Estimating inequality among households from grouped data reported in historical surveys.

Hector Guttierez-Rufrancos[1] and Andrew Newell[2]

[1] University of Stirling, [2] University of Sussex and IZA.

Abstract

We study the biases of six different estimators in estimating four measures of inequality from grouped data derived from historical household expenditure surveys. All that remains of many historical household surveys are the tables of grouped data in the survey reports. In addition, historical surveys, especially those from before the 1930s, were mainly not designed in ways informed by the progress in statistical methods of that era. Interest continues to build in the study of historical data sets. Our results give timely support the idea that parametric methods, in particular Kakwani's (1980) Beta-Lorenz, are the least biased.

JEL codes: C13, N30

Keywords: inequality, grouped data, historical household surveys

Corresponding Author:

Prof. Andrew Newell
Department of Economics
University of Sussex, RIGHTON, BN1 9SL, UK
a.t.newell@sussex.ac.uk

1 Introduction

This short paper reports experiments to find the best method of estimating inequality measures using only tables of grouped results from historical household survey reports. The research is part of the *Global Income Inequality Project*, see http://www.sussex.ac.uk/globalincomeinequality/index. The project aims to improve the data base from which long-run trends in global income inequality can be estimated. To do this the project collects household level data from household income or expenditure surveys and, if the original data are not available, tables of data grouped by income or expenditure using reports of surveys. The aim is to perform this search all over the world, going as far back in time as possible. In many countries, the historical record starts in the early 20th century, but for some countries we have sources from deep into 19th century. The vast majority of the sources located by the project are tables of grouped averages.

Estimating inequality from grouped data is not a new topic. For instance, Gastwirth (1975) developed upper and lower bounds for some measures of inequality from grouped data and Cowell (1991) generalised these. Minoiu and Reddy (2009, 2014) provided very useful evaluations of both parametric and kernel density approaches to estimation. This note is closest to Minoiu and Reddy (2014) as, like them, we use Monte Carlo methods to test biases in alternative ways of inferring inequality from grouped data, and like them we come down on the side of preferring parametric estimation. However we differ from them in two important ways.

Firstly, we compare a broader range of non-parametric and parametric approaches. Secondly, and more importantly, we compare the estimation methods against actual samples, rather than against samples generated from parametric distributions. This is because our historical data sources present special issues for estimation. In particular, modern sampling techniques only came into generalised practise around the time of World War Two. Before then, surveys were often more limited in scope, for instance searching out what were taken to be typical households, or only households whose head worked in certain industries, for instance. These early surveys were also limited by their sampling techniques. They used what today we might describe as *ad-hoc* or snowball sampling. As a consequence our project can

be thought of as having two estimation problems to solve: first, how to generate reliable estimates of inequality in a particular sample, from grouped data, and secondly, how to make inference about what, if anything, our samples tell us about inequality in the broader population of households. The first problem is what concerns us here, and this is why we take a different approach to Minoiu and Reddy *op. cit*., and compare our estimates with actual estimates from original individual data, rather than generated samples.

Figure 1 gives part of a typical table from a historical source. There are six income groups and the table lists the numbers of households in each group, and well as average weekly income and the numbers of children living in the household. Note that the survey was of urban workmen's families, so that it covers only a subset, albeit a large subset, of British households in the period. Given the income distribution of this subset may have an idiosyncratic distribution, it seems best to us to test the estimation of distributional characteristic relative to actual distributions rather than hypothetical distributions.

Figure 1: Grouped household survey data for 1904 from the UK Board of Trade.

AVERAGE WEEKLY COST and QUANTITY of CERTAIN ARTICLES of FOOD CONSUMED by URBAN WORKMEN'S FAMILIES in 1904.

Limits of Weekly Income	Under 25*s.*	25*s.* and under 30*s.*	30*s.* and under 35*s.*	35*s.* and under 40*s.*	40*s.* and above.	All Incomes.
Number of returns	261	289	416	382	566	1,944
Average weekly family income	*s.* *d.* 21 4½	*s.* *d.* 26 11¾	*s.* *d.* 31 11¼	*s.* *d.* 36 6¼	*s.* *d.* 52 0½	*s.* *d.* 36 10
Average number of children living at home.	3·1	3·3	3·2	3·4	4·4	3·6

Source: British Parliamentary Papers, (1905) '*Consumption and the Cost of Food in Workmen's Families in Urban Districts of the United Kingdom*' Cd 2337.

The tables we find vary, in terms of what is recorded, in many important respects, and we need estimation methods that are robust to variations of available information. The primary

source of variation is in the number of income/expenditure groups. There is a well-established result (Gastwirth, ?) that the accuracy of estimating inequality is increasing in the number of groups. In early experimentation we also found this to be so. As a consequence we only report results for the case of eight groups, which is a typical number of groups in our historical sources. It is also well-established (Gastwirth,?) that accuracy improves if there are multiple observations per group. This comes about, for instance, when some tables, in addition to income groups, give mean incomes by, *inter alia*, industry, occupation, or region. This we investigate below. *A final extension is to judge the performance of the various methods of estimation when we are presented with data grouped by characteristics other than income/expenditure. One would expect lower estimates of inequality in those cases, and we investigate the extent of that underestimation.*

Our main finding is that estimation via Kakwani's (op. cit.) Beta-Lorenz curve usually gives the least biased results for cases where we given the mean income/expenditure. For case where we do not have the mean, we find that Beta-Lorenz estimation on interpolated data and interval estimation of lognormal parameters are the best options.

2 Methods

We will focus on the following measures of inequality: the Gini coefficient, and the 90/10, 90/50 and 50/10 percentile ratios. This a minimal set, but it allows us to study biases in the estimation of overall inequality as well as separately in upper and lower tail inequality.

To assess our estimators, we carry out Monte Carlo-style experiments on two individual household-level datasets: the Ministry of Labour 1953–54 survey of 12,854 households in the United Kingdom (Gazeley *et al.*, 2015) and the 1853 survey of 197 Belgian working class households (Ducpétiaux, 1855).

In both cases the data were resampled using bootstrap sampling with replacement. From each sample the measures of the Gini and decile ratios were obtained. We then 'binned' the data into equal size income-based bins (groups) and the data were collapsed to resemble group data such as that presented in Figure 1. We then employed the various estimators outlined in Section 2. This we repeated five hundred times. In the tables that follow we report only the six best estimators, in terms of mean bias, that we have experimented with. Our six best estimators are as follows.

1. 'Freq. Weighted' simply assigns the mean to every household in a particular income group. This is, therefore, akin to the 'between group' variance estimator. Note that this estimator approaches the true variance, from below, as the number of groups increases. We will see this as we progress through the results.
2. 'OLS logN' estimates a log normal curve on the weighted mean data. The inequality measures are derived from the estimated parameters.
3. 'Beta Lorenz' similarly estimates the parameters of the Lorenz curve generated from the Beta distribution, derived by Kakwani (1980), again estimated using weighted group mean data.
4. 'Interval LogN' uses the frequencies and group boundaries and an interval regression approach to estimation.
5. 'Hermite interpolation' follows Gastwirth and Glaubman (1976) and interpolates between group boundaries using Hermite interpolation, see Burden and Faires (2011,

chapter 3.4 pp. 136-144) and a suggestion by Gastwirth (1972) on how to treat the unbounded highest group.[1]

6. 'Hermite Beta-Lorenz' estimates the parameters of a Beta Lorenz curve using the data generated as in 5 above.

In practise, techniques 4, 5 and 6 are useful when the within-group mean income/expenditure is not recorded, so only the group boundary incomes and the numbers in each group are given. So our main question is which is overall least biased, but we are also interested in which is best in the case where group means are not recorded.

[1] This interpolation technique chooses a cubic function that fits between two points in two-dimensional space, given (a) these points are known and (b) the slope of the relationships at the two point are also known. In practice, we use the Cox (2012) implementation in Stata (StataCorp, 2015).

3 Results for income/expenditure groups.

Table 1 gives our four inequality measures calculated with the individual data from the two surveys. As with all surveys of workers' households, overall inequality is lower than might be expected from the population of all households. The results are, however, very typical of the results we find from similar samples of working class households. Figure 2 gives the Lorenz curve for the 1953/4 data set. The curve for the 1855 data is very similar.

Table 1 Household expenditure inequality statistics from the individual household data of two data sets.

	Ministry of Labour 1953/4	Belgium 1855
Gini (%)	31.7	25.4
90/10 decile ratio	4.71	3.08
90/50 decile ratio	1.93	1.70
50/1 decile ratio	2.44	1.81

Figure 2: The Lorenz Curve for household income per equivalent adult in the Ministry of Labour 1953/4 survey.

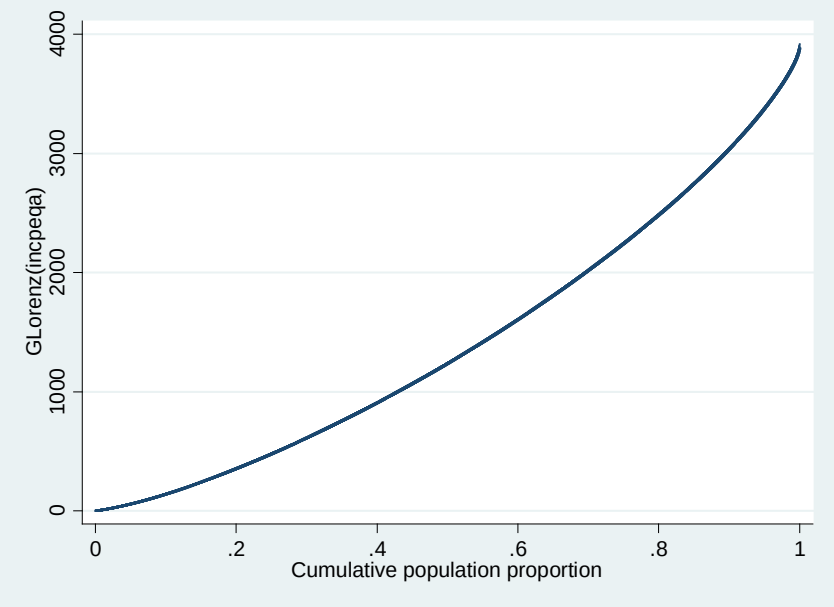

Table 2: Mean bias of inequality measures using eight income groups, Ministry of Labour 1953/4 Household expenditure survey.

Estimator	Gini	P90/10	P90/50	P50/10	
	Mean bias (inverse rank)	Mean bias (inverse rank)	Mean bias (inverse rank)	Mean bias (inverse rank)	Sum of ranks (inverse rank)
Freq. Weighted	-0.015 (4)	-0.206 (2)	0.199 (3)	-0.324 (3)	12 (3)
OLS LogN	-0.045 (6)	-3.528 (6)	-0.846 (6)	-1.344 (6)	24 (6)
Beta-Lorenz	-0.007 (3)	-0.686 (4)	-0.046 (1)	-0.303 (2)	10 (1)
Interval LogN	-0.042 (5)	0.193 (1)	0.283 (4)	-0.224 (1)	11 (2)
Hermite Interpolation	-0.005 (2)	1.633 (5)	0.337 (5)	0.357 (4)	16 (5)
Hermite-Beta-Lorenz	-0.003 (1)	0.544 (3)	-0.054 (2)	0.359 (5)	11 (2)

Table 2 gives our key results. Each cell reports the mean bias derived from the 500 replications described above. We choose not to report results from different numbers of groups. Early experimentation showed, as might be expected, a general tendency for all estimators to exhibit smaller biases as the number of groups increases. This a well-known (Gastwirth, 197?) and intuitive result that we do not test here. In order to render the results simpler to digest each cell also records, in brackets, the rank across estimators of the mean bias, from smallest to largest. To be clear, OLS LogN designates estimation via a log normal regression that consistently generates the worst (highest ranked) mean biases across the 6 methods on all 4 measures.

One way to summarise the results is to sum the ranks for each estimator across the four inequality measures. This is reported in the final column, together with an inverse ranking of those sums. On this criterion, the Beta-Lorenz curve tends to offer the lowest biases. As Beta-Lorenz requires the estimation of three parameters, it seems this extra flexibility compared to the more restrictive lognormal, for instance, reduces bias. Note this superiority is driven by low biases to 90/50 and 50/10 estimation, so the method is good at capturing the shape of the Lorenz curve at each end of the distribution. Of course there are distortions that rankings induce. If we note, for instance that in rankings the Beta-Lorenz performs less well

on Gini bias, coming 3rd from six. However, the size of the mean biases for the best three estimators are all less that one Gini point.

Of the other estimators, the Beta Lorenz estimated from Hermite-interpolated data and the interval lognormal also perform very well. This suggests that we might recommend the use of Beta-Lorenz for the case were the group mean incomes are given and both Hermite- Beta Lorenz and interval lognormal for the cases where the group means are not given. However, before we come to any conclusion, we turn to the results of performing the same set of tests on the Belgian 1855 data. See Table 3.

Table 3 Mean bias of inequality measures based upon eight income bins Belgium 1855 Household expenditure survey.

Estimator	Gini	P90/10	P90/50	P50/10	
	Mean bias (inverse rank)	Mean bias (inverse rank)	Mean bias (inverse rank)	Mean bias (inverse rank)	Sum of ranks (inverse rank)
Freq. Weighted	-0.002 (2)	-0.100 (3)	-0.028 (2)	-0.030 (4)	11 (2)
OLS LogN	-0.011 (4)	-1.857 (6)	-0.618 (6)	-0.692 (6)	22 (6)
Beta-Lorenz	0.005 (3)	0.059 (2)	0.045 (3)	-0.013 (2)	10 (1)
Interval LogN	-0.014 (6)	0.046 (1)	0.054 (4)	-0.029 (3)	14 (4)
Hermite Interpolation	-0.002 (1)	0.562 (5)	0.303 (5)	0.010 (1)	12 (3)
Hermite-Beta-Lorenz	-0.012 (5)	0.225 (4)	0.007 (1)	0.123 (5)	15 (5)

We see from Table 3 that Beta-Lorenz is again, marginally, the estimator of lowest bias across the four inequality measures. However, the rankings of the other estimators are different from the results for the 1953/4 data in Table 2. In particular, the choice of estimator for the cases where the group mean is not reported is no longer clear. To try to find a clearer result, we turn to experiments that reflect cases that are often reported, where results of a survey are grouped not only by income, but by additional characteristics, such as region or occupation.

4. Results by income/expenditure and other characteristics.

For the 1953/4 Ministry of Labour data we now perform the Monte Carlo analysis by: income and household size (HH size); income and occupation; income and region; income and industry, and finally, income, region and industry. For the Belgian 1855 data we report a small set of extended results, adding in, respectively, groups by household size, occupation and region. The reporting is simplified for ease of digestion. In Tables 4 (1953/4) and Table 5 (1855) we report only the final rank, so the ranking by lowest sum of ranks, as in the final columns of Tables 2 and 3.

Table 4 Rank by sum of ranks 53/4

Groups:	Income only	+ HH Size	+ Occupation	+Region	+ Industry	+ occupation and region
No. groups	8	72	104	104	204	1352
Freq. Weighted	4	1	2	2	2	2
OLS LogN	6	6	6	6	6	6
Beta-Lorenz	1	1	1	1	1	1
Interval LogN	2	3	3	4	4	5
Hermite Interpolation	5	5	5	5	5	4
Hermite-Beta-Lorenz	2	3	3	3	3	3

Table 4 results show us something that we mentioned earlier. It is that as the number of groups increases, the 'Freq. Weighted' estimator improves its ranking. These extra observations are giving more information about the shape of the within-group distribution, and this reduces the relating bias of that estimator especially. Overall, though, in these data, Beta-Lorenz remains the best-ranked estimator and Hermite-Beta-Lorenz is the best-ranked estimator for data tables where the group means are not given.

Table 5 Ranks by sum of ranks, 1855.

Groups:	Income only	+ HH Size	+ Occupation	+Region
No. groups	8	24	40	104
Freq. Weighted	2	2	1	1
OLS LogN	6	6	6	6
Beta-Lorenz	1	1	3	2
Interval LogN	4	4	4	3
Hermite Interpolation	3	3	2	5
Hermite-Beta-Lorenz	5	4	4	3

Table 5 gives the results of a similar set of experiments on the Belgium 1855 data set. For smaller numbers of groups, Beta-Lorenz is again the least-biased estimator. Here the predictable reduction in bias for the 'freq. weighed' estimator means that, for large numbers of groups, it becomes the best-ranked estimator. In the case of estimators where group means are unavailable, as in Table 3, Interval Lognormal and Hermite-interpolated Beta-Lorenz are the best ranked.

5 Conclusions

This methodological note has outlined various approaches at estimating measures of income inequality from grouped tables, in historical cases where the aims of the surveys and the sampling methods may make the sample quite idiosyncratic. The biases of each of these approaches were assessed through a bootstrap sampling experiment. We take two data sets, collected almost a century apart, *names here*

Across both datasets it was found that the least biased estimator is the Beta-Lorenz first suggested by (Kakwani, 1980). It characterises the decile ratios very well. It does not, however, provide the best estimate of the Gini coefficient. Where the data only provides interval information, the best estimator is the combination of the Beta-Lorenz and the

Hermite interpolation suggested by Gastwirth and Glauberman (1976). However, in some extreme cases this fails to numerically resolve the non-linear least squares. If this is the case then the suggested second-best performer is the interval regression based lognormal estimator.

6. References

Burden, R. L. and Faires, T., (2011) *Numerical Analysis*, 9th edition, Brook/Cole Cengage Learning, Boston.

Cowell, F. A., (1991) Grouping bounds for inequality measures under alternative informational assumptions, *Journal of Econometrics*, 48, 1-14.

Ducpetiaux, E. A., (1855), *Budgets économiques des classes ouvrières en Belgique: subsistances, salaires, population*, M. Hayez, Bruxelles.

Gastwirth, J.L. (1972), Robust Estimation of the Lorenz Curve and Gini Index. *Review of Economics and Statistics*, 54, 306-316.

Gastwirth, J.L. (1975), The Estimation of a Family of Measures of Economic Inequality, *Journal of Econometrics*, 3, 61-70.

Gastwirth, J.L. and Glauberman, M. (1976), On the Interpolation of the Lorenz Curve and Gini Index from Grouped Data, *Econometrica*, 44, 479-483.

Gazeley, I., Gutierrez Rufrancos, H., Newell, A., Reynolds, K. and Searle, R. (2017), The poor and the poorest, 50 years on: evidence from British Household Expenditure Surveys of the 1950s and 1960s. *Journal of the Royal Statistical Society*: Series A (Statistics in Society), 180: 455–474. doi: 10.1111/rssa.12202

Kakwani, N., (1980) *Income inequality and poverty: methods of estimation and policy applications*, World Bank, Oxford University Press, Oxford.

Minoiu, C. and Reddy, S.G. (2009), Estimating Poverty and Inequality from Grouped data: How Well Do Parametric Methods Perform? *Journal of Income Distribution* Vol. 18,

Minoiu, C. and Reddy, S.G., (2014) Kernel density estimation on grouped data: the case of poverty assessment, *Journal of Economic Inequality* 12: 163. doi:10.1007/s10888-012-9220-9

Appendix

The Gini coefficient

For a population of individuals *i* of size *N* with income *y* the unweighted Gini coefficient is given by:

$$\text{Gini} = \frac{2}{N(N-1)} \sum_{i=1}^{N} (y_i - \bar{y}) \tag{1}$$

The analogous estimator for a weighted Gini coefficient (with weights w_i) is given by the following expression:

$$\text{Gini}_w = \sum_{i=1}^{N} \left(\frac{w_i}{\sum w} \cdot \frac{2}{\bar{y}} \cdot \frac{2 \sum w - w_i + 1}{2 \sum w} |y_i - \bar{y}| \right) \tag{2}$$

Percentile Ratios

If incomes *y* are ordered from lowest to highest over *n* population such that the rank may then be calculated as $R=(n+1)q/100$, with the parts that are integer *r* and fractional portion *f*, thus the q^{th} percentile is given by (Mood and Graybill, 1963):

$$c_q = x_r + f \times (x_{r+1} - x_r) \tag{3}$$

The percentile ratios of interest will be given by c_{90}/c_{10}, c_{90}/c_{50} and c_{50}/c_{10}. In practice we implement this using the Jenkins (1999) implementation in Stata (StataCorp, 2015). However, it is well known that direct application to grouped data creates a downward bias of the estimates of inequality because the procedure ignores intergroup (or within-bin) inequality (Lerman and Yitzhaki, 1989; Pyatt et al., 1980). There are numerous approaches which attempt to overcome this downwards bias.

.2 Parametric Estimators

Other approaches to the group data issue have been considered in the literature. Often these methods rely on a parametric characterisation of the data. Below we will explore two such approaches.

2.4.1 The lognormal approach

It is well known that income often follows a lognormal distribution (Aitchison and Brown, 1963). Thus, for the density:

$$f_{y;\mu,\sigma} = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\ln y-\mu)^2}{2\sigma^2}}, \quad x > 0 \tag{5}$$

The parameters μ and σ may be estimated using the following log-likelihood function for the case of having a continuous variable for mean income:

$$\ln \ell_j = \ln \phi(\frac{y_j - \mu_j}{\sigma_j}) - \ln \sigma_j \tag{6}$$

Where group data only have income bands then interval regression with the following likelihood function will yield the lognormal parameters (Wooldridge, 2010, pp.783):

$$\begin{aligned}
\ln \ell_j = & -\frac{1}{2} \sum_{j \in \mathcal{C}} w_j \left\{ \left(\frac{y_j - x\beta}{\sigma} \right)^2 + \ln 2\pi\sigma^2 \right\} \\
& + \sum_{j \in \mathcal{L}} w_j \ln \Phi \left(\frac{y_{\mathcal{L}j} - x\beta}{\sigma} \right) \\
& + \sum_{j \in \mathcal{R}} w_j \ln \left\{ 1 - \Phi \left(\frac{y_{\mathcal{R}j} - x\beta}{\sigma} \right) \right\} \\
& + \sum_{j \in \mathcal{X}} w_j \ln \left\{ \Phi \left(\frac{y_{2j} - x\beta}{\sigma} \right) - \Phi \left(\frac{y_{1j} - x\beta}{\sigma} \right) \right\}
\end{aligned} \tag{7}$$

Here $\Phi(.)$ is the standard cumulative normal and w_j is the weight for the observation. the expression (6) will yield the estimates of the parameters μ and σ. With these two parameters it is straightforward to obtain the inequality measures of interest. The Lorenz curve of the lognormal distribution is:

$$L(p) = \Phi \left(\frac{\ln z - \mu - \sigma^2}{\sigma} \right) \tag{8}$$

From there it can be shown that the Gini coefficient under the assumption of lognormality can be estimated by (Aitchison and Brown, 1963):

$$\text{Gini}_{LN} = 2\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1 \tag{9}$$

Given that both parameters of the lognormal distribution are known, it is straightforward to estimate the percentiles using the expression:

$$c_p = e^{\mu + \Phi^{-1}(p)\sigma} \tag{10}$$

As both parameters are obtained through estimation on the data it is therefore straightforward to obtain analytical standard errors for these measures.

2.4.2 Beta Lorenz Curve

An alternate approach is to directly estimate the Lorenz curve given the group data. There are various functional forms for this approach. However, one which has been adopted in practice in the literature is the Beta-Lorenz curve (Kakwani, 1980). One of the benefits of this particular functional form, is that in all instances this functional form will yield a valid Lorenz curve (Datt, 1998). This curve can be fit using non-linear least squares on the following functional form:

$$L(p) = p - \theta p^{\gamma}(1-p)^{\delta} \tag{11}$$

The parameters *Θ, γ* and *δ* are then utilised to estimate the Gini coefficient using the following expression due to Datt (1998):

$$\text{Gini}_{BL} = 2\theta B(1, 1+\gamma, 1+\delta) \tag{12}$$

Here *B*(.) is the cummulative function of the incomplete beta distribution. The selected percentiles are obtained by evaluating the first difference of the Lorenz curve at the desired percentile as follows:

$$L'(p) = c_p = -\gamma\theta(1-p)^{\delta}p^{(\gamma-1)} + \delta\theta(1-p)^{(\delta-1)}(p^{\gamma}) + 1 \tag{13}$$

As with the lognormal distribution, it is therefore straightforward to compute the relevant standard errors of all of the parameters.