Is inequality underestimated in Egypt? Evidence from house prices

Roy van der Weide

Joint with Christoph Lakner and Elena Ianchovichina

ABCDE Conference, Washington DC
June, 2016
Objective, challenge, and solution

The objective is to estimate income (or consumption) inequality with an application to (urban) Egypt

- Income inequality is conventionally estimated using household income survey data
- Top incomes tend to be under-covered in household surveys
- Popular solution is to identify top incomes using tax data
- Tax data ideal for this purpose BUT often not available
- The solution proposed here is to estimate top incomes using a database of predictors (such as house prices)
- The idea being that data on predictors of top incomes are more readily available
- This approach introduces new challenges which we will address...
Combining top- and bottom income distributions

\[ F(y) = \begin{cases} 
(1 - \lambda)F_1(y) & y \leq \tau \\
(1 - \lambda) + \lambda F_2(y) & y > \tau
\end{cases} \]

Total inequality can be estimated using \( F \) or using a sub-group decomposition:

\[ T = \sum_k S_k T_k + \sum_k S_k \log \left( \frac{S_k}{P_k} \right), \]

where \( P_k, S_k \) and \( T_k \) denote sub-group \( k \)'s population share, income share and Theil index, respectively. (Similar decompositions can be obtained for other inequality measures, such as the MLD and Gini.)
At least two challenges when estimating top distribution using predictors of top incomes

- Challenge 1: We need to carefully account for the fact that we observe predictors of top incomes, not actual top incomes
- Challenge 2: The database with predictors of top incomes will often not constitute a representative sample
Working with a model for top-incomes using a database of predictors (such as house prices)

Assume a model for (top) household income $Y_h$ given data on predictor $x_h$:

$$
\log (Y_h) = m(x_h; \beta) + \varepsilon_h \\
= a + b \log (x_h) + \varepsilon_h, \quad \text{say}
$$

- Estimate top tail of distribution for $x_h$ denoted by $G_2$, for example by fitting a Pareto distribution with tail parameter $\alpha$
- An estimate of $F_2$ can then be obtained from the estimate of $G_2$ combined with an estimate of the model that links $x_h$ to $Y_h$
- For our log-linear model, if $G_2$ is Pareto with parameter $\alpha$, then $F_2$ is Pareto with parameter $\alpha/b$
The database with predictors of top incomes will often not constitute a representative sample

By definition:

\[ p_{2,d} = Pr[Y > \tau, \text{district } d] \]
\[ = Pr[Y > \tau | \text{district } d] Pr[\text{district } d] \]
\[ = \lambda_d \pi_d \]

Assumption

The target population can be divided into D districts, with \( d = 1, \ldots, D \) indicating the district. It is assumed that:

- The share of the total population residing in each of the districts is known, which will be denoted by \( \{\pi_d\} \).
- For any district \( d \), the database permits a consistent estimator for \( F_{2,d}(y) = Pr[Y \leq y | Y > \tau, \text{district } d] \)
The database with predictors of top incomes will often not constitute a representative sample

Assumption

Let $f_d(y)$ denote the probability density function corresponding to $F_d(y)$. It is assumed that $f_d(y)$ is a continuous function of $y$.

Proposition

Let $\hat{f}_{k,d}(y)$ denote the consistent estimator for $f_{k,d}(y)$ for $k = 1, 2$. Under these assumptions, $\hat{\lambda}_d$ provides a consistent estimator for $\lambda_d$ (for any given district):

$$\hat{\lambda}_d = \frac{\hat{f}_{1,d}(\tau)}{\hat{f}_{1,d}(\tau) + \hat{f}_{2,d}(\tau)}$$
## Number of observations by District and Database

<table>
<thead>
<tr>
<th>Sub-group</th>
<th>Betak-online</th>
<th>Bezaat</th>
<th>HIECS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cairo</td>
<td>5772</td>
<td>8475</td>
<td>1289</td>
</tr>
<tr>
<td>Alexandria</td>
<td>1293</td>
<td>2012</td>
<td>767</td>
</tr>
<tr>
<td>Urban Egypt</td>
<td>6935</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Number of observations used
Pareto tail index: household expenditure (per cap)
Pareto tail index: house prices
Regression model: log household expenditure (per cap) vs. log rent
Regression model: estimates of $b$
Estimates of $b$, $\alpha_{rent}$, and $\alpha_{income}$

<table>
<thead>
<tr>
<th>Sub-group</th>
<th>$\hat{b}$</th>
<th>$\hat{\alpha}_{house}$</th>
<th>$\hat{\alpha}_{house}/\hat{b}$</th>
<th>$\hat{\alpha}_{svy}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cairo</td>
<td>0.66</td>
<td>1.13</td>
<td>1.71</td>
<td>2.22</td>
</tr>
<tr>
<td>Alexandria</td>
<td>0.51</td>
<td>1.14</td>
<td>2.27</td>
<td>2.96</td>
</tr>
</tbody>
</table>

**Table**: Estimates of $b$, $\alpha_{mix}$, and $\alpha_{svy}$
Estimates of $Pr[Y > \tau | sub\text{ }-\text{ }group]$ and pop. share (top income & sub-group)

<table>
<thead>
<tr>
<th>Sub-group</th>
<th>$\pi$</th>
<th>$\hat{\lambda}_{prop}$</th>
<th>$\hat{\lambda}_{svy}$</th>
<th>$\hat{p}_{prop}$</th>
<th>$\hat{p}_{svy}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cairo</td>
<td>0.251</td>
<td>0.116</td>
<td>0.101</td>
<td>0.029</td>
<td>0.025</td>
</tr>
<tr>
<td>Alexandria</td>
<td>0.130</td>
<td>0.079</td>
<td>0.048</td>
<td>0.010</td>
<td>0.006</td>
</tr>
<tr>
<td>Other Urban</td>
<td>0.619</td>
<td>0.028</td>
<td>0.028</td>
<td>0.017</td>
<td>0.017</td>
</tr>
</tbody>
</table>

Table: Estimates of $Pr[Y > \tau | sub\text{ }-\text{ }group]$ and pop. share (top income & sub-group)
### Estimates of top income shares and top inequality levels

<table>
<thead>
<tr>
<th>Sub-group</th>
<th>$\hat{s}_{svy}$</th>
<th>$\hat{s}_{mix}$</th>
<th>$MLD_{svy}$</th>
<th>$MLD_{mix}$</th>
<th>$Theil_{svy}$</th>
<th>$Theil_{mix}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cairo</td>
<td>0.118</td>
<td>0.159</td>
<td>0.087</td>
<td>0.295</td>
<td>0.107</td>
<td>0.532</td>
</tr>
<tr>
<td>Alexandria</td>
<td>0.036</td>
<td>0.042</td>
<td>0.038</td>
<td>0.141</td>
<td>0.041</td>
<td>0.208</td>
</tr>
<tr>
<td>Other Urban</td>
<td>0.066</td>
<td>0.063</td>
<td>0.082</td>
<td>0.082</td>
<td>0.097</td>
<td>0.097</td>
</tr>
</tbody>
</table>

**Table:** Estimates of top income shares and top inequality levels
### Table: Estimates of total inequality for urban Egypt

<table>
<thead>
<tr>
<th>Database</th>
<th>Tot. Inequality (urb. Egypt)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gini</td>
</tr>
<tr>
<td>HIECS</td>
<td>0.364</td>
</tr>
<tr>
<td>HIECS + House prices</td>
<td>0.470</td>
</tr>
</tbody>
</table>

Estimates of total inequality for urban Egypt
Concluding remarks

- Our evidence suggests that inequality is being under-estimated in Egypt
- First assessment suggests that estimates of the Pareto parameter $\alpha$ are somewhat sensitive to the choice of house rental value threshold
- The same applies to estimates of $b$ (the model that links household income and house rent value), but to a lesser extent
- It would take a very conservative estimate of $\alpha$ and $b$ to reproduce the survey-only estimates of inequality
- We wish to replicate the exact same analysis for a country for which we also have tax record data (such as the U.S.)