Histogram vs Lorenz

Reza Hajargasht
Swinburne Business School

Bill Griffiths
Melbourne University
Introduction

• For large scale inequality and poverty analyses [involving many countries over many periods] or historical analyses, grouped data is often the only source.

• Grouped data are provided at least in two forms.
## Type of Data

### Histogram Data

<table>
<thead>
<tr>
<th>Population proportions ($c_i$)</th>
<th>Bounds ($z_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0479</td>
<td>10</td>
</tr>
<tr>
<td>0.0509</td>
<td>20</td>
</tr>
<tr>
<td>0.0988</td>
<td>50</td>
</tr>
<tr>
<td>0.1048</td>
<td>75</td>
</tr>
<tr>
<td>0.1018</td>
<td>100</td>
</tr>
<tr>
<td>0.0938</td>
<td>150</td>
</tr>
<tr>
<td>0.0988</td>
<td>200</td>
</tr>
<tr>
<td>0.1018</td>
<td>250</td>
</tr>
<tr>
<td>0.1038</td>
<td>300</td>
</tr>
<tr>
<td>0.0978</td>
<td>400</td>
</tr>
<tr>
<td>0.0499</td>
<td>500</td>
</tr>
<tr>
<td>0.0499</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

### Lorenz Data

<table>
<thead>
<tr>
<th>Population proportions ($c_i$)</th>
<th>Mean Income ($\bar{y}_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0479</td>
<td>17.812</td>
</tr>
<tr>
<td>0.0988</td>
<td>22.657</td>
</tr>
<tr>
<td>0.1976</td>
<td>26.482</td>
</tr>
<tr>
<td>0.3024</td>
<td>30.567</td>
</tr>
<tr>
<td>0.4042</td>
<td>34.612</td>
</tr>
<tr>
<td>0.4980</td>
<td>38.571</td>
</tr>
<tr>
<td>0.5968</td>
<td>42.989</td>
</tr>
<tr>
<td>0.6986</td>
<td>48.497</td>
</tr>
<tr>
<td>0.8024</td>
<td>56.277</td>
</tr>
<tr>
<td>0.9002</td>
<td>69.185</td>
</tr>
<tr>
<td>0.9501</td>
<td>89.266</td>
</tr>
<tr>
<td>1.0000</td>
<td>174.67</td>
</tr>
</tbody>
</table>

### Unit Record Data

- Many Surveys and Censuses
- World Bank POVCAL
Notation

- $T$ observations
- grouped into $N$ income classes
- Bounds: $(z_0, z_1), (z_1, z_2), \ldots, (z_{N-1}, z_N)$ with $z_0 = 0$ & $z_N = \infty$

- $c_1, c_2, \ldots, c_N$ proportion of observations in each class
- $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_N$ mean incomes for each of the $N$ classes
Let $f(y; \theta)$ be the PDF of the underlying parametric income distribution:

Define

$$k_i(\phi) = \int_{z_{i-1}}^{z_i} f(y; \phi) dy \quad i = 1, 2, \ldots, N$$

and

$$\mu_i(\phi, z) = \frac{\int_{z_{i-1}}^{z_i} y f(y; \phi) dy}{\int_{z_{i-1}}^{z_i} f(y; \phi) dy} \quad i = 1, 2, \ldots, N$$
Moment conditions

\[ H_1(\phi) = c - k(\phi;z) = \begin{bmatrix} c_1 - k_1(\phi;z) \\ \vdots \\ c_N - k_N(\phi;z) \end{bmatrix} \]

We use GMM

\[ \min_{\phi} H(\phi)' W H(\phi) \]

optimal weight matrix

\[ W = D(1/k) \]

\[ \text{Var}(\hat{\phi}) = \frac{1}{T} \left\{ \frac{\partial k'}{\partial \phi} D(1/k) \frac{\partial k}{\partial \phi'} \right\}^{-1} \]
Estimation for Lorenz

- Moment conditions
  \[ H(\theta) = \begin{bmatrix}
    c_1 - k_1(\theta) \\
    \vdots \\
    c_N - k_N(\theta) \\
    \bar{y}_1 - \mu_1(\theta) \\
    \vdots \\
    \bar{y}_N - \mu_N(\theta)
  \end{bmatrix} = \begin{bmatrix}
    c - k \\
    \bar{y} - \mu
  \end{bmatrix} \]

- MD Estimation
  \[ \min_\theta H(\theta)'WH(\theta) \]
  optimal weight
  \[ W = \begin{bmatrix}
    D(1/k) & 0 \\
    0 & D(k/\nu)
  \end{bmatrix} \]

\[ \text{var}(\hat{\theta}) = \frac{1}{T} \left( \begin{bmatrix}
    \frac{\partial k'}{\partial \theta} & \frac{\partial \mu'}{\partial \theta} \\
    \frac{\partial k'}{\partial \theta} & \frac{\partial \mu'}{\partial \theta}
  \end{bmatrix} \begin{bmatrix}
    D(1/k) & 0 \\
    0 & D(k/\nu)
  \end{bmatrix} \begin{bmatrix}
    \frac{\partial k}{\partial \theta'} \\
    \frac{\partial \mu}{\partial \theta'}
  \end{bmatrix} \right)^{-1} \]
Estimation for Unit Record

• Maximum Likelihood:

$$\hat{\phi} = \arg \max_\phi \sum_{t=1}^{T} \ln f(y_t; \phi)$$

• Variance:

$$Var(\hat{\phi}_1) = \frac{1}{T} \left\{ \sum_{t=1}^{T} \frac{\partial^2 \ln f(y_t; \phi)}{\partial \phi \partial \phi'} \right\}^{-1}$$
Efficiency

(i) Histogram

\[ \text{Var}(\hat{\phi}) = \frac{1}{T} \left\{ \frac{\partial k'}{\partial \phi} D(1/k) \frac{\partial k}{\partial \phi'} \right\}^{-1} \]

(ii) Lorenz

\[ \text{Var}(\hat{\phi}) = \frac{1}{T} \left\{ \frac{\partial k'}{\partial \phi} D(1/k) \frac{\partial k}{\partial \phi'} + \frac{\partial \mu'}{\partial \phi} D(k/v) \frac{\partial \mu}{\partial \phi'} - \left[ \frac{\partial k'}{\partial \phi} D(1/k) \frac{\partial k}{\partial z'} + \frac{\partial \mu'}{\partial \phi} D(k/v) \frac{\partial \mu}{\partial z'} \right] \right\}^{-1} \]

\[ \left[ \frac{\partial k'}{\partial z} D(1/k) \frac{\partial k}{\partial z'} + \frac{\partial \mu'}{\partial z} D(k/v) \frac{\partial \mu}{\partial z'} \right]^{-1} \left[ \frac{\partial k'}{\partial \phi} D(1/k) \frac{\partial k}{\partial \phi'} + \frac{\partial \mu'}{\partial \phi} D(k/v) \frac{\partial \mu}{\partial \phi'} \right] \]

(iii) Histogram and Lorenz \( \{c_i, z_i, \bar{y}_i\} \)

\[ \text{Var}(\hat{\phi}) = \frac{1}{T} \left\{ \frac{\partial k'}{\partial \phi} D(1/k) \frac{\partial k}{\partial \phi'} + \frac{\partial \mu'}{\partial \phi} D(k/v) \frac{\partial \mu'}{\partial \phi'} \right\}^{-1} \]

(iv) ML

\[ \text{Var}(\hat{\phi}) = \frac{1}{T} \left\{ \sum_{t=1}^{T} \frac{\partial^2 \ln f(y; \phi)}{\partial \phi \partial \phi'} \right\}^{-1} \]
## Efficiency

### Histogram vs Lorenz vs Unit Record Data

<table>
<thead>
<tr>
<th></th>
<th>PLN with 5 Group</th>
<th>PLN with 20 Group</th>
<th>PLN with 100 Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min</td>
<td>median</td>
<td>max</td>
</tr>
<tr>
<td>m</td>
<td>0.0627</td>
<td>0.0832</td>
<td>0.9411</td>
</tr>
<tr>
<td>s</td>
<td>0.0979</td>
<td>0.1217</td>
<td>0.4726</td>
</tr>
<tr>
<td>a</td>
<td>0.0535</td>
<td>0.0628</td>
<td>0.9453</td>
</tr>
<tr>
<td>Gini</td>
<td>0.2291</td>
<td>0.4178</td>
<td>0.5920</td>
</tr>
<tr>
<td>G</td>
<td>0.0536</td>
<td>0.0769</td>
<td>1.0844</td>
</tr>
<tr>
<td>H</td>
<td>0.0563</td>
<td>0.6676</td>
<td>0.9923</td>
</tr>
<tr>
<td>Tail</td>
<td>2.0052</td>
<td>3.0397</td>
<td>3.9951</td>
</tr>
<tr>
<td></td>
<td>0.0535</td>
<td>0.0628</td>
<td>0.9453</td>
</tr>
</tbody>
</table>

\[
Var(\theta_{Lorenz})/Var(\theta_{Hist})
\]

For a Pareto-Lognormal Distn
## Efficiency

<table>
<thead>
<tr>
<th></th>
<th>PLN with 5 Group</th>
<th></th>
<th>PLN with 20 Group</th>
<th></th>
<th>PLN with 100 Group</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min</td>
<td>median</td>
<td>max</td>
<td>min</td>
<td>median</td>
<td>max</td>
</tr>
<tr>
<td>m</td>
<td>0.7183</td>
<td>0.9717</td>
<td>0.9957</td>
<td>0.9069</td>
<td>0.9957</td>
<td>0.9994</td>
</tr>
<tr>
<td>s</td>
<td>0.7792</td>
<td>0.9551</td>
<td>0.9835</td>
<td>0.9360</td>
<td>0.9945</td>
<td>0.9983</td>
</tr>
<tr>
<td>a</td>
<td>0.6943</td>
<td>0.9730</td>
<td>0.9953</td>
<td>0.8929</td>
<td>0.9951</td>
<td>0.9992</td>
</tr>
<tr>
<td>Gini</td>
<td>0.6834</td>
<td>0.9963</td>
<td>1.0000</td>
<td>0.8913</td>
<td>0.9948</td>
<td>1.0000</td>
</tr>
<tr>
<td>H</td>
<td>0.7624</td>
<td>0.9888</td>
<td>1.0000</td>
<td>0.9630</td>
<td>0.9988</td>
<td>1.0000</td>
</tr>
<tr>
<td>Tail</td>
<td>0.6943</td>
<td>0.9730</td>
<td>0.9953</td>
<td>0.8929</td>
<td>0.9951</td>
<td>0.9992</td>
</tr>
</tbody>
</table>

\[ \text{Var}(\theta_{Hist-Lorenz})/\text{Var}(\theta_{Lorenz}) \] For a Pareto-Lognormal Distn
(v) Data on \( \{z_i, \bar{y}_i\} \)

\[
H(\phi; z) = \begin{bmatrix}
\bar{y}_1 - \mu_1(\phi; z) \\
\vdots \\
\bar{y}_N - \mu_N(\phi; z)
\end{bmatrix} = \bar{y} - \mu(\phi; z)
\]

- MD Estimation

\[
Min_\phi H(\phi; z)' WH(\phi; z)
\]

- But this way of data collection is not a good idea; the estimation becomes worse as the number of groups (moments) increases.
GMM & Grouped Data

Grouped Data can provide good motivating examples for almost all the developments in GMM and MD:

- How to formulate an MoM estimation
- How to derive an optimal weight matrix
- Two-step, repeated two step, CUE
- Singular weight matrix (happens with heavy tails)
- Inference when the parameter is on the boundary
- Various forms of weak identification
- GMM with many, infinite and continuum of moments
- Testing inequalities (e.g. for stochastic dominance)
- GMM with nonparametric components
- Robust GMM
- Bayesian GMM

Histogram vs Lorenz vs Unit Record Data

Inequality Conference, June 2019
Robustness

- Efficiency is not the only criteria, another criteria (sometimes even more important) is robustness. One approach to robustness is through using Influence Function.

- Suppose $F(y)$ is the CDF; let’s introduce a perturbation to $F$ and write the mis-specified distribution as

$$F_{\varepsilon} = (1 - \varepsilon)F + \varepsilon\delta_y$$

- Any estimator $\hat{\theta}$ can be thought of as a functional $\hat{\theta}(F)$ and

$$IF(\hat{\theta}; y) = \lim_{\varepsilon \to 0} \frac{\hat{\theta}(F_{\varepsilon}(y)) - \hat{\theta}(F(y))}{\varepsilon}$$
Influence Function

- For a GMM estimator
  \[ \text{Min}_\varphi \ M(y; \varphi)'WM(y; \varphi) \]

- It has been shown that
  \[ IF(\varphi, F; y) = E \left\{ \frac{\partial M'}{\partial \varphi} W \frac{\partial M}{\partial \varphi'} \right\}^{-1} \frac{\partial M'}{\partial \varphi} WM(y; \varphi) \]

- Histogram:
  \[ IF_{Hist}(\varphi; y) = \left( \frac{\partial k'}{\partial \varphi} D(1/k) \frac{\partial k}{\partial \varphi'} \right)^{-1} \frac{\partial k'}{\partial \varphi} D(1/k) [g(z_{-1} < y < z)] \]

- Lorenz:
  \[ IF_{Lor}(\varphi; y) = (V)^{-1} \begin{bmatrix} \frac{\partial k'}{\partial \varphi} & \frac{\partial \mu'}{\partial \varphi} \end{bmatrix} \begin{bmatrix} 1/k & 0 \\ 0 & k/\mu \end{bmatrix} \begin{bmatrix} g(z_{-1} < y < z) - k \\ yg(z_{-1} < y < z) - \mu \end{bmatrix} \]

- ML:
  \[ IF_{ML}(\varphi; y) = - \left\{ E \left( \frac{\partial^2 \ln f(y; \varphi)}{\partial \varphi \partial \varphi'} \right) \right\}^{-1} \frac{\partial \ln f(y; \varphi)}{\partial \varphi} \]
Figure-1: Influence Function for Mean

\[ f(y) = PL(y;[m,s,a]) = PL(y;[4,0.5,4]) \]

\[ \text{Mean} = 82.49 \quad \text{Gini} = 0.3146 \]
Figure 2- Influence Function of Gini
Bias in Mean

Figure 4 - Error in Mean with Number of Groups = 5

\[ f(y) = PL(y; [4, 0.5, 4]) \]

Mean = 82.49  Gini = 0.3146

0.5% of data is multiplied by 10
Bias in Gini

Figure 3- Error in Gini with Number of Groups = 5

Bias_G_Lor  Bias_G_Hist  Bias_G_ML
Bias in Mean

Figure 7 - Error in Mean with Number of Groups = 100

Bias_M_Lor  Bias_M_Hist  Bias_M_ML
Bias in Gini

Figure 5- Error in Gini with Number of Groups=100

- Bias_G_Lor
- Bias_G_Hist
- Bias_G_ML
Robust Estimation of Lorenz Curves

• The above results indicate that standard GMM for Lorenz Data is more efficient but less robust.

• Is it possible to find another estimator for Lorenz case that is more efficient but as robust?

• Ronchetti and Trojani [J.o.E (2001)] have shown how to estimate GMM in a robust way. Their method can be used to robustly estimate Lorenz curves

• The implementation is a bit more complicated and we haven’t done it yet.
Can we do nonparametric inference with grouped data?
Sample statistics from each group can be considered as a nonparametric estimate of the population equivalent.
Inference for these estimates often needs extra information.
It is also not clear how to estimate PDF or Lorenz for every point.
One thing one can do is to use (linear) interpolation to estimate a Histogram or a Lorenz curve. One can also calculate Gini coefficient and other quantities of interest based on these interpolations.
It is also possible to construct lower and upper bound for Histogram and Lorenz and compare the bounds.
Conclusion

• Our results indicate that standard estimation based on Lorenz Data is more efficient but less robust than Histogram Data
• Robustness might be a more important criteria than Efficiency in the context of income distribution

To do list:
• Compare robust Lorenz curve estimation with Histogram case
• Extending the analysis to nonparametrics in a formal manner if it is possible