Income and Wealth Inequality: Drivers and Consequences

Finding fortunes: a new methodology to tackle differential response bias in wealth survey data

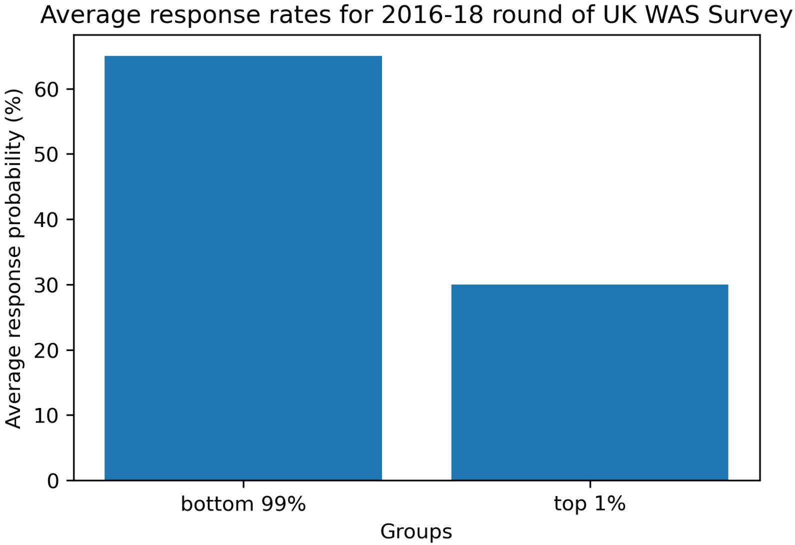
Ben Tippet and Rafael Wildauer, University of Greenwich

28th September 2023

Motivation

- How much do the wealthiest really own?
- Survey data underestimates true level of wealth inequality due to **differential unit response bias**
- Methods to correct for this assume that the true top tail of the wealth distribution follow a Pareto distribution:
  - Maximum likelihood approach (Eckerstorfer et al., 2014, 2016) - biased estimates
  - Rich list approach (Vermeulen, 2018) - poor quality or lack of coverage (Capehart, 2014, Kopczuk, 2015)

Evidence of differential response bias



Overview and findings

- Derive a new methodology **Missingness Maximum Likelihood (MML)** that:
  - Does not use rich list data
  - Like the maximum likelihood approach estimates Pareto distribution
  - Explicitly model the differential non-response process: how much does the probability of response increase with wealth?
- Monte Carlo Simulation:
  - MML corrects for the bias in the standard ML approach
  - MML performs as well as the rich list approach
- Application to ONS Wealth and Assets Survey 2008 to 2020

Modelling Pareto tails: Maximum Likelihood approach

- Key assumption: complete distribution of wealth Y above y_{min} follows a Type 1 Pareto distribution (Wildauer and Heck, 2023) with PDF:

$$Pr(Y = y_i) = y_i^{1-\theta} y_{min}^{\theta} \theta$$

- Estimates the Pareto shape parameter θ that maximises likelihood function given the observed data, weights w_i and imposing some value for y_{min}

$$\ell_{ML}(\theta \mid y_0, y_{min}, w_i) = \sum_{i=y_{min}}^{r} w_i * log(y_i^{1-\theta} y_{min}^{\theta} \theta)$$

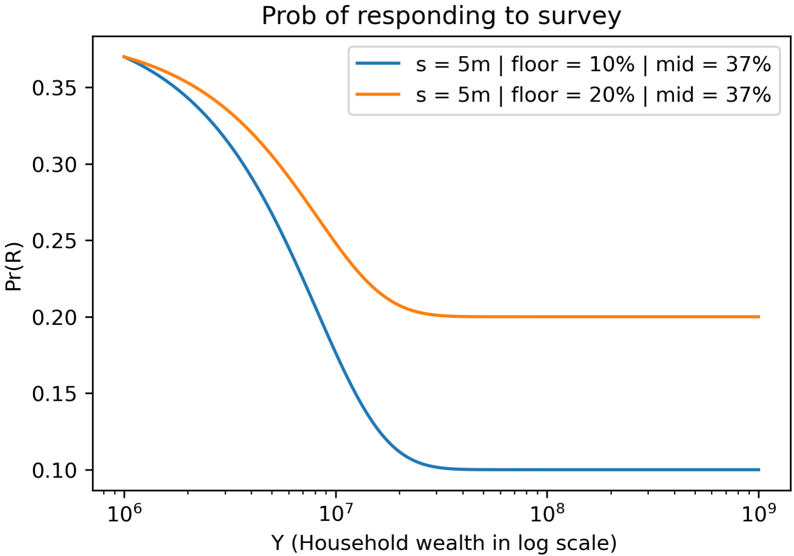
- A larger Pareto shape parameter θ means less concentration of wealth y_{min}

Modelling Pareto tails: New Approach 1

- Includes a "response function" - the probability that a household responds to the survey (R) given its wealth
- Little and Rubin (2019:351) Missing Not At Random process as response probability is conditional on wealth
- Assume Generalised Logit function with four parameters:
  - s is the slope parameter
  - ψ_{floor} is lowest probability of responding (i.e. the floor)
  - y_{min} is the Pareto threshold
  - $\psi_{y_{min}}$ is the response probability y_{min}

$$Pr(R \mid Y, s, \psi_{y_{min}}, \psi_{floor}) = \frac{2 * (\psi_{y_{min}} - \psi_{floor})}{1 + e^{\frac{y - y_{min}}{s}}} + \psi_{floor}$$

Response function



Modelling Pareto tails: New Approach 2

- With this response function we derive a new likelihood function following general framework set out in Little and Rubin (2019:351)

$$\ell_{ML}(\theta, s, \psi_{floor} \mid y_0, y_{min}, \psi_{y_{min}}, n - r) =$$

$$\sum_{i=y_{min}}^{r} w_i * log[(y_i^{1-\theta} y_{min}^{\theta} \theta) * \frac{2 * (\psi_{y_{min}} - \psi_{floor})}{1 + e^{\frac{y - y_{min}}{s}}} + \psi_{floor}]$$

$$+(n - r) * log \int_{y_{min}}^{\infty} [(y_i^{1-\theta} y_{min}^{\theta} \theta) * \frac{2 * (\psi_{y_{min}} - \psi_{floor})}{1 + e^{\frac{y - y_{min}}{s}}} + \psi_{floor})]$$

Modelling Pareto tails: New Approach 3

- Estimate:
  - θ: Pareto shape parameter
  - ψ_{floor}: Minimum probability of response
  - s: Response function slope
- Get from the data:
  - n - r: Number of non-responding households above y_{min}
  - $\psi_{y_{min}}$: Response probability at y_{min}
- Impose: Pareto threshold
  - y_{min}

Monte Carlo

θ

- Estimate:
  - $\theta = 1.2$:Pareto shape parameter
  - ψ_{floor}: Minimum probability of response
  - s: Response function slope
- Get from the data:
  - n - r: Number of non-responding households above y_{min}
  - $\psi_{y_{min}}$: Response probability at y_{min}
- Impose: Pareto threshold
  - y_{min}

Monte Carlo Results

- Estimate a Monte Carlo for both standard ML and new method with 1000 runs
- Synthetic population has $\theta = 1.2$, $\psi_{floor} = 0.1$, $s = 3.8 * 10 * *6$
- standard ML estimates of are upwards biased and therefore underestimate the extent of wealth concentration
- New method estimates are upwards biased and therefore underestimate the extent of wealth concentration θ , ψ_{floor} , s

Monte Carlo

θ **(true)**	θ **(ML)**	θ **(MML)**	**s**	ψ_{floor}
1.3	1.71	1.30	3903716	0.16
	(0.10)	(0.15)	(1973061)	(0.05)
1.6	2.10	1.63	3783406	0.16
	(0.16)	(0.17)	(1576366)	(0.06)

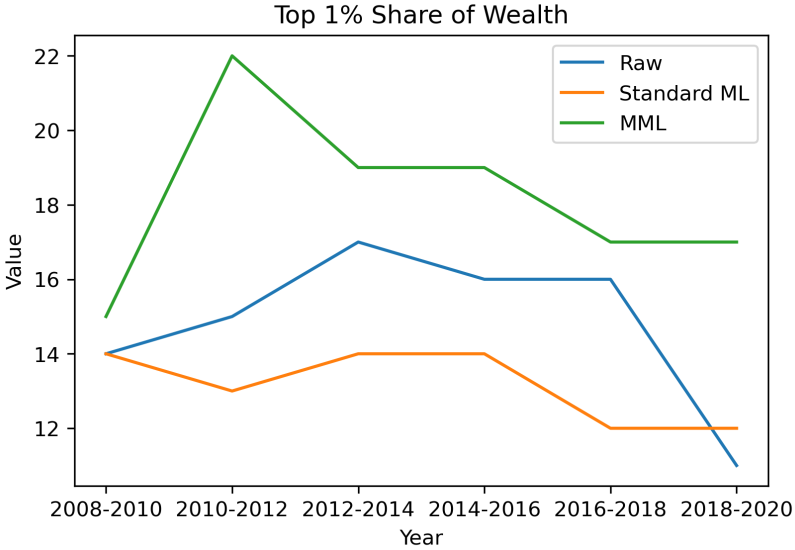
Application to UK Wealth Survey Data

- Apply method to estimate missing wealth in UK Wealth and Assets Survey 2008 to 2020
- Impose y_{min} at 99th percentile and derive n-r and $\psi_{y_{min}}$ from data for each wave
- Estimate θ , ψ_{floor} and s which maximise likelihood function
- Adjust top 1% wealth share with Pareto distribution with estimated θ

UK Wealth Survey: how much wealth is missing

Year	Missing wealth from top 1% (£bn)
2008-2010	425
2010-2012	1276
2012-2014	538
2014-2016	262
2016-2018	227
2018-2020	1075

UK Wealth Survey: adjusting top 1% wealth share



Conclusion

- Differential non-response bias is a known unknown - we know surveys tend to suffer from it but we do not know to what extent
- Designed a new method for estimating the degree of differential non-response bias which:
  - Explicitly models the response function and estimates its key parameters
  - Monte Carlo estimations –¿ unbiased Pareto estimates
  - Does not rely on rich list data so can be applied to wide range of LWS countries

Appendix: UK Application Imposed or Measured Parameters

Year	y_{min} **(£mn)**	$\psi_{y_{min}}$	**n-r**
2008-2010	3.25	0.42	518
2010-2012	3.27	0.39	572
2012-2014	3.68	0.4	536
2014-2016	4.18	0.4	493
2016-2018	3.57	0.39	817
2018-2020	4.04	0.37	944

Appendix: UK Application Estimated Parameters

Year	theta	s	psi_floor
2008-2010	1.52	1.77	0.19
2010-2012	1.24	5.02	0.12
2012-2014	1.38	2.34	0.18
2014-2016	1.3	1.99	0.19
2016-2018	1.5	2.05	0.18
2018-2020	1.57	2.48	0.1

Evidence of differential non-response in ONS

Deriving n-r from the oversampling strategy of ONS

Modelling Pareto tails: New Approach 2

$$Pr(R = r_i) = f_{R|Y}(R \mid Y, \psi)$$

$$Pr(R = r_i) = f_{R|Y}(R \mid Y, \psi)$$

Where R denotes the vector of binary responding indicators with $r_i = 1$ if y_i responds to the survey and $r_i = 0$ if y_i does not respond and ψ are the parameters of this model.We assume that the response function can be modelled as a logit function.

$$L_{ML}(\theta, \psi \mid y_0, m) = \sum_{i=y_{min}}^{r} w_i * log[(y_i^{1-\theta} y_{min}^{\theta} \theta) * f_{M|Y}(M \mid Y, \psi)]$$

$$+(n-r) * log \int_{y_{min}}^{\infty} [(y_i^{1-\theta} y_{min}^{\theta} \theta) * f_{M|Y}(1 - (M \mid Y, \psi))]$$