Nowcasting the household income distribution

Fabrice Murtin
OECD Statistics and Data Directorate

February 4th 2020
UNCTAD Technical Workshop
Background and motivations

• Since 1998, regular OECD data collection on income distribution and poverty (OECD IDD) based on national sources and comparable definitions

• Strong internal and external demand for IDD (e.g. COPE & its reports, *How’s Life?*, Inclusive Growth, Economic Surveys, G20, etc.)

• However, despite annual collection, information is not timely: average lag is 2-3 years...

• This limits the possibility to use distributive information in macro-economic assessments where timeliness is key requirement (e.g. *Economic Outlook, Going for Growth*)
Background and motivations

- Project seeks to *nowcast* household income by decile (in year $T$) in as many OECD countries as possible based on contemporaneous information through reduced-form econometrics.

- Once methodology has been thoroughly tested, estimates could be released regularly by the OECD in various forms (NAD household dashboard, MDLS, *How’s Life?*, ad hoc statistical briefs, G20 documents).

- In the very short term: working paper and feedback from experts in a variety of fora.
Nowcasting: basic principles

- **Objective**: Construct a predictive model that can be evaluated by out-of-sample (OOS) performance.

- **Parsimony**: A complicated model increases in-sample fit (R2) but may decrease OOS.

- **Credibility**: Meaningful coefficients.

- **Specificity**: The model must be decile-specific and possibly country-specific.
The dependent variable

- Average equivalised household disposable income per decile from IDD
- We consider two income series per country (waves 6 and 7)
- Linear interpolations used to cover gaps up to 3 years for countries lacking annual surveys and for earlier (pre-2000s) periods
- We also considered and tested a model to predict individual income’s components for each decile (i.e. wage, self-employment income, capital income, transfers received, taxes paid) but model performed less well than model for total income
Explanatory variables

- We created a group of 30+ predictors, drawn from national accounts and other official sources, that are timely and available for most countries.

- Examples: GDP, unemployment rate, mean net household disposable income (SNA), self-employment rate, wage rate, hours worked per worker, long-term interest rates, house prices, property income, share prices, current transfer received by households, taxes on business and on different kinds of households etc...
The predictive models

- We consider several predictive algorithms routinely used in AI:
  
  Random forest, Gradient boosting, Neural network, SVM

- We compare the results with those obtained from a log-linear model with variable selection (LASSO)

- LASSO model: For each decile we predict the growth rate of real household disposable income (defl. PCD):

  \[
  \Delta_{t,t-1} \log y = \Delta_{t,t-1} X \beta_1 + \Delta_{t-k,t-k-1} \log y \beta_2 + \varepsilon
  \]

- Performance: we evaluated 1 year-ahead out-of-sample performance against observed growth rates and a naive ‘random walk’ model (forecasted growth=last observed growth)
<table>
<thead>
<tr>
<th>Model</th>
<th>Decile 1</th>
<th>Deciles 2 to 9</th>
<th>Decile 10</th>
<th>All deciles</th>
</tr>
</thead>
<tbody>
<tr>
<td>LASSO</td>
<td>0.59</td>
<td>0.79</td>
<td>0.17</td>
<td>0.60</td>
</tr>
<tr>
<td>ANN</td>
<td>-0.12</td>
<td>0.19</td>
<td>0.59</td>
<td>0.09</td>
</tr>
<tr>
<td>SVM</td>
<td>-0.04</td>
<td>-0.05</td>
<td>0.30</td>
<td>0.00</td>
</tr>
<tr>
<td>DRF</td>
<td>-0.39</td>
<td>0.00</td>
<td>0.17</td>
<td>-0.10</td>
</tr>
<tr>
<td>GBM</td>
<td>0.29</td>
<td>0.34</td>
<td>-0.21</td>
<td>0.25</td>
</tr>
</tbody>
</table>
• All coefficients have the same sign across all income deciles (and all variables are ‘correctly’ signed)

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>D5</th>
<th>D6</th>
<th>D7</th>
<th>D8</th>
<th>D9</th>
<th>D10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Control for lags</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Change in/growth of:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GDP</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Wage rate</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Average tax wedge</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Disbursement of government</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Current transfers</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Net current household receipts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-employment rate</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share price</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disposable income</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Estimated model

- Average OOS correlation in 2014 (across all deciles in 13 countries) is 0.59
- Fails to capture tails (and negative growth)
Out-of-sample performance (growth rates)

The estimated model easily beats the random walk. But D1 and D10 are the hardest to predict.
Out-of-sample performance (growth rates)

The chart compares the mean prediction error, growth rate, and random walk error for various countries. The x-axis represents the countries (Poland, United Kingdom, Spain, Finland, Portugal, Belgium, Norway, Czechia, Canada, Slovenia, Ireland, Italy, Austria), and the y-axis represents the mean values. The chart visually distinguishes between prediction error, growth rate, and random walk error for each country.
Conclusions

• Nowcasting real changes in household income for various deciles is difficult because:
  – A complicated model is unstable
  – A simplistic model is inaccurate

• More research is needed to:
  – Better model the tails of the distribution
  – Better capture ‘regime changes’ (large deviations)
  – Better account for country heterogeneity

• On country heterogeneity:
  – other methodologies (microsimulations) may outperform regression-models but are difficult to implement in a consistent way and are much more demanding in terms of information
  – predicting the distribution from NA totals
Thank you!