LIS
Technical Working Paper Series

No. 4

Censoring and Top-Coding in LIS Data

Karin Hederos Eriksson

Revised November 2011

L I S

CROSS-NATIONAL
DATA CENTER
in Luxembourg

Luxembourg Income Study (LIS), asbl

Report: Topcoding Project

Abstract

Variations between years and across countries in topcoding practices of income variables may bias trend and cross-national analyses of income inequality. In this report we take a first step towards analyzing the magnitude of these potential biases by investigating if the income variables in the LIS database have been topcoded by the national statistical offices providing the data. We sent out a question about topcoding practices to 30 national statistical offices. Among the 21 answering countries, five stated that they systematically topcode their income variables. These countries are Australia, Canada, Finland, Switzerland and the United States. The procedure used and the share of observations affected by topcoding seems to vary considerably across these countries. We also performed an empirical analysis of seven income variables in the LIS database. The results from this analysis indicated that Australia and the United States topcode gross wages and salaries.

Introduction

There are several reasons to censor extreme values at the top of the distribution when working with micro level income data. Firstly, one may want to protect the confidentiality of the individuals with the highest incomes. Secondly, it may be reasonable to impose an upper limit on the impact of a few extreme values on the national level of inequality. Finally, extreme values at the top of the distribution may result from errors (income figures may have been erroneously reported or coded). Censoring may therefore reduce the number of incorrect income figures.

Censoring at the very top of the distribution, topcoding, can be carried out in different ways. For example, the values above a certain upper cutoff point, the topcode, can be replaced by the value of the topcode or by the mean of all values larger than the topcode. If the point at which the topcode hits the income distribution varies across years, analyses of changes in income inequality over time may be biased. This issue has been thoroughly investigated for the Current Population Survey (CPS) data from the United States. These data have been subject to considerable changes in the topcoding procedure over the years. For example, between income years 1994 and 1995, the topcode for earnings was increased by 50% and socioeconomic cell means were introduced. This means that starting from income year 1995, all values above the topcode are replaced by the mean of the topcoded values for individuals with similar characteristics. Feng, S., Burkhauser, R. V. and Butler, J.S. (2006) conclude that important changes in the topcodes of the CPS data, like the one between 1994 and 1995, imply that the series are not comparable over time. Consequently, researchers must be cautious when making inference on trends in income inequality derived from these data. Topcoding may also complicate cross-national income inequality comparisons. As long as countries do not censor extreme values at the top of the distribution in the same way, cross-national comparisons will be biased.

The Luxembourg Income Study (LIS) database includes income micro data from thirty countries at multiple points in time. LIS does not apply any topcoding to the datasets received by the national statistical offices. The datasets may, however, have already been topcoded before being sent to LIS. One can thus encounter both types of issues related to topcoding described in the previous paragraph when conducting analyses of income inequality using LIS data. That is, if the topcoding of the LIS data varies over time and/or across countries, income inequality comparisons based on these data will be biased. A first step in order to assess the magnitude of this potential bias, is to investigate if and how the income data that LIS has received from national statistical offices have been topcoded.

To find out to what extent the income data received by LIS have been topcoded we have performed an empirical analysis on a subsample of seven LIS income variables. In addition, we have contacted the national statistical offices to get information on whether they topcode the income variables before sending them to LIS.

Description of Empirical Analysis

To get an idea of to what extent the LIS data have been subject to topcoding, we have examined seven LIS household income variables:

(1)	Gross wages and salaries – head	v39
(2)	Gross wages and salaries – spouse	v41
(3)	Net wages and salaries – head	v39net
(4)	Net wages and salaries – spouse	v41net
(5)	Cash property income	v8
(6)	Factor income	fi
(7)	Net disposable income	dpi

For each dataset from waves I-VI, we have investigated how many observations there are on the maximum value of each of the above income variables (it should however be noted that the datasets rarely contain all seven income variables). If, for a certain country, there is systematically more than one observation on the maximum value of a particular income variable, this indicates that the income variable has been subject to topcoding (that is, the data provider has topcoded the data before sending it to LIS). LIS income variables are often aggregates of several income variables received from the data provider. To ascertain that none of the income variables received from the data provider have been subject to topcoding, it would be necessary to examine all of them separately. Since we do not do that in this analysis, we might miss that some of them have been topcoded. Another issue is that cash property income, factor income and net disposable income are aggregates over individuals whereas topcoding normally takes place at the individual level. Thus, the aggregation over individuals further complicates the detection of topcoding.

All numbers have been in/deflated by consumer price index (CPI) to 2004 prices and converted into international dollars by using the 2004 purchasing power parity (PPP) index. Both the PPP and the national CPI data are primarily from the Source OECD database. When no data were available from the OECD, the International Monetary Fund statistical database was used.

For some datasets (Israel 1979, Poland 1986, Romania 1995 and 1997 and Russia 1992 and 1995) the PPP-adjusted and deflated values are not of reasonable size. This should primarily be due to excessive inflation and the exact figures for the levels of income from these datasets should thus not be given any importance.

Results from Enquiry and Empirical Analysis

Australia

Enquiry

According to the Australian Bureau of Labour Statistical, upper cutoff points are set for all base level income variables for confidentiality reasons. Each variable is considered separately and the decision made on appropriate cutoff point is based on the distribution of the values of this variable. Values above the upper cutoff point are then replaced by the weighted mean of the upper cutoff value and the values above it. Moreover, all income variables are perturbed. This means that between the upper and lower cutoffs, the values are grouped into clusters of observations. Each value within a cluster is perturbed to a new value by adding or subtracting an adjustment that is derived using random numbers, the size of the difference between the highest and lowest values within the cluster, and the mean value of the cluster.

Empirical analysis

See table 1 for detailed results. For all years except 1989 and 1995, the number of observations on the maximum value is larger than one for the gross wages and salaries of the head. It is, however, not possible to distinguish any trend in the development of either the number or the percentage of the observations on the maximum value. For years 1981 and 2003 the number of observations on the maximum value for the gross wages and salaries of the spouse is also larger than one.

Canada

Enquiry

Statistics Canada informs that prior to income year 1999, all income variables were topcoded in the sense that the four largest values within province and sex were replaced by the weighted mean of these values. Then a random rounding method was applied to add perturbation. Starting from 1999, all income variables are topcoded in the sense that the four largest values within province, size of area of residence, sex and age groups are replaced by the weighted mean of these values. Then a random rounding method is applied to add perturbation.

Empirical analysis

In the empirical analysis, there was however not possible to detect the above described topcoding. The gross wages and salaries of the head as well as of the spouse are neither aggregates of original variables nor over individuals. However, the only time the number of observation was larger than one for Canadian income variable, was in 1998 when there were two observations on the maximum value of the gross wages and salaries of the head. Most probably, the perturbation can explain why there is otherwise only one observation on the maximum value of the two mentioned income variables.

Finland

Enquiry

Statistics Finland only topcodes income variables when there is risk of identification. Topcoding has so far been applied twice, and only to capital gains (variable named tmyynt). In 2002 one observation was replaced by the second largest value and in 2004 three observations were replaced by the weighted mean of these three observations.

Empirical analysis

In the empirical analysis, there were no signs of topcoding of the Finnish data. This was expected since the only original income variable that has been subject to topcoding is not included in any of the LIS income variables evaluated in the empirical analysis.

Switzerland

Enquiry

Starting from income year 2002, household income values exceeding 0,5% of the income of all households are replaced by a value calculated on the basis of the income of the top quantile of the household income distribution. Topcoding in Switzerland is thus applied at the household level.

Empirical analysis

In the empirical analysis, there was nevertheless never more than one observation on the maximum value for any Swiss income variable. We do not know why the empirical analysis does not give any indication of topcoding. One possible explanation is that no observed income values were larger than the topcodes. The fact that the LIS income variables are aggregates of several original variables, each of which might have been topcoded, is another possible explanation.

United States

Enquiry

Income years 1979 – 1994

Prior to income year 1995, the values larger than the topcode were truncated. That is, if the topcode was $50,000 and the observed value was $60,000, the value was set to $50,000. All four basic earnings variables (ERN-VAL, WS-VAL, SE_VAL and FRM_VAL) as well as the fourteen non-earnings and non-government-related income variables (SUR-VAL1, SUR-VAL2, DIS-VAL1, DIS-VAL2, RET-VAL1, RET-VAL2, INT-VAL, DIV-VAL, RNT-VAL, ED-VAL, CSP-VAL, ALM-VAL, FIN-VAL and OI-VAL) were subject to the same topcodes. The non-earnings variables were however not affected by the topcodes since the maximum values of these variables were not larger than the topcodes.

Income years 1995-1997

Starting from income year 1995, the values larger than the topcode of the four basic earnings variables (ERN-VAL, WS-VAL, SE_VAL and FRM_VAL) are replaced by the mean of the topcoded values for individuals with similar characteristics (a separate mean is calculated for twelve different socioeconomic groups, defined by sex, race/origin and work status). Thus, even though the topcodes remain unchanged over the years, the maximum values of all earnings variables vary. Regarding the non-earnings income variables, values subject to top-coding are replaced by the average amount across all topcoded values for that income source. That is, when topcoding the non-earnings income variables, in contrast to the earnings variables, the individuals' socioeconomic background is not taken into consideration.

Income years 1998-2006

Starting from income year 1998, new and considerably lower topcodes for the non-earnings-income variables are introduced. There is, however, no change in the method used for topcoding neither earnings nor non-earnings income variables.

Empirical analysis

See table 2 for detailed results. For all years except 1991 and 2004 the number of observations on the maximum value is larger than one for the gross wages and salaries of the head as well as of the spouse. Both the number and the share of observations on the maximum value seem to decrease after income year 1994. This is not surprising since, as described above, prior to income year 1995, the values of the original variables constituting the base of the LIS-variable gross wages and salaries larger than the topcode were truncated. Since there are now 12 different values used for replacement instead of only one, everything else being equal, the number of observations on the maximum value should decrease. Note that since 12 different values are used for replacement, the share of the observations on the maximum value reported in table 2 underestimates the share of the observations affected by the topcoding.

When comparing the number as well as the share of observations on the maximum value of the gross wages and salaries of the head to the corresponding number and share of the spouse, the value is higher for the head for all years except in 1997. This is expected since the head normally has a higher income than the spouse. The difference between head and spouse does however seem to decrease over the years. Note that the maximum value of the wages and salaries of the head and of the spouse are the same for all years except 1991 and 2004. This is not surprising since these two LIS income variables are constituted by the same American original variable. That is, looking at the earnings and salaries of the head and the spouse separately, will underestimate the true number and share of the observations of the underlying original variable affected by the topcoding of that variable.

Apart from the gross wages and salaries of the head and the spouse, no other variables had repeatedly more than one observation on the maximum value over the years.

Conclusion

Regarding the empirical analysis, see table 3 for an overview of all datasets that contain at least one income variable for which the number of observations on the maximum value is larger than one. Australia and the United States are the only two countries for which it is possible to discern a recurrent pattern of more than one observation on the maximum value. This applies to the wages and salaries of, first and foremost, the head of the household, but also of the spouse. For the rest of the countries there is either never more than one observation on the maximum value of any income variable, or there are two observations on the maximum value at, at most, two points in time.

In total, we contacted 30 national statistical offices asking whether they topcode the income variables that they provide LIS with. We have received answers from 21 countries. Five of these countries, namely Australia, Canada, Finland, Switzerland and the United States answered that they systematically topcode their income variables. The 16 remaining countries answering the enquiry have not topcoded the data LIS has received (note, however, that starting from income year 2006, the Israeli income variables are subject to topcoding). See table 4 for an overview of the results from both the empirical analysis and the enquiry.

Concerning the five countries that we know topcode their income variables, the procedure used for topcoding, as well as the share of the observations affected by the topcoding, vary considerably. In Finland, a negligible number of observations is affected by the topcoding. Regarding Switzerland, we do not know if any observations are affected. For the United States, we have access to detailed information on how the topcodes as well as the procedure used for topcoding have changed over the years. Furthermore, the impact of these variations on intertemporal inequality analyses has been thoroughly investigated in the literature. For Canada, we do not know the values of the topcodes. However, the Canadian topcoding affects an in advance determined number of observations. Thus, except for between income years 1998 and 1999, when the topcoding procedure was revised, the share of observations affected by the topcoding should remain stable over time. Concerning Australia, we do not know if and how the topcodes have varied over time. The empirical analysis suggests that the gross wages and salaries of the head are topcoded, but perturbation renders the estimation of the exact share of the observations affected by the topcoding difficult.

In conclusion, we have started to investigate if the LIS income data have been subject to topcoding. However, in order to fully assess the magnitude of the potential biases arising when performing intertemporal and/or cross national income inequality analyses using LIS data, more detailed information on topcoding is needed.

References

Bishop, J. A., Chiou, J-R. and Formby J. P., “Truncation Bias and the Ordinal Evaluation of Income Inequality”, *Journal of Business & Economic Statistical*, Vol. 12, No. 1 (Jan., 1994), pp. 123-127

Burkhauser, R. V., Butler J. S., Feng S., Houtenville A. J., “Long term trends in earnings inequality: what the CPS can tell us”, *Economics Letters*, Volume 82, Issue 2, February 2004, Pages 295-299.

Burkhauser, R. V., Feng, S. and Jenkins, S. P., “Using the P90/P10 Index to Measure U.S. Inequality rends with Current Population Survey Data : A View from Inside Census Bureau Vaults”, Census Bureau Center for Economic Studies Paper No. CES-WP-07-17 (June, 2007).

Feng, S., Burkhauser, R. V. and Butler, J.S., "Levels and Long-Term Trends in Earnings Inequality: Overcoming Current Population Survey Censoring Problems Using the GB2 Distribution," *Journal of Business & Economic Statistical*, American Statistical Association, vol. 24 (Jan., 2006) pp. 57-62.

Fichtenbaum, R. and Shahidi, H., “Truncation Bias and the Measurement of Income Inequality”, *Journal of Business & Economic Statistical*, Vol. 6, No. 3 (Jul., 1988), pp. 335-337

Van Kerm, P., “Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC”, IRISS Working Paper Series 2007-01, IRISS at CEPS/INSTEAD.

Tables

Table 1: All income variables from Australia for which the number of observations on the maximum value is larger than one in at least one dataset.

Income variable	Dataset	Mean	Median	Max	No. of obs on max value	% of obs on max value
v39	Australia 1981	31490	30704	140793	2	0.0195
	Australia 1985	32117	30593	257600	6	0.1191
	Australia 1989	30465	28697	425641	1	0.0100
	Australia 1995	32019	28415	451556	1	0.0281
	Australia 2001	33959	29703	220129	13	0.3736
	Australia 2003	34603	30670	271394	4	0.0760
v41	Australia 1981	17222	16397	97565	2	0.0497
	Australia 1985	17856	16172	158895	1	0.0450
	Australia 1989	17983	16782	171262	1	0.0203
	Australia 1995	20739	18943	192178	1	0.0529
	Australia 2001	22959	20751	220129	1	0.0543
	Australia 2003	23924	21434	271394	2	0.0731

All values are expressed in 2004 international dollars.

Table 2: All income variables from the United States for which the number of observations on the maximum value is larger than one in at least one dataset.

Income variable	Dataset	Mean	Median	Max	No. of obs on max value	% of obs on max value
v8	United States 1979	3892	521	517925	1	0.0093
	United States 1986	5273	758	361312	1	0.0119
	United States 1991	5197	694	554927	2	0.0051
	United States 1994	4216	446	323076	1	0.0024
	United States 1997	6228	761	416002	1	0.0032
	United States 2000	5845	736	243774	1	0.0033
	United States 2004	4891	469	258139	1	0.0023
v39	United States 1979	37728	33834	130132	183	1.6469
	United States 1986	39193	34444	172218	86	1.0196
	United States 1991	38197	33296	277464	1	0.0026
	United States 1994	38514	31225	254891	319	0.7117
	United States 1997	42946	33539	520200	2	0.0059
	United States 2000	41474	32912	430049	11	0.0335
	United States 2004	42001	32000	748263	1	0.0020
v41	United States 1979	18876	15928	130132	4	0.0847
	United States 1986	21728	18083	172218	4	0.0973
	United States 1991	23115	19423	144976	1	0.0048
	United States 1994	25009	20392	254891	27	0.1419
	United States 1997	26280	21183	520200	6	0.0334
	United States 2000	37688	27866	430049	2	0.0106
	United States 2004	40462	30000	713263	1	0.0032

All values are expressed in 2004 international dollars.

Table 3: Datasets that contain at least one income variable for which the number of observations on the maximum value is larger than one.

Dataset	Income variable	Mean	Median	Max	No. of obs on max value	% of obs on max value
Australia 1981	v39	31490	30704	140793	2	0.0195
	v41	17222	16397	97565	2	0.0497
Australia 1985	v39	32117	30593	257600	6	0.1191
Australia 2001	v39	33959	29703	220129	13	0.3736
Australia 2003	v39	34603	30670	271394	4	0.0760
	v41	23924	21434	271394	2	0.0731
Belgium 1985	v41net	13960	13424	73221	2	0.1153
Canada 1998	v39	33702	29845	1025910	2	0.0093
Finland 1987	v41	15633	15868	91048	2	0.0325
Greece 2000	v39net	18238	16926	144669	2	0.1579
Luxembourg 1985	v41net	14945	12503	50168	2	0.6211
Luxembourg 1994	v8	6388	2526	75788	2	0.3766
Mexico 2000	v8	4732	1717	82429	2	0.5714
Russia 2000	v39net	4227	2725	54508	2	0.1229
United States 1979	v39	37728	33834	130132	183	1.6469
	v41	18876	15928	130132	4	0.0847
United States 1986	v39	39193	34444	172218	86	1.0196
	v41	21728	18083	172218	4	0.0973
United States 1991	v8	5197	694	554927	2	0.0051
United States 1994	v39	38514	31225	254891	319	0.7117
	v41	25009	20392	254891	27	0.1419
United States 1997	v39	42946	33539	520200	2	0.0059
	v41	26280	21183	520200	6	0.0334
United States 2000	v39	41474	32912	430049	11	0.0335
	v41	37688	27866	430049	2	0.0106

All values are expressed in 2004 international dollars.

Table 4: Summary of results from the enquiry and from the empirical analysis.

Country	Enquiry	Empirical analysis
		LIS income variables for which there are repeatedly more than one observation on the max value
Australia	All income variables are topcoded.	v39 v41
Austria	No answer	-
Belgium	No income variables are topcoded	-
Canada	All income variables are topcoded	-
Czech Republic	No answer	-
Denmark	No answer	-
Estonia	No income variables are topcoded	-
Finland	Extreme values are topcoded	-
France	No answer	-
Germany	No income variables are topcoded	-
Greece	No answer	-
Hungary	No income variables are topcoded	-
Ireland	No income variables are topcoded	-
Israel	No income variables are topcoded (i.e. no topcoding of income variables that LIS has received so far, but starting from income year 2006, the income variables are topcoded)	-
Italy	No answer	-
Luxembourg	No income variables are topcoded	-
Mexico	No answer	-
Netherlands	No income variables are topcoded	-
Norway	No income variables are topcoded	-
Poland	No income variables are topcoded	-
Romania	No income variables are topcoded	-
Russia	No income variables are topcoded	-
Slovak Republic	No answer	-
Slovenia	No income variables are topcoded	-
Spain	No income variables are topcoded	-
Sweden	No income variables are topcoded	-
Switzerland	Starting from 2002, all household income variables are topcoded	-
Taiwan	No answer	-
United Kingdom	No income variables are topcoded	-
United States	All income variables are topcoded	v39 v41