EU-SILC and the potential for synthetic panel estimates

Brian Colgan

John Stuart Mill College, VU Amsterdam

b.p.colgan@vu.nl

7th European User Conference for EU-Microdata

March 26, 2021

In the absence of panel data

- Retrospective questions - relies on recall
- Pseudo panel techniques (Deaton, 1985) - focusses on inter-cohort dynamics

The two main ‘synthetic” panel approaches are:

- **Dang, Lanjouw, Luto, & McKenzie (DLLM) 2014 - produces parametric bounds for transitions**
- Bourguignon & Moreno 2015

How does it work?

DLLM synthetic panel approach

Two key elements:

- Income models - which link households over time on the basis of time invariant characteristics
- Residual autocorrelation - how to map the unexplained portion of income over time

Income models

A1:

the underlying population sampled must be the same in survey round 1 and survey round 2.

$$y_{i1} = \beta_1^{'} x_{i1} + \epsilon_{i1} \tag{1}$$

$$y_{j2} = \beta_2^{'} x_{j2} + \epsilon_{j2} \tag{2}$$

1. Using the data in survey round 1 obtain predicted coefficients $\hat{\beta}_1$ and predicted residuals $\hat{\epsilon}_{i1}$ from the linear income model (1)
2. For each household in round 2 predict round 1 income using the predicted coefficient $\hat{\beta}_1$

Residual Autocorrelation

A2:

ϵ_{i1} and ϵ_{i2} have a bivariate normal distribution with (partial) correlation coefficient ρ and standard deviations $\sigma_{\epsilon 1}$ and $\sigma_{\epsilon 2}$.

3. Estimate probability of dynamics using equation 3

$$P(y_{i1} \sim z_1 \text{ and } y_{i2} \sim z_2) = \Phi_2 \left(d_1 \frac{z_1 - \beta_1' x_{i2}}{\sigma_{\epsilon 1}}, d_2 \frac{z_2 - \beta_2' x_{i2}}{\sigma_{\epsilon 2}}, \rho_d \right) \tag{3}$$

DLLM suggest estimating parametric bounds with $\rho = 0$ and $\rho = 1$

DL approximating ρ - return to pseudo panel techniques

Assume household income follows a simple linear dynamic data generating process (AR(1)) given by:

$$y_{i2} = \alpha + \delta y_{i1} + \eta_{i2} \tag{4}$$

Replace individual level observations with cohort level averages:

$$\tilde{y}_{c(t),2} = \alpha + \delta \tilde{y}_{c(t-1),1} + \tilde{\eta}_{c(t),2} \tag{5}$$

The simple correlation coefficient $\rho_{y_{i1}y_{i2}}$ can then be approximated by:

$$\rho_{y_{c1}y_{c2}} = \sqrt{\frac{var(y_{c1})}{var(y_{c2})}}\delta \tag{6}$$

$$\rho = \frac{\rho_{y_{c1}y_{c2}}\sqrt{var(y_{i1})var(y_{i2})} - \beta_1' var(x_i)\beta_2}{\sigma_{\epsilon 1}\sigma_{\epsilon 2}} \tag{7}$$

Does it work in practice?

Overview of the empirical validation of the DL method

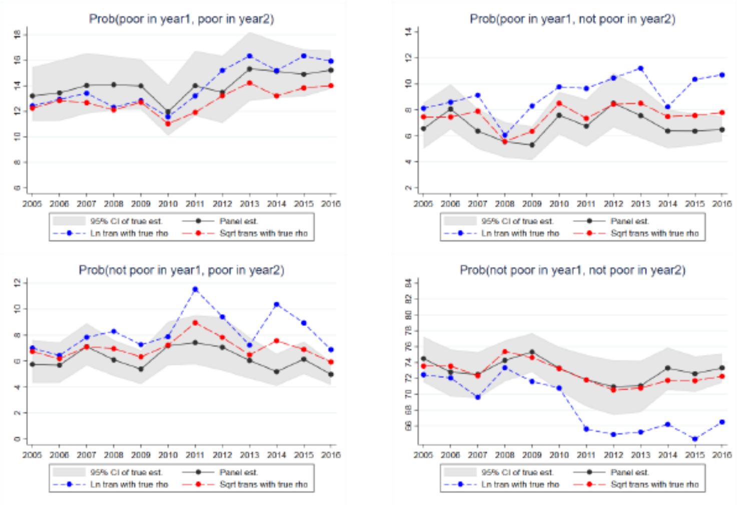
	Country	Y/C	Cohort ρ	Accuracy
Dang & Lanjouw (2018)	5	C & Y	yob(1)	High
Urzanqui (2017)	Thailand	Y	yob(3)*region	High
Herault & Jenkins (2019)	UK	Y	yob(5)*sex	Low
	Australia	Y	yob(5)*cob	Low

Validation approach - EU-SILC

- Countries: France, Poland and Greece
- Within panel validation
- Age of household head: 25-75
- Income: household disposable income
- Income model includes: Sex, 5 year birth cohort, country of birth, education level, children, and interaction terms between Sex and both education and birth cohort

Poverty dynamics mw60 - Greece

(a) Joint Probabilities



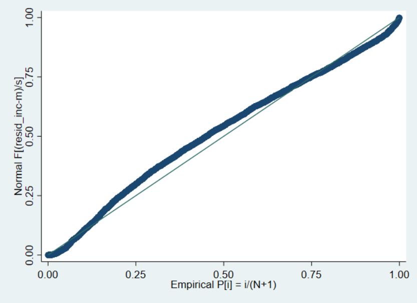
Prob(poor in year1, poor in year2)

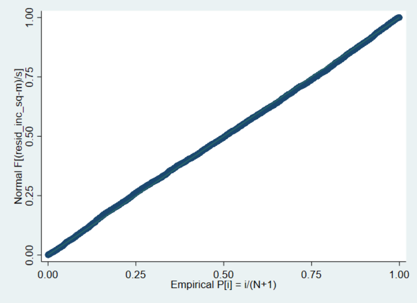
Prob(poor in year1, not poor in year2)

Prob(not poor in year1, poor in year2)

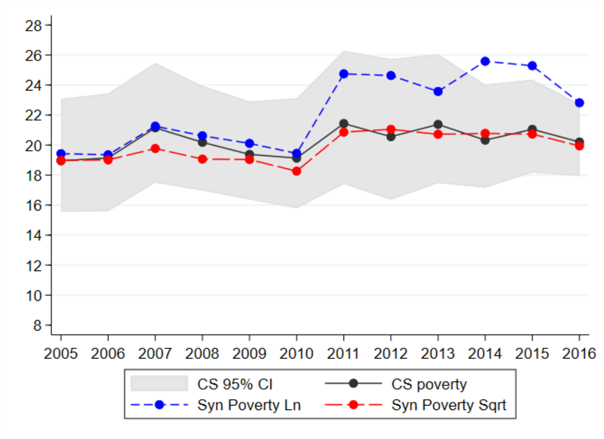
Prob(not poor in year1, not poor in year2)

Normality of residuals - Greece 2014





Predicted poverty - Greece



An additional step

1. Using the data in survey round 1 obtain predicted coefficients $\hat{\beta}_1$ and predicted residuals $\hat{\epsilon}_{i1}$ from the linear income model (1)
2. **Examine the normality of residuals and accuracy of predicted poverty. Explore alternative data transformations and/or the exclusion of outliers to improve normality.**
3. For each household in round 2 predict round 1 income using the predicted coefficient $\hat{\beta}_1$
4. Estimate probability of dynamics using equation 3.

Summary of findings with true ρ

After incorporating the new step:

1. Aggregate estimates are found to be highly accurate using most standard poverty lines - nearly all estimates lie within the 95% confidence interval of the true panel estimate
2. Insensitive to choice of income model
3. Insensitive to choice of age range
4. Accurate for sub-populations - urban-rural, education level, children and adults
5. Sensitive to poverty line chosen - higher poverty lines are associated with less accurate estimates.

What to do about ρ?

DL approximation - France

	Age 25-75				
	ρ	δ	Adj R^2	Size	Coh
Panel	0.71	0.72			
	(0.71, 0.75)	(0.70, 0.74)			
yob(2)	0.58	0.51	0.022	129	25
yob(3)	0.66	0.60	0.022	189	17
yob(4)	0.74	0.69	0.021	247	13
yob(5)	0.76	0.73	0.020	321	10
yob(3)*Sex	0.59	0.53	0.027	94.5	34
yob(4)*Sex	0.65	0.59	0.027	124	26
yob(5)*Sex	0.71	0.67	0.025	161	20
yob(10)*Sex	0.83	0.86	0.023	321	10
yob(1)*Ed	0.74	0.67	0.11	21.1	152
yob(2)*Ed	0.85	0.79	0.11	42.9	75

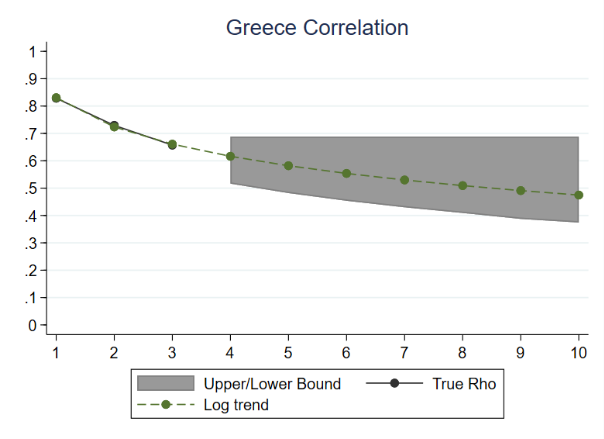
Examining ρ for EU-SILC - NB characteristics

1. ρ is decreasing as the time period considered is extended
2. The percentage decline in ρ declines as the time period considered is extended

Implications for extending ρ:

- Given (1), the three year ρ estimate can serve as an upper bound
- Given (2), applying the percentage decline in ρ to the three year ρ estimate can be used to produce a lower bound estimate

Predicted ρ - Greece



Concluding remarks on ρ

1. the DL approximation is highly sensitive to the cohort definition
2. in the absence of panel data there is no one statistic or combination of statistics which indicates the optimal cohort definition
3. the DL approximation also exhibits much greater volatility over time compared to the true ρ.
4. Extending what is known from the longitudinal element of EU-SILC it is possible to produce practically useful bounds

EU-SILC and the potential for synthetic panel estimates

1. Can be used to generate medium to long run income dynamics
2. Can be used to link ad hoc modules over time. For example income dynamics by sub-populations defined by parental background.
3. Can be used to produce alternative estimates for countries with high rates of attrition or large discrepancies between poverty measured using cross-sectional data and longitudinal data.

Thank you for you attention!

Questions and comments are much appreciated

References

- Dang, Hai-Anh, Peter Lanjouw, Jill Luoto, and David McKenzie. (2014). “Using Repeated Cross-Sections to Explore Movements in and out of Poverty”. Journal of Development Economics, 107: 112-128.
- Dang, Hai-Anh, and Peter Lanjouw. "Measuring poverty dynamics with synthetic panels based on cross-sections." (2013).
- Hérault, Nicolas, and Stephen P. Jenkins. "How valid are synthetic panel estimates of poverty dynamics?." The Journal of Economic Inequality 17.1 (2019): 51-76.
- Garce´s Urzainqui, D.: Poverty transitions without panel data? An appraisal of synthetic panel methods. Paper presented at the ECINEQ Conference, New York City (2017)
- Bourguignon, F., Moreno, H.: On the construction of synthetic panels. Paper Presented at the North East 788 Universities development consortium annual conference, Brown University, Providence RI (2015)

a) France

	Age 25-75				
	ρ	δ	Adj R^2	Size	Coh
Panel	0.71	0.72			
	(0.71, 0.75)	(0.70, 0.74)			
yob(2)	0.58	0.51	0.022	129	25
yob(3)	0.66	0.60	0.022	189	17
yob(4)	0.74	0.69	0.021	247	13
yob(5)	0.76	0.73	0.020	321	10
yob(3)*Sex	0.59	0.53	0.027	94.5	34
yob(4)*Sex	0.65	0.59	0.027	124	26
yob(5)*Sex	0.71	0.67	0.025	161	20
yob(10)*Sex	0.83	0.86	0.023	321	10
yob(1)*Ed	0.74	0.67	0.11	21.1	152
yob(2)*Ed	0.85	0.79	0.11	42.9	75

Approximation of ρ

1. Pseudo panel techniques can produce accurate approximates of ρ on average.
2. Approximates do not capture underlying trends in ρ and display greater volatility
3. Approximates are sensitive to cohort definition
4. There is not one summary statistic or combination of summary statistics which identify the best performing cohort
5. Cohorts defined by 1 year birth cohort interacted with education perform best