# 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 challanges 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 $x_h$:

$$\begin{aligned} \log(Y_h) &= m(x_h;\beta) + \varepsilon_h \\ &= a + b\log(x_h) + \varepsilon_h, \quad \textit{say} \end{aligned}$$

- 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:

$$\begin{aligned} p_{2,d} &= Pr[Y > \tau, \textit{ district } d] \\ &= Pr[Y > \tau | \textit{district } d]\, Pr[\textit{district } d] \\ &= \lambda_d \pi_d \end{aligned}$$

## 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 $F_{2,d}(y) = Pr[Y \leq y | Y > \tau, \textit{ district } d]$*

# 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):*

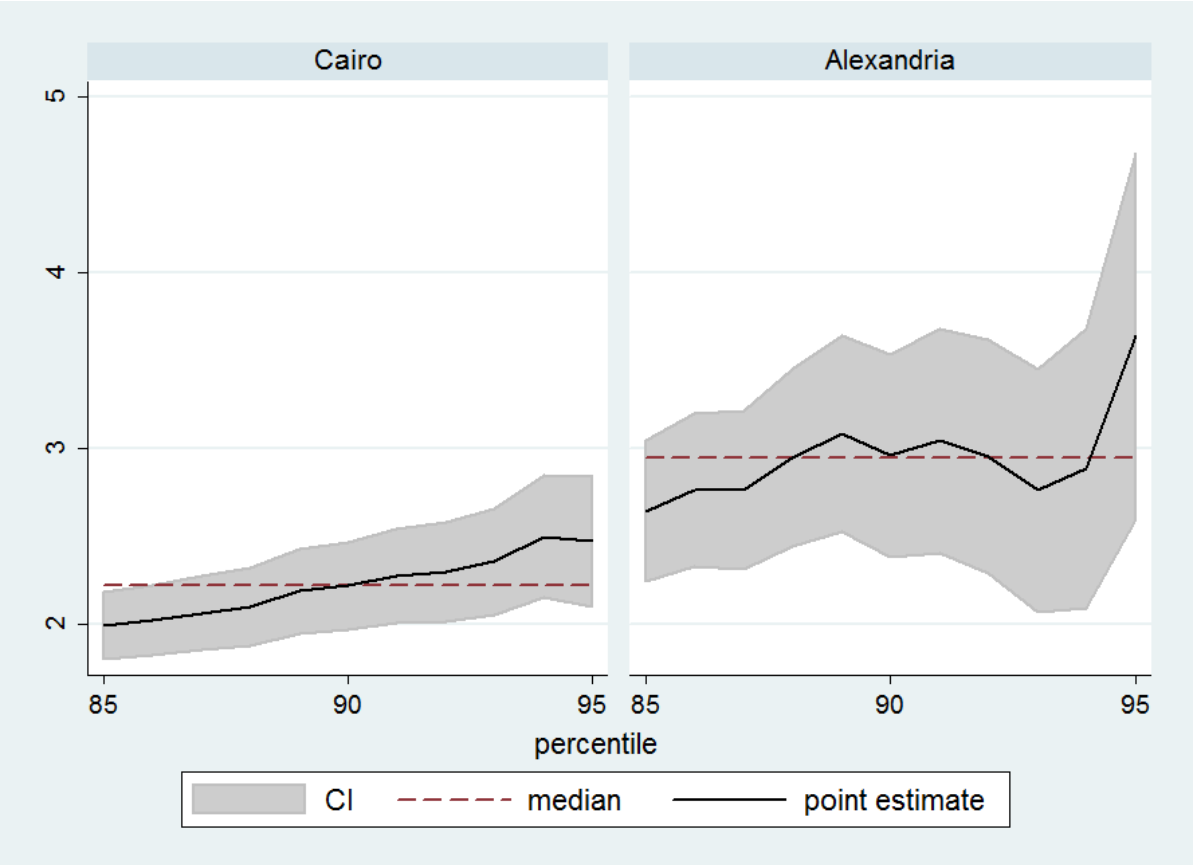
$$\hat{\lambda}_d = \frac{\hat{f}_{1,d}(\tau)}{\hat{f}_{1,d}(\tau) + \hat{f}_{2,d}(\tau)}$$

# Number of observations by District and Database

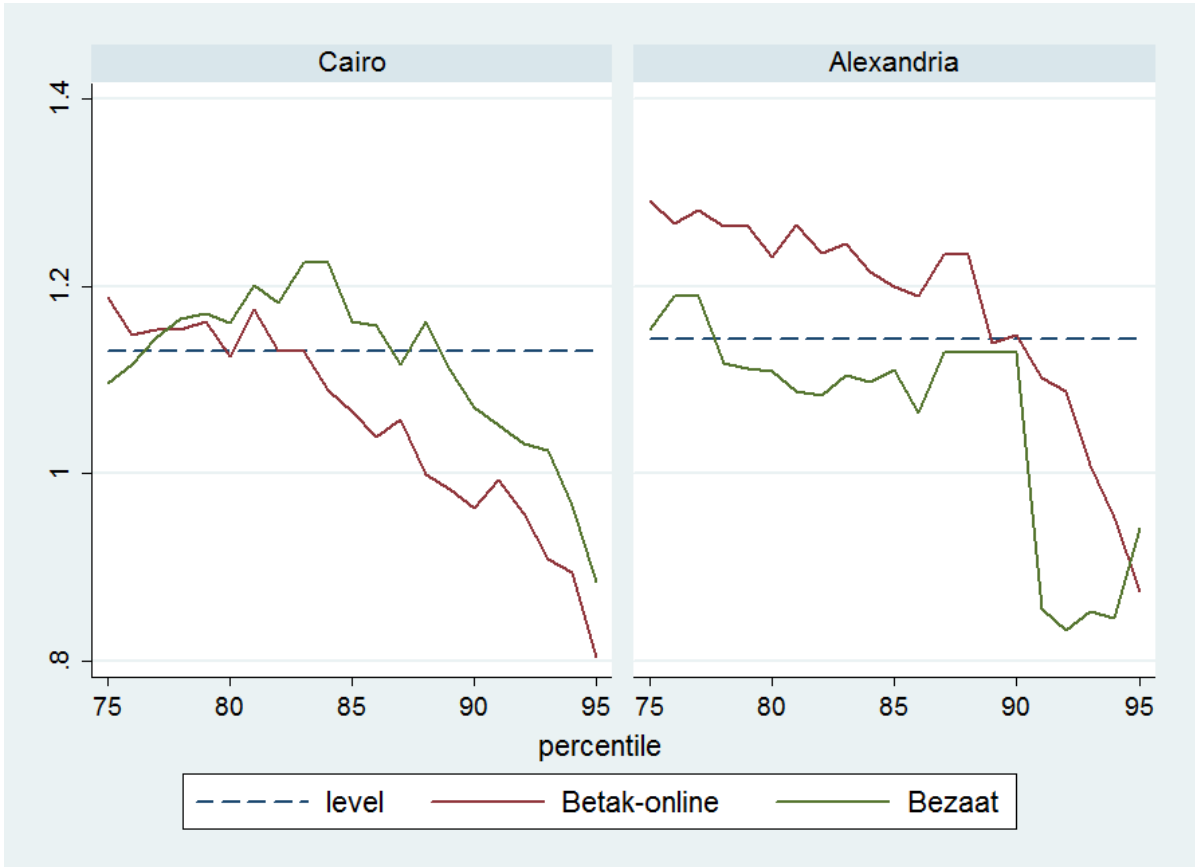
| | Database | | |
|---|---|---|---|
| sub-group | Betak-online | Bezaat | HIECS |
| Cairo | 5772 | 8475 | 1289 |
| Alexandria | 1293 | 2012 | 767 |
| Urban Egypt | | | 6935 |

Table: Number of observations used

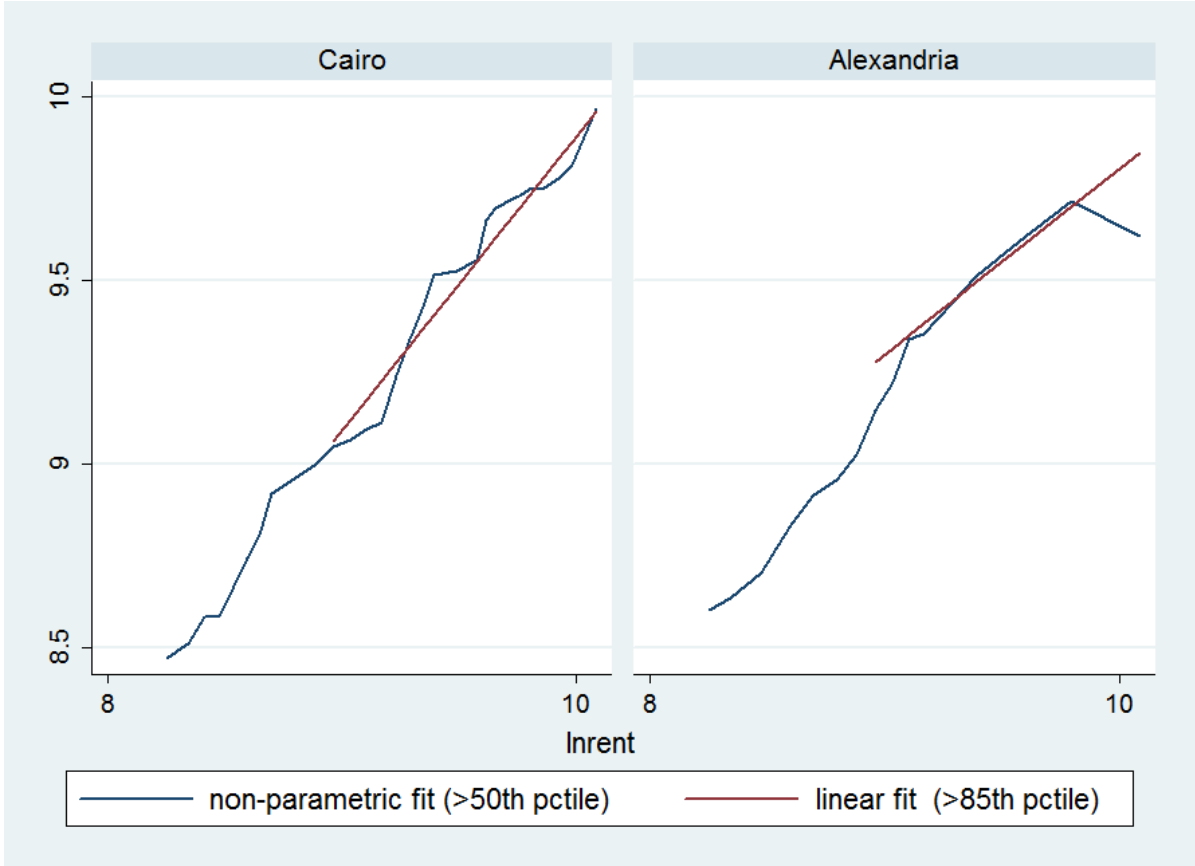
# Pareto tail index: household expenditure (per cap)



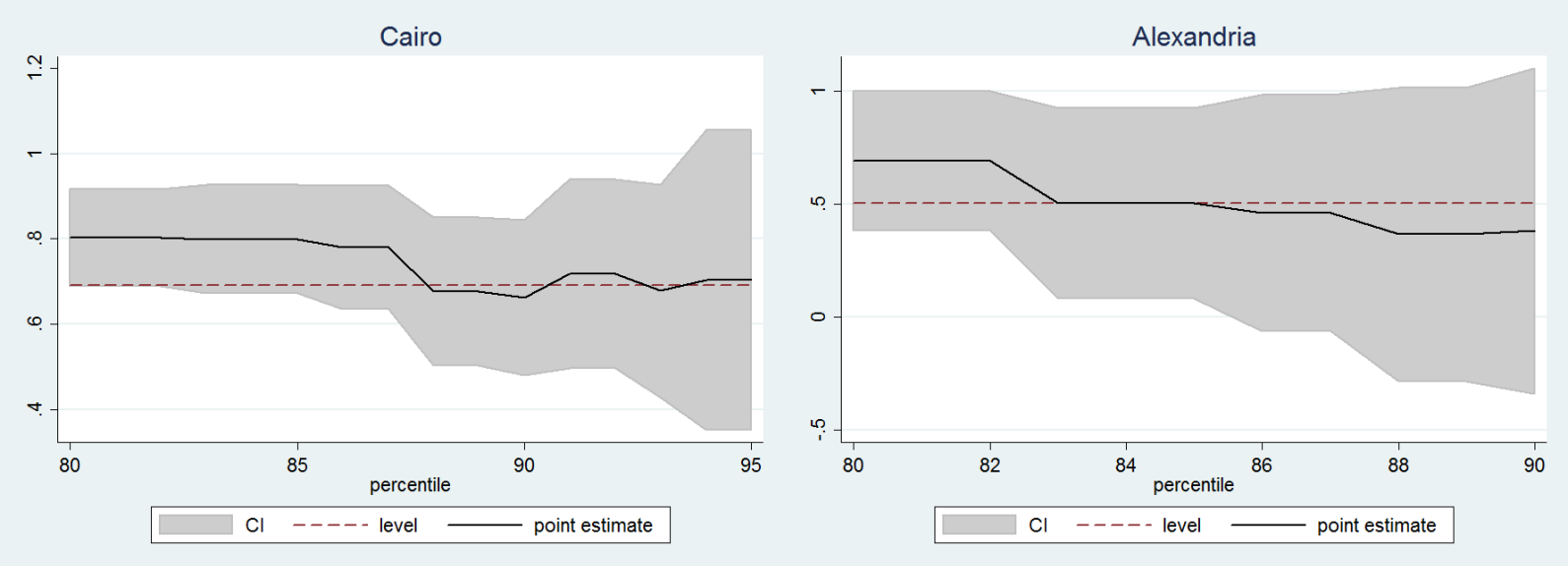
# Pareto tail index: house prices



# Regression model: log household expenditure (per cap) vs. log rent



# Regression model: estimates of $b$



# Estimates of $b$, $\alpha_{rent}$, and $\alpha_{income}$

| Sub-group | $\hat{b}$ | $\hat{\alpha}_{house}$ | $\hat{\alpha}_{house}/\hat{b}$ | $\hat{\alpha}_{svy}$ |
|---|---|---|---|---|
| Cairo | 0.66 | 1.13 | 1.71 | 2.22 |
| Alexandria | 0.51 | 1.14 | 2.27 | 2.96 |

Table: Estimates of $b$, $\alpha_{mix}$, and $\alpha_{svy}$

# Estimates of $Pr[Y > \tau | sub - group]$ and pop. share (top income & sub-group)

| Sub-group | $\pi$ | $\hat{\lambda}_{prop}$ | $\hat{\lambda}_{svy}$ | $\hat{p}_{prop}$ | $\hat{p}_{svy}$ |
|---|---|---|---|---|---|
| Cairo | 0.251 | 0.116 | 0.101 | 0.029 | 0.025 |
| Alexandria | 0.130 | 0.079 | 0.048 | 0.010 | 0.006 |
| Other Urban | 0.619 | 0.028 | 0.028 | 0.017 | 0.017 |

Table: Estimates of $Pr[Y > \tau | sub - group]$ and pop. share (top income & sub-group)

# Estimates of top income shares and top inequality levels

| | Top income shares and top inequality | | | | | |
|---|---|---|---|---|---|---|
| Sub-group | $\hat{s}_{svy}$ | $\hat{s}_{mix}$ | $MLD_{svy}$ | $MLD_{mix}$ | $Theil_{svy}$ | $Theil_{mix}$ |
| Cairo | 0.118 | 0.159 | 0.087 | 0.295 | 0.107 | 0.532 |
| Alexandria | 0.036 | 0.042 | 0.038 | 0.141 | 0.041 | 0.208 |
| Other Urban | 0.066 | 0.063 | 0.082 | 0.082 | 0.097 | 0.097 |

Table: Estimates of top income shares and top inequality levels

# Estimates of total inequality for urban Egypt

| | Tot. Inequality (urb. Egypt) | | |
|---|---|---|---|
| Database | Gini | MLD | Theil |
| HIECS | 0.364 | 0.217 | 0.258 |
| HIECS+House prices | 0.470 | 0.278 | 0.420 |

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.)