Economic Modelling 35 (2013) 892–899

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.elsevier.com/locate/ecmod

Parametric and nonparametric income distribution estimators in CGE micro-simulation modeling

Dorothée Boccanfuso, Patrick Richard *, Luc Savard

Université de Sherbrooke — GRÉDI, Canada

A R T I C L E   I N F O

Article history:
Accepted 6 July 2013

JEL classification:
D31
C13
C68
I32

Keywords:
CGE modeling
Income distribution
Econometric
Impact analysis

A B S T R A C T

We consider the issue of income distribution modeling in the context of poverty analysis impact based on computable general equilibrium micro-simulation models. The empirical distribution function (EDF) is by far the most commonly used estimator in practice. It is, however, not the only available consistent estimator and there may be situations in which a different estimator would be able to provide more accurate results. An alternative is to use a smooth estimator of the population income distribution. Two types of such estimators are available: parametric and nonparametric ones. In the first case, one has to chose a particular parametric form for the distribution function and estimates its parameters. The main drawback is the difficulty associated with the selection of the functional form. The nonparametric approach sidesteps this functional form issue by using kernel density estimators that only impose mild restrictions on the distribution function. This is obviously an important advantage, but its cost is that the accuracy of these estimators typically depends to a large extent on the bandwidth used in the kernel function. Another advantage is that it nests the EDF as a special case. We propose to extend the work of Boccanfuso et al. (2008) in two ways. First, we consider a larger set of parametric functions, including the 5 parameter generalized beta distribution and some of its special cases. Second, we use non-parametric kernel estimators and study their accuracy under different bandwidth selection schemes. Lastly, we provide Monte Carlo comparisons of the accuracy of these methods with the widely used EDF.

1. Introduction

Traditionally, computable general equilibrium (CGE) models have been used to simulate the impact of exogenous shocks or policy changes on the socio-economic system. In the late seventies, some modelers (Adelman and Robinson, 1978; Taylor and Lysy, 1979) attempted to use CGE models for distributional analysis, but limited the exercise to decomposing households in the CGE model into income quintiles. Later, in the early nineties, Chia et al. (1994) analyzed the impact of structural adjustment programs and were the first to use the Foster, Greer and Thorbecke (FGT) metric (Foster et al., 1984) to measure poverty changes in the Ivory Coast. Because CGE models are calibrated on the basis of a social accounting matrix (SAM) for a reference period characterized by a set of consistent initial conditions, the SAM does not contain any information on the income distribution of socioeconomic household groups. Therefore, conventional CGEs can only simulate the effect of a scenario on the representative households specified in the model. Using this representative household approach implies a very strong assumption that the distribution of income within groups of households is exogenous to the model. This method can then lead to misleading conclusions, as demonstrated in Savard (2005). As income and expenditure surveys become more readily available, it becomes relatively easy to integrate large sample of households or entire survey samples into a SAM to calibrate the CGE model in order to take into account intra-group distributional changes. In this context, two main approaches have been used to link macroreforms to changes in income distribution and poverty.[1]

The integrated multi-household (IMH) approach first proposed by Decaluwé et al. (1999), consists in including all households from the household survey in a CGE model.[2] However, it can raise some difficulties at the implementation and resolution stage such as data reconciliation (Hertel and Reimer, 2004), each household account needs to be balanced and aggregated to the level found in the social accounting matrix. Another problem is related to the challenges of finding a numerical solution, as raised by Chen and Ravallion (2004). Finally, the modeler is constrained to use functional forms that respect the standard conditions of general equilibrium. The second approach is referred to as the CGE micro-simulation sequential method (MSS) and was formalized in Chen and Ravallion (2004). The general idea of the MSS approach is that a CGE module feeds market and factor price changes into a micro-simulation household model. The approach offers more flexibility with

* Corresponding author.
E-mail address: patrick.richard2@usherbrooke.ca (P. Richard).

[1] For an interesting review and discussion on the value of the CGE macro–micro approach to analyze poverty and inequality impact, the reader can consult Hertel and Reimer (2004) and Bourguignon and Spadaro (2006).

[2] Some authors refer to this approach as a CGE micro-simulation application.

0264-9993/$ – see front matter
http://dx.doi.org/10.1016/j.econmod.2013.07.002

respect to household behavior modeling, as standard constraints imposed by CGE models need not be respected. In this article, we use the MSS approach applied to Mali.

Since the end of the nineteenth century, researchers look to mathematically approximate the true income distribution. Pareto (1897) empirically demonstrated that incomes are lognormally distributed and that the skewness to the right had a flat tail, meaning unequal distribution. Since Pareto, various functional forms have been proposed that can be grouped into three main categories. The first category consists of forms describing an income distribution generated by a stochastic process (Champernowne, 1953a,b; Parker, 1999; Rutherford, 1955). The second category consists of the functional forms that provide a good fit to empirical data but have no theoretical basis (McDonald, 1984; Salem and Mount, 1974). In the last group, functional forms are found as a solution to specified differential equations such that the theoretical foundation is developed on the basis of empirical evidence (Dagum, 2008; Pareto, 1897; Singh and Maddala, 1976). The current literature offers many alternative forms of probability density functions to approximate true income distribution. It is acknowledged, however, that even if lognormal and Pareto distributions are easy to estimate and interpret, other distributions can improve the fit. Champernowne (1953b,a) suggested a three-parameter distribution, which provides a better fit than the two-parameter one. According to McDonald (1984), McDonald and Xu (1995), Bordley et al. (1996), Bandourian et al. (2003), and Reed and Wu (2008) even if beta distributions are flexible and can take a variety of shapes, it is a two-parameter distribution, and its accuracy for fitting data is limited. Over the past few years, these researchers have contributed to generalizing the beta function. This more complex model seems to reflect more appropriately the impact of economic fluctuations. Better fits may be obtained with two members of the Burr family: Singh and Maddala (1976) (Burr 12), and Dagum (2008) (Burr 3). Singh–Maddala is a generalization of Pareto and Weibull distributions, and in terms of goodness-of-fit, this model outperforms both lognormal and gamma distributions (Salem and Mount, 1974). Dagum (2008) proposed a theoretical description based on regular income-elasticity characteristics observed in income distribution. Tadikamalla (1980), Kleiber (1996) and Bandourian et al. (2003) demonstrate that the Dagum usually exhibit a better fit than the Singh–Maddala. However, Boccanfuso et al. (2008) find that the Singh–Maddala give results as good as Dagum but conclude that there is no single best fitting functional form for all groups and that the most flexible ones appear to be more efficient in most cases. McDonald (1984) develops two four-parameter generalizations of the beta distribution, GB1 and GB2. As shown by McDonald and Xu (1995), every two and three parameter distributions previously used to model income distributions are special cases of either the GB1 or the GB2. Empirically, Bandourian et al. (2003) find that the GB2 distribution fits the data better than the GB1 and any three parameter distribution in 44% of 82 cases considered.

In the literature, other four parameter distributions have been used to estimate income distribution. Among these are the type II Dagum (Dagum, 2008) and double Pareto-lognormal (dPlN) proposed by Reed and Jorgensen (2004). Many authors went further to propose five parameter distributions such as McDonald and Xu (1995) who proposed the generalized beta (GB) from which one can derive the GB1 and GB2 under certain assumptions. Bandourian et al. (2003), Dastrup et al. (2007), and McDonald and Ransom (2008) all illustrate that this distribution performs well without being able to systematically arrive at the conclusion that it is the best distribution. Reed (2007) as well as Reed and Wu (2008) proposed the Generalized Normal-Laplace (GNL), which is a generalization of the five parameter double Pareto-lognormal (dPlN) and find that the GNL is more appropriate compared to other distributions including the GB of McDonald and Xu (1995). However, Reed and Wu (2008) question the usefulness of using five parameter functions of model income distribution. More recently, Kaniadakis (2002, 2005) developed the three parameter κ-generalized distribution, a non-Gaussian distribution with asymptotic power-law tails (Clementi et al., 2010). Clementi et al. (2010) find that in a satisfactory number of cases the performance of this distribution is not to be considered inferior from the statistical point of view to that of the Singh–Maddala, Dagum and GB2 distributions. Given that a great deal of literature has been produced on modeling income distribution over the past 60 years, we believe that it has not been fully exploited in the context of CGE micro-simulation analysis.

DeJanvry et al. (1991) applied Pareto distributions to model the income distribution of different subgroups in Ecuador, the best suited to represent the income distribution for higher income groups. Chia et al. (1994) and Montaud (2003) used lognormal distribution for groups in the Ivory Coast and Burkina Faso respectively. Dervis et al. (1991) also chose this two parameter distribution assuming that the change in mean is equal to the change in income of the representative household generated by the CGE model. Adelman and Robinson (1978) also performed a statistical test on lognormal that proved unsatisfactory in a few cases when testing for skewness and kurtosis. To solve this problem, they eliminated a socio-economic group by aggregation. Stifel and Thorbecke (2003), Agénor et al. (2005), Bazlul et al. (2006), Decaluwé et al. (2009) and Oum (2009) among others, applied the beta distribution arguing that the income distribution modeling approach and statistical literature provide evidence that income distribution may be represented by more appropriate and flexible functional forms (Bordley et al., 1996). Boccanfuso et al. (2008) verify whether the distributional impact analysis is affected by the choice of various income distribution functions used in the CGE context such as lognormal, gamma or beta. They illustrated that using more flexible functional forms to model income distribution is more appropriate but that there is no single best fitting functional form for all groups. Another approach used to model income distribution in this context was first proposed by Cockburn (2002). In his CGE–IMH model applied to Nepal, he used an empirical method to compare the income distribution of different household groups. However, Boccanfuso et al. (2008) showed that when sample sizes are relatively small, this empirical approach is found not to be as sensitive as the continuous functional forms.

However, beyond the choice of the distribution through the scope of its flexibility characterized by its number of parameters to estimate, some authors reveal another problem for parametric estimation of distributions. Dastrup et al. (2007) find that the best fitting parametric distribution depends on the definition of income (earnings, total income, disposable income, expenditures, …) and may change following a policy (ex: tax increases). This means that whenever attempting to fit a parametric distribution to a sample of observed income variables, one is forced to estimate several different specifications, which may be computationally cumbersome, as it proves to be with our Malian data set. Clementi et al. (2010) expressed the same observation in their conclusion.

Furthermore, even when maximum likelihood estimation proceeds well, one must face some difficulties at the model selection step. For instance, the GB1 and GB2 distributions are not nested, which renders testing somewhat more complicated than simply using a likelihood ratio statistic. Indeed, even when the distributions are nested the null hypothesis may impose a set of constraints which lie on the border of the parametric space. For instance, a test of the GB2 against the GB imposes that $c = 1$ under H_0 while $0 \leq c \leq 1$ under H_1. It is well known that common statistics such as the likelihood ratio do not follow their usual χ^2 distribution under this type of null hypothesis (see e.g. Gourieroux et al., 1982) and that the bootstrap fails to deliver valid inferences (Andrews, 2000). Other, more complicated, inference methods such as subsampling are available, but their finite sample properties often are somewhat poor.

Considering the importance of modeling the impact of price variations on income distribution to analyze the impact of poverty changes in the context of CGE microsimulation modeling, the difficulties associated with the choice of the good distribution that can change according to the characteristics of the target groups based on the effects of simulations are shown in Boccanfuso et al. (2008) or again given a welfare indicator as in Dastrup et al. (2007), and Clementi et al. (2010), and