China Economic Review 30 (2014) 383–396

Contents lists available at ScienceDirect

China Economic Review

Income distribution in urban China: An overlooked data inconsistency issue

Hailong JIN [a,*], Hang QIAN [b], Tong WANG [c], E. Kwan CHOI [b,d]

[a] CIGI Global Economy, Canada
[b] Iowa State University, USA
[c] Texas A&M AgriLife Research, USA
[d] City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 28 May 2013
Received in revised form 17 February 2014
Accepted 19 February 2014
Available online 26 February 2014

JEL classification:
D31
C32
C46

Keywords:
Income distribution
Sampling inconsistency
Unscented Kalman filter
China

ABSTRACT

The Urban Household Income and Expenditure Surveys, conducted by the National Bureau of Statistics, are extensively explored in income distribution studies. However, we find that a survey coverage expansion that includes migrant residents in the urban sample may induce serious data inconsistency before and after the year 2002. To further unveil the inconsistency, we construct a random walk hierarchical beta-2 distribution model, estimated by the unscented Kalman filter, to investigate the magnitude of the structural break. Results show that the gaps of Gini coefficients and the shape of the distribution can be bridged by a counterfactual time series that coherently measures the urban China income distribution.

1. Introduction

The China economic reform in 1978 ushered in an era of unprecedented economic growth accompanied by dramatic income inequality. As pointed out by Lee and Selden (2009), the recent decades witness China's transition from "one of the most egalitarian societies" to one of the most polarized economies in the world. As a key issue to the socioeconomic analysis, income inequality has received considerable attention from the literature.[1]

Unlike income studies of industrialized countries that boast high quality data sources, lack of representative income data of China poses a huge challenge to empirical studies. Among the available income data sources, the data set compiled by the National Bureau of Statistics (NBS) is "the most comprehensive and nationally representative data" (Cai, Chen, & Zhou, 2010, pp. 386). Each year NBS conducts two types of income surveys, namely the urban and rural household surveys that cover all of the provincial units, to gather the latest national income information. NBS survey data have been widely adopted in studies of income distributions (e.g., Cai et al., 2010; Chotikapanich, Rao, et al., 2007; Han, Liu, & Zhang, 2012; Wu & Perloff, 2005).

* Corresponding author.
E-mail addresses: jinhl03@hotmail.com (H. Jin), matlabist@gmail.com (H. Qian), wangtong03@gmail.com (T. Wang), kchoi@iastate.edu (E.K. Choi).

[1] See, for example, Cai et al. (2010), Chi (2012), Chotikapanich, Rao, and Tang (2007), Han et al. (2012), Huang (2008), Kanbur and Zhang (1999), Khan, Griffin, and Riskin (1999), Khan and Riskin (1998), Knight and Li (2006), Ravallion and Chen (2007).

http://dx.doi.org/10.1016/j.chieco.2014.02.004
1043-951X/

Despite its popularity, an unfrequented side of the NBS survey is its sampling history. Little attention is paid to the drastic definition change in the year 2002 when NBS significantly widened its urban sample frame. Prior to 2002, the urban survey only covered residents with formal urban residency certificates (*hukou*). Those city dwellers without urban *hukou*, hereafter referred to as migrant residents or migrant households, were excluded. However, starting from 2002, NBS embarked a program in collaboration with the World Bank to expand the definition of urban residents: all living in urban areas are covered in the sample. Therefore, the urban surveys before and after 2002 differ in the coverage of migrant households.

To date, the data inconsistency problem receives insufficient attention in the literature. Some papers (e.g., Chotikapanich, Rao, et al., 2007; Han et al., 2012) are unaware of it. Some (e.g., Wu & Perloff, 2005) resort to the subsample prior to 2002. Other studies (e.g., Cai et al., 2010; Chi, 2012) notice this issue, but claim its insignificance due to the small fraction of migrant population in the sample.

Our paper is among the first attempts to fill in this gap. Specifically, based on the aggregated urban income data set,[2] we propose a two-step dynamic income distribution model to address the sampling consistency. First, we evaluate the structural break by constructing a random walk hierarchical beta-2 distribution model to depict the evolvements of urban China's income distributions. Fitted by the publicly available aggregated data by income percentiles, the model well characterizes the income dynamics before and after the structural break. In addition, the model has good out-of-sample forecasting power. Second, we bridge the gap by constructing a counterfactual time series that coherently measures the urban income distribution. Our model facilitates the disentanglement of native urban resident income from the observed mixture income. The observed and counterfactual series can be interpreted as the income distributions with and without migrants respectively, if the survey coverage of migrants is the cause of the identified structural break.

Results suggest that the survey coverage change may induce serious data inconsistency before and after the year 2002. The coverage change uplifted urban China's Gini coefficient by approximately 1/4 since 2002. Considering that migrant households are substantially underrepresented in NBS's urban dataset (see Cai et al., 2010; Chi, 2012), the true Gini coefficients in urban China may be much higher than the officially reported figures. Likewise, the latent Gini coefficients of the whole China may also be underestimated.[3]

The remainder of the paper is organized as follows. Section 2 briefly discusses the background of the NBS urban database and illustrates the data inconsistency. Section 3 develops a dynamic model and outlines estimation strategies. Section 4 presents the fitted income distributions as well as the conjectured distribution if the survey coverage were unchanged. Section 5 checks the robustness of the model by forecast validation, alternative data series and model specifications. Section 6 concludes the paper.

2. Background

2.1. Data source

Each year, NBS deploys tremendous resources conducting surveys on urban and rural households. The former is referred to as Urban Household Income and Expenditure Survey (UHIES). In UHIES, sample urban areas are selected as follows. First, according to population sizes, NBS divides urban areas of all provincial units (provinces, autonomous regions and municipalities directly under the central government) into three strata: large and medium-sized cities, county cities, and county towns. NBS decides the sample size of each stratus according to its provincial population share. Then it ranks cities and towns by average earnings of the employed. Finally, sample cities and towns are selected by a systematic sampling scheme.

Individual incomes collected in UHIES are not published. The publicly available data are aggregated statistics, retrievable in China Statistical Yearbook (CSY), an annual publication of NBS. In the chapter “Basic Conditions of Urban Households by Income Percentile”, households are classified into seven groups by the per capita disposable incomes: lowest income (10%), low income (10%), lower middle income (20%), middle income (20%), upper middle income (20%), high income (10%) and highest income (10%), where numbers in parentheses indicate household number shares. In addition, the lowest 5% of households are referred to as the “poor” households.[4] To extract more information from the available data, this group is singled out from the lowest 10% income group. Hence the data used in this paper contain eight groups: Group 1 denotes the poorest group while Group 8 denotes the richest group. CSY also provides statistics on the number of households surveyed and the average household sizes, which enables conversion from the household number shares to the population shares of each group.

2.2. Data inconsistency issue

In 2002, to better collaborate with the World Bank, NBS widened its urban sample frame to include migrant residents who lived in urban areas without *hukou*. Before that, UHIES data set only covered residents with the formal urban *hukou*, who are institutionally protected in the urban labor market against migrant workers and enjoy a wage premium.[5] Therefore, UHIES is likely to contain a higher proportion of low income people after 2002.

[2] Clearly, results would be more convincing if the individual level incomes, or the original data, were available. However, due to the living conditions of migrant households and inequality are sensitive topics in China, obtaining permission from the authority to use the original data to evaluate this data inconsistency seems less feasible.

[3] As the present analysis is based on the aggregated data, all results here can only be viewed as suggestive rather than conclusive.

[4] In the column of the lowest income group, CSY provides two types of information: one is for the lowest 10% households (named as “lowest income” group), the other is for the lowest 5% households (named as “poor” group).

[5] See “China Floating Population Development Report 2012”issued by the Chinese National Population and Family Planning Commission.