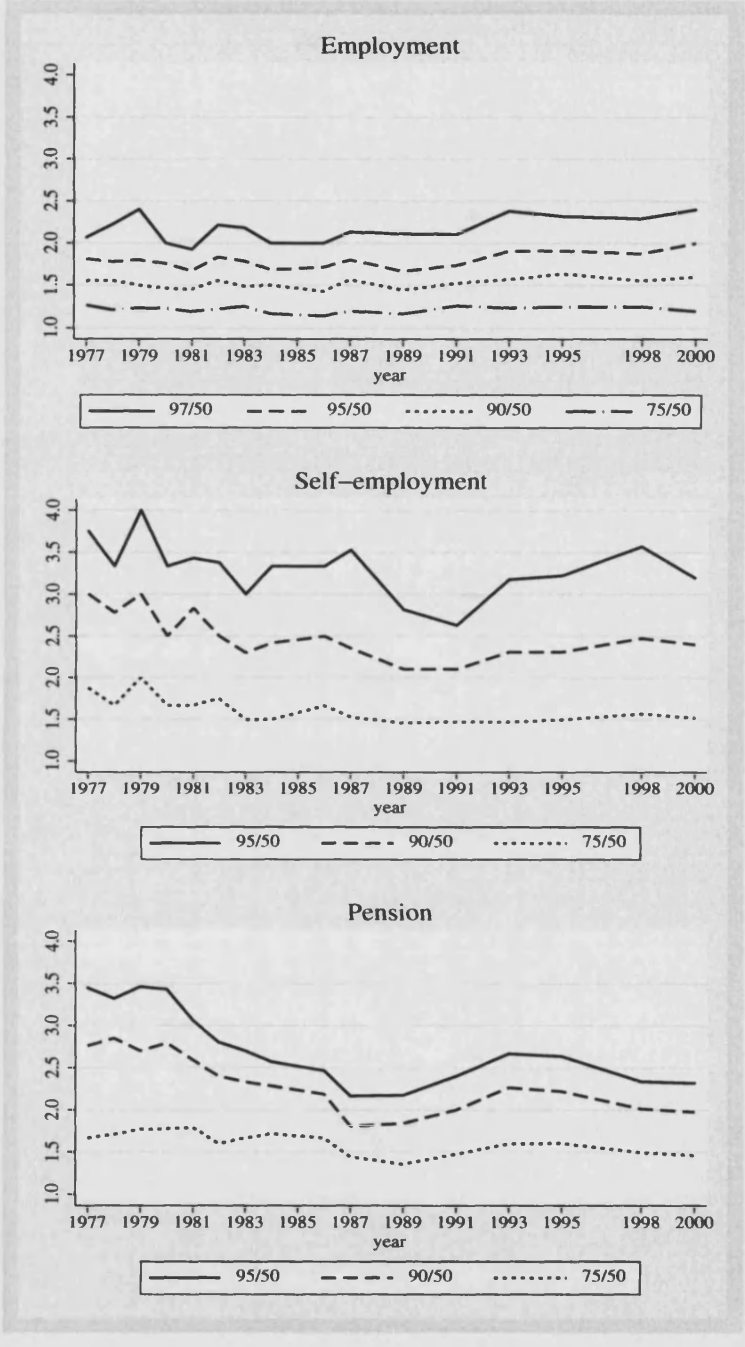


Figure 6-7: Inequality indices, individual (monthly) incomes

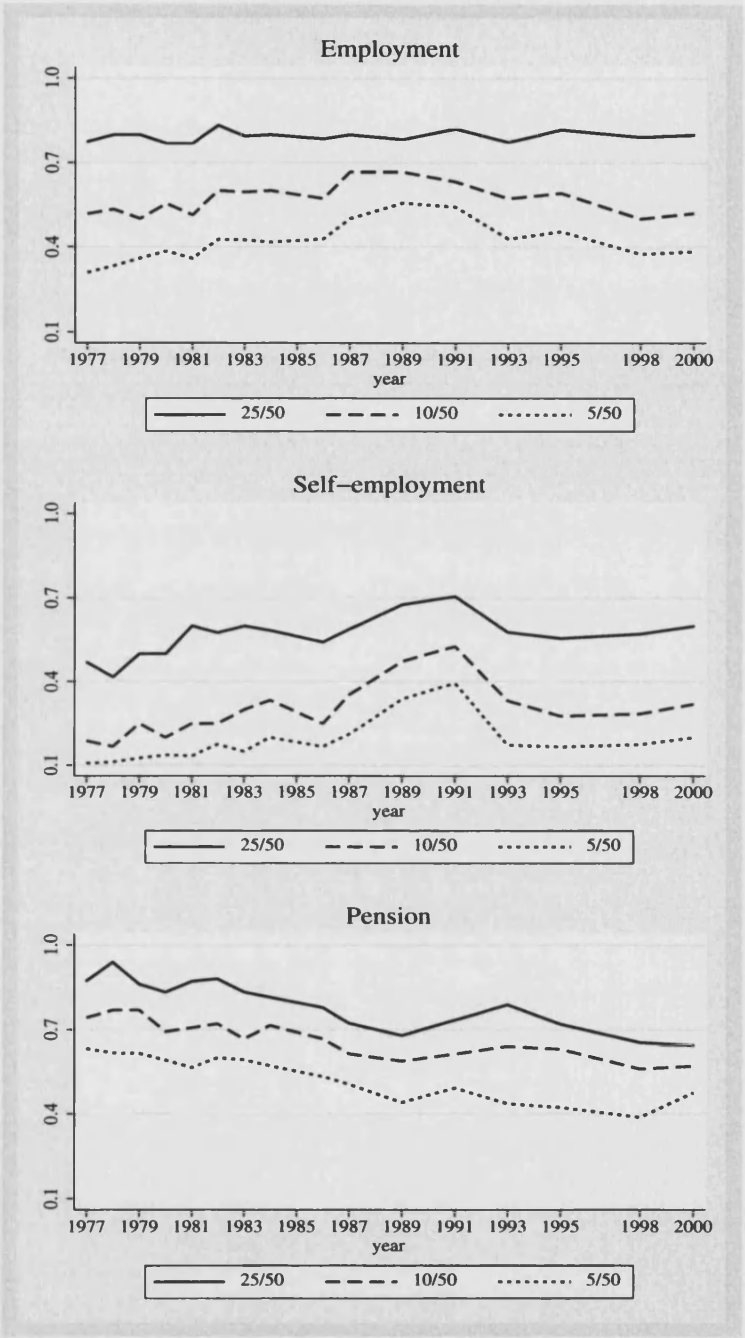


Figure 6-8: Inequality indices, individual (monthly) incomes

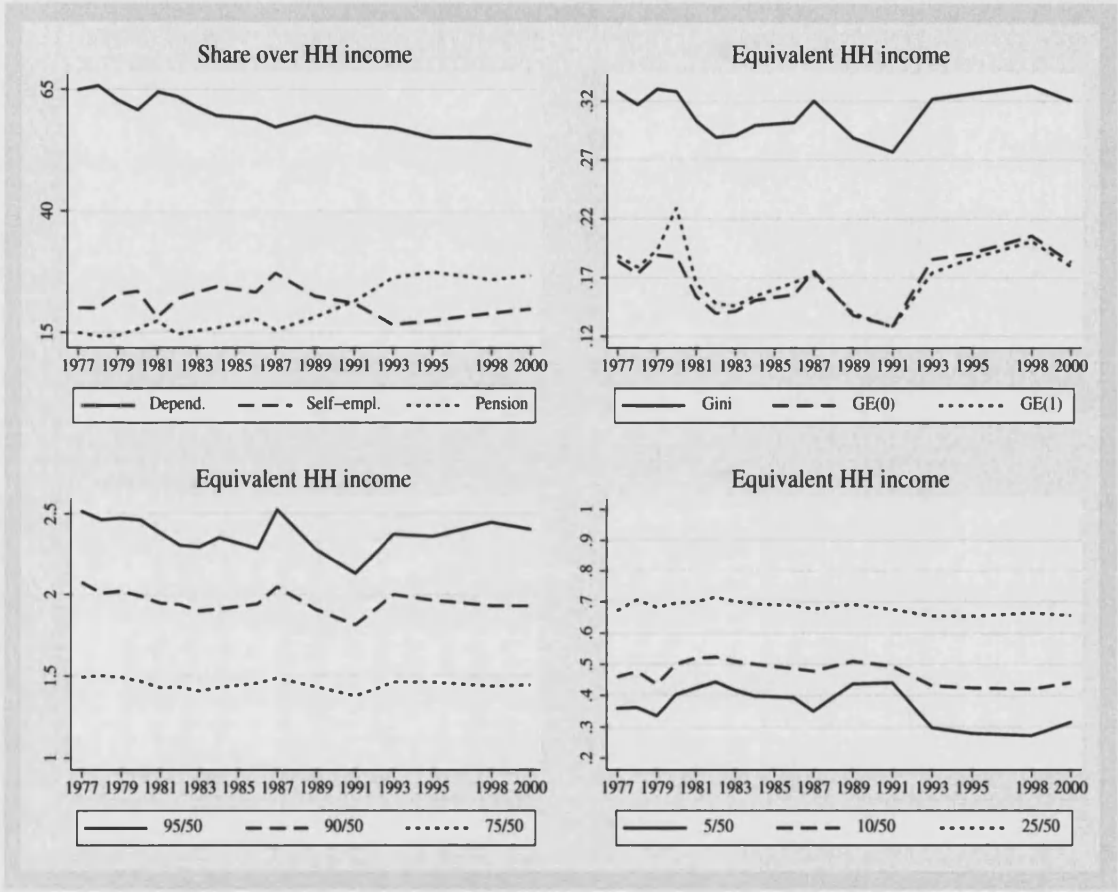


Figure 6-9: Frequency of income earners in av. HH or population

Monthly employment income

Year	N	GINI	GE0	GE1	GE2	95/50	90/50	75/50	25/50	10/50	5/50
1977	2,831	0.239	0.119	0.110	0.144	1.810	1.552	1.267	0.776	0.517	0.310
1978	3,293	0.243	0.123	0.116	0.152	1.778	1.556	1.213	0.800	0.533	0.333
1979	3,011	0.232	0.113	0.106	0.131	1.800	1.500	1.236	0.800	0.500	0.360
1980	3,035	0.213	0.095	0.083	0.087	1.754	1.462	1.231	0.769	0.554	0.385
1981	4,169	0.225	0.111	0.103	0.139	1.667	1.449	1.192	0.769	0.513	0.359
1982	4,128	0.216	0.094	0.086	0.096	1.833	1.556	1.222	0.833	0.600	0.427
1983	4,153	0.213	0.089	0.081	0.089	1.786	1.488	1.255	0.796	0.595	0.425
1984	3,869	0.211	0.089	0.084	0.097	1.682	1.500	1.167	0.800	0.600	0.417
1986	7,023	0.202	0.078	0.073	0.079	1.714	1.429	1.143	0.786	0.571	0.429
1987	7,219	0.205	0.076	0.077	0.089	1.800	1.567	1.200	0.800	0.667	0.500
1989	7,066	0.194	0.065	0.070	0.085	1.667	1.444	1.167	0.783	0.667	0.556
1991	6,802	0.195	0.067	0.069	0.087	1.737	1.526	1.263	0.821	0.632	0.541
1993	6,457	0.240	0.107	0.110	0.143	1.905	1.571	1.238	0.774	0.571	0.429
1995	6,472	0.236	0.101	0.104	0.132	1.909	1.636	1.250	0.818	0.591	0.455
1998	5,931	0.243	0.120	0.115	0.147	1.875	1.556	1.250	0.792	0.500	0.375
2000	6,439	0.244	0.126	0.114	0.141	2.000	1.600	1.200	0.800	0.520	0.384

Monthly self-employment income

Year	N	GINI	GE0	GE1	GE2	95/50	90/50	75/50	25/50	10/50	5/50
1977	791	0.481	0.484	0.419	0.643	3.750	3.000	1.875	0.469	0.188	0.107
1978	1,020	0.466	0.472	0.392	0.586	3.333	2.778	1.667	0.417	0.167	0.111
1979	975	0.472	0.459	0.408	0.621	4.000	3.000	2.000	0.500	0.250	0.125
1980	925	0.503	0.504	0.559	1.763	3.333	2.500	1.667	0.500	0.200	0.137
1981	1,028	0.444	0.403	0.365	0.562	3.429	2.833	1.667	0.600	0.250	0.133
1982	1,158	0.441	0.387	0.356	0.579	3.375	2.500	1.750	0.575	0.250	0.175
1983	1,254	0.406	0.337	0.311	0.490	3.000	2.300	1.500	0.600	0.300	0.150
1984	1,212	0.411	0.325	0.305	0.435	3.333	2.417	1.500	0.583	0.333	0.200
1986	2,111	0.446	0.413	0.392	0.851	3.333	2.500	1.667	0.542	0.250	0.167
1987	2,101	0.408	0.318	0.292	0.396	3.529	2.353	1.529	0.588	0.353	0.212
1989	2,387	0.381	0.255	0.347	1.322	2.817	2.106	1.459	0.676	0.473	0.338
1991	2,161	0.325	0.179	0.198	0.301	2.632	2.105	1.474	0.705	0.526	0.395
1993	2,035	0.413	0.329	0.309	0.451	3.177	2.310	1.470	0.578	0.332	0.173
1995	2,213	0.439	0.374	0.375	0.699	3.222	2.309	1.500	0.556	0.278	0.167
1998	1,898	0.451	0.408	0.401	0.762	3.571	2.476	1.571	0.571	0.286	0.175
2000	2,031	0.411	0.320	0.321	0.569	3.195	2.399	1.520	0.600	0.320	0.200

Monthly pension income

Year	N	GINI	GE0	GE1	GE2	95/50	90/50	75/50	25/50	10/50	5/50
1977	1,804	0.345	0.199	0.254	0.486	3.448	2.759	1.667	0.874	0.743	0.632
1978	1,716	0.345	0.205	0.284	0.748	3.321	2.847	1.708	0.939	0.769	0.617
1979	1,595	0.337	0.186	0.229	0.444	3.462	2.698	1.769	0.862	0.769	0.615
1980	1,773	0.330	0.181	0.188	0.234	3.432	2.790	1.775	0.832	0.692	0.592
1981	2,441	0.343	0.201	0.275	0.886	3.062	2.591	1.790	0.871	0.707	0.565
1982	2,254	0.282	0.132	0.139	0.169	2.800	2.400	1.600	0.880	0.720	0.600
1983	2,385	0.282	0.128	0.133	0.156	2.700	2.333	1.667	0.833	0.667	0.593
1984	2,337	0.281	0.127	0.134	0.165	2.571	2.286	1.714	0.814	0.714	0.571
1986	4,537	0.284	0.133	0.137	0.183	2.467	2.189	1.667	0.778	0.667	0.533
1987	4,239	0.263	0.115	0.114	0.129	2.166	1.805	1.444	0.722	0.614	0.505
1989	4,558	0.270	0.120	0.120	0.135	2.176	1.838	1.357	0.679	0.588	0.441
1991	5,030	0.282	0.130	0.131	0.154	2.400	2.000	1.477	0.733	0.613	0.492
1993	5,756	0.304	0.154	0.159	0.207	2.667	2.267	1.600	0.788	0.640	0.437
1995	5,774	0.307	0.155	0.158	0.190	2.636	2.221	1.604	0.719	0.630	0.424
1998	4,707	0.308	0.161	0.165	0.211	2.338	2.014	1.496	0.655	0.561	0.388
2000	5,444	0.298	0.147	0.153	0.194	2.320	1.976	1.460	0.644	0.571	0.476

Source: own calculation on SHIW-HA

Table 6.5: Inequality by different type of incomes

year	N	GINI	GE0	GE1	GE2	95/50	90/50	75/50	25/50	10/50	5/50
1977	2915	0.328	0.184	0.189	0.257	2.515	2.075	1.494	0.671	0.460	0.359
1978	3044	0.317	0.173	0.179	0.246	2.464	2.008	1.504	0.707	0.477	0.362
1979	2886	0.330	0.189	0.194	0.267	2.473	2.020	1.493	0.683	0.436	0.335
1980	2980	0.328	0.188	0.230	0.519	2.462	1.990	1.469	0.696	0.500	0.403
1981	4091	0.303	0.154	0.165	0.225	2.382	1.945	1.430	0.700	0.520	0.424
1982	3967	0.289	0.140	0.148	0.212	2.306	1.936	1.433	0.716	0.524	0.444
1983	4107	0.291	0.142	0.146	0.189	2.295	1.898	1.410	0.702	0.509	0.417
1984	4172	0.300	0.151	0.154	0.194	2.353	1.909	1.431	0.693	0.501	0.398
1986	8022	0.302	0.156	0.166	0.267	2.286	1.941	1.458	0.687	0.486	0.393
1987	8027	0.320	0.176	0.173	0.211	2.528	2.049	1.489	0.676	0.478	0.350
1989	8274	0.289	0.138	0.140	0.169	2.278	1.907	1.436	0.692	0.510	0.437
1991	8188	0.277	0.129	0.128	0.153	2.131	1.813	1.379	0.675	0.495	0.441
1993	8089	0.322	0.186	0.174	0.205	2.375	2.000	1.464	0.655	0.432	0.298
1995	8135	0.327	0.191	0.187	0.243	2.362	1.964	1.464	0.653	0.425	0.279
1998	7147	0.333	0.206	0.201	0.291	2.448	1.930	1.440	0.664	0.421	0.272
2000	8001	0.321	0.183	0.180	0.237	2.406	1.929	1.447	0.656	0.441	0.315

Source: own calculation on SHIW-HA

Table 6.6: Inequality of equivalent household income

6.3 Description of the methodology

In this section, two different microsimulation methods are combined: that of DiNardo et al. (1996) and that of Burtless (1999) (henceforth DFL and B, respectively). The re-weighting method introduced by DFL allows one to disentangle the impact of demographic changes on equivalent household income inequality. The B method allows one to determine the relative importance of the change of the distribution of either self-employment income, employment income, or pension income on total inequality.

6.3.1 Effects of individual and household characteristics on household inequality

The $N \times 1$ vector of weighted equivalent household income, $\mathbf{y}$, for a sample of N individuals and H households is obtained as follows. Let $\mathbf{z}$ be the $N \times 1$ vector of

individual incomes ordered by household, $\mathbf{w}$ be the $N \times 1$ vector of corresponding sampling weights, $\mathbf{E}$ be the $N \times N$ matrix of equivalence scale and $DIAG(\mathbf{w})$ the diagonal matrix whose diagonal elements are the elements of $\mathbf{w}$, then

$$\underset{N\times 1}{\mathbf{y}} = DIAG(\mathbf{w}) \cdot \mathbf{E} \cdot \mathbf{z}$$

$\mathbf{E}$ is a block diagonal matrix, with H blocks on the diagonal. The blocks have dimension $b_h \times b_h$, $(h = 1, 2..., H)$, all the elements of each block are the same and equal to $1/b_h^{0.5}$, where b_h is the dimension of household h, $h \in 1, 2, ..., H$. For example, if the first household has 3 members, the second 2, etc, the matrix $\mathbf{E}$ is as follows:

$$\underset{N\times N}{\mathbf{E}(b)} = \begin{pmatrix} \begin{pmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{pmatrix} & 0 & 0 \\ 0 & \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} & 0 \\ & & \ddots \end{pmatrix}$$

As in any microsimulation analysis a base year had to be picked and 1991 was chosen for two reasons: (i) the sample size in SHIW data sets was enlarged from 1989 keeping sampling methodology unchanged. Hence post-1989 data sets represent a more reliable picture of the underlying population; (ii) in 1991 equivalent income inequality reached its lowest point and results are easily interpretable using 1991 as

reference year[11].

The DFL methodology was described in Section 5.3. For the current application it is useful to interpret each observation as a vector (y, X, t) coming from the CDF $G(y, X, t)$, where y records equivalent household income, X is a vector of individual and household characteristics (some of which are discrete variables), and t is a date. The CDF of income and attributes at time t is the conditional distribution $G(y, X|t_y = t, t_X = t)$. The density of income at a point in time, $g(y|t_y = t)$, can be seen as the integral of the density of equivalent household incomes conditional on a set of individual and household characteristics and on a date $t_y = t$, $g(y|X, t_y = t)$ over the distribution of individual and household characteristics, $G(X|t_X = t)$, at date $t_X = t$:

$$g(y|t_y = t, t_X = t) = \int_{X \in \Omega_X} g(y|X, t_y = t) dG(X|t_X = t) \tag{6.1}$$

where Ω_X is the space of all possible values of the individual and household characteristics. For example, $g(y|t_y = 2000, t_X = 2000)$ represents the actual density of equivalent household income in 2000, $g(y|t_y = 2000, t_X = 1991)$ represents the density of equivalent household income that would have prevailed in 2000 had the distribution of individual and household characteristic been as in 1991. Hence the

[11]For economy of space the analysis with 1991 as base year will only be presented, without reverse order decomposition. However, the analysis has been performed using different base years after year 1989 and using reverse order decomposition: results change interpretation but do not change substance.

counterfactual density $g(y|t_y = 2000, t_X = 1991)$ is:

$$\begin{aligned} g(y|t_y = 2000, t_X = 1991) &= \int g(y|X, t_y = 2000) dG(X|t_X = 1991) \\ &= \int g(y|X, t_y = 2000) \cdot \psi_X \cdot dG(X|t_X = 2000) \end{aligned} \tag{6.2}$$

Clearly, (6.2) differs from (6.1) only by the factor ψ_X, where:

$$\psi_X = \frac{dG(X|t_X = 1991)}{dG(X|t_X = 2000)} = \frac{Pr(X|t_X = 1991)}{Pr(X, t_X = 2000)} \tag{6.3}$$

Hence, following DFL, the counterfactual density can be computed as a weighted version of the actual one.

Here the vector X is a set of individual and household characteristics, which comprises (a) the number of income receivers, (b) the number of members in the household, (c) the number of pension receivers in the household, (d) and the female labor force participation. In X other individual and household characteristics, such as area of residence (if either North, Center or South), size of town of residence, age, age squared, years of study, years of study squared are also recorded. The counterfactual inequality indices have been computed on the counterfactual distribution. The probabilities in (6.3) are estimated either using standard logit (when the outcome is binary) or with ordered logit models (when possible outcomes of the dependent variables are ordered).

Four counterfactual weights $\psi_i, i = 1, ..., 4$ are defined as the ratio of :

(a) the probability of having R ($R = 1, ..., 4, 5+$) income receivers in the household in year 1991 over the probability of having R income receivers in year t (t from 1977 to 2000, available years);

(b) the probability of having a household of N members (N=1,...,5,6+) in year 1991 over the probability of having a household of N members in year t;

(c) the probability of receiving a pension (regardless of the amount) in 1991 over the probability of receiving a pension in year t;

(d) the probability that a working age (15-65) woman is in the labor force in 1991 over the probability she is in the labor force in year t.

The application of the DFL methodology is aimed at assessing the effects on household income distribution of (a) number of income receivers, (b) number of member in the household, (c) number of pension receivers in the household, (d) female labor force participation. It allows the estimation of counterfactuals with an easy interpretation.

The traditional decomposition was not used since such an analysis only made a limited contribution to understanding Italian household inequality trends as reviewed in Section 6.1. Because of the limitations of the regression-based methodology pointed out in Section 5.2, I looked at microsimulation methodologies. The Bourguignon et al. (2001) was discarded mainly because it would not help in explaining the questions under consideration (i.e., the effect of individual income dispersion on household inequality); it gives a nice decomposition of inequality into a limited number of effects at the expense of developing a complicated structural model with debatable assump-

tions (recall Section 5.3). Such a procedure becomes even more cumbersome once you try to compare a full range of years, rather than just two years.

6.3.2 Effects of changing dispersion of individual incomes on household incomes

The application of the B methodology investigates the importance of the trend of employment inequality for household inequality. However, the analysis can also be extended to self-employment and pension income to see whether and to what extent, they caused an effect on the distribution of equivalent household income.

For the B methodology the counterfactual analysis is applied on the vector of individual incomes, $\mathbf{z}$. The individual income variable is constructed as the sum of her employment (z_i^{empl}), self-employment (z_i^{self}) and pension income (z_i^{pen}), i.e. $z_i = z_i^{empl} + z_i^{self} + z_i^{pen}$; each income variable is then split into number of months the income was received times the average income received in each month, dividing annual income by number of months it was received, by income type. The rank-dependent transformation is based on holding the distribution of certain kind of income constant through time and then calculating how much household inequality would change under this assumption.

The B methodology was presented in Section 5.3 and implemented here as follows. For instance, assuming that monthly wage inequality changed between year 1991 and 2000, i.e. the distribution of the $\mathbf{y}^{empl}$ vector was different in the two years, the basic idea is to assign to each 2000 employee the wage the employee at her

rank would have received according to the 1991 wage distribution. This preserves the exact 2000 earning distribution of wages but ignores the change in the average wage between the two years. To keep the sum of the 2000 wages constant, I simply assigned to 2000 employees the 1991 wage corresponding to their rank in the wage distribution, multiplied by the ratio of total year-round wages in 2000 divided by total year-round wages in 1991. This procedure is straightforward if the number of employees is the same in the two years but this can happen only by pure coincidence, and does not happen in this case. Hence, the empirical distribution function using the same number of quantiles is computed, properly weighted to take into account sampling weights. Then the median within each quantile is calculated. For each individual in the 2000 data set the median income of the wage quantile distribution she belongs to is subtracted and replaced by the median income of the same quantile in the normalized 1991 quantile wage distribution. The individual wages are then summed up to other individual incomes. All individual incomes of each household are then summed together and equivalized using the LIS equivalence scale, as described in Section 6.2.3. In the empirical application a distribution by centiles is used, i.e. a quantile distribution with 100 quantiles, but even with 500 quantiles the results do not change significantly. For the centile distributions, only incomes greater then Lit 1000 (€0.52) are considered, i.e. if an individual did not have any wage income in 2000, the replacement based on normalized 1991 wage distribution would still have left her with zero wage income. In the same way the 1991 income vectors are also replaced with normalized 2000 income vectors. An analogous analysis was performed for self-employment and pension incomes.

This procedure aims at accounting for the importance of the dispersion of income from different sources for household inequality. However, it is not an exact decomposition of inequality and a residual is expected to come mainly from the covariance between different incomes accruing to the same individual or between different individuals in the same household.

6.3.3 Testing the change of inequality

All inequality estimates for the GE class are accompanied with their asymptotic standard errors, as in Cowell (1989); Cowell and Jenkins (2003); Biewen and Jenkins (2003). A large standard error compared to the estimate would mean that the inequality index is not significantly different from zero and that the data set is unsuitable for inequality analysis. This is never an issue for our data sets. Asymptotic standard errors are then used to perform a test for the significance of the difference between inequality for different data sets. Given an inequality index belonging to the $GE(a)$ class with $a = 0, 1, 2$, computed on two different independent data set, say I^a_{1991} and I^a_{2000}, the asymptotically normal statistic,

$$\tau^a = \frac{I^a_{1991} - I^a_{2000}}{\sqrt{varI^a_{1991}} + \sqrt{varI^a_{2000}}} \tag{6.4}$$

tests the hypothesis "H_0 : there is no difference in inequality according to index $GE(a)$ between year 1991 and year 2000".

The test is performed on differences between actual figures, to test if there is a statistically significant change in inequality in different years. Whenever the difference

in inequality is significantly different from zero, and a counterfactual distribution is computed, I tested whether the difference between the counterfactual and actual distribution is still statistically different from zero. If not, this is prima facie evidence that the simulation exercise explains most of the change in inequality that actually occurred.

6.4 Results of the analysis

The combination of the DFL and B methodology allows one to put into a single framework the analysis of the effects of socio-demographic trends and income factors dispersion on inequality. In the traditional analysis of income decomposition the effects of socio-demographic changes are assessed by "population subgroup decomposition", and the effects of income sources dispersion by "factor source decomposition". However, these two approaches cannot be easily integrated. In the present framework, socio-demographic changes and effects of income sources inequality on total inequality are assessed by DFL and B methodologies, respectively and eventually put together.

The main results are presented using graphs of the actual and counterfactual inequality measures using Gini, $GE(0)$, $GE(1)$ and $GE(2)$ indices. The results of the significance tests on the changes for the most interesting cases, limited to the GE indices, are also presented.

The DFL methodology applied to Italian household income inequality shows that demographic changes had a limited effect for the trend of overall inequality, as other researchers had found using traditional decomposition analysis (recall Section 6.1).

Socio-economic variables, such as female labor force participation and number of income receivers, had a greater impact, although it does not exhaust the total change of inequality relative to 1991 (Figures 6-10 to 6-12).

However, the DFL re-weighting methodology is effective in identifying the effect of some odd figures obtained from the sample and their effect on the picture of overall distribution. For instance, if we were to accept the idea that the marked decrease of income receivers in the household recorded in 1987 data set was not a reliable picture of the population due to data contamination (see Section 6.2.4), the DFL methodology allows one to re-estimate the inequality indices and shows that Italian household inequality, at least according to Gini, $GE(0)$ and $GE(1)$ indices, did have a trend, decreasing at first (up to 1991) and then increasing (Figure 6-10).

Summing up, using the DFL methodology we may conclude that:

- The decrease of the average household size does not have much effect on income distribution. The only visible effect induced by conditioning the equivalent income distribution to 1991 distribution of household size was to reduce the odd spikes in 1980, especially for the $GE(1)$ and $GE(2)$ indices (Figure 6-10).

- The increased percentage of income receivers per household (Figure 6-5) had a negative effect on income distribution. In fact, holding the distribution of income receivers per household at 1991 levels, makes the spikes in 1987 and 1980 decrease for all indices, correcting for the bias due to data contamination (see Section 6.2.4). It also reduces household inequality in the post-1991 period (Figure 6-10).

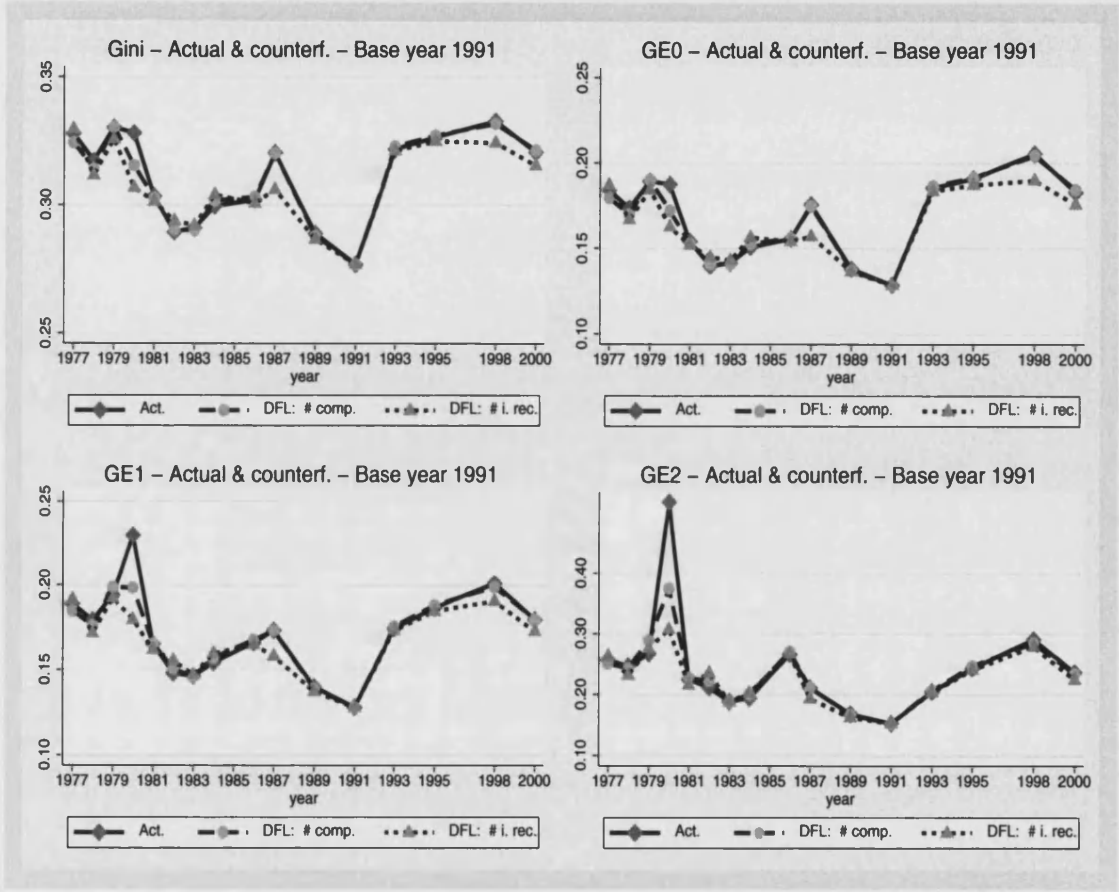


Figure 6-10: DFL methodology: number of income receivers and number of components in HH

Act.		Actual figures
DFL: # comp.	DFL:	conditioning on number of household components
DFL: # i.rec.	DFL:	conditioning on number of income receivers in the household
DFL: # pens.	DFL:	conditioning on probability of being a pensioner
DFL: fem. LFP	DFL:	conditioning on female in the labor force
DFL: i.rec.+f.LFP	DFL:	conditioning on number of income receivers in the household and on female in the labor force
B: s.-e. inc. ineq.	B:	holding self-employment inequality constant at base year
B: dep. inc. ineq.	B:	holding employment inequality constant at base year
B: lab. inc. ineq.	B:	holding employment and self-employment (work income) inequality constant at base year
B: pens. ineq.	B:	holding pension inequality constant at base year
B: lab. + pens. ineq.	B:	holding work and pension inequality constant at base year

Table 6.7: Abbreviations used in tables and figures

- The increased number of pensioners in the average household has only a small effect on income distribution. In particular, had the probability of being a pensioner remained at the 1991 level, inequality would have been higher in 1987, using Gini and $GE(0)$ indices, and in 1980 and 1998 using the $GE(2)$ index (Figure 6-11). In 1980, 1987 and 1998 pension income inequality was higher than in 1991 but the average number of pensioners increased steadily since the beginning of the period. However, modifying the weights of pensioners without changing the distribution of pension income had the same inequality-increasing effect.

- Conditioning on the female labor force participation shows a larger effect on the post 1991 period. From this it is clear that if the female labor force participation had remained as in 1991 the household inequality measures would have significantly decreased in the following decade. The effect would have been less

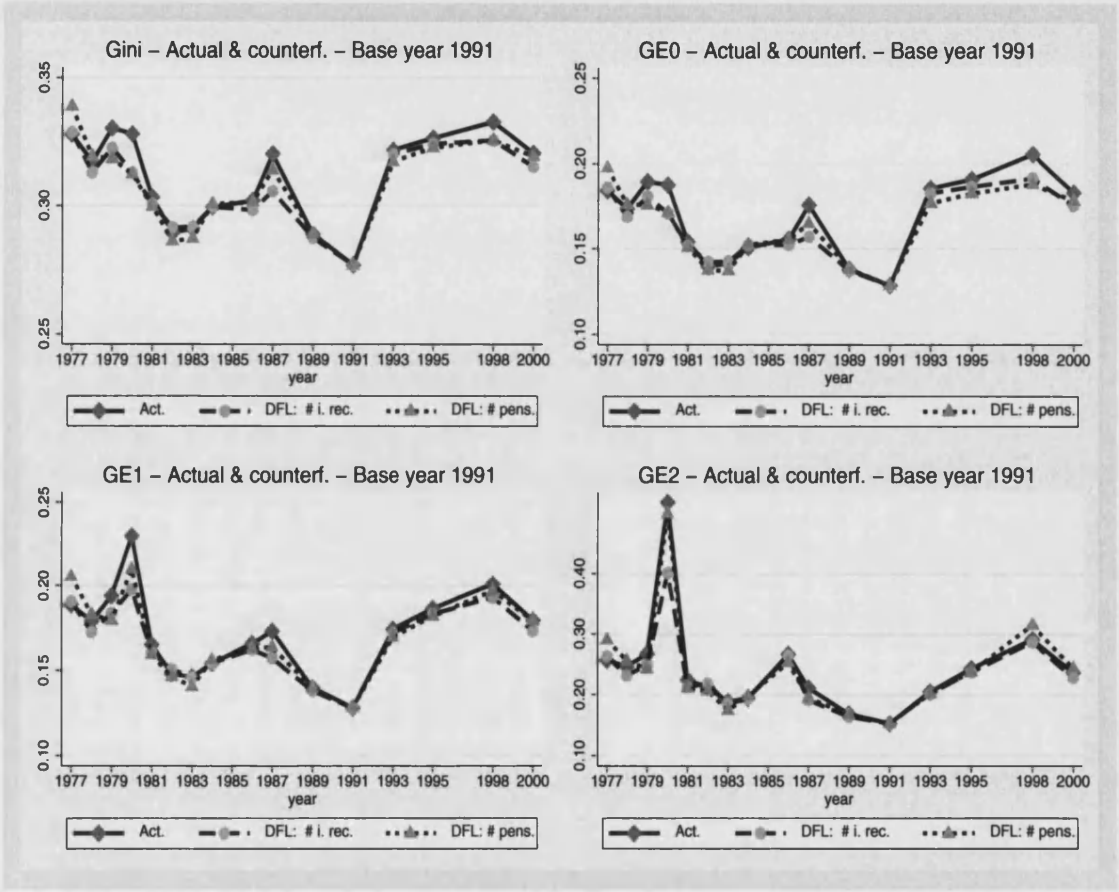
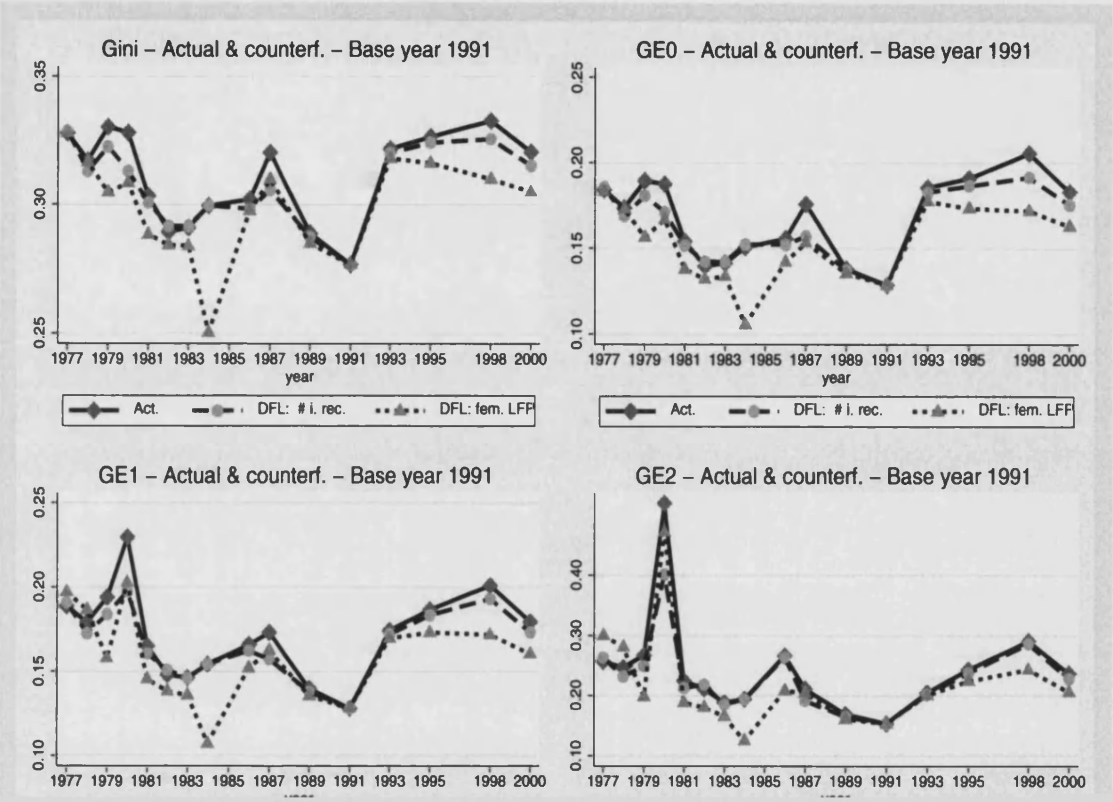


Figure 6-11: DFL methodology: number of income receivers in HH and probability of being pensioners

clear in the period before 1991, moreover with an odd spike at 1984 (Figure 6-12), because of the reduced sample size and of data contamination (see also Figure 6-13 for the non-standard over-sampling of females in 1984).



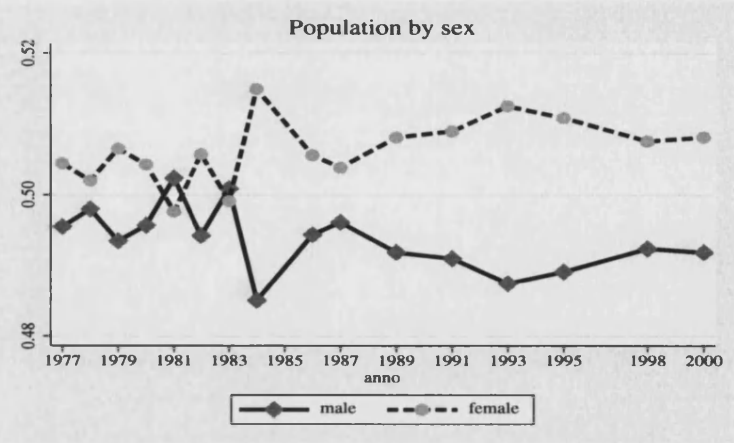


Figure 6-13: Decomposition of the population by sex

the 5% significance level. The same conclusion applies to the $GE(2)$ except for 1989, where a larger point estimate of .102 is not significantly different from the base year at the 10% significance level. The change of inequality indices after conditioning on the number of income receivers and, cumulatively, on female labor force participation is tested, showing that counterfactual figures are no longer significantly different from the base year for 1989, for main part of the 1980s and all GE indices used. As for the 1990s, if the number of income receivers and labor force participation (conditional on household size and other individual and household characteristics) had remained as in 1991, the inequality indices considered would have shown a reduction in their value of about 15-20%. However, the differences between the counterfactual and the actual 1991 inequality figures are still significantly different from zero for the whole post-1991 period.

The B methodology applied to Italian household income inequality shows that much of the dynamics in inequality is due to the changed distribution of income sources. In particular, individual self-employment and employment income dynamics,

GE(0)

	Actual figure		Act. - base year		Count.1 Cond. on # inc. rec.		Cond. on # fem. LFP		Count. 1 - base year	
year	*index*	*s.e.*	*%*	*p-val*	*index*	*s.e.*	*index*	*s.e.*	*%*	*p-val*
1977	0.184	0.004	43.1	0.000	0.186	0.005	0.183	0.010	42.3	0.000
1978	0.173	0.004	34.9	0.000	0.169	0.004	0.173	0.008	34.2	0.000
1979	0.189	0.005	47.3	0.000	0.181	0.005	0.156	0.006	21.6	0.000
1980	0.188	0.009	45.8	0.000	0.171	0.007	0.167	0.011	29.6	0.001
1981	0.154	0.004	19.7	0.000	0.152	0.004	0.138	0.005	7.1	0.090
1982	0.140	0.003	8.6	0.012	0.142	0.004	0.132	0.004	2.7	0.455
1983	0.142	0.003	10.0	0.001	0.142	0.003	0.134	0.004	4.0	0.333
1984	0.151	0.003	17.2	0.000	0.152	0.003	0.106	0.016	-17.9	0.145
1986	0.156	0.003	21.0	0.000	0.152	0.003	0.142	0.013	10.4	0.317
1987	0.176	0.003	36.7	0.000	0.157	0.003	0.153	0.010	19.3	0.017
1989	0.138	0.002	7.3	0.008	0.137	0.002	0.135	0.002	5.3	0.054
1991	0.129	0.003	0.0	1.000	0.129	0.003	0.129	0.003	0.0	1.000
1993	0.186	0.003	44.3	0.000	0.183	0.003	0.177	0.003	38.0	0.000
1995	0.191	0.003	48.5	0.000	0.187	0.003	0.173	0.004	34.7	0.000
1998	0.206	0.005	59.9	0.000	0.192	0.005	0.172	0.005	33.5	0.000
2000	0.183	0.003	42.3	0.000	0.175	0.003	0.163	0.003	26.6	0.000

GE(1)

	Actual figure		Act. - base year		Count.1 Cond. on # inc. rec.		Cond. on # fem. LFP		Count. 1 - base year	
year	*index*	*s.e.*	*%*	*p-val*	*index*	*s.e.*	*index*	*s.e.*	*%*	*p-val*
1977	0.189	0.007	47.4	0.000	0.191	0.008	0.197	0.018	53.9	0.000
1978	0.179	0.007	39.8	0.000	0.173	0.006	0.186	0.014	45.6	0.000
1979	0.194	0.009	51.5	0.000	0.184	0.008	0.158	0.008	23.3	0.001
1980	0.230	0.021	79.5	0.000	0.198	0.015	0.203	0.026	58.4	0.004
1981	0.165	0.006	29.1	0.000	0.161	0.005	0.145	0.006	13.5	0.027
1982	0.148	0.006	15.6	0.012	0.151	0.007	0.138	0.005	7.9	0.132
1983	0.146	0.004	14.2	0.002	0.146	0.003	0.136	0.005	6.1	0.267
1984	0.154	0.004	20.5	0.000	0.154	0.004	0.107	0.017	-16.3	0.230
1986	0.166	0.005	29.7	0.000	0.162	0.006	0.152	0.015	18.6	0.128
1987	0.173	0.003	35.4	0.000	0.157	0.003	0.163	0.010	27.0	0.003
1989	0.140	0.003	9.3	0.032	0.138	0.003	0.135	0.003	5.8	0.172
1991	0.128	0.005	0.0	1.000	0.128	0.005	0.128	0.005	0.0	1.000
1993	0.174	0.003	36.2	0.000	0.172	0.003	0.170	0.004	32.6	0.000
1995	0.187	0.004	45.8	0.000	0.183	0.004	0.173	0.005	35.4	0.000
1998	0.201	0.008	57.1	0.000	0.193	0.008	0.172	0.007	34.3	0.000
2000	0.180	0.004	40.3	0.000	0.173	0.004	0.161	0.004	25.4	0.000

GE(2)

	Actual figure		Act. - base year		Count.1 Cond. on # inc. rec.		Cond. on # fem. LFP		Count. 1 - base year	
year	*index*	*s.e.*	*%*	*p-val*	*index*	*s.e.*	*index*	*s.e.*	*%*	*p-val*
1977	0.257	0.021	68.0	0.000	0.263	0.023	0.300	0.060	96.1	0.016
1978	0.246	0.020	61.2	0.000	0.232	0.016	0.281	0.041	83.6	0.003
1979	0.267	0.030	74.8	0.001	0.249	0.027	0.198	0.017	29.4	0.036
1980	0.519	0.094	239.3	0.000	0.401	0.068	0.472	0.123	208.4	0.010
1981	0.225	0.015	46.9	0.000	0.213	0.012	0.188	0.013	23.0	0.055
1982	0.212	0.026	38.3	0.047	0.218	0.029	0.180	0.013	17.6	0.143
1983	0.189	0.009	23.3	0.026	0.185	0.008	0.165	0.008	7.8	0.448
1984	0.194	0.008	27.1	0.009	0.194	0.009	0.125	0.022	-18.3	0.277
1986	0.267	0.022	74.3	0.000	0.264	0.025	0.208	0.023	36.0	0.037
1987	0.211	0.006	38.0	0.000	0.191	0.006	0.205	0.014	33.9	0.008
1989	0.169	0.006	10.2	0.284	0.164	0.005	0.160	0.005	4.9	0.606
1991	0.153	0.013	0.0	1.000	0.153	0.013	0.153	0.013	0.0	1.000
1993	0.205	0.006	33.8	0.000	0.202	0.006	0.200	0.007	31.1	0.001
1995	0.243	0.009	58.9	0.000	0.237	0.009	0.224	0.011	46.4	0.000
1998	0.291	0.028	90.2	0.000	0.287	0.030	0.243	0.022	59.1	0.001
2000	0.237	0.014	54.8	0.000	0.228	0.014	0.206	0.011	34.5	0.002

Source: own calculation on SHIW-HA data

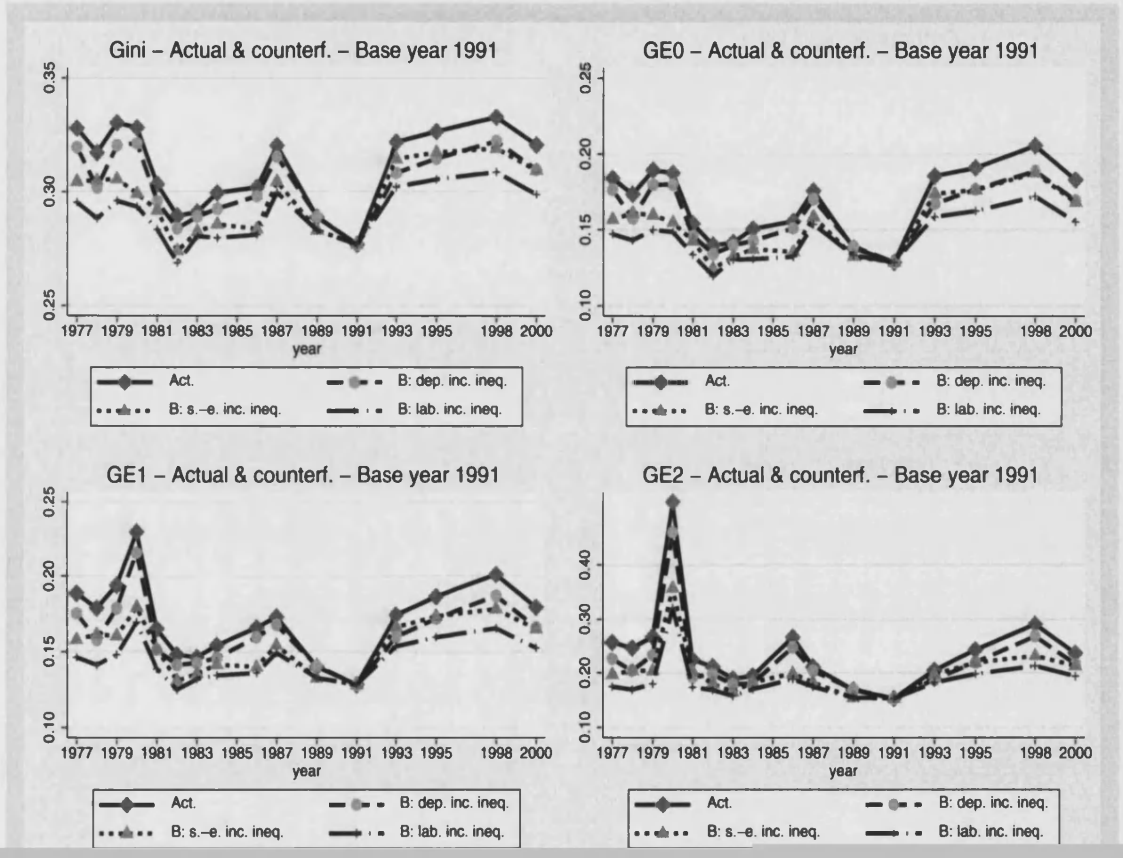
Table 6.8: Counterfactuals using DFL methodology - Base year is 1991

which were both decreasing up until 1991 and then increasing, had a strong influence on the evolution of equivalent household income. If the distribution of labor income is kept constant at 1991, there would be no trend before 1991 and a very slight increase in post-1991 period (Figure 6-14). The higher incomes would have been more affected by the transformation on self-employment income (note 1998 for $GE(2)$[12]), while the jump registered from 1991 to 1993 seems due to employment rather than to self-employment increase of inequality. These pictures also show the importance of the change in dispersion of self-employment, which was at least as important as the change in dispersion of employment income, besides the self-employed workers being only a quarter of the labor force.

Holding the distribution of individual pension income as in 1991 instead, had virtually no effect for the period 1987-2000. For the previous period, had distribution of individual pension income been as in 1991, household inequality would have been between 5 and 25% lower (Figure 6-15).

Putting together the B decomposition on different incomes individually, it is possible to account for a much larger share of income dynamics (Figure 6-16). Holding labor income dispersion at 1991 levels decreases household inequality by about a half, and even more for the period before 1991. As noted before, holding pension income dispersion fixed as well would make no substantial difference for the post 1991 period, while it would induce an overshooting of the decomposition, causing the counterfactual inequality to be even lower than in 1991. These results suggest that concern

[12]As explained in Section 6.6 on page 182, the $GE2$ is more sensitive to income differences at the top of the distribution than $GE1$ and $GE0$: a larger GE parameter a means more sensitivity to high incomes in the data.


Gini – Actual & counterf. – Base year 1991
GE0 – Actual & counterf. – Base year 1991
GE1 – Actual & counterf. – Base year 1991
GE2 – Actual & counterf. – Base year 1991
year
Act.
B: dep. inc. ineq.
B: s.–e. inc. ineq.
B: lab. inc. ineq.

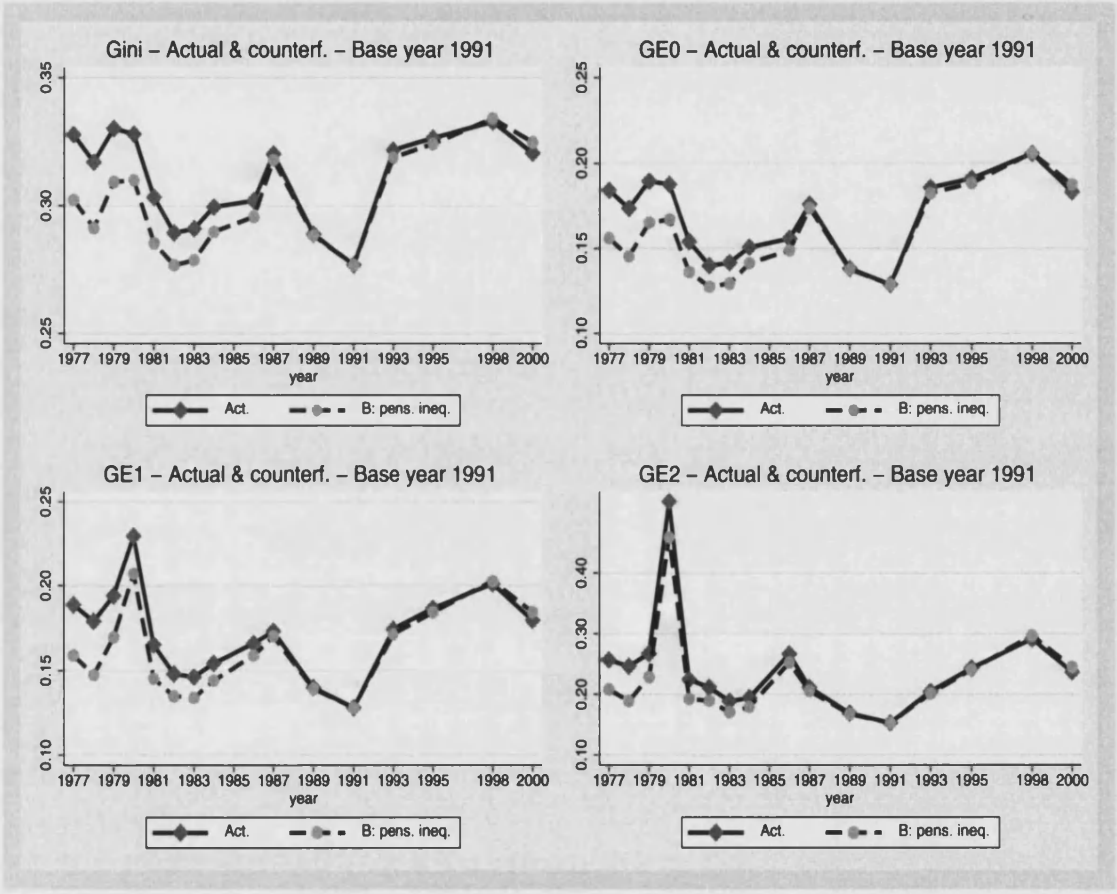
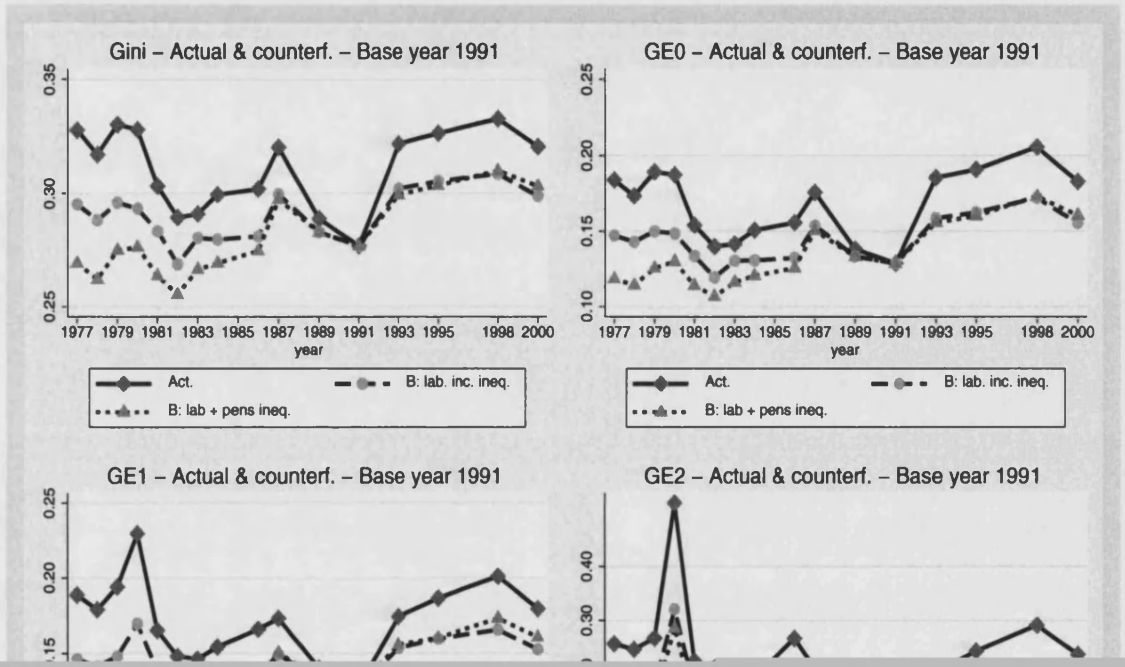


Figure 6-15: Counterfactuals using Burtless methodology

about pension income is indeed misplaced. While the trend in pension income is often taken to be a major cause of increasing inequality in the 1990s, it actually had only a limited effect during those years. By contrast, household inequality during the 1980s would have been significantly lower had pension income been distributed as in 1991.



Counterfactual 2 is not statistically different from zero for $GE(2)$, though the results for lower incomes (i.e. $GE(0)$ and $GE(1)$)) are less conclusive. In other words, combining both DFL and B methodologies the change of inequality in top incomes (as stated by $GE(2)$) is mainly due to the increased dispersion of employment and self-employment income but also to the increased female participation in the labor force and the increased number of income receivers in the average household. However, there is still much to explain in the case of $GE(0)$ and $GE(1)$ inequality measures.

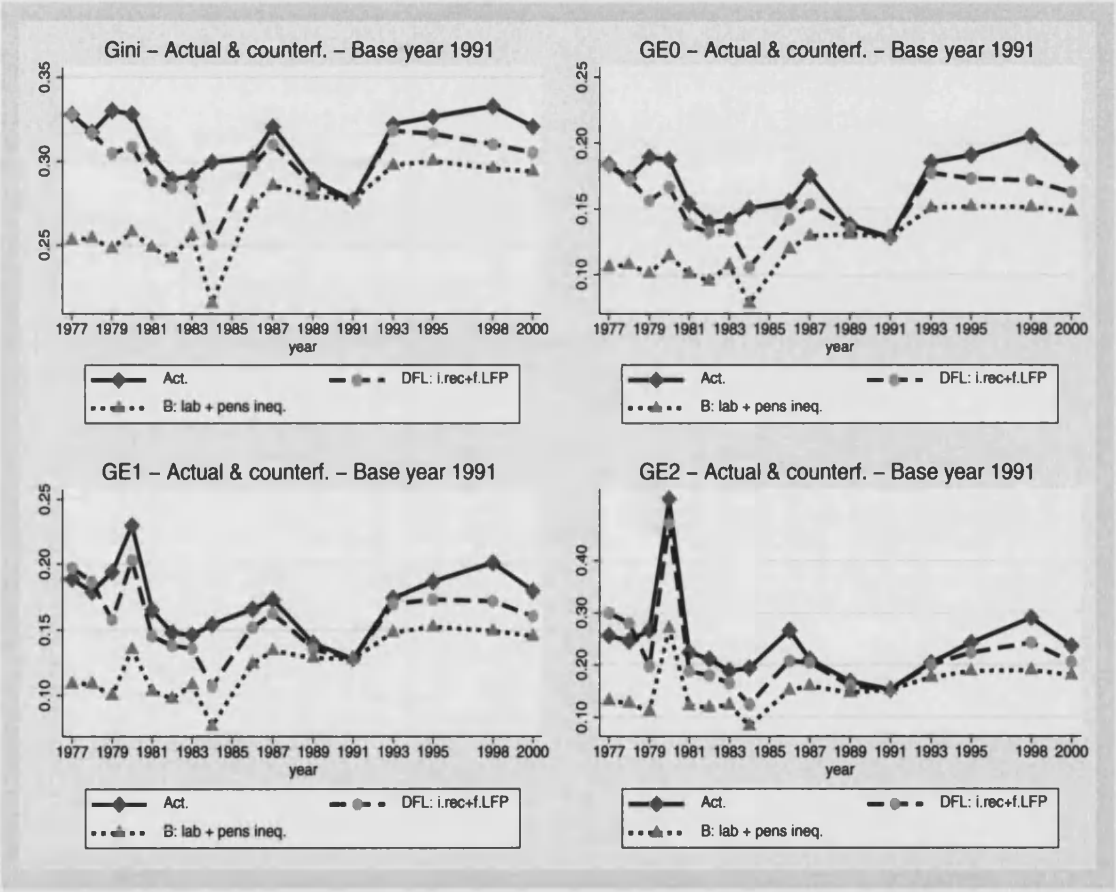


Figure 6-17: Counterfactuals using DFL+Burtless methodology

GE0

	Actual figure		Holding dep. Inc. const.		Holding s-e inc. const		Count. 1: Holding lab inc const		Count. 2: lab +pens const		Diff. of Actual wrt base year		Diff. Count. 1 wrt base year		Diff. Count. 2 wrt base year	
Year	index	*s.e.*	index	*s.e.*	index	*s.e.*	index	*s.e.*	%	p-val	%	p-val	%	p-val	%	p-val
1977	0.184	0.004	0.182	0.009	0.155	0.006	0.152	0.006	0.106	0.004	0.431	0.000	0.180	0.000	-0.176	0.000
1978	0.173	0.004	0.159	0.007	0.160	0.007	0.145	0.006	0.108	0.004	0.349	0.000	0.126	0.011	-0.162	0.000
1979	0.189	0.005	0.159	0.006	0.142	0.006	0.143	0.006	0.101	0.004	0.473	0.000	0.112	0.022	-0.214	0.000
1980	0.188	0.009	0.161	0.010	0.140	0.009	0.135	0.008	0.115	0.008	0.458	0.000	0.053	0.440	-0.109	0.092
1981	0.154	0.004	0.135	0.004	0.130	0.004	0.126	0.004	0.101	0.003	0.197	0.000	-0.017	0.661	-0.217	0.000
1982	0.140	0.003	0.130	0.004	0.119	0.003	0.118	0.003	0.095	0.003	0.086	0.012	-0.086	0.010	-0.262	0.000
1983	0.142	0.003	0.135	0.004	0.126	0.004	0.127	0.004	0.107	0.004	0.100	0.001	-0.010	0.797	-0.166	0.000
1984	0.151	0.003	0.104	0.015	0.094	0.013	0.095	0.013	0.078	0.011	0.172	0.000	-0.265	0.011	-0.391	0.000
1986	0.156	0.003	0.144	0.012	0.131	0.012	0.134	0.011	0.120	0.010	0.210	0.000	0.040	0.650	-0.070	0.396
1987	0.176	0.003	0.152	0.010	0.139	0.010	0.138	0.010	0.129	0.008	0.367	0.000	0.074	0.350	0.006	0.924
1989	0.138	0.002	0.137	0.002	0.130	0.002	0.131	0.002	0.131	0.002	0.073	0.008	0.018	0.492	0.017	0.515
1991	0.129	0.003	0.129	0.003	0.129	0.003	0.129	0.003	0.129	0.003	0.000	1.000	0.000	1.000	0.000	1.000
1993	0.186	0.003	0.161	0.003	0.168	0.003	0.155	0.003	0.151	0.003	0.443	0.000	0.202	0.000	0.173	0.000
1995	0.191	0.003	0.164	0.004	0.165	0.004	0.156	0.004	0.152	0.003	0.485	0.000	0.214	0.000	0.181	0.000
1998	0.206	0.005	0.161	0.004	0.160	0.004	0.150	0.004	0.152	0.004	0.599	0.000	0.167	0.000	0.178	0.000
2000	0.183	0.003	0.153	0.003	0.152	0.003	0.142	0.003	0.148	0.003	0.423	0.000	0.104	0.001	0.151	0.000

GE1

	Actual figure		Holding dep. Inc. const.		Holding s-e inc. const		Count. 1: Holding lab inc const		Count. 2: lab +pens const		Diff. of Actual wrt base year		Diff. Count. 1 wrt base year		Diff. Count. 2 wrt base year	
Year	index	*s.e.*	index	*s.e.*	index	*s.e.*	index	*s.e.*	%	p-val	%	p-val	%	p-val	%	p-val
1977	0.189	0.007	0.189	0.016	0.159	0.010	0.152	0.008	0.109	0.007	0.474	0.000	0.188	0.013	-0.150	0.018
1978	0.179	0.007	0.165	0.012	0.167	0.010	0.147	0.009	0.109	0.005	0.398	0.000	0.147	0.058	-0.150	0.005
1979	0.194	0.009	0.155	0.007	0.140	0.006	0.138	0.006	0.100	0.005	0.515	0.000	0.076	0.206	-0.218	0.000
1980	0.230	0.021	0.191	0.023	0.163	0.021	0.155	0.019	0.135	0.018	0.795	0.000	0.212	0.173	0.057	0.692
1981	0.165	0.006	0.138	0.005	0.136	0.006	0.128	0.005	0.103	0.004	0.291	0.000	0.002	0.968	-0.192	0.000
1982	0.148	0.006	0.135	0.005	0.122	0.004	0.120	0.004	0.096	0.003	0.156	0.012	-0.062	0.198	-0.236	0.000
1983	0.146	0.004	0.136	0.005	0.127	0.005	0.127	0.005	0.108	0.004	0.142	0.002	-0.012	0.824	-0.154	0.002
1984	0.154	0.004	0.105	0.016	0.094	0.013	0.093	0.013	0.077	0.011	0.205	0.000	-0.270	0.011	-0.399	0.000
1986	0.166	0.005	0.153	0.014	0.137	0.013	0.139	0.012	0.124	0.011	0.297	0.000	0.084	0.409	-0.030	0.757
1987	0.173	0.003	0.160	0.011	0.145	0.010	0.143	0.010	0.134	0.009	0.354	0.000	0.119	0.176	0.048	0.536
1989	0.140	0.003	0.136	0.003	0.129	0.002	0.129	0.002	0.129	0.002	0.093	0.032	0.006	0.878	0.004	0.921
1991	0.128	0.005	0.128	0.005	0.128	0.005	0.128	0.005	0.128	0.005	0.000	1.000	0.000	1.000	0.000	1.000
1993	0.174	0.003	0.157	0.003	0.163	0.003	0.152	0.003	0.148	0.003	0.362	0.000	0.184	0.000	0.157	0.000
1995	0.187	0.004	0.163	0.005	0.166	0.005	0.156	0.005	0.152	0.004	0.458	0.000	0.219	0.000	0.191	0.000
1998	0.201	0.008	0.163	0.007	0.156	0.005	0.148	0.005	0.149	0.005	0.571	0.000	0.153	0.006	0.166	0.003
2000	0.180	0.004	0.150	0.004	0.149	0.004	0.140	0.003	0.145	0.003	0.403	0.000	0.091	0.047	0.135	0.004

GE2

	Actual figure		Holding dep. Inc. const.		Holding s-e inc. const		Count. 1: Holding lab inc const		Count. 2: lab +pens const		Diff. of Actual wrt base year		Diff. Count. 1 wrt base year		Diff. Count. 2 wrt base year	
Year	index	*s.e.*	index	*s.e.*	index	*s.e.*	index	*s.e.*	%	p-val	%	p-val	%	p-val	%	p-val
1977	0.257	0.021	0.268	0.050	0.206	0.023	0.188	0.020	0.132	0.014	0.680	0.000	0.231	0.137	-0.138	0.280
1978	0.246	0.020	0.232	0.034	0.225	0.024	0.187	0.020	0.127	0.009	0.612	0.000	0.223	0.156	-0.171	0.101
1979	0.267	0.030	0.185	0.014	0.164	0.011	0.155	0.009	0.111	0.006	0.748	0.001	0.014	0.897	-0.272	0.005
1980	0.519	0.094	0.419	0.107	0.348	0.097	0.310	0.084	0.271	0.075	2.393	0.000	1.024	0.067	0.769	0.121
1981	0.225	0.015	0.166	0.010	0.173	0.011	0.153	0.009	0.122	0.008	0.469	0.000	0.001	0.993	-0.203	0.044
1982	0.212	0.026	0.172	0.012	0.153	0.010	0.147	0.010	0.119	0.008	0.383	0.047	-0.036	0.738	-0.223	0.027
1983	0.189	0.009	0.161	0.008	0.148	0.007	0.145	0.007	0.124	0.006	0.233	0.026	-0.052	0.591	-0.191	0.046
1984	0.194	0.008	0.121	0.020	0.106	0.016	0.104	0.015	0.084	0.014	0.271	0.009	-0.319	0.014	-0.451	0.000
1986	0.267	0.022	0.205	0.021	0.173	0.018	0.172	0.017	0.152	0.015	0.743	0.000	0.124	0.377	-0.006	0.964
1987	0.211	0.006	0.201	0.014	0.176	0.013	0.172	0.013	0.160	0.012	0.380	0.000	0.126	0.305	0.045	0.697
1989	0.169	0.006	0.162	0.006	0.147	0.004	0.148	0.004	0.147	0.004	0.102	0.284	-0.034	0.710	-0.039	0.675
1991	0.153	0.013	0.153	0.013	0.153	0.013	0.153	0.013	0.153	0.013	0.000	1.000	0.000	1.000	0.000	1.000
1993	0.205	0.006	0.185	0.007	0.194	0.006	0.180	0.006	0.176	0.006	0.338	0.000	0.176	0.066	0.150	0.118
1995	0.243	0.009	0.207	0.011	0.207	0.009	0.192	0.009	0.188	0.009	0.589	0.000	0.255	0.015	0.230	0.028
1998	0.291	0.028	0.228	0.021	0.199	0.014	0.187	0.014	0.190	0.014	0.902	0.000	0.221	0.076	0.240	0.061
2000	0.237	0.014	0.190	0.010	0.187	0.010	0.172	0.010	0.180	0.010	0.548	0.000	0.127	0.237	0.178	0.106

Source: our calculation on SHIW-HA data

Table 6.9: Counterfactuals using DFL + Burtelss methodology - Base year is 1991

GE0

	Actual figure		Holding dep. Inc. const.		Holding s-e inc. const		Count. 1: Holding lab inc const		Count. 2: lab +pens const		Diff. of Actual wrt base year		Diff. Count. 1 wrt base year		Diff. Count. 2 wrt base year	
Year	index	s.e.	index	*s.e.*	index	*s.e.*	index	*s.e.*	%	p-val	%	p-val	%	p-val	%	p-val
1977	0.184	0.004	0.176	0.004	0.157	0.003	0.147	0.003	0.119	0.002	0.431	0.000	0.144	0.000	-0.078	0.005
1978	0.173	0.004	0.157	0.004	0.161	0.003	0.143	0.003	0.114	0.002	0.349	0.000	0.111	0.000	-0.113	0.000
1979	0.189	0.005	0.180	0.005	0.160	0.004	0.150	0.003	0.126	0.003	0.473	0.000	0.165	0.000	-0.024	0.446
1980	0.188	0.009	0.180	0.009	0.155	0.007	0.149	0.007	0.130	0.006	0.458	0.000	0.155	0.006	0.010	0.856
1981	0.154	0.004	0.146	0.004	0.143	0.004	0.134	0.003	0.114	0.003	0.197	0.000	0.038	0.257	-0.114	0.000
1982	0.140	0.003	0.134	0.003	0.124	0.003	0.119	0.003	0.107	0.003	0.086	0.012	-0.071	0.021	-0.170	0.000
1983	0.142	0.003	0.140	0.002	0.132	0.002	0.130	0.002	0.116	0.002	0.100	0.001	0.013	0.635	-0.097	0.000
1984	0.151	0.003	0.143	0.003	0.137	0.003	0.131	0.003	0.120	0.003	0.172	0.000	0.016	0.607	-0.064	0.029
1986	0.156	0.003	0.151	0.003	0.135	0.002	0.132	0.002	0.125	0.002	0.210	0.000	0.027	0.323	-0.025	0.361
1987	0.176	0.003	0.170	0.003	0.159	0.002	0.154	0.002	0.150	0.002	0.367	0.000	0.194	0.000	0.169	0.000
1989	0.138	0.002	0.139	0.002	0.132	0.002	0.133	0.002	0.133	0.002	0.073	0.008	0.035	0.188	0.035	0.195
1991	0.129	0.003	0.129	0.003	0.129	0.003	0.129	0.003	0.129	0.003	0.000	1.000	0.000	1.000	0.000	1.000
1993	0.186	0.003	0.167	0.003	0.174	0.003	0.159	0.002	0.156	0.002	0.443	0.000	0.234	0.000	0.211	0.000
1995	0.191	0.003	0.177	0.003	0.176	0.003	0.162	0.002	0.160	0.002	0.485	0.000	0.263	0.000	0.245	0.000
1998	0.206	0.005	0.189	0.004	0.188	0.004	0.172	0.004	0.173	0.004	0.599	0.000	0.338	0.000	0.346	0.000
2000	0.183	0.003	0.169	0.003	0.168	0.003	0.155	0.002	0.161	0.002	0.423	0.000	0.208	0.000	0.248	0.000

GE1

	Actual figure		Holding dep. Inc. const.		Holding s-e inc. const		Count. 1: Holding lab inc const		Count. 2: lab +pens const		Diff. of Actual wrt base year		Diff. Count. 1 wrt base year		Diff. Count. 2 wrt base year	
Year	index	s.e.	index	*s.e.*	index	*s.e.*	index	*s.e.*	%	p-val	%	p-val	%	p-val	%	p-val
1977	0.189	0.007	0.176	0.006	0.158	0.004	0.146	0.004	0.120	0.003	0.474	0.000	0.141	0.003	-0.060	0.173
1978	0.179	0.007	0.158	0.006	0.161	0.005	0.141	0.004	0.114	0.003	0.398	0.000	0.102	0.035	-0.113	0.008
1979	0.194	0.009	0.179	0.008	0.161	0.006	0.148	0.005	0.127	0.005	0.515	0.000	0.157	0.005	-0.009	0.870
1980	0.230	0.021	0.216	0.019	0.179	0.016	0.170	0.015	0.152	0.014	0.795	0.000	0.325	0.007	0.186	0.101
1981	0.165	0.006	0.151	0.005	0.152	0.006	0.138	0.005	0.119	0.004	0.291	0.000	0.082	0.128	-0.068	0.186
1982	0.148	0.006	0.141	0.006	0.130	0.006	0.125	0.005	0.113	0.005	0.156	0.012	-0.026	0.641	-0.121	0.022
1983	0.146	0.004	0.142	0.003	0.135	0.003	0.132	0.003	0.119	0.002	0.142	0.002	0.029	0.492	-0.071	0.087
1984	0.154	0.004	0.146	0.004	0.141	0.004	0.134	0.004	0.124	0.004	0.205	0.000	0.048	0.327	-0.029	0.541
1986	0.166	0.005	0.160	0.005	0.140	0.004	0.136	0.003	0.129	0.003	0.297	0.000	0.060	0.189	0.012	0.797
1987	0.173	0.003	0.168	0.003	0.154	0.003	0.149	0.003	0.146	0.003	0.354	0.000	0.165	0.000	0.143	0.001
1989	0.140	0.003	0.140	0.003	0.132	0.002	0.132	0.002	0.132	0.002	0.093	0.032	0.035	0.405	0.033	0.434
1991	0.128	0.005	0.128	0.005	0.128	0.005	0.128	0.005	0.128	0.005	0.000	1.000	0.000	1.000	0.000	1.000
1993	0.174	0.003	0.160	0.003	0.166	0.003	0.153	0.003	0.151	0.003	0.362	0.000	0.199	0.000	0.179	0.000
1995	0.187	0.004	0.172	0.004	0.174	0.003	0.160	0.003	0.158	0.003	0.458	0.000	0.247	0.000	0.232	0.000
1998	0.201	0.008	0.187	0.008	0.178	0.006	0.165	0.006	0.167	0.006	0.571	0.000	0.293	0.000	0.303	0.000
2000	0.180	0.004	0.166	0.004	0.165	0.004	0.152	0.004	0.157	0.004	0.403	0.000	0.191	0.000	0.226	0.000

GE2

	Actual figure		Holding dep. Inc. const.		Holding s-e inc. const		Count. 1: Holding lab inc const		Count. 2: lab +pens const		Diff. of Actual wrt base year		Diff. Count. 1 wrt base year		Diff. Count. 2 wrt base year	
Year	index	s.e.	index	*s.e.*	index	*s.e.*	index	*s.e.*	%	p-val	%	p-val	%	p-val	%	p-val
1977	0.257	0.021	0.226	0.017	0.196	0.009	0.175	0.008	0.142	0.006	0.680	0.000	0.142	0.159	-0.073	0.446
1978	0.246	0.020	0.204	0.016	0.204	0.011	0.169	0.009	0.129	0.004	0.612	0.000	0.107	0.303	-0.154	0.095
1979	0.267	0.030	0.233	0.025	0.203	0.018	0.180	0.015	0.153	0.013	0.748	0.001	0.175	0.180	0.000	0.999
1980	0.519	0.094	0.462	0.082	0.357	0.073	0.321	0.063	0.286	0.056	2.393	0.000	1.100	0.009	0.869	0.022
1981	0.225	0.015	0.192	0.012	0.203	0.013	0.173	0.011	0.148	0.010	0.469	0.000	0.134	0.233	-0.035	0.747
1982	0.212	0.026	0.197	0.024	0.180	0.022	0.168	0.019	0.151	0.017	0.383	0.047	0.102	0.511	-0.015	0.919
1983	0.189	0.009	0.179	0.008	0.165	0.006	0.157	0.005	0.142	0.005	0.233	0.026	0.029	0.754	-0.074	0.420
1984	0.194	0.008	0.180	0.008	0.182	0.010	0.170	0.009	0.156	0.009	0.271	0.009	0.112	0.297	0.023	0.824
1986	0.267	0.022	0.248	0.020	0.199	0.015	0.188	0.013	0.178	0.013	0.743	0.000	0.231	0.063	0.167	0.165
1987	0.211	0.006	0.204	0.006	0.182	0.005	0.175	0.005	0.170	0.005	0.380	0.000	0.142	0.132	0.114	0.224
1989	0.169	0.006	0.170	0.006	0.154	0.005	0.154	0.005	0.154	0.005	0.102	0.284	0.009	0.919	0.005	0.958
1991	0.153	0.013	0.153	0.013	0.153	0.013	0.153	0.013	0.153	0.013	0.000	1.000	0.000	1.000	0.000	1.000
1993	0.205	0.006	0.188	0.006	0.197	0.006	0.182	0.006	0.179	0.006	0.338	0.000	0.190	0.047	0.170	0.076
1995	0.243	0.009	0.220	0.008	0.219	0.007	0.198	0.007	0.196	0.007	0.589	0.000	0.294	0.003	0.279	0.004
1998	0.291	0.028	0.269	0.026	0.232	0.018	0.213	0.017	0.215	0.017	0.902	0.000	0.392	0.005	0.409	0.004
2000	0.237	0.014	0.216	0.013	0.213	0.013	0.194	0.012	0.200	0.013	0.548	0.000	0.266	0.024	0.307	0.011

Source: our calculation on SHIW-HA data

Table 6.10: Counterfactuals using Burtless methodology - Base year is 1991

6.5 Conclusions

This chapter combined the DiNardo et al. (1996) and Burtless (1999) microsimulation methodologies for decomposing income inequality indices. The purpose of this combination was to provide a unifying framework for inequality decomposition analysis that corresponds to decomposition by "population subgroups" and by "factor sources". It does not provide a picture of inequality evolution with zero residual, however it shows where the main changes came from.

The combination of the DFL and B methodologies was applied to Italian household income distribution across the period 1977-2000. Results show that socio-demographic factors, such as the reduction of average household size and increased probability of receiving pension income had a negligible effect on household income distribution. The increased participation of women in the labor market as well as the increased proportion of income earners in the household were more effective in increasing inequality in the period after 1991. Results also showed that socio-demographic factors are less relevant in determining inequality dynamics than effects of dynamics of dispersion of income sources. In fact the changed dispersion of employment and self-employment income played the major role in explaining where the increase of household inequality came from. The increasing dispersion of self-employment and of employment income had a major role in explaining the increase in overall inequality after 1991. The increased dispersion of individual incomes together with the evolution of female labor force participation and number of income receivers explains most of the increase in inequality recorded by $GE(2)$ but most of the inequality by $GE(0)$

and $GE(1)$ still remains partly unexplained. Results also suggest that the concern about pension income is often misplaced. While the pension income trend is often taken to be a major cause of increasing inequality in the 1990s, it actually had only a limited effect during those years. By contrast, household inequality during the 1980s would have been significantly lower had pension income been distributed as in 1991.

The approach taken in this chapter builds on that of Daly and Valletta (2002) but differs from that in three main respects. First, the concern of Jenkins (1995) that analysis often changes because different years are compared is taken seriously: this microsimulation study is extended to each and every year available in the data set and then the overall trend is discussed. Secondly, the B methodology is extended to all labor income receivers, regardless of their sex and their role in the family, while Daly and Valletta (2002) applied it to male household heads only. Basically, my extension is motivated by the assumption that household heads and non-household heads do not have different income distributions and by the fact that male household head is becoming less important across time (see Section 6.2.1, page 140). Finally, the B methodology is extended to pension income and the effect of labor income divided into employment and self-employment. The different propensity to work or receive a pension are considered holding the number of months of income each individual received constant and then replacing only the monthly income vector rather than the yearly income. Structural changes in the economy, that would induce a change in income distribution, are not considered. The increased dispersion of income, for example, could be due to an increasing importance of specialized and skill-intensive industries that pay high skill premia. The effect of these changes as well as differing

employment probabilities in different years at different level of incomes are not considered. These factors should account for changes in inequality that are not explained by the present analysis.

Finally, it was shown that data quality is an issue. Some authors have recently written that the Italian distribution shows large fluctuations but no trend. Our analysis suggest instead that there was a trend in household income inequality, at first decreasing up until 1991, then increasing. Large fluctuations, especially in the pre-1991 period, are mainly due to data contamination. These problems are of course larger the more important are issues such as non-response and under-reporting. Results also show that no matter what concern we may have about the reliability of self-employment income data, if we are interested in household equivalent income we cannot neglect the role of self-employment dynamics and should instead think of possible improvements in survey data collection.

6.6 Appendix C: The Generalized Entropy class of inequality indices

In this chapter three different inequality indices are considered: the Generalized Entropy (GE) indices, with $a = 0, 1, 2$. They are known as the mean logarithmic deviation ($GE(0)$), the Theil index ($GE(1)$) and half the square of the coefficient of variation ($GE(2)$). In order to incorporate the sample weights the GE indices considered can be formalized as follows. Given a vector of incomes $\mathbf{y}$ of dimension N,

its arithmetic mean, $\overline{y}$, and a vector of weights, $\mathbf{w}$, of the same dimension as $\mathbf{y}$, and such that $\sum_{i=1}^{N} w_i = N$, the GE class of inequality indices is given by

$$GE(a) \equiv I_a = \frac{1}{a(a-1)} \left[\left[\sum_{i=1}^{N} \frac{w_i}{N} \left(\frac{y_i}{\overline{y}} \right)^a \right] - 1 \right], a \neq 1, a \neq 0 \tag{6.5}$$

$$GE(0) \equiv I_0 = \sum_{i=1}^{N} \frac{w_i}{N} \log \left(\frac{\overline{y}}{y_i} \right) \tag{6.6}$$

$$GE(1) \equiv I_1 = \sum_{i=1}^{N} \frac{w_i}{N} \frac{y_i}{\overline{y}} \log \left(\frac{y_i}{\overline{y}} \right) \tag{6.7}$$

These indices are chosen because they should provide a broad picture of the distribution. In fact, these inequality indices differ in their sensitivity to difference in various parts of the distribution: the more positive the parameter a of the GE class is the more $GE(a)$ is sensitive to income differences at the top of the distribution, the smaller a is the more $GE(a)$ is sensitive to differences at the bottom of the distribution (Cowell, 1995). Cowell and Flachaire (2002) examined the sensitivity of estimates of some inequality indices to extreme values, in terms of their robustness properties and of their statistical performance. Their analysis is performed using the influence function, a tool taken from the theory of robust estimation. They find that GE indices are not robust when the tail of the distribution is heavy, and in particular that GE measures with $a > 1$ are very sensitive to high incomes in the data. Extreme values can appear in the income distribution because of data contamination but also because the true distribution is thick-tailed (in Chapter 7 the issue of performing inference with thick tail distributions will be discussed).

Chapter 7

Inference issues with thick tail distributions

As was discussed in Chapter 6, Section 6.4 - with an application to Italian data - and in Section 6.6, inequality indices can be highly non-robust in the presence of extremely large income values. There are two main reasons why it is possible to find high values in an income data set. The first is because data could have been contaminated, due to, for instance, measurement error or misreporting; the second is because income distribution may truly be thick-tailed, i.e. it is possible that some people are very much richer than the average.

The robustness of inequality indices has been addressed in a number of recent papers. Cowell and Victoria-Feser (1996) study the effect of data contamination on income distribution inequality and suggest adopting a parametric approach to income distribution analysis and inequality measurement, estimating inequality from robust estimates of the parameters of the income distribution model. Cowell and Flachaire

(2002) address the sensitivity of estimates of inequality to extreme values and find that these measures are very sensitive to the properties of the income distribution. They show that estimation and inference can be dramatically affected when the tail of the distribution is thick. As a possible solution, they suggest making use of the semi-parametric approach introduced by Cowell and Victoria-Feser (2001): apply an appropriate functional form to the tail of the income distribution and estimate the parameters of this functional form robustly. However the issue of correct estimation of the tail of the income distribution remains: the assumption that the tail is parametrically distributed does not eliminate all the troubles. For instance, assuming that the tail of the distribution is distributed as a Pareto distribution reduces the problem to the robust estimation of the tail and location parameters but there is no clear results about which is the robust estimation to be preferred. In fact, the literature on robust estimation of the location parameter is large and still evolving (among others, see Victoria-Feser and Dupuis (2003); Hsieh (1999); Beirlant et al. (1996)).

This chapter shows that Pareto distributions are also affected by other kind of problems. For instance, given an income data set that was extracted from a Pareto distribution, a standard test to verify that the average income is larger than some given number can be seriously misleading if the tail parameter is $1 < \alpha \leq 2$. Analogous problems arise with a two-sample test of difference in mean when the samples are both drawn from a Pareto distribution, and the location parameter of at least one of the two is $1 < \alpha \leq 2$. The analysis that follows focuses on the t-ratio of thick tail distributions. In fact, it is well known that the t-ratio of any sequence $\{X_i\}$ of i.i.d. random variables converges towards a standard normal distribution, where the

t-ratio is defined as

$$t = \frac{\overline{X}}{S} \tag{7.1}$$

$\overline{X} = N^{-1}\sum_i^N X_i$ and $S = N^{-2}\sum_{i=1}^N X_i^2$, $i = 1, ...N$. However, for the t-ratio to converge to a standard normal it is required that $EX_i^2 < +\infty$.

Phillips and Hajivassiliou (1987) studied the t-ratio of a random sample $X_1, ..., X_N$ from a Cauchy distribution, where the r-moments are all infinite for $r = 1, 2, ...$ They showed that the statistic t does converge towards a stable distribution, which is bimodal with modes around ± 1.

This chapter provides a simulation-based study of the asymptotic distribution of the classical t-ratio for distributions with no finite variance, it discusses how classical testing is affected and then proposes an alternative way to perform inference with such distributions. As reviewed in Section 7.1, Pareto and symmetric Pareto distributions have finite first moments provided the tail-thickness parameter, α, is larger than 1. They have finite second moments provided the parameter $\alpha > 2$. Section 7.2 replicates simulation results about the distribution of t-ratios with Cauchy distributions of Phillips and Hajivassiliou (1987) and extends them to the Pareto and symmetric Pareto distributions with $0 < \alpha \leq 2$. Section 7.3 discusses the issue of naive testing with Pareto or symmetric Pareto distributions with finite mean and infinite variance: it discusses the error that a researcher would commit assuming that the standard t-ratio test is normally distributed when in fact it is not. A solution is suggested in Section 7.4 and an illustration using Italian household income data is proposed in Section 7.5.

7.1 The t-ratio distributions of some infinite variance distributions

Distribution functions with infinite first moment belong to the family of thick tail (TT) distributions. In the literature there is no universally accepted definition of a TT distribution. In general, a random variable from a TT distribution presents a non negligible probability of assuming very large values (extremal events). In other words, TT distributions have more weight in the tails than some reference distribution. When it is assumed that the reference distribution is the normal whose tails decay as the square of an exponential, it implies that distributions with power-law decay, as well as with exponential decay, are considered to be TT distributions. Other, more complete definitions, consider as TT a distribution whose exponential moments are infinite, $E(e^{tX}) = \infty, \forall t \geq 0$, which implies that the moment-generating function does not exist. Since different distributions have different degrees of thick-tailedness, a number of quantitative indicators for evaluating the probability of extremal events have been developed, such as the *extremal claim index* to assign weights to the tails and thus the probability of extremal events (Embrechts et al., 1999). Finally, some cruder though widely used definitions consider as TT a distribution with an infinite variance, or kurtosis larger than 3 (leptokurtic) (Bryson, 1982).

In what follows I focus on some aspects of some distributions with thick tails, namely the Cauchy, the Pareto and the symmetric Pareto with infinite mean or with finite mean but infinite variance. Phillips and Hajivassiliou (1987) studied in detail the t statistic defined in (7.1) from a Cauchy random sample, providing theoretical

and simulation results about its asymptotic distribution. They showed that when $X_1, X_2, ..., X_N$ is a random sample from a Cauchy (0,1) population, the numerator and the denominator of t_1 converge weakly to random variables, which are dependent, as $n \rightarrow \infty$. Hence, asymptotically the t-statistic is a ratio of random variables. It is well known that the ratio of two random variables gives rise to a random variable with a possibly bimodal distribution. Such a distribution is derived in Fieiller (1932) and its density in Phillips (1982). Marsaglia (1965) shows the conditions under which the ratio of two independent normal random variables with variance 1 and different means has a bimodal rather than a unimodal distribution.

Phillips and Hajivassiliou (1987) show that the phenomenon of bimodality can also occur with the classical t-ratio test statistic for populations with undefined second moments as is the case of the standard Cauchy (0,1). In the classical case the numerator and denominator statistics in the t-ratio are independent and, as $n \rightarrow \infty$, the denominator, properly scaled, converges in probability to a constant. They argue that the dependence of the numerator and denominator in the t-statistic is the main factor that induces the bimodality in the distribution. The fact that the modes are at ± 1 comes from simulation evidence that the numerator and denominator of the t-statistic are identical up to the sign. They suggest studying the distribution of the t-statistic focusing on the dependence between the numerator and denominator statistics. Such dependency remains even in the limit. In fact they showed that S^2 converges weakly towards a stable random variate with exponent $\alpha = 1/2$ and that the numerator and the denominator of the t-statistic follow a jointly stable distribution.

Stable distributions are not in general a subset of TT distribution as they also include the normal distribution. There are four different and equivalent ways to define stable distributions (Samorodnitsky and Taqqu, 1994; Focardi, 2001). A key property of a stable distributions is that a random variable is said to have a stable distribution if it has the same distribution of the (normalized) independent sum of any number of identical replicas of the same variable. This property involves that the entire distribution is equal and not only the tail. Note that in this context equal distribution means that distributions have the same functional form but they can have different parameters[1].

This chapter contributes on the study of the t-ratio statistics when first moments do not exist. In particular, it will focus on the case of the t-ratios from random samples extracted from distributions with finite mean and infinite variance and on the implications this has for hypothesis testing.

For distributions with infinite first moment the t-ratio statistic will be defined as:

$$t_1 = \frac{\overline{X}}{S_X} = \frac{\sum_1^N X_i/N}{\sqrt{(\sum_1^N (X_i - \overline{X})^2/N^2)}} \tag{7.3}$$

[1]Formally, a random variable X is said to have a stable distribution if, for any positive number a_i, there exist a positive number c and a real number d such that

$$\sum_{i=1}^{n} a_i X_i \xrightarrow{d} cX + d \tag{7.2}$$

where X_i are independent copies of X, and $\xrightarrow{d}$ denotes convergence in distribution.

and for distributions with finite first moment, it will be defined as:

$$t_2 = \frac{\overline{X} - \mu}{S_X} = \frac{\sum_1^N X_i/N - \mu}{\sqrt{(\sum_1^N (X_i - \overline{X})^2/N^2)}} \tag{7.4}$$

where $X_1, X_2, ..., X_N$ is a random sample from some distribution and μ is the true mean. It can be proved that the difference between S and S_X is negligible as $N \to \infty$ as well as the difference between t and t_1, hence asymptotically they give the same results[2]: t_2 differs from t_1 only in the location factor, $-\mu/S_X$.

The first TT distribution considered is the standard Cauchy (0,1). It has density function (DF)

$$\frac{1}{\pi(1+x^2)} \tag{7.5}$$

and all its moments are infinite.

The Pareto distribution (type I) has DF

$$f(x) = \alpha\beta^\alpha x^{-\alpha-1} \tag{7.6}$$

The Pareto cumulative distribution function (CDF) is

$$F(x) = 1 - \left(\frac{\beta}{x}\right)^\alpha, x \geq \beta, \alpha > 0, \beta > 0 \tag{7.7}$$

The first moment, $E(x)$, exists if $\alpha > 1$ and the second central moment, $V(x)$, exists if $\alpha > 2$:

[2]Formally, $S^2 - S_X^2 = O_p(N^{-1})$ and $t - t_1 = O_p(N^{-1})$ (Phillips and Hajivassiliou (1987), Lemma 1, p. 5.)

$$E(x) = \frac{\alpha}{\alpha - 1}\beta \tag{7.8}$$

$$V(x) = \frac{\alpha}{(\alpha - 1)^2(\alpha - 2)}\beta^2 \tag{7.9}$$

The Pareto distribution belongs to the family of the exponential distribution since its DF can be written as:

$$p_\alpha(x) = C(\alpha) \times e^{\left(\sum_{i=1}^{k} Q_{i(}(\alpha)t_i(x)\right)h(x)} \tag{7.10}$$

with $C(\alpha) = \alpha\beta^\alpha$, $Q_i(\alpha) = -(\alpha + 1), t_i(x) = \ln x, h(x) = 1$ (Silvey, 1975). Analogously to the bilateral exponential (see, Feller, 1971), p. 49), the CDF of the symmetric Pareto distribution, with $|x| > \beta$, is:

$$F(x) = 1 - \frac{1}{2}\left(\frac{\beta}{|x|}\right)^\alpha, |x| > \beta, \beta > 0 \tag{7.11}$$

The DF can be written as:

$$f(x) = \frac{1}{2}\alpha\beta^\alpha |x|^{-\alpha-1} \tag{7.12}$$

and can be seen as the convolution of the Pareto density $\alpha\beta^\alpha x^{-\alpha-1}$ $(x \geq \beta, \alpha > 0, \beta > 0)$ with the mirrored density $\alpha\beta^\alpha(-x)^{-\alpha-1}$ $(x \leq -\beta, \alpha > 0, \beta > 0)$. In other words, the symmetric Pareto is the density of $X_1 - X_2$ when X_1 and X_2 are independent and have the common exponential density $\alpha\beta^\alpha x^{-\alpha-1}$ $(x > \beta, \beta > 0, \alpha > 0)$.

Its first two centered moments are (see Appendix D, Section 7.7):

$$\begin{aligned} E(x) &= 0, \alpha > 1 \\ V(x) &= 2\frac{\alpha}{(\alpha-1)^2(\alpha-2)}\beta^2, \alpha > 2 \end{aligned}$$

7.2 Simulation results

The results that follow have been obtained via Monte Carlo simulations from random samples of dimension N using the method of inverted CDF, i.e. a random sample of dimension N is extracted from a unit rectangular variate, $U(0,1)$, and then it is mapped into the sample space using the inverse CDF. The number of simulations M has been set to 10,000. This study allows one to disentangle some differences about the asymptotic distribution of the t-ratio statistic when either one or both first two moments do not exist.

The Cauchy and the symmetric Pareto distribution with $\alpha \leq 1$ are both symmetric and with infinite mean. For these distributions, as sample size increases, the statistic t_1 converges towards a stable distribution which is symmetric and bimodal. The convergence is fairly rapid, even for samples as small as 10, and the two modes are located at ± 1. As for the symmetric Pareto, the t-ratio distribution does depend on α: the lower is α, the higher is the concentration around the two modes (Figure 7-1).

For $1 < \alpha < 2$ the t-ratio, t_2, is not always clearly bimodally distributed. The more α departs from 1 the less evident is the bimodal distribution of the t-ratio and the clearer the convergence towards a standard normal distribution (Figure 7-2). This

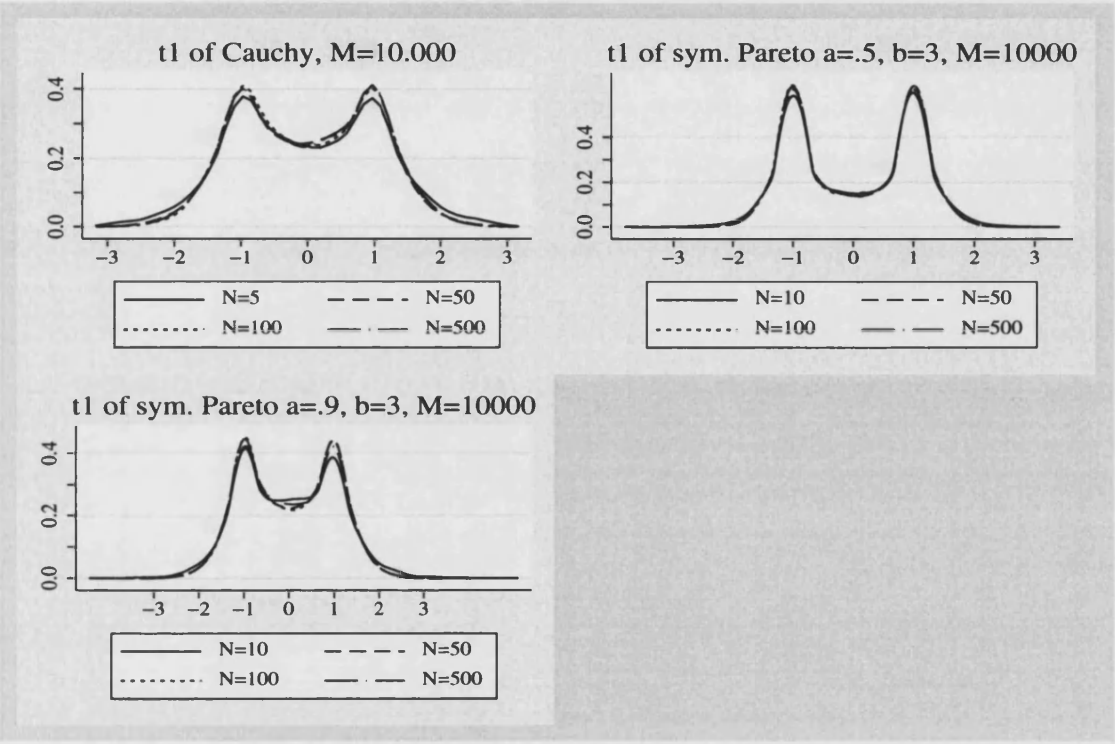


Figure 7-1: t-ratio of Cauchy and infinite-first-moment symmetric Pareto distributions

result applies for any value of $\beta > 0$, since β is simply a threshold parameter that does not affect the t_1 statistic behavior.

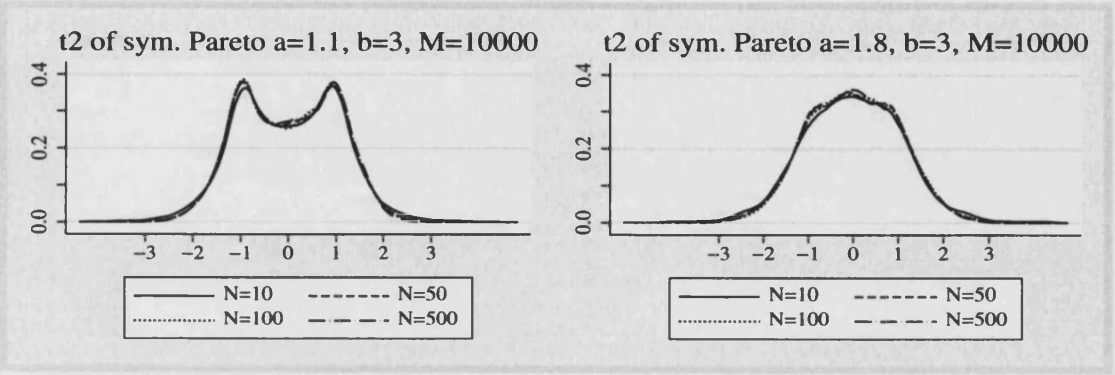


Figure 7-2: t-ratio of symmetric Pareto distributions with $1 < \alpha \leq 2$

Moving to a Pareto distribution defined on a positive support[3] with $\alpha \leq 1$, the t-ratio t_1 is clearly non-normal. However, the convergence towards a unimodal distribution with mode located just above 1, is clearer the smaller is α. The closer α gets to 1, the more dispersed the distribution becomes (Figure 7-3).

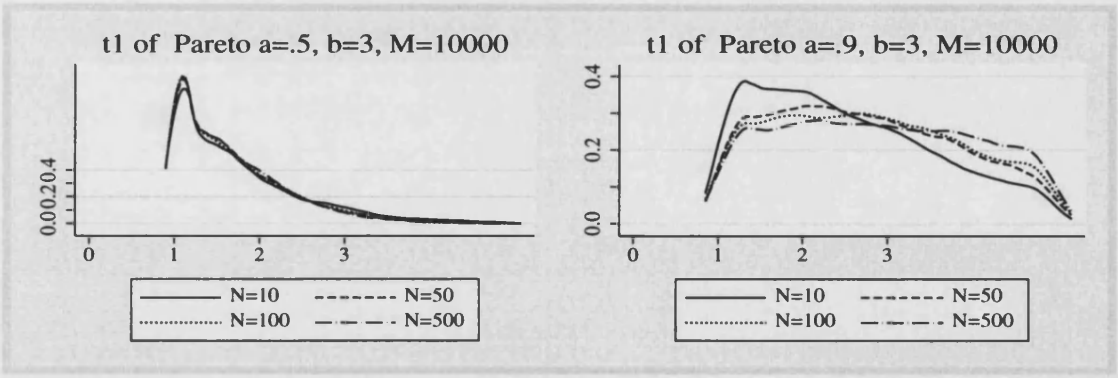


Figure 7-3: t_1 Pareto distributions with $0 < \alpha \leq 1$

When a sample is randomly drawn from a Pareto distribution with α grater than

[3]Results for the negative Pareto distribution are symmetric to those for the positive Pareto and are not presented here.

1 and less than 2, the t-ratio, t_2, is still clearly non-normally distributed. Due to the occurrence of large values, the t_2 distribution is asymmetric and biased towards negative values. The closer α gets to 2, the clearer the convergence to a standard normal appears. With α=1.8, the distribution is still clearly non-normal with strong skewness to the left (Figure 7-4).

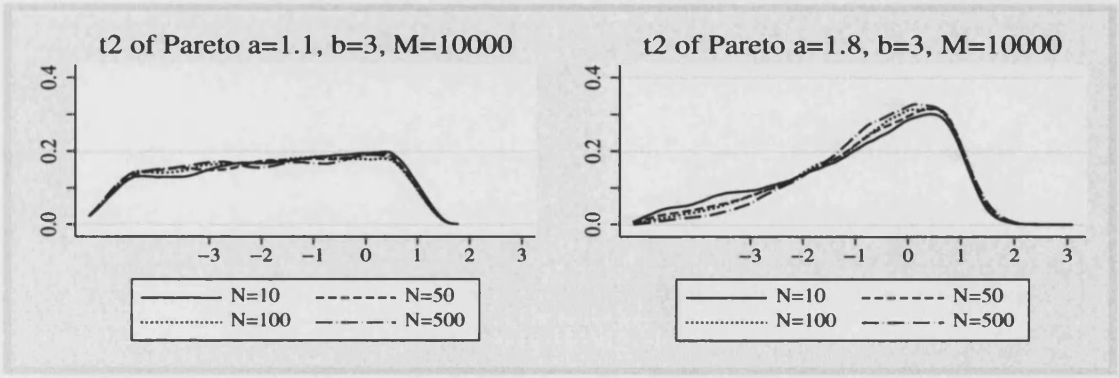


Figure 7-4: t_2 of Pareto distributions with $1 < \alpha \leq 2$.

The regularity in the t_1 distribution leads us to investigate the relationship between the first and the second centered moment, respectively in the numerator and denominator of t_1. Phillips and Hajivassiliou (1987) noted that if the distribution is

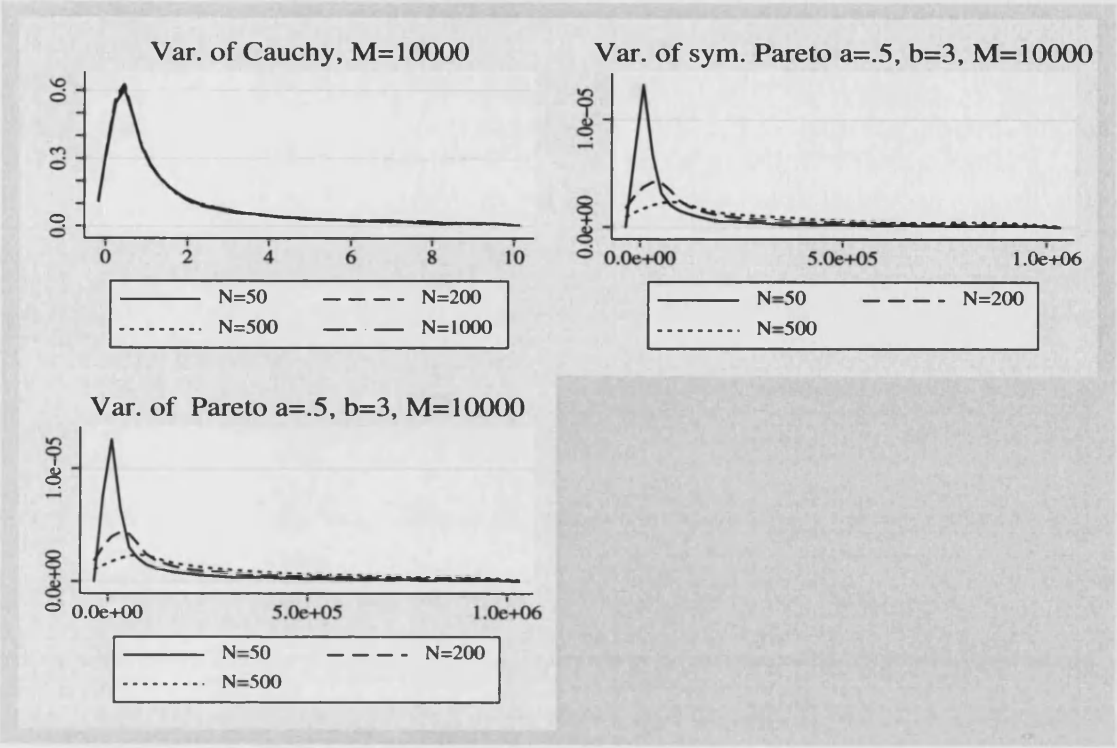


Figure 7-5: Distribution of the variance of some distributions with infinite mean.

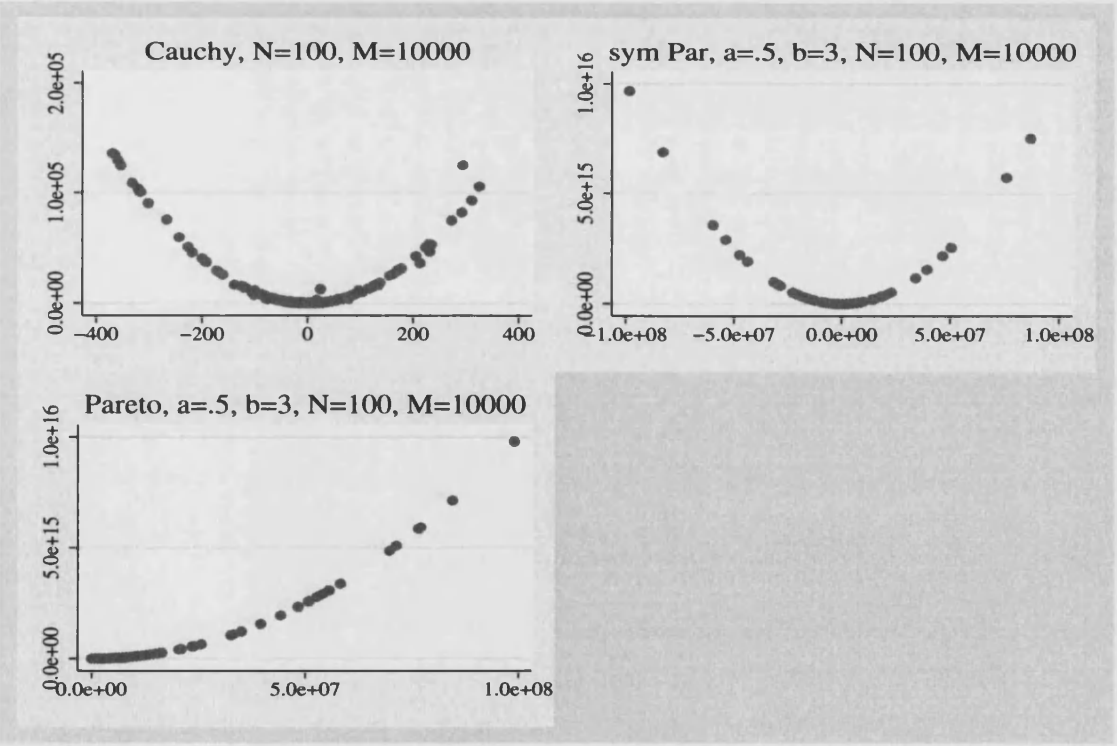


Figure 7-6: Distribution of the variance of some distributions with infinite mean.

A simple OLS estimate for the coefficient b of the parabolic relationship ($S_X^2 = a + b\overline{X}^2$) is always very close to 1 and highly significant using the Cauchy, the Pareto or the symmetric Pareto with $\alpha \leq 2$. However, the coefficient a is not significantly different from zero for any value of the sample size[4]. In other words, the average of the squared deviation from the sample mean can be well approximated by the square of the sample mean. This property is a direct consequence of the fact that the Pareto distribution with infinite variance belongs to the class of subexponential distribution, which is characterized by two properties: the convolution closure property and the property of the sum (Embrechts et al., 1999). The first property states that the shape of the tail is preserved after the summation of a random sample from a given subexponential distribution. The second property states that in a sum of observations from a random sample, the largest value will be of the same order of magnitude as the sum itself[5]. The latter property implies that the deviation from the mean will be of the same order of magnitude as the mean, hence the ratio between the mean of the squared deviation from the mean and the squared mean will be of the same order of magnitude. The fact that the modes of the bimodal distribution for the t_1 statistic are at ± 1 comes from this property and the fact that the sample mean can be negative whereas its standard error cannot.

[4]Phillips and Hajivassiliou (1987) found a b coefficient between .570 and .376 for the Cauchy distribution and different sample sizes. However, in their regression the dependent variable was the uncentered second moment while the centered one is considered here.

[5]Formally, for any sample size N, if $S_N(x) = \sum_{i=1}^{N} X_i$ is the sum of i.i.d. random variables and M_N is their maxima, it is verified that

$$\lim_{x \to \infty} \frac{P(S_N > x)}{P(M_N > x)} = 1 \tag{7.13}$$

7.3 Testing with TT distribution

The preceding results are of relevance for hypothesis testing in regressions with error terms that are independent and identically distributed as a Pareto, with $1 < \alpha \leq 2$. It is of interest also for testing the hypothesis of difference in means or other statistics of two samples when either or both come from a TT distribution.

What happens in these cases if the classical t-ratio test statistic is compared with the critical values of a $N(0,1)$ distribution? This problem is illustrated using the $p-$value discrepancy plot (Davidson and MacKinnon, 1998). The $p-$value discrepancy plot is based on the empirical distribution function (EDF) of the $p-$values of some test statistic τ, generated via Monte Carlo simulation using a data-generating process (DGP) that is a special case of the null hypothesis. The simulation is usually carried out for a large number of M replications obtaining simulated values $\tau_j, j = 1, 2, ...M$. The $p-$value of the τ_j is the probability of observing a value of τ more extreme than τ_j, according to some distribution $F(\tau)$. This distribution could be the asymptotic distribution of τ, derived numerically or theoretically, as well as other distributions such as an approximation derived by bootstrapping. The $p-$value is a function of τ_j, $p_j \equiv p(\tau_j)$. Assuming τ is asymptotically distributed as a standard normal with DF $\phi(z)$ and CDF $\Phi(z)$, then $p_j = 1 - \Phi(\tau_j)$[6].

The EDF of the p_j is an estimate of the CDF of $p(\tau)$. At any point x_i in the $(0,1)$ interval, it is defined by

$$\widehat{F}(x_i) \equiv \frac{1}{m} \sum_{j=1}^{m} I(p_j \leq x_i) \tag{7.14}$$

[6]For a two-sided test, the $p-$value is $p_j \equiv p(|\tau_j|) = 2(1 - \Phi(\tau_j))$.

where $I(p_j \leq x_i)$ is a Boolean operator that takes the value 1 if the argument is true and 0 if not true. Although the function (7.14) can be evaluated at every data point, when m is large it is unnecessary in order to produce a reasonable picture of the $(0,1)$ interval or one of its portions. In these applications 1000 equally spaced data points are considered, $x_i, i = 1, 2, ...1000$. The simplest graph that can be analyzed is the plot of $\widehat{F}(x_i)$ against x_i. However, for dealing with test statistics that are well behaved, it is more revealing to plot the $p-$value discrepancy plot, namely $\widehat{F}(x_i) - x_i$ against x_i.

The $p-$value discrepancy plot of the t-ratio statistic, for the Pareto and the symmetric Pareto with different values of α was constructed as in (7.14), where the $p-$value is derived both using the standard normal and the distributions derived previously by simulation. The $p-$value discrepancy plot allows one to distinguish at a glance among test statistics that systematically over-reject, test statistics that systematically under-reject and test statistics that reject about the right proportion of times at each desired level of x_i: in the first case the plot will be over, in the second below, in the third around the zero line.

Let us now assume that we have a random sample from a symmetric Pareto distribution with $1 < \alpha \leq 2$ and we run a test $H_0 : \mu = \mu_0$ against the alternative $H_A : \mu \neq \mu_0$, where μ is the true mean and μ_0 some value on the real line. The sample mean is used to estimate μ. Performing such a test using the standard normal rather than the correct distribution causes the null hypothesis to be under-rejected by quite a small amount, not larger than 5% for tests of size 5%, and even less for tests of size 1% or 10%. This conclusion would often lead us to ignore the caveat

of having a systematic error in rejection probability using the standard normal for testing two-sided hypothesis with a symmetric Pareto distribution with $1 < \alpha \leq 2$. However, two important points should be noted.

First, the "ignore argument" can be an acceptable policy if the size of the test is smaller than 10%. If the test has a larger size - for instance 40% - the ERP can be larger than 10 and is obviously more difficult to tolerate[7]. Clearly, the former policy corresponds to minimize the type II error as opposed to minimize the type I error, as it is typically performed in economics and several other disciplines. In such cases it is common to find confidence intervals with about 60% coverage probability (see for instance Karlen, 2002).

Secondly, the "ignore argument" cannot be extended to the Pareto distribution. The ERP for a two sided test about the mean of a Pareto distribution with $1 \leq \alpha < 2$ can be quite larger than 10%. For instance if $\alpha = 1.1$, the test will over-reject H_0 about 60% of times (Figure 7-7), even for tests of size 5%. This result clearly comes from the non standard distribution of t_2 (Figure 7-4). The same concerns apply to one-sided tests: standard testing is highly unreliable. For instance, a test of the hypothesys $H_0 : \mu = \mu_0$ vs. $H_A : \mu > \mu_0$, for the Pareto distribution with $1 < \alpha \leq 2$, assuming asymptotic normality, will seriously under-reject with an ERP that increases with the size up to the 40% level. For test of the hypothesis $H_0 : \mu = \mu_0$ vs. $H_A : \mu < \mu_0$ the test will dramatically over-reject with an ERP which can be larger than 60%, even for tests of size 5% (Figure 7-8).

[7]Although tests of size larger than 10% are rather unusual in economics it is much less so in other disciplines, such as physics, where the main point is often to maximize the power of the test, rather than to minimize its size.

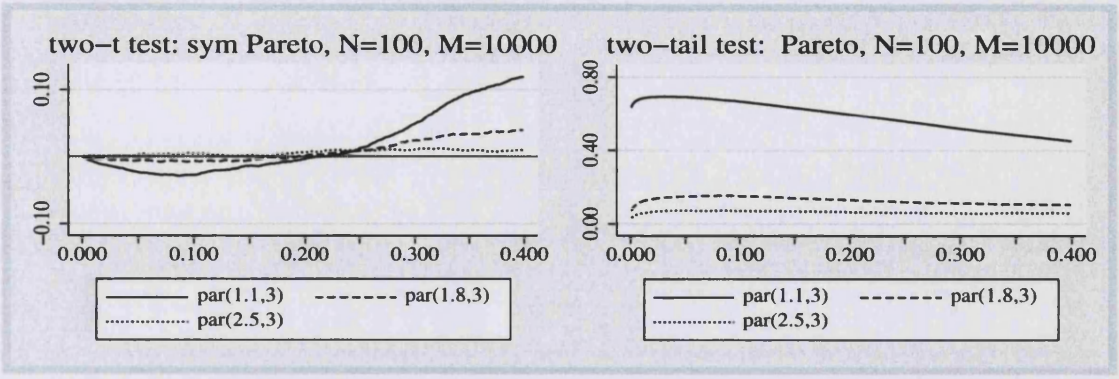


Figure 7-7: ERP for two-tail test with a sample from a symmetric or a positive definite Pareto with $1 \leq \alpha$ 2.

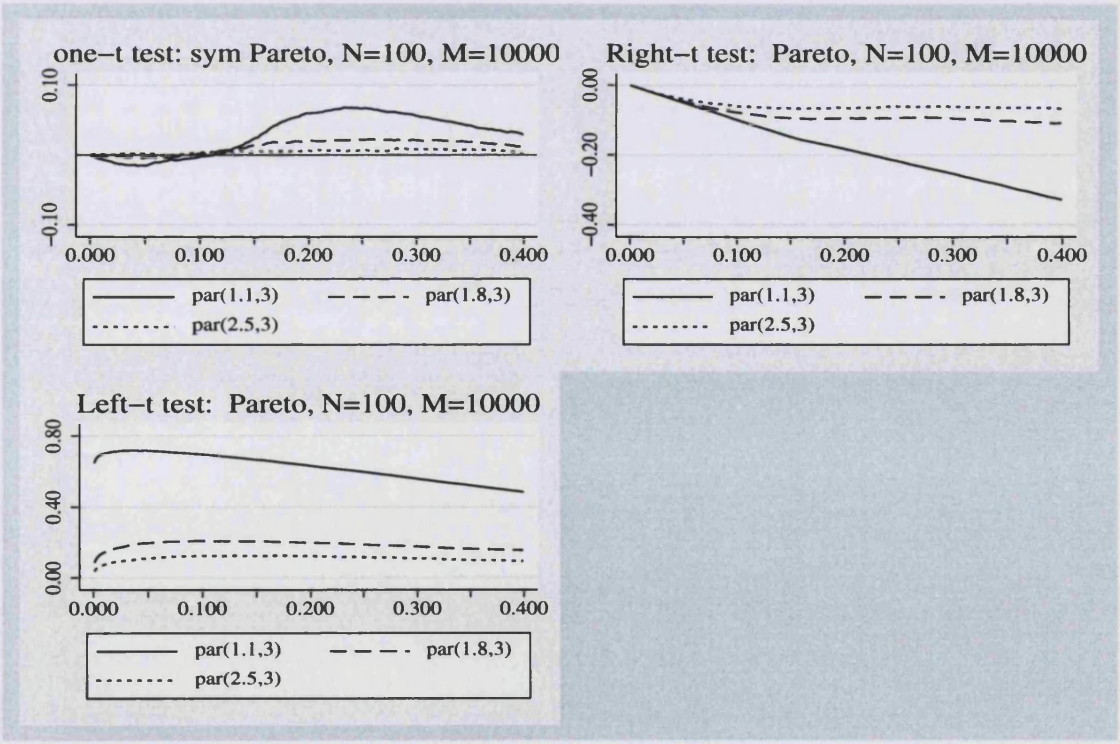


Figure 7-8: ERP for one-tail test with a sample from a symmetric or a positive definite Pareto with $1 \leq \alpha$ 2.

Obviously, the non standard distribution of the t-ratio with infinite second moment does also affect the two-sample test of difference of means. Let us assume that we have two independent samples from two different distributions, one of which is a Pareto distribution with infinite first or second moment. Call the two distributions A and B. We want to test whether the mean of the first is different from the mean of the second using the t-ratio

$$t_2^D = \frac{\mu_A - \mu_B}{\sqrt{(S_X^A/N^A) + (S_X^B/N^B)}} \tag{7.15}$$

where μ_A, μ_B are the true means, and S_X^A, S_X^B are the sample variance of A and B, respectively. The distribution of t_2^D is again non-standard. Moreover, in many cases it does not look to converge to a stable distribution as the sample size increases. Figure 7-9 shows via Monte Carlo simulations the distributions of the t-ratio, t_2^D, for testing the difference in means of two Pareto distributions that may differ in α but are constrained to have the same location parameter, $\beta = 3$, on the assumption that a sample of the same size has been drawn from each. Clearly, there is no point in using the sample t-ratio and comparing it with the normal critical values.

7.4 Solutions

The main problem with testing with TT distribution, such as the Pareto and symmetric Pareto with $1 < \alpha \leq 2$, comes from the fact that the t-ratio distribution converges towards a distribution which is clearly not the standard normal. It also

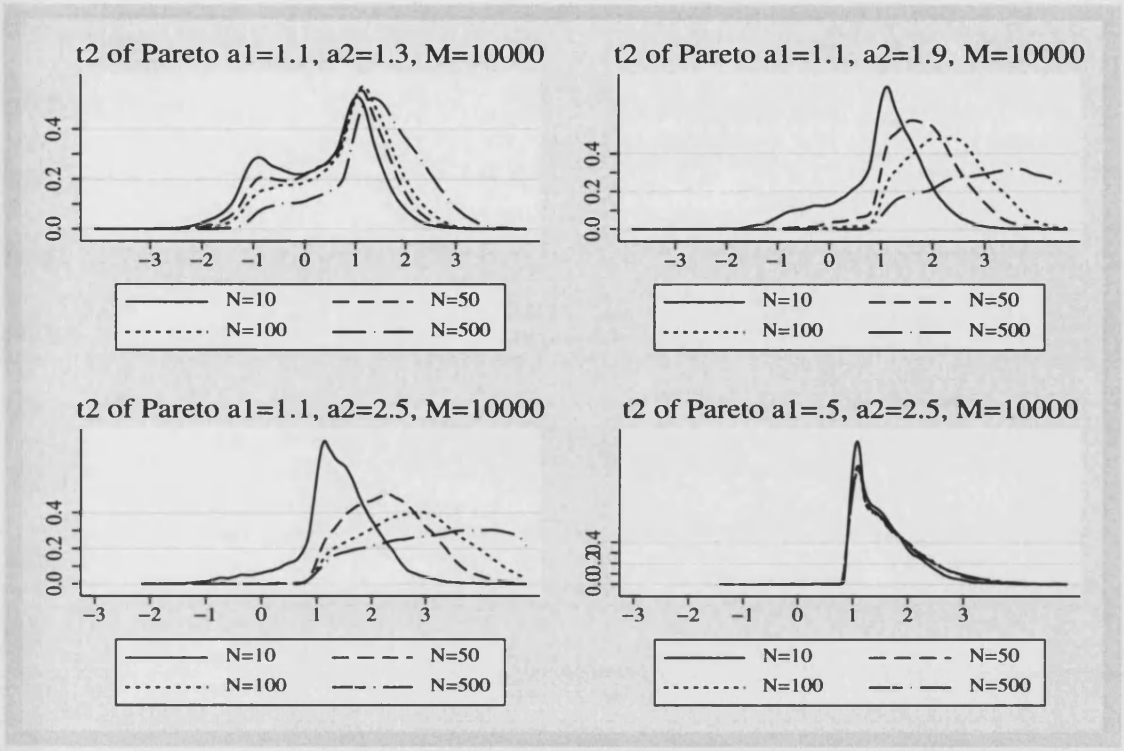


Figure 7-9: Test of difference in mean of Pareto distributions with $\beta = 3$.

depends on unknown parameters, such as the α in the Pareto and symmetric Pareto distributions. Provided we had full knowledge of the true α parameter of the Pareto distribution from which the sample has been drawn, the first solution would be to

complication derives from the fact that the parameter α is rather difficult to estimate with confidence even in large samples: the Maximum Likelihood estimation of α, being based on the sample mean, which is highly non-robust, is highly non-robust itself and presents a large variability (Rytgaard, 1990); the Hill estimator (Hill, 1975), which is based on ordered statistics and in its simplest form produces a plot to identify the α parameter, in many cases is totally unhelpful, producing what have been defined "Hill horror plots" (Embrechts et al., 1999). Hence, it is not possible to derive the true distribution of the t-ratio using Monte Carlo simulation since it changes quite significantly for different values of α, as we saw in Section 7.2.

An alternative solution is then to consider a more robust statistic than the mean, such as the median (Amemiya, 1985, among others).

Let $X_1, ..., X_N$ be a sample from a continuous distribution F, defined on the real line, and $X_1 < X_2 < ... < X_N$ be the order statistics, obtained arranging the observations in increasing orders without ties. The p-th quantile of F is defined as $x_p = F^{-1}(p)$, and the p-th sample quantile is defined as X_k where $k = [pN]$ is the smallest integer greater than or equal to pN. Provided the DF $f(x)$ exists and is continuous and positive in a neighborhood of some quantile, then the joint distribution of the corresponding sample quantile is asymptotically normal. For the median, $x_{.5}$, it can be proved (Ferguson, 1996, Ch. 13) that:

$$\sqrt{N}(X_{[.5N]} - x_{.5}) \xrightarrow{d} N\left(0, \frac{1}{4f(x_{.5})^2}\right) \tag{7.16}$$

and the t-ratio statistic is

$$t_3 = \frac{X_{.5} - x_{.5}}{S_X} \tag{7.17}$$

where $X_{.5}$ and $x_{.5}$ are the sample and true median, respectively, and $S_X = \frac{1}{2\sqrt{N}\widehat{f}(x_{.5})}$ where $\widehat{f}(x_{.5})$ is a consistent estimate of $f(x_{.5})$. The asymptotic normality of t_3 can be also seen in Figure 7-10 and 7-11, where $\widehat{f}(x_{.5})$ has been estimated using a kernel density estimator with fixed bandwidth.

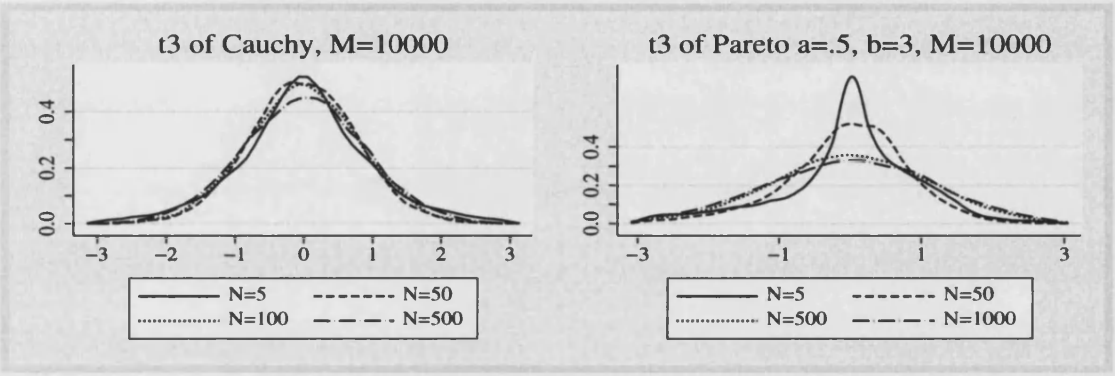
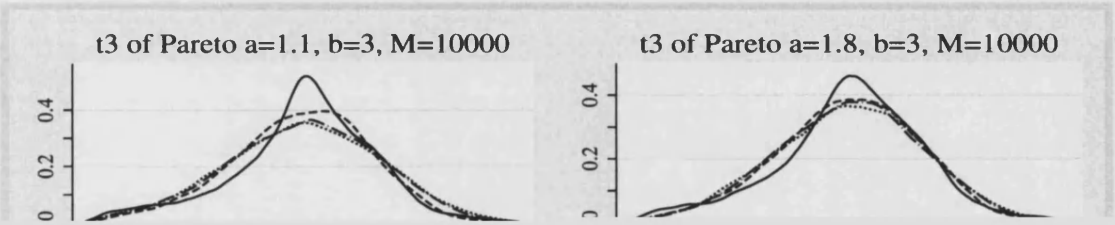


Figure 7-10: t_3 with infinite first moment distributions



line. In these cases, using a sample size $N = 100$, the ERP is never larger than 4% for tests of size less than 10%. The ERP is always smaller than 5% even with larger test sizes for two-tail and left-tail tests. It is negligible for left-tail tests (Figure 7-12).

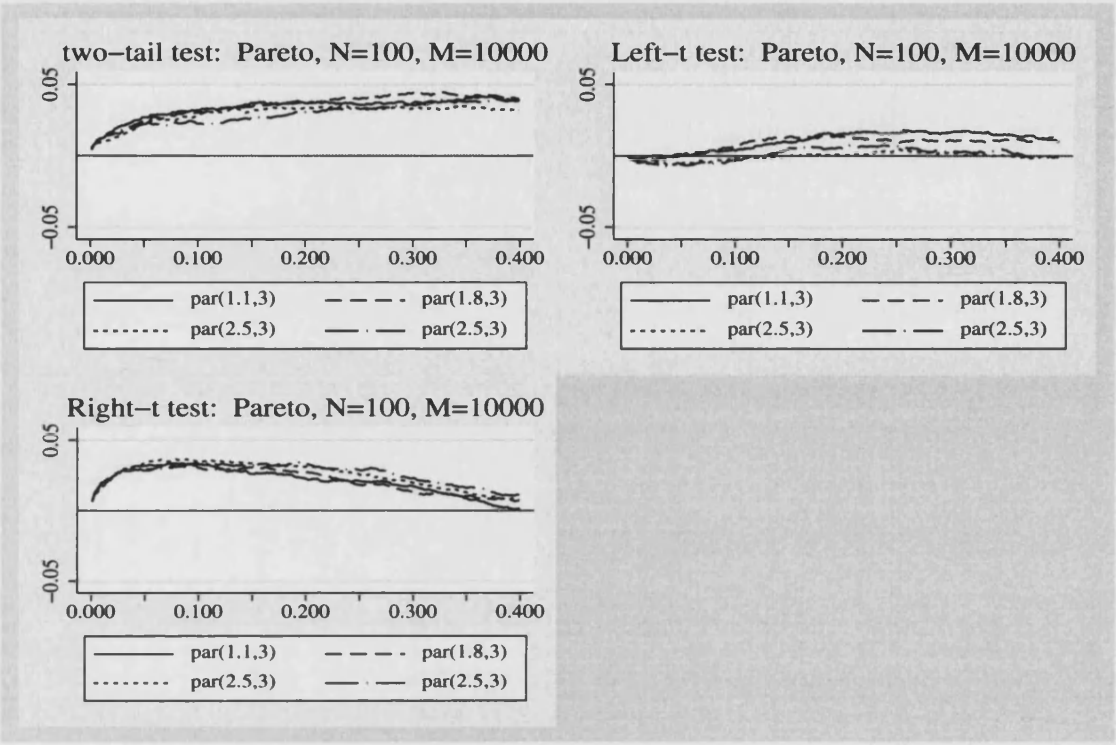


Figure 7-12: ERP with t_3 from a Pareto distribution with $1 < \alpha \leq 2$

7.5 An application to income data

'emely large
ions with a
)ecomes the
)ution. As a
M described

Income distributions typically present a non-negligible probability of extr
incomes. It is common practice to estimate the tail of income distribut
Pareto distribution. The estimation of the tail of the distribution then b
estimate of α and β, the two unknown parameters of the parametric distrib
simple illustration, I use the BT income as it was simulated using the MSI

income presents a larger probability of large incomes since the progressive personal income taxation compresses the income range. In this simple illustration, the problem of the consistent estimation of the β parameter is neglected. The variable is truncated at Lit 30 millions (about €15.000). The sample is split into two groups: the employed and the others. The incomes in the former group belongs to people who declare to be employed; the latter is the residual group. The average BT income over Lit 30 millions of the employed is 53.8 millions and maximum value is about 10 time larger than β. The average BT income of the residual group is on average about Lit 2 millions larger than employment income and the maximum value is about 50 times larger than β. The second sample is nearly twice the size of the first (Table 7.1). Testing the hypothesis that the average income over β of employment income is different from the average income over β of the non-employment income, the t-ratio statistic, t_2^D, is -1.739. A naive researcher, who does not check for the possibility that $\alpha > 2$ for both distributions, would compare it with the standard normal critical values and conclude that the two distributions are different and that the mean of the BT income of the employed group is smaller than the income of other group as the $p-$value is $Pr(t < -1.739) = 0.041$. Since β is given, she would also conclude that the α parameter is larger for the employed than for the others. However, as shown previously, the mean cannot be used if the samples are distributed as Pareto distributions with $\alpha < 2$. Moreover the mean is a highly non robust estimate and, for these type of data, it is likely that $\alpha < 2$.

Using the median instead, the t-ratio statistic, t_3, is 6.044. The results of Section

Variable	Obs	Mean	Median	Min	Max
employed	2093	53830.44	46138.3	30150.38	384154.1
others	3485	55718.96	42935.3	30075.19	1449541

Table 7.1: Summary statistics for BT income over Lit 30 millions

7.4, lead us conclude that the two medians are different and the samples come from two different Pareto distributions. Moreover, it also suggests that the median of the employed sample is larger that the the median of the residual group, i.e. the α parameter is larger for the latter group.

7.6 Conclusions

This chapter has investigated the issue of performing inference with TT distributions, which are often found in various fields of economics, including the income inequality literature[8]. It has been shown that when the distribution is TT, and the first moment is finite while the second is not, the standard t-ratio does not asymptotically converge to a standard normal distribution. Hence, it was discussed when inference is invalidated and how relevant the ERP can be. A simple road map is suggested to the careful researcher: whenever she suspects that the sample could come from a Pareto distribution with infinite variance, either she does not perform any inference at all or she uses the median. In the latter case it is then possible to compute a t-ratio statistic with it, and compare it with the critical values of a standard normal distribution.

This solution is superior to the classical t-ratio based on the sample mean and

[8]For a short review of the relevance of the TT distributions in the economics literature, see Appendix E, Section 7.8.

would be correct in many situations. However, the median-based t-ratio statistic could still present problems in particular samples: as studied in Cowell and Victoria-Feser (2002), the median can also be nonrobust, especially in the tails, where dead intervals are more likely to appear.

7.7 Appendix D: Moments of the symmetric Pareto distribution

The first moment of the symmetric Pareto distribution can be obtained using convolution:

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} t \left[\int_{\beta}^{\infty} (\alpha\beta^{\alpha})^2 (x+t)^{-\alpha-1} x^{-\alpha-1} dx \right] dt = \\
&= (\alpha\beta^{\alpha})^2 \int_{\beta}^{\infty} x^{-\alpha-1} \left[\int_{-\infty}^{\infty} t(x+t)^{-\alpha-1} dt \right] dx = \\
&= (\alpha\beta^{\alpha})^2 \int_{\beta}^{\infty} x^{-\alpha-1} \left[\int_{\beta-x}^{\infty} t(x+t)^{-\alpha-1} dt \right] dx = \\
&= (\alpha\beta^{\alpha})^2 \int_{\beta}^{\infty} x^{-\alpha-1} \left[\int_{\beta}^{\infty} (z-x)(z)^{-\alpha-1} dz \right] dx = \\
&= (\alpha\beta^{\alpha})^2 \int_{\beta}^{\infty} x^{-\alpha-1} \left[\int_{\beta}^{\infty} z^{-\alpha} dz - x \int_{\beta}^{\infty} z^{-\alpha-1} dz \right] dx = \\
&= (\alpha\beta^{\alpha})^2 \int_{\beta}^{\infty} x^{-\alpha-1} \left[\frac{1}{-\alpha+1} z^{-\alpha+1} \Big|_{\beta}^{\infty} dz + x \frac{1}{\alpha} z^{-\alpha} \Big|_{\beta}^{\infty} \right] dx = \qquad \alpha > 1 \\
&= (\alpha\beta^{\alpha})^2 \int_{\beta}^{\infty} x^{-\alpha-1} \left[\frac{\beta^{-\alpha+1}}{\alpha-1} - \frac{x\beta^{-\alpha}}{\alpha} \right] dx = \\
&= (\alpha\beta^{\alpha})^2 \int_{\beta}^{\infty} \frac{\beta^{-\alpha+1}}{\alpha-1} x^{-\alpha-1} - \frac{\beta^{-\alpha}}{\alpha} x^{-\alpha} dx = \\
&= \left(\alpha\beta^{2\alpha}\right) \left[\frac{\beta^{-\alpha+1}}{\alpha-1} \frac{1}{-\alpha} x^{-\alpha} \Big|_{\beta}^{\infty} - \frac{\beta^{-\alpha}}{\alpha} \frac{1}{-\alpha+1} x^{-\alpha+1} \Big|_{0}^{\infty} \right] = \\
&= 0
\end{aligned}
$$

The second central moment of the symmetric Pareto distribution is:

$$
\begin{aligned}
V(X) &= (\alpha\beta^{\alpha})^2 \int_{-\infty}^{\infty} t^2 \int_{\beta}^{\infty} x^{-\alpha-1}(x+t)^{-\alpha-1}dxdt = \\
&= (\alpha\beta^{\alpha})^2 \int_{\beta}^{\infty} x^{-\alpha-1} \int_{\beta-x}^{\infty} t^2(x+t)^{-\alpha-1}dtdx = \\
&= (\alpha\beta^{\alpha})^2 \int_{\beta}^{\infty} x^{-\alpha-1} \int_{\beta}^{\infty} (z-x)^2(z)^{-\alpha-1}dzdx = \\
&= (\alpha\beta^{\alpha})^2 \int_{\beta}^{\infty} x^{-\alpha-1} \left[\int_{\beta}^{\infty} z^{-\alpha-1}dz - 2\int_{\beta}^{\infty} xz^{-\alpha}dz + \int_{\beta}^{\infty} x^2 z^{-\alpha-1}dz \right] dx = \qquad \alpha > 2 \\
&= (\alpha\beta^{\alpha})^2 \int_{\beta}^{\infty} x^{-\alpha-1} \left[\frac{-\beta^{-\alpha+2}}{-\alpha+2} + \frac{2x\beta^{-\alpha+1}}{-\alpha+1} - \frac{x^2\beta^{-\alpha}}{-\alpha} \right] dx = \\
&= (\alpha\beta^{\alpha})^2 \left[\int_{\beta}^{\infty} \frac{\beta^{-\alpha+2}}{\alpha-2} x^{-\alpha-1}dx - \int_{\beta}^{\infty} \frac{2\beta^{-\alpha+1}}{\alpha-1} x^{-\alpha}dx + \int_{\beta}^{\infty} \frac{\beta^{-\alpha}}{\alpha} x^{-\alpha+1}dx \right] = \\
&= (\alpha\beta^{\alpha})^2 \left[\frac{\beta^{-\alpha+2}}{\alpha-2} * \frac{\beta^{-\alpha}}{\alpha} - \frac{2\beta^{-\alpha+1}}{\alpha-1} * \frac{\beta^{-\alpha+1}}{\alpha-1} + \frac{\beta^{-\alpha}}{\alpha} * \frac{\beta^{-\alpha+2}}{\alpha-2} \right] = \\
&= \alpha^2 \left[\frac{\beta^2}{\alpha(\alpha-2)} - \frac{2\beta^2}{(\alpha-1)^2} + \frac{\beta^2}{\alpha(\alpha-2)} \right] = \\
&= \frac{2\alpha\beta^2 \left[(\alpha-1)^2 - \alpha(\alpha-2) \right]}{(\alpha-2)(\alpha-1)^2} = \\
&= 2 * \frac{\alpha\beta^2}{(\alpha-2)(\alpha-1)^2}
\end{aligned}
$$

7.8 Appendix E: TT in the economics literature

The results established above relate to the Cauchy and the symmetric Pareto distribution with infinite first moments as important elements of the class of TT distributions. In public economics the Pareto distribution was introduced in Pareto (1896) to describe the distribution of incomes. Pareto distribution has proved important for modelling top-income distribution or wealth distribution and the coefficient α has often been estimated around 1.5 (for various reference see Cowell, 1995). In economic geography TT distributions are of interest for the distribution of cities by size. The interest springs from "Zipf's law", which says that for most countries the size distribution of cities fits a power law, i.e. the number of cities with population greater that a given threshold C is proportional to $1/C$ (Zipf, 1949). Zipf's law is a discrete form of the Pareto distribution. Given the size in population of a given city, X, its density function is:

$$Pr(X > C) = 1 - aX^{-\alpha} \tag{7.18}$$

In this model α can be estimated by regression:

$$log(y) = \mu + \alpha log(x) \tag{7.19}$$

where x and y are size and rank, respectively. In many empirical studies α is found to be not significantly different from 1 (for a discussions and reference, see among others Gabaix, 1999). In industrial economics TT distributions are of interest for the distribution of firms by size (Hart and Prais, 1956; Steindl, 1965). Gibrat (1931)

proposes a stochastic growth model to explain the rank-order relationship, primarily referring to firms[9]. Let x_t be the size[10] of a firm x at time t, the evolution of x over time is expressed by:

$$x_t = \mu + \lambda_t x_{t-1} \tag{7.20}$$

where λ_t is a random growth factor, and μ is a constant. In general, λ_t is does not depend on firm size, x_t, but can depend on other economic variables, including the size of the other firms in the industry. Hence, the rank-order relationship within an industry is conditional on their relative size. The implied distribution of random variables generated by an equation as (7.20) is a Pareto distribution.

The Gibrat and Zipf relationships are referred as "laws" because of the striking constancy of the estimate of the α parameter, found to be not significantly different to 1. This means that none of the standard moments of the underling Pareto distributions exist.

In finance TT distributions are of relevance since it is well established that returns of financial data, as well as size of corporate bankruptcies, are non-normal. These issues arise also policy matters regarding the regulation of markets where such extremes occurs (Embrechts, 2001; Danielsson and de Vries, 1997; Loretan and Phillips, 1994).

In the economics of information technology there is a growing literature on heavy tailed distributions. The design of robust and reliable networks has become an in-

[9]For a survey and an embedding of this law into economic optimization framework with a model of market structure see Sutton (1997).

[10]Size may be any measure of firm size, e.g., market capitalization.

creasingly important issue in today's Internet world. It has been established that Web traffic often presents heavy-tailed distributions, i.e. that the distribution of file sizes in some systems declines with power law (Arlitt and Williamson, 1996) and the parameter α has been estimated to be approximately equal to 1 (Crovella and Bestavros, 1996). This literature suggests that there is a direct link between the self-similar nature[11] of measured aggregate network traffic and the underlying heavy-tailed distributions of the Web traffic at the source level (Zhu et al., 2001). For example the lengths of bursts in network traffic and the sizes of files in some systems are well described by distributions with non-negligible probability of extremely large events. Additional evidence of power-law behavior is present in same data on transmission lengths of network transfers. Bodnarchuk and Bunt (1991) show that the sizes of reads and writes to an NSF server[12] seem to show power-law behaviour. Paxson and Floyd (1995) found that the upper tail of the distribution of data bytes in FTP bursts was well fit to a Pareto distribution with $0.9 \leq \alpha \leq 1.1$. These are relevant element for designers of computing and telecommunication systems who need to use heavy tail distributions for simulating workloads (Crovella and Lipsky, 1997).

[11]Self-similar process are stochastic processes that are invariant in distribution under suitable scaling of time and space (Embrechts and Maejima, 2000). Self-similarity refers to the condition in which the autocorrelation of a time-series declines like a power-law, leading to positive correlations among observations widely separated in time.

[12]Originally developed by Sun Microsystems, the Network File System (NFS) is a TCP/IP application that has since been implemented on most DOS and Unix systems. NFS allows one to graft remote filesystems - or portions of them - onto your local namespace. Directories on the remote systems appear as part of the local filesystem and all the utilities used for listing and managing files (e.g. ls, cp, mv) operate on the remote files exactly as they do on local files.

Chapter 8

Conclusions

This thesis focuses on income inequality in Italy but it raises a number of points that are of more general interest. It proceeds along three main themes: (a) The value of microsimulation and its relevance in terms of public policy assessment, analysis of socio-economic trends and data analysis. (b) Practical methods for analyzing the causes of inequality. (c) Special methods required for analysis of income distribution data.

8.1 The value of microsimulation

The term microsimulation is used here in a broad sense. In the Public Finance literature microsimulation is often considered to be a contraction of tax-benefit microsimulation. However I call microsimulation any technique based on a process of imitation of complex systems given a set of information provided by micro data sets (see Chapter 1). Microsimulation can then be used for simulating the effect of tax-

benefit policies as well as the effect of socio-demographic changes through the development of counterfactuals. Conterfactuals are constructed to answer "what happens if" questions and can be used to forecast ex-ante the effect of a set of interventions on a population or to evaluate ex-post what would have happened had a different policy been introduced.

The first part of the thesis is mainly focused on tax-benefit microsimulation. Tax-benefit microsimulation models (MSMs) are powerful tools for fiscal policy analysis: they are developed using a representative micro-data set and, when compared to representative household analysis, they provide a more reliable picture of the effects of changes in tax-benefit policies on individual and social welfare. The new TABEITA98 model (Chapter 2) plays a special role as an MSM in the context of Italian personal income taxation. As any Italian MSM it is necessary to model before-tax income and assess the importance of taxation in altering income distribution since available data for Italy only come as disposable income, with no information about taxes paid and (some of the) benefits received. TABEITA98 improves on existing Italian MSMs for addressing the grossing-up of the sample to population totals, both for redistributive analysis and the forecasting of effects of fiscal reforms on public accounts. It is shown that if the grossing-up procedure is overlooked results can be seriously biased since some population sub-groups are not correctly represented in the sample. TABEITA98 is also used to provide an updated estimate of tax evasion in Italy. The general point of tax-benefit microsimulation models is that they are a powerful but demanding task and their aims should not be just the production of numbers: proper attention should be paid to issues such as validation of the results, grossing-up to population totals

and, in particular, reliability of their estimates.

As with other comparable MSMs, TABEITA98 allows one to build counterfactuals through the modification of tax-benefit parameters. In Chapter 3.3 TABEITA98 is used to analyze the effect of the 1998 Italian personal income tax reform on household income distribution. However, the main contribution of Chapter 3.3 is to suggest analyzing MSMs' output using nonparametric density estimation to complement analysis through standard summary statistic. Nonparametric density estimation allows one to detect peculiarities of the income density and to have a visual representation of the effect of personal income tax on before-tax incomes. Although other contributions had already noted that the 1998 Italian personal income tax reform had reduced inequality, the use of nonparametric density estimation was successful to show what others had overlooked: reduction of inequality was not evenly spread among population subgroups and some elements of the reform where more effective than others to improve income distribution. Moreover, it was shown that personal income taxation has a major role in explaining the bimodality of disposable income distribution. The analysis was performed developing counterfactuals such as "what would have occurred if a given feature of the 1998 reform was as it was in 1991".

8.2 Practical methods for analyzing the causes of inequality

Chapter 6 develops a microsimulation methodology to build counterfactuals to analyze inequality trends. It does not use a tax-benefit MSM as it is performed only on after-tax incomes and it is applied to Italy. The main reason why microsimulation is used is because traditional methods for inequality decomposition are unable to explain but a small part of the change of inequality across time (Brandolini and D'Alessio, 2001). Regression-based methods for inequality decomposition were not considered since they are not convincing for a number of reasons, and in particular because a one-equation regression model is unable to explain the most of household income formation (see Chapter 5). The combination of two microsimulation methodologies, namely the DiNardo et al. (1996) and the Burtless (1999), made it possible to show that among the socio-economic factors that played the major role for the evolution of household inequality there are the increased female labor force participation and the increased dispersion of employed and, in particular, self-employment income. Contrary to received wisdom, changes in the probability of receiving a pension and changes in pension income dispersion had only a minor effect.

8.3 Special methods required for analysis of income distribution data

A recurrent point in the thesis is that data quality is an issue for inequality analysis. As revised in Chapter 6, Section 6.1, various authors noted a U-shaped trend in employment inequality since the late 1970s but failed to find any similar trend on household income inequality. Brandolini and Sestito (1993); D'Alessio and Signorini (2000); Brandolini and D'Alessio (2001) believe instead that household income distribution is mainly characterized by fluctuations rather than by a clear trend. Section 6.4 shows that these fluctuations are mainly due to an over-sampling of household with small number of income receivers. Counterfactual analysis can contribute to the removal of such a bias: the resulting picture is that of an household income inequality trend similar to the individual income inequality trend, decreasing until the end of the 1980 and then increasing. Moreover, the importance of the dispersion of self-employment income for the evolution of household income inequality shows that no matter what concern we may have about the reliability of self-employment income data, if we are interested in household income we cannot neglect the role of self-employment dynamics and should instead think of possible improvements in survey data collection.

Data quality seriously affects the reliability of estimates. In Chapter 4 the issue is analyzed in the context of MSMs. In particular, using the bootstrap to compute confidence intervals, we can see that it is not the tax-benefit microsimulation per se that increases variability. On the contrary, since tax-benefit simulation is highly

non-linear it can sometimes even have beneficial effects on the variability of estimates. The main factors that seem to affect reliability of estimates are data contamination or extreme values, which are a typical feature of income distributions. In Chapter 7 the issue of thick tail distributions is analyzed. It is shown that simple inference on samples drawn from Pareto distributions with tail indices typically found in income distribution analysis can be seriously misleading. A possible solution lies in using a modified t-ratio statistics that is well-behaving as the sample size increases.

This thesis highlights the common-sense point that empirical analysis of public policies is a complex task. Institutions interested in forecasting and analyzing effects of public policies on income distribution should devote great effort to the production of good quality data. Researchers should be careful to use them: even when data are reliable and representative of a population, available methodologies can have unsatisfactory performance, also because income distributions typically present a positive probability of very high incomes compared to the average. However, alternative procedures can be considered. I here suggest the construction of counterfactuals using microsimulation analysis. It can indeed be very informative, provided major complication with data are properly handled.

Bibliography

Abramson, I. S. (1982). On bandwidth variation in kernel estimates - a square root law. *Annals of Statistics*, 10:1217–1223.

Amemiya, T. (1985). *Advanced econometrics.* Harvard University Press, Cambridge, Massachusetts.

Andreassen, L., Fredricksen, D., and Ljones, O. (1996). The future burden of public pensions benefits: a microsimulation study. In Harding (1996), chapter 15, pages 329–360.

Arlitt, M. F. and Williamson, C. L. (1996). Web server workload characterization: The search for invariants. In *Measurement and Modeling of Computer Systems*, pages 126–137.

Atella, V., Coromaldi, M., and Mastrofrancesco, L. (2001). Euromod country report. italy. *mimeo.*

Atkinson, A. B., Gomulka, J., and Sutherland, H. (1988). Grossing-up FES data for Tax-Benefit Models. In Atkinson and Sutherland (1988a), pages 223–253.

Atkinson, A. B. and Micklewright, J. (1983). On the Reliability of Income Data in the Family Expenditure Survey 1970-1977. *Journal of the Royal Statistical Association*, 146(1):33–61.

Atkinson, A. B., Rainwater, L., and Smeeding, T. M. (1995). In *Income Distribution in OECD Countries. Evidence from the Luxembourg Income Study*, Paris. OCDE.

Atkinson, A. B. and Sutherland, H. (1983). Hypothetical families in the DHSS Tax/Benefit model and families in the Family Expenditure Survey 1980. SSRC Program on Taxation, Incentives and the Distribution of Income, Research Note No. 1. STICERD, London School of Economics.

Atkinson, A. B. and Sutherland, H., editors (1988a). *Tax-Benefit Models*, number 10 in STICERD Occasional Paper, London. London School of Economics, STICERD.

Atkinson, A. B. and Sutherland, H. (1988b). TAXMOD. In Atkinson and Sutherland (1988a), pages 32–61.

Baldini, M. (1996). Una scomposizione della disuguaglianza tra le famiglie italiane per fonti di reddito, 1987-1993. *Politica Economica*, 2(XII):175–203.

Baldini, M. (1998). Mapp98: un Modello di Analisi delle Politiche Pubbliche. *CAPP, Materiali di discussione, Modena*, (331).

Banca d'Italia (1999). *Relazione del Governatore sull'esericizio 1998.* Assemlea Generale Ordinaria dei Partecipanti. Banca d'Italia, Roma.

Banca d'Italia (2000). *I bilanci delle famiglie italiane nell'anno 1998*, volume X of

Supplementi al Bollettino Statistico. Note metodologiche e informazioni statistiche. Banca d'Italia, Roma.

Beirlant, J., Vynckier, P., and Teugels, J. L. (1996). Tail Index Estimation, Pareto Quantile Plots, and Regression Diagnostics. *Journal of the American Statistical Association*, 91(436):1659–1667.

Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83:687–697.

Biewen, M. and Jenkins, S. P. (2003). Estimation of generalized entropy and atkinson inequality indices from complex survey data. *DIW Berlin - Discussion Papers*, (343).

Birindelli, L., Inglese, L., Proto, G., and Ricci, L. (1998). Gli effetti redistributivi della politica economica e sociale. In Rossi, N., editor, *Il lavoro e la sovranità sociale, 1996-1997. Quarto rapporto CNEL sulla distribuzione e redistribuzione del reddito in Italia*, pages 147–200. Il Mulino, Bologna.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4):436–455.

Blow, L., Hawkins, M., Clemm, A., McCrae, J., and Simpson, H. (2002). Budget 2002: business taxation measures. *IFS Briefing Notes*, (24).

Blundell, R., Duncan, A., and Meghir, C. (1992). Taxation in Emprical Labor Supply Models: Lone Mothers in the U.K. *Economic Journal*, 102:265–278.

Blundell, R., Duncan, A., and Meghir, C. (1998). Estimating labor supply responses using tax reforms. *Econometrica*, 66(4):827–861.

Blundell, R. and Lewbel, A. (1991). The information content of equivalence scales. *Journal of Econometrics*, 50:49–68.

Bodnarchuk, R. R. and Bunt, R. B. (1991). A synthetic workload model for a distributed sustem file server. In *Proceedings of the 1991 SIGMETRICS Conference on Measurement and Modeling of computer Systems*, pages 50–59.

Bollinger, C. and David, M. (1997). Modelling discrete choice with response error: Food stamp participation. *Journal of the American Statistical Association*, 92:827–835.

Bosi, P., Mantovani, D., and Matteuzzi, M. (1999). Analisi degli effetti redistributivi della riforma IRAP-IRPEF - nota di lavoro. Technical Report 2, Prometeia, Bologna.

Bourguignon, F. (1979). Decomposable income inequality measures. *Econometrica*, 47:901–20.

Bourguignon, F., Chiappori, P.-A., and Sastre-Descals, J. (1988). SYSIFF: A Simulation Program of the French Tax-Benefit System. In Atkinson and Sutherland (1988a), pages 32–61.

Bourguignon, F., Ferreira, F., and Leite, P. (2002). Beyond Oaxaca-Blinder: Accounting for Differences in Household Income Distributions Across Countries. mimeo.

Bourguignon, F., Fournier, M., and Gurgand, M. (2001). Fast development with a stable income distribution: Taiwan, 1979-94. *Review of Income and Wealth*, 47(2):139–163.

Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations.* Oxford University Press, Oxford.

Brandolini, A. (1999). The distribution of personal income in post-war italy: Source description, data quality, and the time pattern of income inequality. *Giornale degli Economisti e Annali di Economia*, 58:183–239.

Brandolini, A., Cipollone, P., and Sesisto, P. (2001). Earnings dispersion, low pay and household poverty in italy, 1977-1998. Rome. Temi di discussione del Servizio Studi, N. 427.

Brandolini, A. and D'Alessio, G. (2001). Household Structure and Income Inequality. Luxemburg Income Study Working Paper No. 254.

Brandolini, A. and Sestito, P. (1993). La distribuzione dei redditi familiari in Italia:1977-1991. In Rossi, N., editor, *La crescita ineguale, Primo rapporto CNEL sulla Distribuzione e Redistribuzione del Reddito in Italia*, pages 335–382. Il Mulino, Bologna.

Bryson, M. (1982). Heavy-tailed distributions. In Kotz, N. and Read, S., editors, *Encyclopedia of Statistical Sciences*, volume 3, New York.

Burkhauser, R. V., Crews, A. D., Daly, M. C., and Jenkins, S. P. (1999). Testing the significance of income distribution changes over the 1980s business cycle: a crossnational comparison. *Journal of Applied Econometrics*, 14(3):253–272.

Burtless, G. (1999). Effects of growing wage disparities and changing family composition on the u.s. income distribution. *European Economic Review*, 43:853–865.

Calzaroni, M. (2000). L'occupazione come strumento per la stima esaustiva del PIL e la misura del sommerso . Seminario "La nuova contabilità nazionale, 12-13 January, ISTAT, Rome.

Cannari, L. and D'Alessio, G. (1992). Mancate interviste e distorsione degli stimatori. *Temi di discussione del Servizio Studi*, (172).

CER (1998a). La manovra IRAP-IRPEF. effetti microeconomici sui percettori di reddito e sui bilanci familiari. Technical report, Centro Europa Ricerche, Rome.

CER (1998b). La riforma fiscale: una simulazione con il modello macroeconomico del CER. Technical report, Centro Europa Ricerche, Rome.

Chantreuil, F. and Trannoy, A. (1999). Inequality decomposition values: the trade-off between marginality and consistency. *mimeo.*

CNEL (2004). *Popolazione e forze di lavoro. Rilevazione Istat Forze di Lavoro ed elaborazioni CNEL su dati Istat.* CNEL, Consiglio Nazionale dell'Economia e del Lavoro, Rome. Data available at http://www.cnel.it/archivio/banche_dati.asp.

Coulter, F., Heady, C., Lawson, C., Smith, S., and Stark, G. (1998). A Microsimulation Model of Personal Tax and Social Security Benefits in the Czech Republic. In Spahn, P. and Pearson, M., editors, *Tax Modelling for Economies in Transition*, pages 163–189. Macmillan Press Ltd.

Cowell, F. A. (1980). On the structure of additive inequality measures. *Review of Economic Studies*, 47:521–31.

Cowell, F. A. (1989). Sampling variance and decomposable inequality measures. *Journal of Econometrics*, 42(1):27–41.

Cowell, F. A. (1995). *Measuring Inequality.* Harvester Wheatsheaf, Hemel Hempstead, second edition.

Cowell, F. A. and Flachaire, E. (2002). Sensitivity of inequality measures to extreme values. Distributional Analysis Discussion Paper, 60, STICERD, LSE.

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

Cowell, F. A. and Jenkins, S. P. (2003). Estimating welfare indices: household weights and sample design. In Amiel, Y. and Bishop, J. A., editors, *Inequality, welfare and poverty: theory and measurement*, pages 147–172, The Netherlands. Elsevier Science Ltd.

Cowell, F. A., Jenkins, S. P., and Litchfield, J. (1996). The changing shape of the uk income distribution: Kernel density estimates. In Hill, J., editor, *New inequalities.* Cambridge University Press, Cambridge.

Cowell, F. A. and Mercader-Prats, M. (1997). Equivalence of scales and inequality - distributional analysis discussion paper. Technical Report 27, STICERD, London School of Economics, London.

Cowell, F. A. and Victoria-Feser, M. P. (1996). Robustness properties of inequality measures. *Econometrica*, 64:77–101.

Cowell, F. A. and Victoria-Feser, M. P. (2001). Robust Lorenz Curves: A semi-parametric approach. Number 50 in Distributional Analysis Discussion Papers, London School of Economics, WC2A 2AE, London. STICERD.

Cowell, F. A. and Victoria-Feser, M. P. (2002). Welfare rankings in the presence of contaminated data. *Econometrica*, 70(3):1221–1233.

Crovella, M. and Bestavros, A. (1996). Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. In *Proceedings of SIGMETRICS'96: The ACM International Conference on Measurement and Modeling of Computer Systems.*, Philadelphia, Pennsylvania. Also, in Performance evaluation review, May 1996, 24(1):160-169.

Crovella, M. and Lipsky, L. (1997). Long-lasting transient conditions in simulations with heavy-tailed workloads. In *Winter Simulation Conference*, pages 1005–1012.

D'Alessio, G. and Signorini, L. F. (2000). Disuguaglianza dei redditi individuali e ruolo della famiglia in Italia. Roma. Temi di discussione del Servizio Studi, N. 390, Banca d'Italia, December.

Daly, M. C. and Valletta, R. G. (2002). Inequality and poverty in the united states: The effects of rising wage dispersion and changing family behavior. *mimeo.*

D'Ambrosio, C. (2001). Household characteristics and the distribution of income in italy: an application of social distance measures. *Review of Income and Wealth*, 47(1):43–64.

Danielsson, J. and de Vries, C. (1997). Extreme Returns, Tail Estimation, and Value-at-Risk. *FMG Discussion Paper*, 273.

Davidson, R. and MacKinnon, J. (1998). Graphical Methods for Investigating the Size and Power of Hypothesis Tests. *The Mancester School*, 66(1):1–26.

Davidson, R. and MacKinnon, J. (1999a). Bootstrap testing in nonlinear models. *International Economic Review*, 40:487–508.

Davidson, R. and MacKinnon, J. (1999b). The size distorsion of bootstrap tests. *Econometric Theory*, 15:361–376.

Davidson, R. and MacKinnon, J. (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews*, 19:55–68.

Davidson, R. and MacKinnon, J. (2001). Improving the Reliability of Bootstrap Tests. *mimeo.*

De Santis, G. (1998). Le misure della povertà in italia: scale di equivalenza e aspetti demografici. In *Commissione di indagine sulla povertà e sull'emarginazione.* Presidenza del Consiglio dei Ministri, Dipartimento per gli affari sociali, Rome.

Devicienti, F. (2003). The rise of wage inequality in Italy. Observable factors or unobservables? LABORatorio Riccardo Revelli. Torino.

Dilnot, A., Stark, G., and Webb, S. (1988). The IFS Tax and Benefit model. In Atkinson and Sutherland (1988a), chapter 4, pages 62–96.

DiNardo, J. (2002). Propensity score reweighting and changes in wage distributions. mimeo.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor Market Insititutions and the Distribution of Wages, 1973-1992: a Semiparametric Approach. *Econometrica*, 64:1001–1044.

Duclos, J. (1995). Modelling the take-up of state support. *Journal of Public Economics*, 58(391-415).

Dufour, J. and Kiviet, J. (1998). Exact inference methods for first-order autoregressive distributed lag models. *Econometrica*, 66:79–104.

Duncan, A. (2003). *A Short Course in Microsimulation Methods.* Cemmap Course Notes CCN01/03. The Institute of Fiscal Studies, Department of Economics, UCL, London.

Duncan, A. and Giles, C. (1996). Labour Supply Incentives and Recent Family Credit Reforms. *Economic Journal*, 106:142–155.

Eason, R. (1996). Microsimulation of direct taxes and fiscal policy in the United Kingdom. In Harding (1996), pages 23–46.

Eason, R. (2000). Modelling Corporation Tax in the United Kingdom. In Gupta and Kapur (2000), chapter 8, pages 133–151.

Efron, B. and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals and Other Measures of Statistical Accuracy. *Statistical Science*, 1:54–77.

Eklind, B., Eriksson, I., Hussénius, J., and Müller, M. (1996). Pension analysis in a static model with lifetime income distribution: initial results. In Harding (1996), chapter 14, pages 313–328.

Eliasson, G. (1978). *A micro-to-macro model of the Swedish economy.* Papers on the Swedish model from the symposium on microsimulation methods, September 1977. Almquist & Wicksell, Stockholm.

Embrechts, P. (2001). Extremes in economics and the economics of extremes. *paper presented at SemStat meeting on Extreme Value Theory and Applications.*

Embrechts, P., Klupperlberg, C., and Mikosch, T. (1999). *Modelling Extremal Events for Insurance and Finance.* Springer Verlag, Berlin.

Embrechts, P. and Maejima, M. (2000). An introduction to the theory of selfsimilar stochastic processes. *mimeo.*

Erickson, C. L. and Ichino, A. (1995). Wage differentials in italy:market forces, institutions, and inflation. In Freeman, R., L.F. Kats (Eds), D., and in Wage Structures, C., editors, *Differences and Changes in Wage Structures*, pages 265–305. University of Chicago Press, Chicago, IL.

Falkingham, J. and Harding, A. (1996). Poverty alleviation vs social insurance system: a comparison of lifetime redistribution. In Harding (1996), chapter 11, pages 233–266.

Favreault, M. M. and Caldwell, S. B. (2000). Assessing Distributional Impacts of Social Security Using Microsimulation. In Gupta and Kapur (2000), chapter 20, pages 395–426.

Fei, J. C. H., Ranis, G., and Kuo, S. W. Y. (1978). Growth and family distribution of income by factor components. *Quarterly Journal of Economics*, 92:17–53.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications, Vol. II, Second Edition.* John Wiley and Sons, Inc., Canada.

Ferguson, T. S. (1996). *A course in large sample theory.* Text in Statistical Science. Chapman & Hall, London.

Fieiller, E. (1932). The distribution of the index in a noraml bivariate population. *Biometrika*, (24):428–440.

Fields, G. S. (2002). Accounting for income inequality and its change: a new method with application to distribution of earnings in the united states. Forthcoming in Research in Labour Economics.

Fields, G. S. and Mitchell, J. C. O. (1999). Changing income inequality in taiwan: A decomposition analysis.

Fiorio, C. (2004). Confidence intervals for kernel density estimation. *The Stata Journal*, 4(2):103–114.

Focardi, S. M. (2001). Fat tails, scaling and stable laws: A critical look at modeling extremal events in economic and financial phenomena. *The Intertek Group Discussion Paper*, (02).

Fournier, M. (2001). Inequality decomposition by factor component: a rank correlation approach illustrated on the taiwanese case. *Recherches Économiques de Louvain / Louvain Economic Review*, 67(4):381–403.

Frey, B. S. and Weck-Hanneman, H. (1984). The hidden economy as an 'unobserved' variable. *European Economic Review*, 26:33–53.

Fry, V. and Stark, G. (1993). The Take-up of Means-Tested Benefits, 1984-90. Institute for Fiscal Studies, London.

Gabaix, X. (1999). Zipf's law for cities: an explanation. *The Quarterly Journal of Economics.*

Galler, H. P. (1996). Microsimulation of pension reform proposals: modelling the earnings of couples. In Harding (1996), chapter 13, pages 293–309.

Giannini, S. and Guerra, M. C. (1999). Il sistema tributario verso un modello di tassazione duale. In Bernardi, L., editor, *La finanza pubblica italiana: Rapporto 1999.* Il Mulino, Bologna.

Gibrat, R. (1931). *Les inegalites economiques; applications: aux inegalites des*

richesses, a la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel. Librairie du Recueil Sirey, Paris, France.

Godfrey, L. (1998). Tests of non-nested regression models: Some results on small sample behaviour and the bootstrap. *Journal of Econometrics*, 84:59–74.

Gupta, A. and Kapur, V., editors (2000). *Microsimulation in Government Policy and Forecasting.* Contributions to economic analysis. North-Holland, Amsterdam.

Hagenaars, A. (1990). Female labour supply in microsimulation models. In Brunner, J. and Petersen, H.-G., editors, *Prospects and Limits of Simulation Models in Tax and Transfer Policy*, Frankfurt/New York.

Hall, P. (1992a). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

Hall, P. (1992b). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Annals of Statistics*, 20:675–694.

Harding, A. (1990). Dynamic Microsimulation Models: Problems and Prospects. Welfare State Programme. STICERD - London School of Economics.

Harding, A., editor (1996). *Microsimulation and Public Policy.* Selected papers from the IARIW Special Conference on Microsimulation and Public Policy, Camberra, 5-9 December 1993. Elsevier.

Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. (2004). *Nonparametric and Semiparametric Models.* Springer Series in Statistics. Springer, Berlin.

Hart, P. E. and Prais, S. J. (1956). An analysis of business concentration. *Journal of the Royal Statistical Society, A.*

Heckman, J. J. and MaCurdy, T. E. (1980). A lifecycle model of female labour supply. *Review of Economic Studies*, 47(1):47–44.

Hill, D. M. (1975). A Simple Approach to Inference About the Tail of a Distribution. *The Annals of Statistics*, 3:1163–1174.

Hollenbeck, K. (1976). An algorithm for adjusting n-dimensional tabular data to conform to general linear constraints. In *Proceedings of the American Statistical Association*, pages 402–405.

Hope, S. (1988). Model Validation. In Atkinson and Sutherland (1988a), pages 272–292.

Horowitz, J. (1994). Bootstrap-based critical values for the information matrix test. *Journal of Econometrics*, 61:395–411.

Hsieh, P.-H. (1999). Robustness of Tail Index Estimation. *Journal of Computational and Graphical Statistics*, 8(2):318–332.

ISTAT (2004). *Popolazione & statistiche demografiche.* ISTAT, Istituto nazionale di statistica, Rome. Data available at http://demo.istat.it/.

Jenkins, S. P. (1994). Did the middle class shrink during the 1980s? UK evidence from kernel density estimates. *Economics letters*, pages 407–413.

Jenkins, S. P. (1995). Accounting for inequality trends: Decomposition analyses for the uk, 1971-86. *Economica*, 62(245), pages = 29-63, month= Feb.).

Juhász, I. (1998). The Hungarian Personal Income Tax Model. In Spahn, P. and Pearson, M., editors, *Tax Modelling for Economies in Transition*, pages 191–205. Macmillan Press Ltd.

Karlen, D. (2002). Credibility of Confidence Intervals. In *Advanced Statistical Techniques in Particle Physics, Proceedings*, pages 53–57. Grey College, Durham.

Lambert, P. J. (1993). *The distribution and redistribution of income.* Manchester University Press, Manchester.

Leontief, W. (1951). *The Structure of the American Economy.* Oxford University Press, New York.

Loretan, M. and Phillips, P. C. B. (1994). Testing the covariance stationarity of heavy-tailed time series. *Journal of Empirical Finance*, (1):211–248.

MaCurdy, T. E. (1981). An empirical model of labour supply in lifecycle setting. *Journal of Political Economy*, 89(6):1059–1085.

MaCurdy, T. E. (1983). A Simple Scheme for Estimating an Intertemporal Model of Labor Supply and Consumption in the Presence of Taxes and Uncertainty. *International Economic Review*, 24:265–289.

Manacorda, M. (2002). Can the scala mobile explain the fall and rise of earnings

inequality in italy? a semiparametric analysis, 1977-1993. *CEP, London School of Economics, mimeo.*

Mantovani, D. (1998). Manuale DIRIMOD95. *mimeo*, Prometeia, Bologna.

Marenzi, A. (1996). Prime analisi sulla distribuzione dell'evasione dell'irpef per categorie di contribuenti e per livelli di reddito. Rapporto CNEL, pages 305–341. Il Mulino.

Marsaglia, G. (1965). Ratios of normal variables and ratios of sums of uniform variables. *Journal of the American Statistical Association*, 60:193–204.

Meagher, G. (1996). Forecasting Changes in the Distribution of Income: An Applied General Equilibrium Approach. In Harding (1996), pages 362–384.

Merz, J. (1991). Microsimulation - A survey of principles, developments and applications. *International Journal of Forecasting*, 7:77–104.

Ministero delle Finanze (1995). *Analisi del 740. Anno 1991 redditi 1990.* Ministero delle Finanze, Rome.

Ministero delle Finanze (2002). *Analisi delle dichiarazioni dei redditi. Il modello Unico 1999/redditi 1998: le persone fisiche.* Ministero delle Finanze, Rome.

Mitton, L., Sutherland, H., and Weeks, M., editors (2000). *Microsimulation Modelling for Policy Analysis: Challenges and Innovations.* Cambridge University Press, Cambridge.

Mookherjee, D. and Shorrocks, A. (1982). A decomposition analysis of the trend in uk income inequality. *The Economic Journal*, 92:886–902.

Morduch, J. and Sicular, T. (2002). Rethinking inequality decomposition, with evidence from rural china. *The Economic Journal*, 112(476):93–106(14).

Nelissen, J. H. M. (1996). Social security and lifetime income redistribution: a microsimulation approach. In Harding (1996), chapter 12, pages 267–292.

Oaxaca, R. R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review.*

Orcutt, G. (1957). A new type of socio-economic systems. *The Review of Economics and Statistics*, 58:773–797.

Orcutt, G., Caldwell, S., and Wertheimer II, R. (1976). *Policy Exploration Through Microanalitic Simulation.* The Urban Institute, Washington, D.C.

Orcutt, G., Greenberg, J., Korbel, J., and Rivlin, A. (1961). *Microanalysis of Socioeconomic Systems: A Simulation Study.* Harper & Row, New York.

Pareto, V. (1896). *Cours d'Economie Politique.* Droz, Geneva, Switzerland.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076.

Paxson, V. and Floyd, S. (1995). Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244.

Phillips, P. C. B. (1982). Exact small sample theory in the simultaneous equation model. In Intrilligator, M. and Griliches, Z., editors, *Handbook of Econometrics*, Amsterdam. North-Holland.

Phillips, P. C. B. and Hajivassiliou, V. A. (1987). Bimodal t-ratios. *Cowles Foundation Discussion Paper No.842.*

Pittau, M. G. and Zelli, R. (2001). Income distibution in italy: a non parametric analysis. *Statistical methods and applications*, 1-3:175–190.

Proto, G. (2000). Il modello di microsimulazione MASTRICT: struttura e risultati. *mimeo.* ISTAT, Direzione Centrale Imprese e Istituzioni.

Pudney, S. (1990). The estimation of Engel curves. In Myles, G., editor, *Measurement and Modelling in Economics*, Contributions to economic analysis, pages 267–323, Amsterdam. North-Holland.

Pudney, S. (1993). Income and wealth inequality and the life cycle. A non-parametric analysis for China. *Journal of applied econometrics*, 8:248–276.

Pudney, S. (2001). The impact of measurement error in probit models of benefit take-up. *mimeo*, page University of Leicester: Working Paper.

Pudney, S., Hernandez, M., and Hancock, R. (2002). The welfare cost of means-testing: pensioner participation in income support. *mimeo.*

Pudney, S. and Sutherland, H. (1994). How reliable are microsimulation results? an

analysis of the role of sampling error in a u.k. tax-benefit model. *Journal of Public Economics*, 53:327–365.

Pudney, S. and Sutherland, H. (1996). Statistical reliability in microsimulation models with econometrically-estimated behavioural responses. In Harding, A., editor, *Microsimulation and Public Policy*, Amsterdam. North-Holland.

Pyatt, G., Chen, C.-N., and Fei, J. (1980). The distribution of income by factor components. *Quarterly Journal of Economics*, (451-73).

Quah, D. (1997). Empirics for growth and distribution: Stratification, polarization, and convergence clubs. *Journal of Economic Growth*, 2(1):27–59.

Rytgaard, M. (1990). Estimation in the Pareto Distribution. *ASTIN Bulletin*, 20(2):201–216.

Samorodnitsky, G. and Taqqu, M. S. (1994). *Stable Non-Gaussian Random Processes. Stochastic models with infinte variance.* Chapman and Hall, New York.

Sarndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

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

Sastre, M. and Trannoy, A. (2000b). Shapley inequality decomposition by factor components: some methodological issues. *mimeo.*

Schneider, F. (2000). The increase of the size of the shadow economy of 18 OECD countries: some preliminary explanations. *IFO Working papers*, (306).

Shorrocks, A. F. (1980). The class of additively decomposable inequality measures. *Econometrica*, 52:1369–85.

Shorrocks, A. F. (1982). Inequality Decomposition by Factor Components. *Econometrica*, 50:193–211.

Shorrocks, A. F. (1983). The impact of income components on the distribution of family incomes. *The Quarterly Journal of Economics*, 98(2):311–326.

Shorrocks, A. F. (1984). Inequality Decomposition by Population Subgroups. *Econometrica*, 52:1369–1385.

Shorrocks, A. F. (1999). Decomposition procedures for distributional analysis: a unified framework based on Shapley value. *mimeo.*

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Silvey, S. D. (1975). *Statistical Inference.* Chapman and Hall Ltd, London.

Smeeding, T. M. (2000). Changing Income Inequality in OECD Countries: Updated Results from the Luxembourg Income Study (LIS). Syracuse University, New York, 13244-1020. Luxembourg Income Study, Working Paper No. 252, March.

Steindl, J. (1965). *Random Processes and The Growth of Firms.* Griffin, London.

Sutherland, H. (1989). Constructing a Tax-Benefit Model: What Advice Can One Give? In *Taxation, incentive and the distribution of income*, number 141. STICERD, London School of Economics.

Sutherland, H. (1994). Static Microsimulation Models in Europe: a Survey. *The Microsimulation Unit.*

Sutherland, H., editor (2001). *EUROMOD: An integrated European Benefit-Tax Model. Final Report.* EUROMOD Working Paper No. EM9/01. Microsimulation Unit, Cambridge.

Sutton, J. (1997). Gibrats legacy. *Journal of Economic Literature.*

Theil, H. (1967). *Economics and Information Theory.* North-Holland, Amsterdam.

Tinbergen, J. (1939). *Statistical Testing of Business-Cycle Theories.* Society of Nations, Geneva.

van Tongeren, F. W. (1995). *Microsimulation Modelling of the Corporate Firm.* Lecture Notes in Economics and Mathematical Systems. Springer.

Victoria-Feser, M. P. and Dupuis, D. J. (2003). A Prediction Error Criterion for Choosing the Lower Quantile in Pareto Index Estimation. *mimeo.*

Walker, A. (2000). Modelling immigrants to Australia - to Enter a Dynamic Microsimulation Model. In Gupta and Kapur (2000), chapter 16, pages 313–340.

Zhu, X., Yu, J., and Doyle, J. (2001). Heavy-tailed distributions, generalized source coding and optimal web layout design.

Zipf, G. (1949). *Human behavior and the principle of the last effort.* Addison-Wesley, Cambridge, MA.

Zizza, R. (2002). Metodologie di stima dell'economia sommersa: un'applicazione al caso italiano. *Temi di discussione del Servizio Studi*, (463). Banca d'Italia.