Small area consumption estimates combining survey and financial footprints data

Peter Levell [1] Lars Nesheim [2] Gautam Vyas [1]

[1]Institute for Fiscal Studies, ESCoE

[2]University College London

LCS Workshop

Motivation

Growing demand among policy-makers for sub-national statistics on living standards

(Preaching to converted!) Consumption is v. important measure of material well-being

But HH surveys typically do not have the sample sizes to compute reliable measures of local average consumption spending

For example: The **Living Costs and Food Survey (LCFS)** in UK:

- Sample of 5000 households records expenditures on all goods (including durables)

But Great Britain has around 350 local authorities

- $\Longrightarrow$ Some local authorities will have v. low (or zero) sample sizes
- Estimates will be very noisy

Motivation (II)

Growing availability of bank account data **millions** of HHs
Not a substitute for surveys

- Provides a measure of outgoings but does not correspond to natl acc. defs of consumption
- Cannot equivalise for household characteristics
- Cannot impute housing or other service flows

But a complement...

Also : labour market surveys contain useful information on households (tenure status, employment and so on) + larger sample sizes

How can we combine this information to get reliable estimates of local consumption spending?

This paper

1. Shows how to combine naturally occurring bank account data with budget survey data + labour survey to produce small-area consumption estimates
2. Constructs standard errors and evaluates improvement from adding data sources (using simulations)
3. Produces new estimates of local consumption spending for local authorities (municipalities) in GB
   - Compares these with other measures (based on income data)

Approach: intuition

We use small area consumption estimation methods (drawing in particular on Molina and Rao, 2010)

- Intuition: weighted average of regression prediction (using other sources) and raw mean from HH survey
- When N small, weights regression prediction more heavily

Direct estimator

- Goal: estimate the mean consumption for areas $a = 1, \ldots, A$
- Population mean:

$$\bar{c}_a = \mathbb{E}[c_{i,a}]$$

- Only a sample s of size $n_a \leq N_a$ is observed - index by I_s
- The direct (survey-only) estimator:

$$\hat{\bar{c}}_a^{Direct} = \frac{1}{n_a} \sum_{i \in I_s} c_{i,a}$$

- Unbiased under random sampling
- Problems:
  - High variance when n_a is small
  - Undefined when $n_a = 0$

Model-based estimator

Let $\mathbf{c}_a = \left(\mathbf{c}'_{a,r}, \mathbf{c}'_{a,s}\right)$

- $\mathbf{c}'_{a,s}$: observed in the survey
- $\mathbf{c}'_{a,r}$: unobserved
- But observe covariates $\mathbf{X}_{a,r}$ observed (in population census or large survey) and $\mathbf{Z}_a$ (area level info, (e.g. average spending from bank account data, local average energy consumption)
- Can compute $\mathbb{E}[\mathbf{c}_{a,r}|\mathbf{X}_{a,r}, \mathbf{Z}_a, \mathbf{c}_{a,s}]$ using model
- *Empirical Best* estimate (shrinkage estimator)

$$\hat{\bar{c}}_a^{EB} = \hat{\bar{c}}_a^{Direct} + \frac{1}{N_a - n_a} \sum_{i \in I_r} \mathbb{E}\left[c_{i,a}|\mathbf{c}_{a,s}, \mathbf{X}_i, \mathbf{Z}_a\right]$$

- 'best' because minimises MSE

Nested Error Linear Regression Model

Random effects model for log consumption

$$\begin{aligned}
\log \mathbf{c}_{a,r} &= \mathbf{X}_{a,r}\boldsymbol{\beta} + \mathbf{Z}_a\boldsymbol{\pi} + u_a\mathbf{1}_{N_a-n_a} + \boldsymbol{\varepsilon}_{a,r} \\
\log \mathbf{c}_{a,s} &= \mathbf{X}_{a,s}\boldsymbol{\beta} + \mathbf{Z}_a\boldsymbol{\pi} + u_a\mathbf{1}_{n_a} + \boldsymbol{\varepsilon}_{a,s} \\
u_a &\sim_{iid} N(0, \sigma_u^2) \\
\varepsilon_{i,a} &\sim_{iid} N(0, \sigma_\varepsilon^2)
\end{aligned}$$

Integrating over the two error terms, u_a and $\boldsymbol{\varepsilon}_a$ yields

$$\begin{pmatrix} \log \mathbf{c}_{a,r} \\ \log \mathbf{c}_{a,s} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{X}_{a,r}\boldsymbol{\beta} + \mathbf{Z}_a\boldsymbol{\pi} \\ \mathbf{X}_{a,s}\boldsymbol{\beta} + \mathbf{Z}_a\boldsymbol{\pi} \end{pmatrix}, \begin{pmatrix} \sigma_u^2\mathbf{I}_{N_a-n_a} + \sigma_\varepsilon^2\mathbf{I}_{N_a-n_a} & \sigma_u^2\mathbf{1}_{N_a-n_a}\mathbf{1}'_{n_a} \\ \sigma_u^2\mathbf{1}_{n_a}\mathbf{1}'_{N_a-n_a} & \sigma_u^2\mathbf{1}_{n_a}\mathbf{1}'_{n_a} + \sigma_\varepsilon^2\mathbf{I}_{n_a} \end{pmatrix} \right]$$

Conditional distribution of log consumption

This yields conditional distribution of $\log \mathbf{c}_{a,r} | \mathbf{X_a}, \mathbf{Z_a}, \log \mathbf{c}_{a,s}$

Mean

$$\mu_{a,r|s} = \mathbf{X}_{a,r}\boldsymbol{\beta} + \mathbf{Z}_a\boldsymbol{\pi} + \sigma_u^2 \mathbf{1}_{N_a - n_a} \mathbf{1}'_{n_a} \left(\sigma_u^2 \mathbf{1}_{n_a} \mathbf{1}'_{n_a} + \sigma_\varepsilon^2 \mathbf{1}_{n_a} \right)^{-1} \left(\log \mathbf{c}_{a,s} - \mathbf{X}_{a,s}\boldsymbol{\beta} - \mathbf{Z}_a\boldsymbol{\pi} \right).$$

Variance

$$\mathbf{V}_{a,r|s} = \sigma_u^2 \left(1 - \gamma_a\right) \mathbf{1}_{N_a - n_a} \mathbf{1}'_{N_a - n_a} + \sigma_\varepsilon^2 \mathbf{I}_{N_a - n_a}$$

where $\gamma_a = \sigma_u^2 \left(\sigma_u^2 + \sigma_\epsilon^2 / n_a \right)^{-1}$

Empirical Best (EB) Estimator

$exp(\mathbb{E}[\log \mathbf{c}_{a,r}|\mathbf{X_a}, \mathbf{Z_a}, \log \mathbf{c}_{a,s}]) \neq \mathbb{E}[\mathbf{c}_{a,r}|\mathbf{X_a}, \mathbf{Z_a}, \log \mathbf{c}_{a,s}]$

So we use Monte Carlo draws of log consumption (with residuals)

- For each draw, exponentiate to obtain $\hat{c}_{i,a}$
- Area mean:

$$\hat{\bar{C}}_a^{EB} = \frac{1}{N_a} \sum_{i=1}^{N_a} \frac{\hat{c}_{i,a}}{\Omega(X_i)}$$

- $\Omega(X_i)$ is an equivalence scale (we use OECD)
- EB minimises MSE under the assumed model
- If $n_a = 0$: reduces to a *synthetic estimator*

Data

- **Living Costs Food Survey:** Pool samples for calendar years 2018-2019. Use geocoded version.
- **Annual Population Survey:** Labour market survey. N = 387,000 for two years (2018-2019).
- **FINDS data:** Natwest bank account data (large retail bank). Use outgoings from current account. Available below LA level.
- **Domestic Electricity and gas consumption**: Average KWh consumption per domestic user.

Consumption definitions

Different possible consumption definitions

- Nondurable consumption
- Nondurable + durable consumption
- Total consumption including housing (rent + imputed service flows for owner occupiers)
- Deflated consumption including housing

Treatment of housing costs reflects different assumptions about capitalisation of local amenities

- Perfect mobility + homogenous prefs $\implies$ no need to deflate

We also compare our estimates with **mean per capita income from admin data**

Calculating MSEs

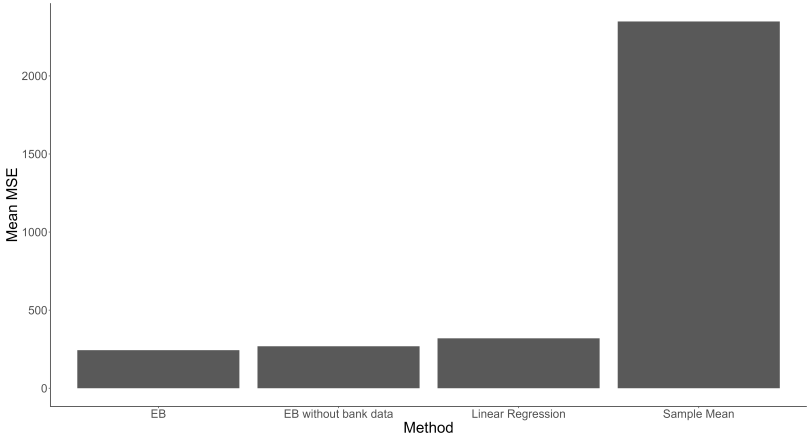
Simulation/model-based MSEs using parametric bootstrap

For draws $b = 1..B$

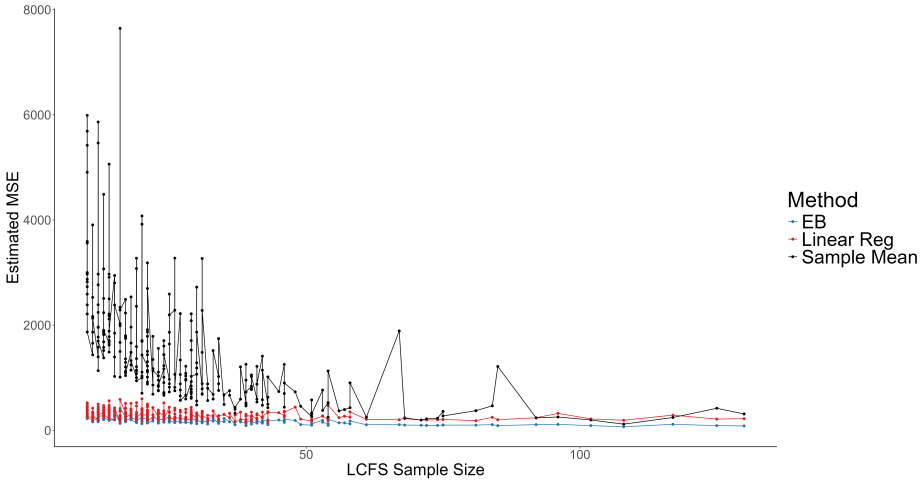
- Take our estimates of $\hat{\beta}, \hat{\pi}$
- Repeatedly re-draw distributions of residuals (u_a, ε_a)
- Calculate new consumption values
- Draw new survey sample
- $\hat{c}_a^{EB*(b)}, \hat{c}_a^{Direct*(b)}$
- Compare this with 'truth'

MSE is average across draws

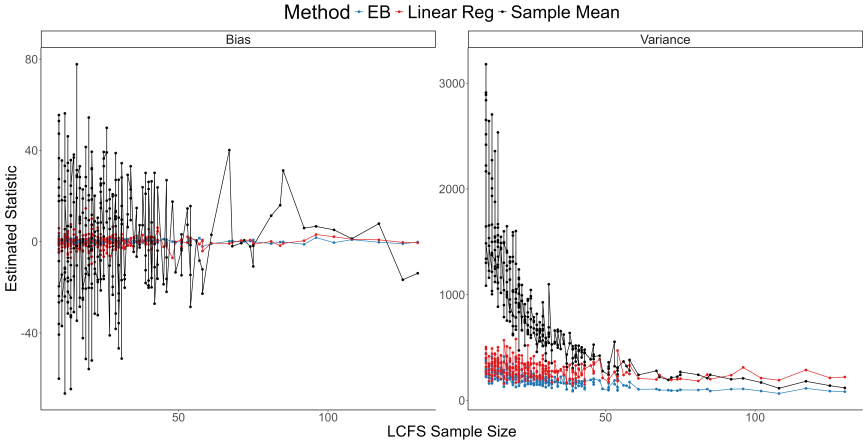
Comparing MSEs across different measures



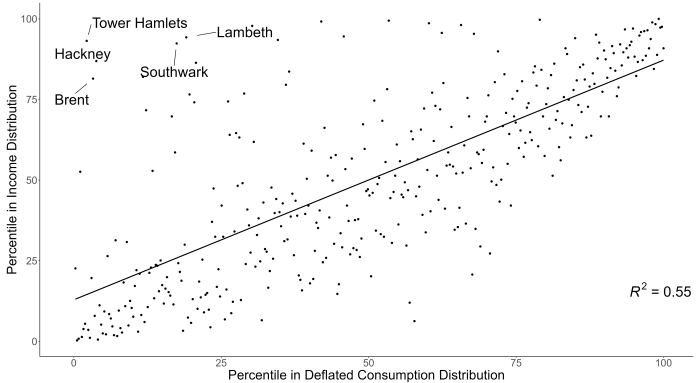
Comparing MSE across different measures by sample size



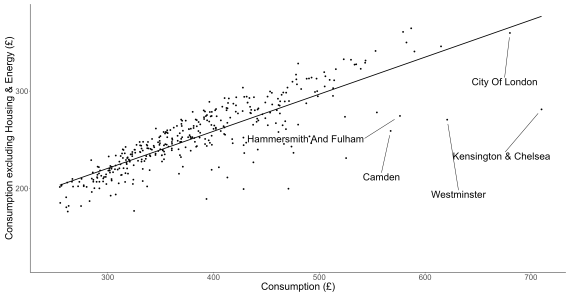
Bias/variance decomposition



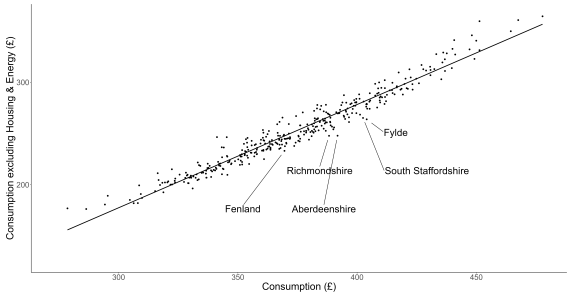
Equivalised consumption vs per-capita income - ranks



EB estimates: including and excluding housing-related costs



(a) Nominal consumption including housing-related expenditures



(b) Deflated consumption including housing-related expenditures

Summary

- Combining household budget data with auxiliary datasets can be used to measure local-area consumption spend
  - Financial footprints data is not a substitute for good survey data, but can be used to complement it
- Simulations show significant improvements in MSE (especially in small areas) relative to taking the sample mean
- Income and consumption can lead to different rankings across areas in terms of living standards

Part II: Responses to questions

- Qs: Which consumption concept did you use?
  - We consider two measures: total spend with and without housing?
  - For comparisons across areas, treatment of housing crucial
    - Include flows to owner-occupiers?
  - Deflate or not deflate according to local house prices? Depends on assumptions about moving costs and preference for locations
    - Rosen-Roback model - no need to deflate (differences in prices reflect differences in amenities)
    - No mobility $\implies$ should deflate
- Similar issues apply to cross-country comparisons

Response to questions (II)

Comments on LCS note (what should be included or adjusted)

- Qs about social transfers in kind: ideal would be to include these but need to value them to recipients
- Including education spending (some countries mainly public, some mainly private) might be problematic

Health spending

- OOP spending not included, but insurance contributions are included
- Gross (of expected payouts) or net insurance contributions?
- Similar issue arises with CPI weights for insurance (HICP) uses net
- Contributions to claims pool are transfers between states of the world

Other comments

- What about deflators? Are same definitions of consumption used for deflators (in say PPP) as in LCS?
- Known issues of reporting biases in consumption surveys relative to national accounts
  - Can these be addressed? (or just discussed?)