Joining the dots and fitting curves:

A *revised* Practicioner's guide to measuring inequality using group data.

Hector Rufrancos and Andrew Newell

January 2016

1 Introduction

One of the challenges that we face is the estimation of inequality measures using summary results. There are many approaches for this class of problems that have been suggested by the literature. The aim of this note is to introduce some of these approaches and assess the suitability of these methods using two datasets using Monte Carlo methods.

2 Methodology

For the purposes of this paper we will focus on the following measures of inequality: the gini coefficient, and the following percentile ratios: 90/10, 90/50 and 50/10. The estimation of these measures is to an extent dependent on the accuracy of the implicitly estimated Lorenz curve. Formally, if one ignores the group data nature of the data the measures may be obtained as follows:

2.1 Individual measures of Inequality

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 can be obtained from 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 Equally if incomes y are ordered from lowest to highest over n population such that $y_1 \leq y_2 \leq \ldots \leq y_n$ the rank may then be calculated such that $R = (n+1)q/100$, with the parts that are integer r and fractional portion f, thus the q percentile is given by (Mood and Graybill, 1963):

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

Thus, the percentile ratios of interest will be given by $\frac{c_{90}}{c_{10}}$, $\frac{c_{90}}{c_{50}}$ and $\frac{c_{50}}{c_{10}}$.

Table 1: Group data derived from MoL 1953/54

Group	Lower bound income	Upper bound income	Number of households	Mean Household Income in pence	Mean Household Expenditure in pence
1	0	1,084	1,960	516.28	4,133.38
2	1,085	2,168	2,981	1,727.50	5,313.20
3	2,169	3,252	3,896	2,670.99	7,275.74
4	3,253	4,335	2,263	3,714.76	9,446.33
5	4,337	5,421	914	4,802.26	11,723.82
6	5,422		837	7,892.87	15,894.52

In practice we implement this using the Jenkins (1999) implementation in Stata (StataCorp, 2015).

This simplistic approach ignores the group nature of the data and it is well known that it creates a downward bias of the estimates inequality because the data ignores intergroup (or within bin) inequality (Lerman and Yitzhaki, 1989, Pyatt et al., 1980). For this reason, this class of estimators will not be tested. There are numerous approaches which attempt to overcome this downwards bias, but we shall implement and assess various parametric and non-parametric approaches.

The data that we will be using to estimate is group data that may or may not contain mean incomes and expenditures. Table 1 shows a derived dataset from Gazeley et al. (2015) and is a typical example of the type of data that we may expect a typical dataset to contain. The data that is often reported in summary data of household surveys may contain all of the elements shown in table 1, or only a few portions of this data.

2.2 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.

2.2.1 Log Normal

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

$$f_{y;\mu,\sigma} = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\ln y-\mu)^2}{2\sigma^2}}, \quad x > 0 \tag{4}$$

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

$$\ln \ell_j = \ln \phi(\frac{y_j - \mu_j}{\sigma_j}) - \ln \sigma_j \tag{5}$$

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

$$\begin{aligned}\ln \ell_j = &-\frac{1}{2}\sum_{j\in\mathcal{C}} w_j \left\{ \left(\frac{y_j - x\beta}{\sigma}\right)^2 + \ln 2\pi\sigma^2 \right\} \\ &+ \sum_{j\in\mathcal{L}} w_j \ln \Phi\left(\frac{y_{\mathcal{L}j} - x\beta}{\sigma}\right) \\ &+ \sum_{j\in\mathcal{R}} w_j \ln \left\{1 - \Phi\left(\frac{y_{\mathcal{R}j} - x\beta}{\sigma}\right)\right\} \\ &+ \sum_{j\in\mathcal{X}} w_j \ln \left\{\Phi\left(\frac{y_{2j} - x\beta}{\sigma}\right) - \Phi\left(\frac{y_{1j} - x\beta}{\sigma}\right)\right\}\end{aligned} \tag{6}$$

where $\Phi(.)$ is the standard cumulative normal and w_j is the weight for the j_th observation. Both expressions (5) and (6) will yield the estimates of the parameters which characterise the lognormal distribution namely, μ and σ. With these two parameters it is then straightforward to obtain the inequality measures of interest. The Lorenz curve of the Lognormal distribution is:

$$L(p) = \Phi\left(\Phi^{-1}(p) - \sigma^2\right) \tag{7}$$

where $\Phi^{-1}(.)$ is the inverse of the standard cumulative normal, and σ^2 is estimated from either (5) or (6). From there it can be shown that the Gini coefficient under the assumption of lognormality can be estimated by (Aitchison and Brown, 1963):

$$\text{Gini}_{LN} = 2\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1 \tag{8}$$

Given that both parameters of the lognormal distribution are known, it is straightforward to estimate the p^{th} percentile using the following:

$$c_p = e^{\mu + \Phi^{-1}(p)\sigma} \tag{9}$$

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

2.2.2 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:

$$L(p) = p - \theta p^{\gamma}(1-p)^{\delta} \tag{10}$$

The parameters θ, γ, δ are then utilised to estimate the gini coefficient using the following (Datt, 1998):

$$\text{Gini}_{BL} = 2\theta B(1, 1 + \gamma, 1 + \delta) \tag{11}$$

where $B(.)$ is the cummulative function of the incomplete beta distribution.

Equally, we can obtain the selected percentiles by evaluating the first difference of the lorenz curve at the desired percentile as follows:

$$L'(p) = c_p = -\gamma\theta(1-p)^{\delta}p^{(\gamma-1)} + \delta\theta(1-p)^{(\delta-1)}(p^{\gamma}) + 1 \tag{12}$$

and similarly to the lognormal distribution as all of the parameters are estimated it is therefore straightforward to compute their relevant standard errors.

BL With Hermite Interpolation The Log Normal characterisation of the data allow us to ignore the issue of banded data, as we are able to employ interval regression to obtain the relevant parameters for inequality estimation. However, when assessing the fit of the Beta Lorenz to banded data we are faced with a problem in the estimation of inequality, namely: Which income should be used for the estimation of inequality? Brittain (1962) suggests that the appropriate approach is to take the mid-point of the bounds should be used for estimation of the gini coefficient using (1). However, this approach may lead to inconsistent estimates of the gini as noted by Gastwirth and Glauberman (1976). Instead we opt to evaluate their proposal of employing the piecewice cubic Hermite interpolator to interpolate a fixed number of extra observations within each group (or bin). Given the data $L(p_i) = L_i$ for $i = 0, 1, \ldots, k+1$ as the lower bound and $L'(p_i) = L'_i$ for $i = 0, 1, \ldots, k$ for the upper bound. The Hermite interpolator would then be given as follows for the bin $[p_i, p_{i+1}]$:

$$H(p) = a_0 + a_1(p - p_i) + a_2(p - p_i)^2 + a_3(p - p_i)^3 \tag{13}$$

where

$$\begin{aligned} a_0 &= L_i \\ a_1 &= L'_i \\ a_2 &= \Delta^{-2}(L_{i+1} + L_i - \Delta L'_i) \\ a_3 &= \Delta^{-3}[\Delta(L'_{i+1} + L'_i) - 2(L_{i+1} - L_i)] \\ \Delta &= (p_{i+1} - p_i) \end{aligned}$$

In practice we implement this using the Cox (2012) implementation in Stata (StataCorp, 2015), which yields the ‘continuous measure’ of income $L(p)$ which is then utilised to estimate the Beta-Lorenz curve.

2.3 Bias of Estimators

In order to determine which estimator should be used by the GII project we decided to carry out a Monte Carlo-style experiment. We took two datasets for which we have microdata, namely

Ministry of Labour 1953–54 survey of 12,854 household in the United Kingdom (Gazeley et al., 2015) and the 1853 survey of 197 Belgian working class households (Ducpétiaux, 1855).

For both datasets the same approach was taken. The data was resampled using boostrap sampling with replacement. From each sample the measures of the Gini and decile ratios were obtained. We then 'binned' the data into equal size wage-based bins and the data was collapsed to resemble group data such as that presented in Table 1. We then employed the various estimators outlined in Section 2. This was repeated 500 times. The bias of an estimator is normally yielded by:

$$\mathrm{Bias}_\theta[\hat{\theta}] = E_\theta[\hat{\theta}] - \theta = E_\theta[\hat{\theta} - \theta] \tag{14}$$

which for our purposes will be obtained by:

$$\mathrm{Bias}_\theta[\hat{\theta}] = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta} - \theta^\star) \tag{15}$$

where $\hat{\theta}$ is the estimate for the measure of interest yielded by the estimator, and $\theta^\star$ is the 'true' theta for the sample obtained prior to collapsing the data.

This exercise was carried out for a number of potential number of bins in order to assess the performance of the estimators with respect to the type of group data obtained.

3 Results

This section reports the results of the experiment outlined in section 2. The procedure was carried out on the number of bins reported in each table in this section. In all instances the bins were determined by income but the means of the sample in each bin were computed for both income and expenditure. Thus, we will be able to determine the performance of measures of inequality using mean expenditures in income ordered bins. This type of data is very similar to what the GII project has found in summary reports of surveys carried out pre 1960.

3.1 MoL 1953–54

3.1.1 Income

Tables 2, 3, 4 and 5 present the results for the MoL 1953/54 survey for the income measures. Each table reports the result of 500 bootstrap sample replications of the bias estimate measure outlined in Section 2. Each procedure was carried out for bin sizes ranging from 5–10 bins. Each table reports the rank of the estimator. This rank is the field rank across all number of bins. Thus, the best performing estimator is that which minimises the absolute bias across the different sizes of bins. For all of the Hermite interpolations the ten extra observations were interpolated per bin.

The results for income on the MoL survey are as follows. For all three decile ratios the least biased estimator is the Beta-Lorenz. The mean bias is alway less than one percentage point. However, if the data that is available only contains the group intervals the suggested least biased approach would be to utilise the combination of the Hermite interpolation and the Beta-Lorenz estimator.

However, with respect to the Gini coefficient the results are quite stark. A naive application of the individual gini coefficient weighted by the number of households per interval obtained by the estimator in expression 2 performs the best amongst all of the estimators presented. However, the clear 'winner' for the banded-only data would be the interval regression based estimator.

Table 2: Bias on Estimates of Gini Coefficient using Income from Group data, UK 1953/4

# of Bins	5	6	7	8	9	10	Rank
Groups Naive Freq. Weighted	0.010	0.014	0.017	0.018	0.019	0.020	1
Hermite Interpolation (bands)	0.041	0.038	0.036	0.034	0.033	0.031	5
Lognormal interval regression (bands)	0.014	0.017	0.019	0.021	0.023	0.024	2
Lognormal OLS	0.046	0.048	0.057	0.060	0.064	0.074	6
Beta-Lorenz	0.027	0.027	0.027	0.026	0.026	0.026	4
Hermite-Beta Lorenz (bands)	0.044	0.033	0.025	0.021	0.018	0.014	3

Table 3: Bias on Estimates of $p90/p10$ ratio using Income from Group data, UK 1953/4

# of Bins	5	6	7	8	9	10	Rank
Groups Naive Freq. Weighted	4.601	6.831	9.943	0.348	0.296	0.999	4
Hermite Interpolation (bands)	15.506	15.965	5.631	4.817	4.252	4.516	6
Lognormal interval regression (bands)	3.895	4.099	4.280	4.442	4.602	4.726	5
Lognormal OLS	-3.359	-3.361	-3.339	-3.339	-3.324	-3.312	3
Beta-Lorenz	0.305	0.322	0.336	0.337	0.356	0.352	1
Hermite-Beta Lorenz (bands)	3.418	2.584	2.021	1.791	1.717	1.486	2

Table 4: Bias on Estimates of $p90/50$ ratio using Income from Group data, UK 1953/4

# of Bins	5	6	7	8	9	10	Rank
Groups Naive Freq. Weighted	-0.140	0.234	-0.082	0.199	-0.275	-0.053	4
Hermite Interpolation (bands)	-0.238	0.044	-0.182	0.039	-0.143	-0.051	3
Lognormal interval regression (bands)	0.993	1.029	1.059	1.086	1.111	1.133	6
Lognormal OLS	-0.789	-0.786	-0.780	-0.779	-0.776	-0.769	5
Beta-Lorenz	-0.058	-0.051	-0.045	-0.044	-0.041	-0.039	1
Hermite-Beta Lorenz (bands)	0.038	-0.010	-0.035	-0.056	-0.065	-0.080	2

Table 5: Bias on Estimates of $p50/p10$ using Income from Group data, UK 1953/4

# of Bins	5	6	7	8	9	10	Rank
Groups Naive Freq. Weighted	2.741	2.877	5.456	-0.063	0.579	0.596	5
Hermite Interpolation (bands)	9.436	7.978	3.450	2.382	2.555	2.453	6
Lognormal interval regression (bands)	0.505	0.536	0.567	0.594	0.622	0.639	2
Lognormal OLS	-1.249	-1.250	-1.240	-1.239	-1.231	-1.227	4
Beta-Lorenz	0.237	0.235	0.235	0.233	0.240	0.235	1
Hermite-Beta Lorenz (bands)	1.678	1.349	1.105	1.022	0.999	0.902	3

4 Conclusions

This methodological note has outlined various approaches at estimating measurments of income inequality when encountering group data. The suitability of each of these estimators was assessed through a bootstrap sampling experiment. This allowed to determine the bias of the estimators.

Across both datasets it was determined that the least biased estimator is the Beta-Lorenz first suggested by (Kakwani, 1980). It characterises the decile ratios very well. It however, does not 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 the suggested second-best performer is the interval regression based lognormal estimator.

References

Aitchison, J. and J. A. C. Brown (1963). *The Lognormal Distribution.* Number 5 in Department of Applied Economics Monograph. Cambridge: Cambridge University Press.

Brittain, J. A. (1962). Interpolation of Frequency Distributions of Aggregated Variables and Estimation of the Gini Concentration Measure. *Metron 22*(1), 98–109.

Cox, N. J. (2012, December). PCHIPOLATE: Stata module for piecewise cubic Hermite interpolation. Statistical Software Components, Boston College Department of Economics.

Datt, G. (1998, October). Computational Tools for Poverty Measurement and Analysis. Food Consumption and Nutrition Division Working Paper 50, International Food Policy Research Institute, Washington DC, USA.

Ducpétiaux, E. (1855). *Budgets Économiques des Classes Ovrières en Belgique. Subsistance, Salaires, Population.* Bruxelles: M. Hayez, Imp. De La Commission Centrale de Statistique.

Gastwirth, J. L. and M. Glauberman (1976, May). The Interpolation of the Lorenz Curve and Gini index from Grouped Data. *Econometrica 44*(3), 479–483.

Gazeley, I., A. Newell, and M. Hawkins (2015). Ministry of Labour & National Service Family Expenditure Household Survey 1953-54. Dataset, University of Sussex, Brighton.

Jenkins, S. P. (1999, January). INEQDECO: Stata module to calculate inequality indices with decomposition by subgroup. Statistical Software Components, Boston College Department of Economics.

Kakwani, N. (1980, March). On a Class of Poverty Measures. *Econometrica 48*(2), 437–446.

Lerman, R. I. and S. Yitzhaki (1989). Improving the Accuracy of Estimates of Gini Coefficients. *Journal of Econometrics 42*(1), 43–47.

Mood, A. M. and F. A. Graybill (1963). *Introduction to the Theory of Statistics* (2 ed.). New York: McGraw-Hill.

Pyatt, G., C.-N. Chen, and J. Fei (1980). The distribution of income by factor components. *The Quarterly Journal of Economics 95*(3), 451–473.

StataCorp (2015). *Stata Statistical Software: Release 14.* College Station, TX: StataCorp LP.

Wooldridge, J. M. (2010). *Econometric Analysis of Panel and Cross Sectional Data* (Second ed.). Cambridge, Mass: MIT Press.