Estimation and Quality of Household Income Data From Cambodia Socio-Economic Survey (CSES)

Nor Vanndy, MBA
Deputy Director Department of National Accounts
National Institute of Statistics (NIS), Phnom Penh, Cambodia
Email: norvanndy@gmail.com

Abstract

The National Institute of Statistics (NIS) of the Ministry of Planning (MOP) in Cambodia has not published household income statistics from the Cambodia Socio-Economic Survey (CSES) from 1999 to 2008 due to the insufficient quality of survey data collection on income. It was published in 2009 and onwards. There are a lot of challenges in estimation of household income data, i.e. households with observed negative income, extreme values, missing values, and so forth. There are many approaches for treatment of survey data required before the publication, first is cleaning CSES data in different sectors/variables, analyzing potential underestimation of incomes, outlier treatment, and a comparative analysis of CSES data, consumption vs. disposable income and national accounts. Second is to measure disposable income per capita and experimental Gini-coefficient of income. In this paper, the estimation of household income data in 2007 to 2011 from the CSES will be used to explain how the Cambodian household income data is compiled.

Key Words: Cleaning Survey Data, Measuring household income per capita, Experimental Gini- coefficient.

1. Introduction

The National Institute of Statistics (NIS) of the Ministry of Planning in Cambodia has conducted the Cambodia Socio-Economic Survey (CSES) since 1993 with different sampling designs. Previous CSES were undertaken in 1993/94, 1996, 1997, 1999, 2004, and then annually between 2007 and 2012. No household income statistics from the CSES was published between 1999 and 2008 due to the insufficient quality of survey data collection on income such as households with observed negative income, extreme values, missing values, changes of sample size, and error in data collection and processing and so forth. Household income data are mainly focused on household income composition and comparative household incomes data from the CSES survey in time series 2007 to 2012. The data quality of household income is a major issue to be improved by using various methods such as checking and cleaning survey data, outliers, a comparative trend analysis of household income survey data with national accounts. The measurement of household disposable income per capita and experimental Gini-coefficient of income are also discussed.

2. Collection of household income data and its use in Cambodia

Both diary and recall methods are used in CSES by using field enumerators (interviewers) and supervisors who are from NIS, MOP and the provincial planning and statistics offices. The sampling design of CSES is a three-stage design with villages, enumeration areas and households as sampling units in each stage. Every fifth year the sample size is bigger allowing for estimates of smaller domains, but during the 'small sample' years the survey is aimed for estimates on national level and large domains such as the capital Phnom Penh and other urban and other rural areas. Field interviewers are required to collect data information using questionnaires that consists of 1)-Household Listing Questionnaire, Form 2)-Village Questionnaire, Form 3)-Household Questionnaire, and Form 4)-Diaries Questionnaire. The questionnaire design has been changed in some items in CSES 2009-2012 from other previous CSES 2004, 2007, and 2008.

Recall vs Diary

CSES data has been collected by interviewers (NIS officials) using both recall and diary methods. Data from these methods are used for measuring household income, an investigation led to the conclusion that the recall data were used for income data and diary data for negative transfers as taxes, transfers to other households and for charity because these expenditures is not captured in recall data. There were different income data in the recall and diary methods. Comparison shows that there are some contradictions in a household's reporting, e.g. a household can report high wage/salary in the recall but low or no value in the diary and vice versa. These concerns also exist when checking recall vs. diary for self-employment income. The current transfers paid by households are not asked about in recall data of CSES2010 and earlier, these transfers were captured in diary, however they are likely to be deleted from the diary in 2012. Instead, it is requested to insert new recall questions to capture these transfers in the section recall non-food expenditure in CSES2011, of which: 1). regular cash transfers to charities, 2). regular inter-households transfers, and 3). income tax.

3. Processing income data (CSES) and improving the quality

3.1. The Quality of household income data

A serious issue of data analysis, is the quality of survey data. Poor quality can affect the report or interpretation of results. The quality of the household income data (CSES) has been an issue; while there having the improvement of survey process it still has its weaknesses. They are resulting from the difficulty of gathering accurate income data about self-employment in small businesses, and an agricultural sector that has no depreciation of investments such as tools and animals, influencing a rather large number of households with negative income as well as causing income trends to fluctuate among the years (NIS (2010)). However, these issues can be resolved to improve the quality of survey data through various techniques such as: data cleaning, dealing with negative income, comparing data of consumption versus disposable income.

Data cleaning and dealing with outliers

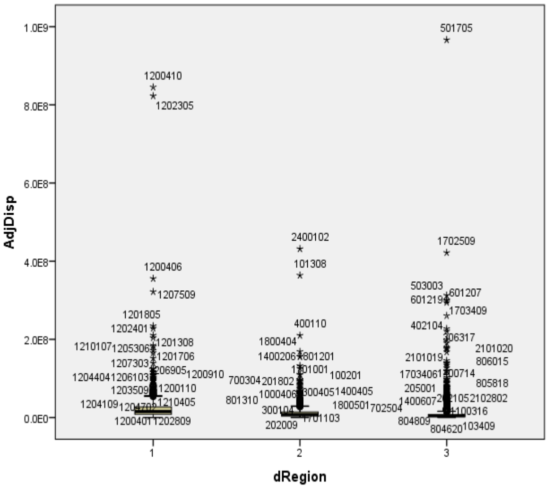
Data cleaning is very important stage in the analysis of data in order to know how the income data distribution is. The identification of the households with outliers values can be made by using SPSS fuctions with following steps: 1). loading dataset (i.e.household income data) to SPSS; 2). Select Analyze --> Descriptive Statistics --> Explore; 3) Move variables into the variable(s) box in explorer window; 4). Click "Statistics", click "Outliers"; 4). Click "Plots", and unclick "Stem-and-leaf"; and 5). Click OK. The output shows:

Figure1.Extreme values and outliers in income data

Extreme Values

			Case Number	Value
hhDispInc	Highest	1	1	966197500
		2	2	845340000
		3	3	822867500
		4	4	431466000
		5	5	421775300
	Lowest	1	11971	0
		2	11970	0
		3	11969	0
		4	11968	0
		5	11967	0[a]

[a] Only a partial list of cases with the value 0 are shown in the table of lower extremes.



Then a manual check of households with very high incomes or negative income (extreme values) is carried out using information from the boxplot and above table which show household income outliers. There are many different ways to handle these extreme values depending on the most important output statistics.

Survey data outliers maybe resulted from both sampling and non-sampling errors. Sampling error, there are not all households are included in the survey that causes to an uncertainty in the survey results that shows the standard error for the estimate. Non-sampling errors in CSES are non-response errors, response errors and data processing errors. CSES2010 reports show "a very low non-response error and its effects can be negligible. Some responses errors (measurement errors) in CSES cannot be detected unless special quality studies are carried out (re-interview studies, register studies, "data confrontation"). This has not been done" (NIS (2012)). Data processing errors are also shown in the outlier values and other values that are clearly inconsistent. Final solution dealing with outliers has still not been determined, we are currently investigating the use of Robust estimation methods; e.g. reverse calibration, regression, the nearest neighbor and numerical evaluations (Chambers & Ren, 2004).

3.2. Dealing with negative incomes

When deriving variables from the survey some households show a negative income, i.e. households with income from agricultures and non-agricultures sectors. These are resulting because no estimation of depreciation was made, since there are no rules for depreciation in Cambodia, when expenditures of investments for several different years are calculated. In this case, household with negative income have been replaced by a small amount of 4000 riels per annum, or around USD1.0 per year for household disposable income (NIS (2010)). To take care of the critical data and variables mentioned above in the forthcoming income report one ought to transform negative income to some positive value just above zero, "replacement with the minimum non-zero value" (Kovacevic, 2010). This is a method used in many income studies to take care of negative values in particular if measures such as the Gini-coefficient are to be computed, see e.g Kovacevic (2010).

4. An illustration of Cambodian income statistics 2007- 2011

The compilation of Cambodian household income statistics is mainly based on the income composition and income distribution for households as defined in the Recommendations on Household Income Statistics from Canberra Expert Group (The Canberra Group (2011). The major components of income are employee income, income from self-employment (agricultures, non-agriculture and owner occupied house), property income, current transfers received, total income, current transfers paid, and disposable income. The formula is defined as (SCB (2011)):

- Total Income = Employee income + income from self-employment + property income + current transfers received
- Disposable income = Total income – current transfers paid

Table 1. Cambodian Household Income Composition, average per month in 2007-2011

Source of income	Value in US Dollars				
	2007	2008	2009	2010	2011*
Cambodia					
Primary income	145	172	176	217	215
Wage and salary	49	58	58	72	85
Self-employment income	95	113	116	144	129
Agriculture	33	40	39	51	52
Non-agriculture	47	52	60	72	56
Owner occupied house	14	21	17	22	21
Property income	1	1	1	1	0
Total transfers received	10	7	5	6	6
Total income	155	179	180	223	221
Total transfers paid	1	3	3	6	4
Disposable income	153	176	178	217	217
Ex-rate (KHR/US$):	4,060	4,060	4,140	4,044	4,016

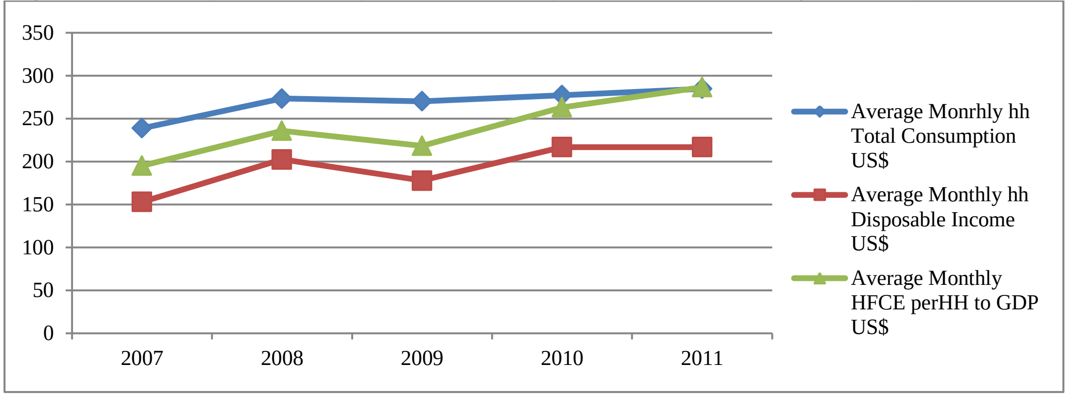
**Preliminary result and value 0 is rounding up.*

Trend analysis shows an annual growth rate of household disposable income and total income from 2009 to 2011, a significant jump. This may be a result from survey errors for both sampling and non-sampling errors, in partial is data collection "in a sample survey like CSES there will always be an inaccuracy in the estimated results as not everyone concerned is asked. When comparing CSES results between different years it is important to recognize the statistical uncertainty in the estimates, e.g. the true average number of rooms per household was in 2009 between 1.1 and 1.7 and in 2010 between 1.4 and 1.8. As these intervals are overlapping we cannot conclude that there is a real change in average room per household between 2009 and 2010" (NIS (2012)).

Total consumption vs. Disposable income

The mean of total consumption is higher than the mean disposable income. "The empirical literature on the relationship between income and consumption has established, for both rich and poor countries, that consumption is not closely tied to short-term fluctuations in income, and that consumption is smoother and less-variable than income. It is found that consumption is less variable over the period of a year and much more stable than income, especially in agricultural economies and therefore easier to estimate in a survey"(Deaton & Zaidi, (2002)). Evidence from other countries show that too little income is captured in surveys, especially this is the case with property income, as households with high income is more unwilling to answer (The Canberra group (2001)). In this case, the CSES also show that consumption is higher than disposable income.

Figure2. Consumption vs. Disposable Income per household, average values per month



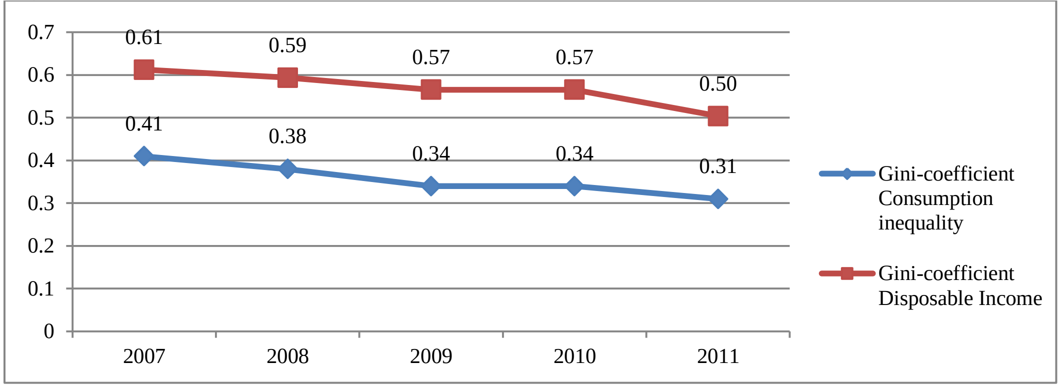
CSES/CSNA2011

Figure2 illustrates that the household disposable income shares to household final consumption expenditure (HFCE) of GDP represented 78%, 86%, 81%,82% and 76% in 2007 to 2011 respectively. So the disposable income shows average discrepancy 19% to HFCE of the GDP. And the discrepancy is higher to total household consumption, in average about 27%. "Generally, Household real adjusted net disposable incomes have risen less quickly than GDP in several countries, except in Norway, Denmark and France. However, it is hard to pinpoint the exact causes that enter the calculation directly (except for relative price changes)" (Dupont (2010)).

Gini-coefficient Consumption vs. Gini-coefficient Disposable income from CSES

The results show the gini-coefficient of disposable income is around 20% higher than the gini-coefficient of consumption. In this connection, it is preferred to use the gini-coefficient of consumption while the quality of survey data is still required to improve in data collection on household income as well as data processing.

Figure3. Gini-coefficient Consumption vs. Gini-coefficient Disposable income



CSES2007-2011 & MOP (2012)

5. Conclusions

Cambodian household income statistics were compiled in according to the concept and methodology as stated in the Canberra Group recommendation on household income statistics and basic applied to the CSES survey method. Its publication has been made by the NIS in accordance with available data collection of CSES. The quality of household income data estimates from Cambodia Socio-Economic Survey (CSES) has been has been some issues in its quality of survey data, such households with observed negative income, extreme values, missing values, changes of sample size, and error in data collection and processing. These issues are reduced by using many different methods and techniques such as data cleaning, treatment of outlier values, comparing household income with national accounts data, comparing household incomes and household expenditures, Gini-coefficient consumption vs. Gini-coefficient disposable income, and trend analysis for income data. However, Cambodian household income statistics from CSES should be improved in survey data quality and required to reduce significant non-sampling errors. Future work includes finding explanations of the shift in disposable income and the seemingly big differences between the Gini measures based on income versus consumption data.

References:

Angus Deaton & Salman Zaidi (2002), *Guidelines for Constructing Consumption Aggregates for Welfare Analysis*, The World Bank, p.12

The Canberra group (2001), *Final Report and Recommendations*, Expert Group on Household Income Statistics, Ottawa, p.54

Julien Dupont (2010), *From GDP to adjusted net disposable income of households–a decomposition analysis*, OECD Working Party on National Accounts, p.4

Kovacevic, M. (2010), *Measurement of Inequality in Human Development – A Review*, Human development research paper 2010/35 UNDP, p.13-23

Ministry of Planning (2012), *2012 Annual progress report on the Implementation of the NSDP update 2009-2013 with an overview of economic and social progress, including in select CMDGs*, p.16 & 39

National Institute of Statistics (2010), *Cambodian Socio-Economic Survey 2009*, p.97-107

National Institute of Statistics (2012), *Cambodia Socio-Economic Survey 2010*, p.104-110

Raymond L. Chambers and Ruilin, (2004) *Outlier Robust Imputation of Survey Data*, Joint Statistical Meetings Proceedings of Survey Research Section pp. 3336-3342

Statistics Sweden International Consulting Office (2011), *Methodological report on Household Income Statistics in Cambodia*, SCB-ICO, p.9-12