A Microsimulation Based Method to Nowcast Household Income Survey Data

Work in Progress!

Cathal O'Donoghue*, Denisa Sologon**

* National University of Ireland, Galway

** Luxembourg Institute of Socio-Economic Research

16TH July 2021

Funded by HRB-IRC

Centre for Microsimulation and Policy Analysis

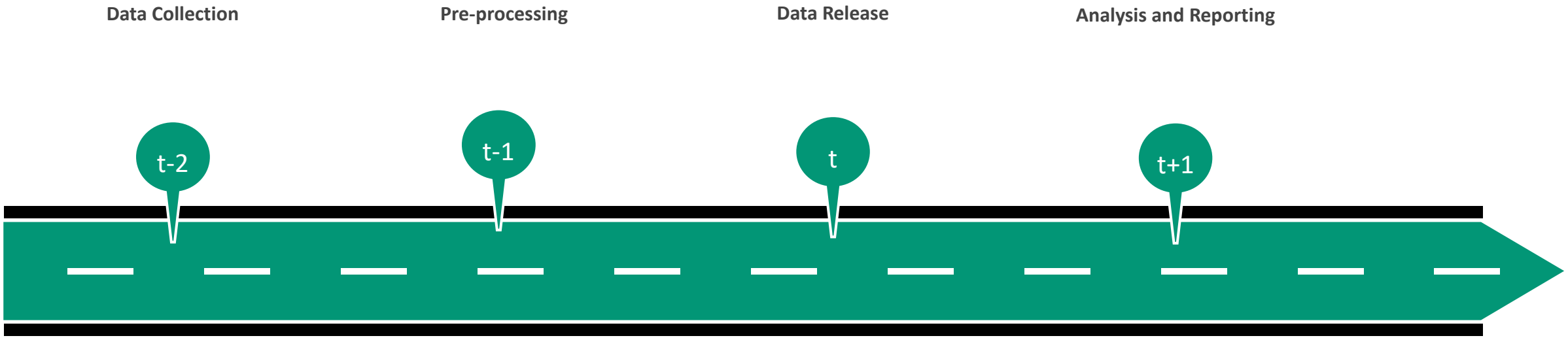
Home of Microsimulation

Expert meeting on Nowcasting and mid-term projections through microsimulation models

Context – Survey Data Collection

Need for Now casting

Data Preparation



Analytical Steps

Survey

- Data Collection takes time
- A year – or more

Pre-Processing

- Cleaning
- Coding
- Validation
- Harmonisation
- Comparative
- Anonyimity

Data Release

- Request Data
- Derived Variables
- Model Adjustment
- Further Validation

Analysis

- Simulation
- Analysis
- Write-up
- Presentation
- Publication

NOWCASTING

- The nowcasting process can be described in a number of ways.
  - Uprating typically refers to issues associated with indexing market income for wage growth.
  - Updating may refer to adjusting the tax-benefit rules in a microsimulation model to account for policy change.
  - Reweighting or static ageing may apply to changing weights to account for changed population structure, while dynamic ageing refers to simulating changes to the population and economic structure.
- Is the risk in making a major change to the distribution worth it?
  - In “normal” volatility, possibly not
  - In super-normal volatility, maybe no other choice

NOWCASTING

- Macro-economic literature (Giannone et al., 2008)
- Directly modelling poverty incidence (Álvarez et al., 2014)
- O'Donoghue and Loughrey (2014) survey article
- EUROMOD Leventi et a., (2014); Navicke et al, (2014) applied Nowcasting to update poverty indicators, calibrating to LFS using transition probabilities of employment
    - Distributional impact of COVID crisis (Figari and Fiorio, 2020; Brewer and Tasseva, 2020; Brewer & Gardiner, 2020; Bronka et al., 2020).
- Addabbo et al., (2016) extend the parametric perspective of EUROMOD by modelling employment transitions using estimated logit equations.
- Carta (2019) Instead of taking labour force status from household surveys, imputes labour incomes into LFS.
- Li et al (2020) used semi-parametric perspective, a la DiNardo, Fortin and Lemieux (1996)

NOWCASTING – CHOICE OF METHOD

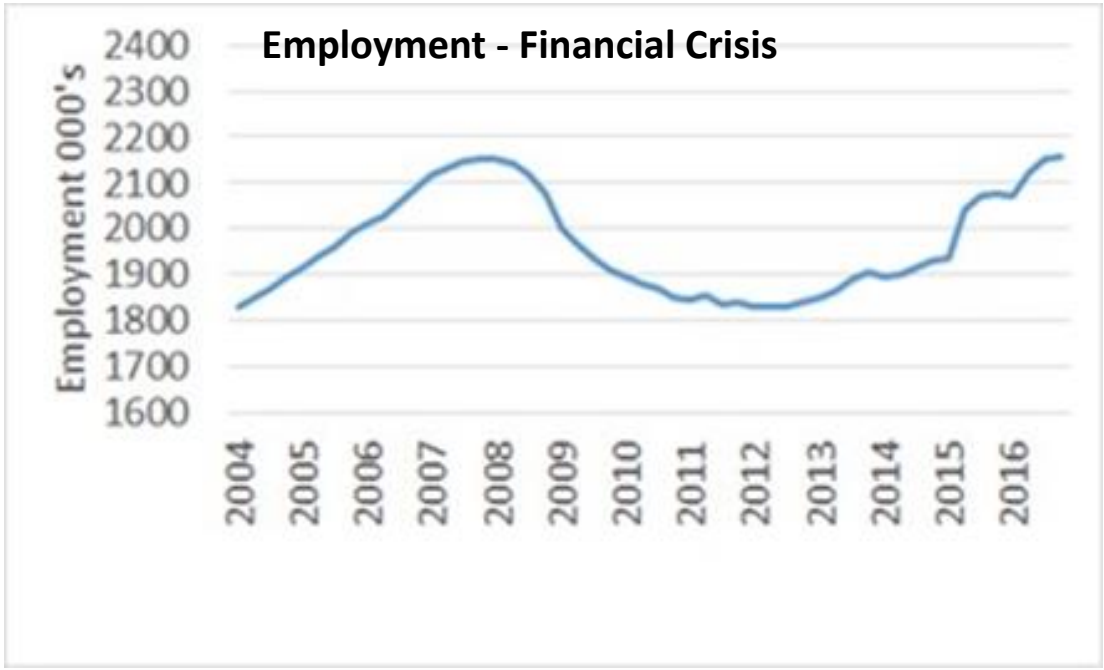
- Use of Monte-Carlo simulations based upon cell-specific probabilities may ignore some of the important heterogeneity exhibited in a crisis. For example, family status may be an important driver.
- Partners within a family may be correlated as identified in Carta (2019).
  - Carta (2019) avoids issues associated with intra-household variation or sectoral biases by using recent labour force survey data.
  - Ignoring the impact of other non-labour force characteristics and the impact of public policy as a insulating mechanism is an issue.
  - Question → is it harder to simulate labour market changes in a model that contains the full range of incomes and policies or vice versa.

NOWCASTING – CHOICE OF METHOD

- Reweighting or semi-parametric approaches are strong in the sense that they avoid distributional assumptions.
  - However like other static-reweighting procedures they rely on sufficient sample sizes and face the risk with very many dimensions of relying for cell weights on small numbers. (Klevmarken 1997).
- In the case where reweighting or semi-parametric approaches prove unfeasible, a parametric approach may be a more pragmatic approach.
- In this paper we consider in more detail a parametric approach akin to the Addabbo et al., (2016) methodology, but in addition drawing upon the alignment aspects of the dynamic microsimulation literature.

THE FINANCIAL CRISIS

- Financial Crisis 2008-2012
- One of the deepest and longest lasting Financial Crises (FC)
  - Worst year 2009. Lowest point 2012
  - Young and Male
  - Big tax loss



Theoretical Framework – Income Generating Process

$$Y_{it}^{H} = \sum_{F} \sum_{I} \sum_{S} Y(a + B_{st}X_{it} + v_{si} + u_{ist}) \cdot I(c + G_{st}Z_{it} + w_{si} + e_{sit})$$

- Estimate system of equations representing
  - Z Demographic and Data Sampling Error
  - Ii() Presence of Income Source I
  - Yi() Level of Income Source I

Correlation between income sources

$$Corr(u_{is_1t,}u_{is_2t,}) = \rho_{s_1s_2}$$

Correlation over time

$$Corr(u_{its_2,}u_{ist_2,}) = \rho_{t_1t_2}$$

Correlation between people in household

$$Corr(u_{i_1st,}u_{ist,}) = \rho_{i_1i_2}$$

We assume they remain the same – in reality they change

Theoretical Framework – Assumptions about process

$$Y_{it}^{H} = \sum_{F} \sum_{I} \sum_{S} Y(a + B_{st}X_{it} + v_{si} + u_{ist}) \cdot I(c + G_{st}Z_{it} + w_{si} + e_{sit})$$

- Calibration means adjusting the constant or at most adjusting the constant for separate population groups for which we have control total
- In addition to the assumptions about the data generating process that are made
- There is sampling variation → so other variation

Assume

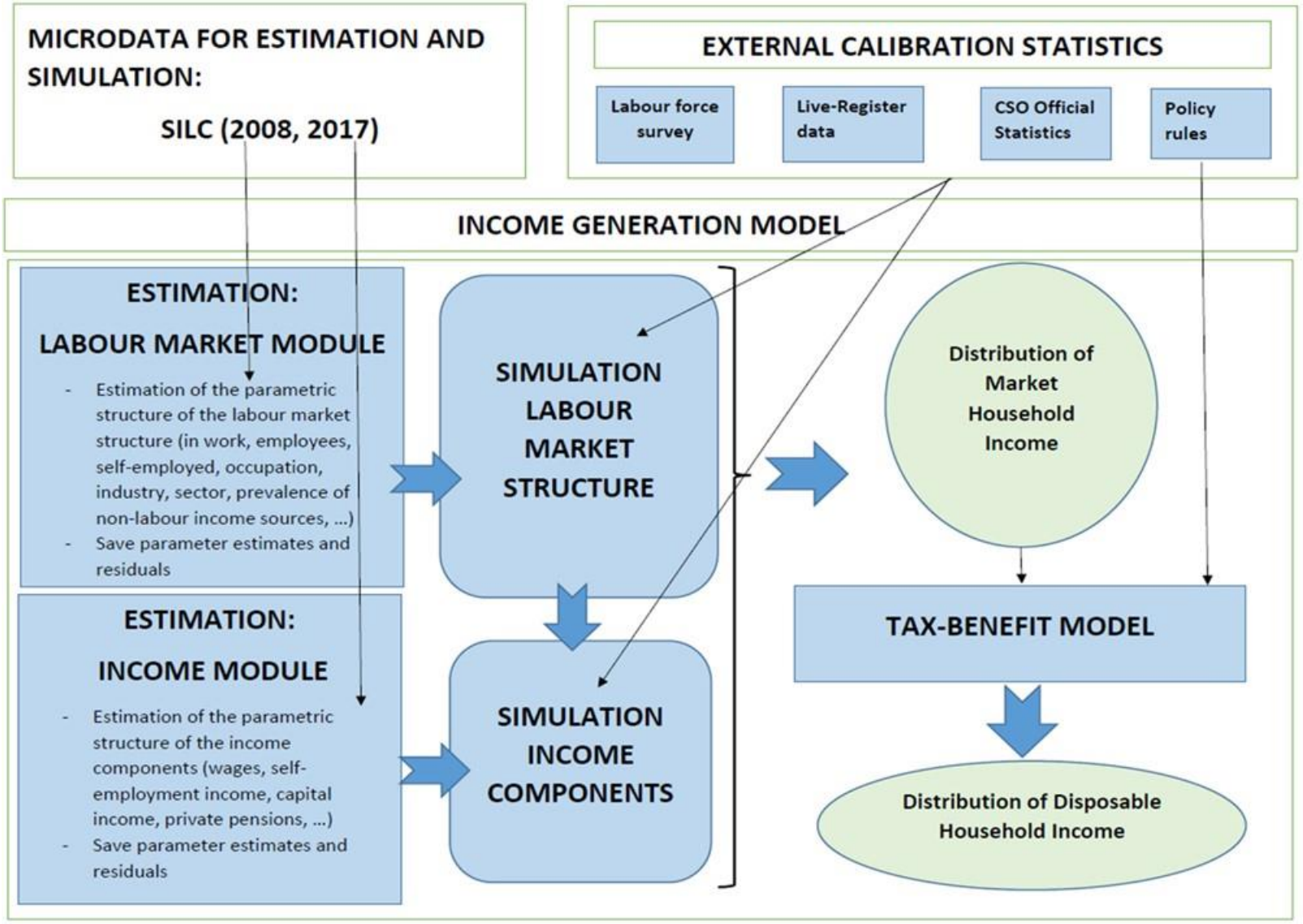
$Var(u_{ist,}) = constant$ $Var(e_{ist,}) = constant$

B_{st}=B_{st+1} G_{st}=G_{st+1}

X_{st}=X_{st+1} Z_{st}=Z_{st+1}

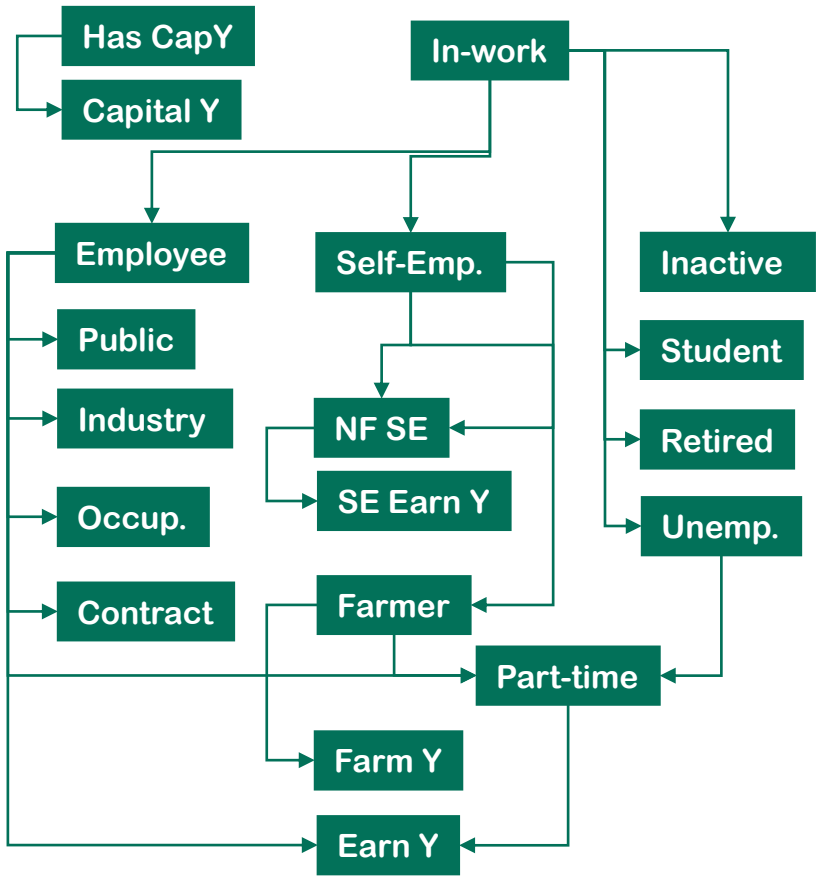
We assume they remain the same – in reality they change

METHODOLOGY



METHODOLOGY – SYSTEM OF EQUATIONS

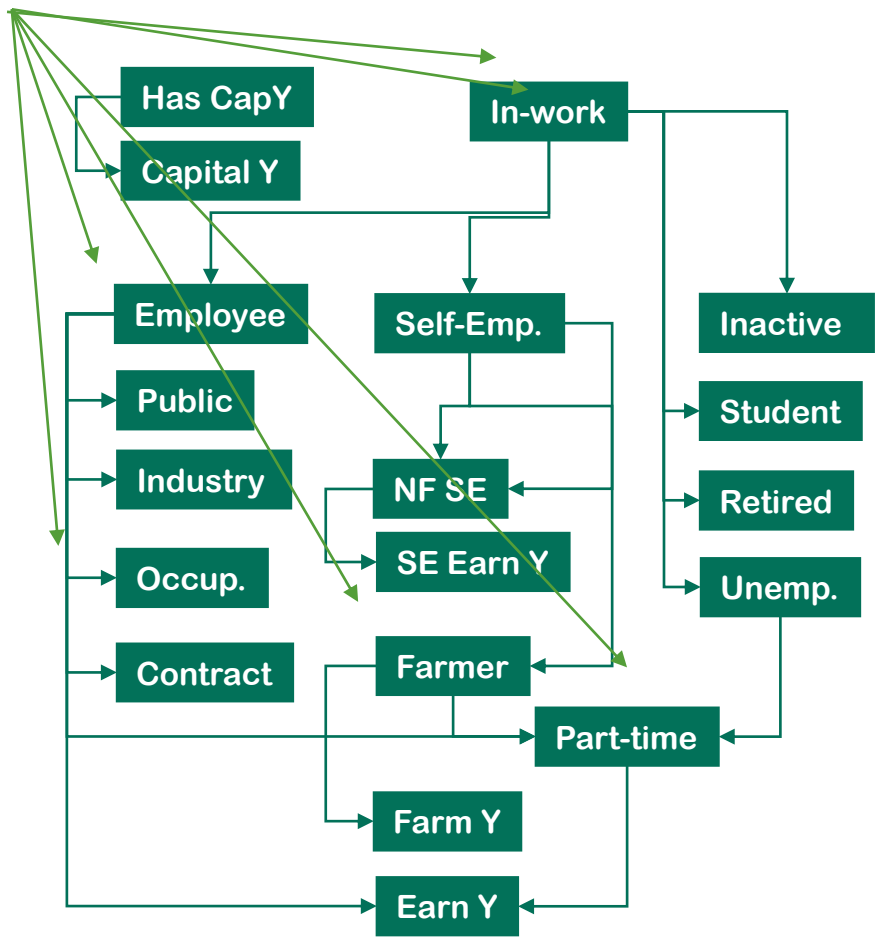
- relies on a system of hierarchically structured, multiple equation models for detailed income sources, combining:
  - a set of personal (individual and household) characteristics, X,
  - parameters describing how employment, the receipt and level of income sources vary with personal characteristics, B ,
  - residuals linking model predictions to observed income sources, e;
  - a tax-benefit simulator for converting market incomes into disposable incomes
  - parametric structure of the IGM
- $\boldsymbol{Y}=\boldsymbol{m}(\boldsymbol{X}, \boldsymbol{e}, \boldsymbol{l}(\boldsymbol{B}), \boldsymbol{r}(\boldsymbol{B}), \boldsymbol{tb})$
- l=labour market models, r=returns models, tb=tax-benefit system



METHODOLOGY – STEPS

$l^*(B_s)$

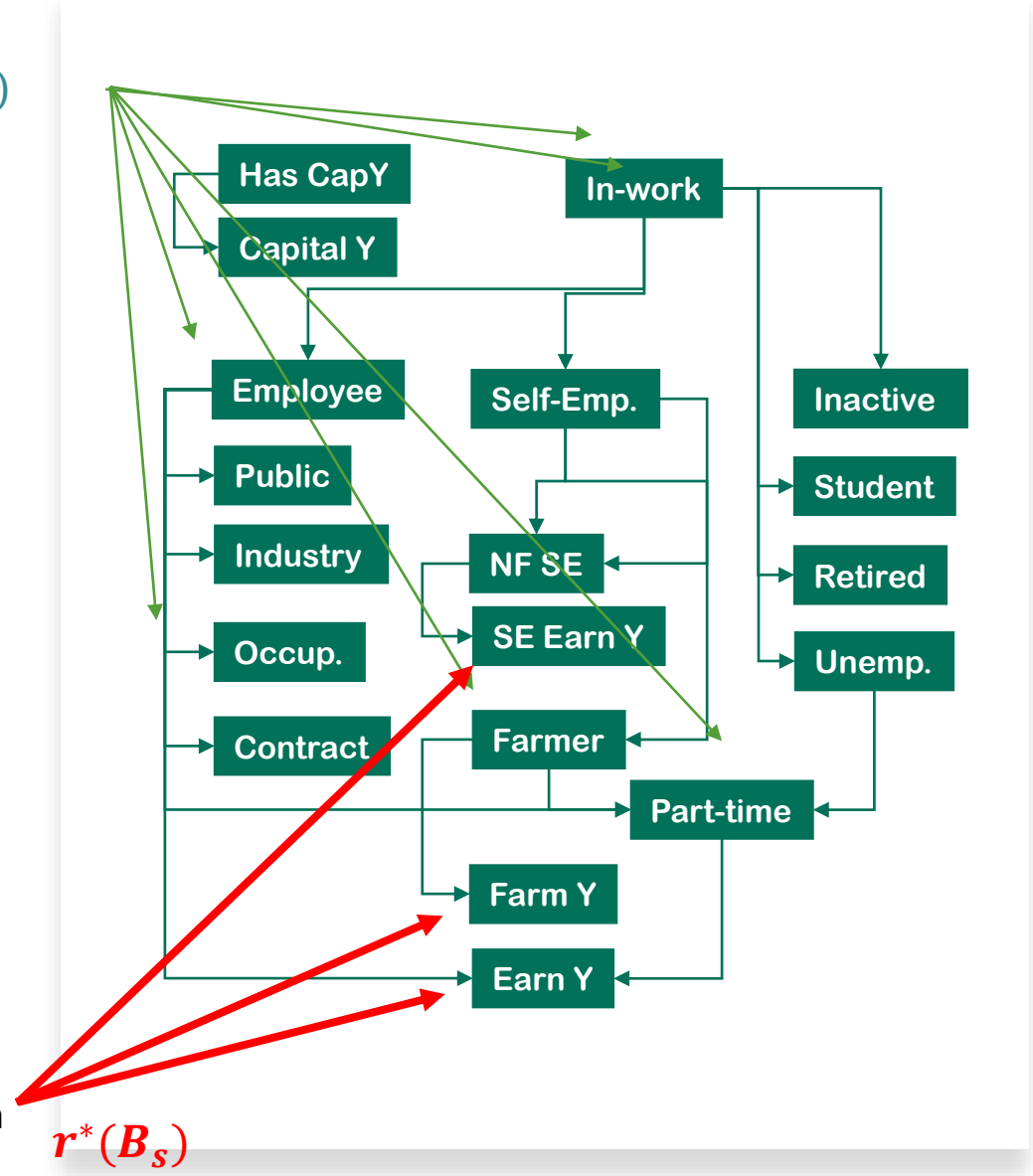
- Step 1: Estimation
  - we use the latest available survey year (s) to estimate the parameters of the IGM
- $Y\_s=m(X\_s, e\_s, l(B\_s), r(B\_s), tb\_s)$
- Step 2: Nowcasting to period t
  - X_s (exogenous characteristic), e_s (error structure), B_s(IGM parameters) are kept the same
  - $l\hat{}*(B\_s)$ – labour market simulations are calibrated to external controls (go to picture)
  - $r\hat{}*(B\_s)$
    - simulate all income sources based on the new labour market structure $l\hat{}*(B\_s)$ (new inwork, new occupation, new industry structure → wage structure)
    - uprate pre-fiscal monetary variables to align them with the policy parameters
- tb_t policy parameters of period t



METHODOLOGY – STEPS

$l^*(B_s)$

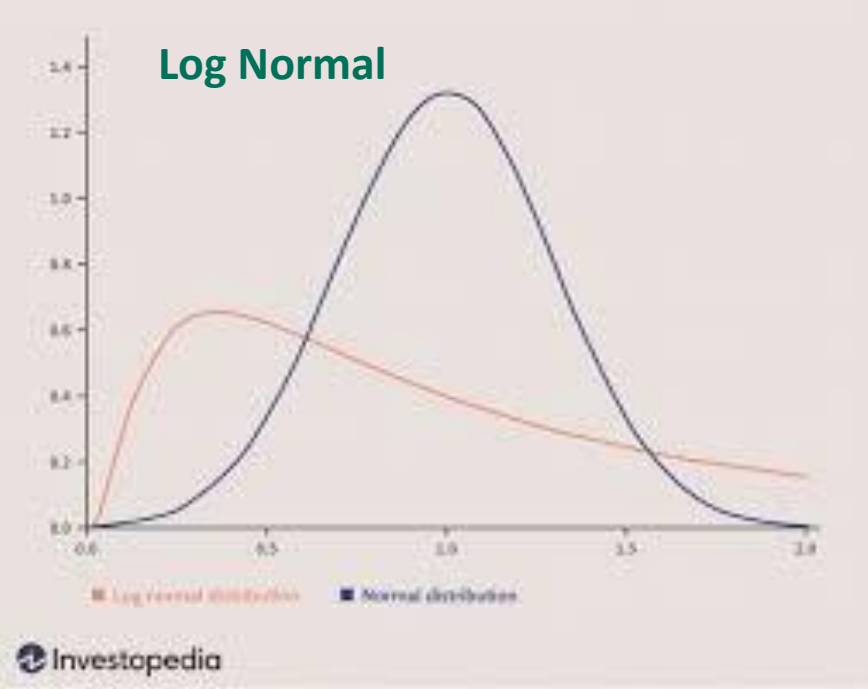
- Step 1: Estimation
  - we use the latest available survey year (s) to estimate the parameters of the IGM
- $Y\_s = m(X\_s, e\_s, l(B\_s), r(B\_s), tb\_s)$
- Step 2: Nowcasting to period t
  - X_s (exogenous characteristic), e_s (error structure), B_s(IGM parameters) are kept the same
  - $l\hat{}*(B\_s)$ – labour market simulations are calibrated to external controls (go to picture)
  - $r\hat{}*(B\_s)$
    - simulate all income sources based on the new labour market structure $l\hat{}*(B\_s)$ (new inwork, new occupation, new industry structure → wage structure)
    - uprate pre-fiscal monetary variables to align them with the policy parameters
- tb_t policy parameters of period t



MAIN DIFFERENCES BETWEEN THE TWO NOWCASTING APPROACHES

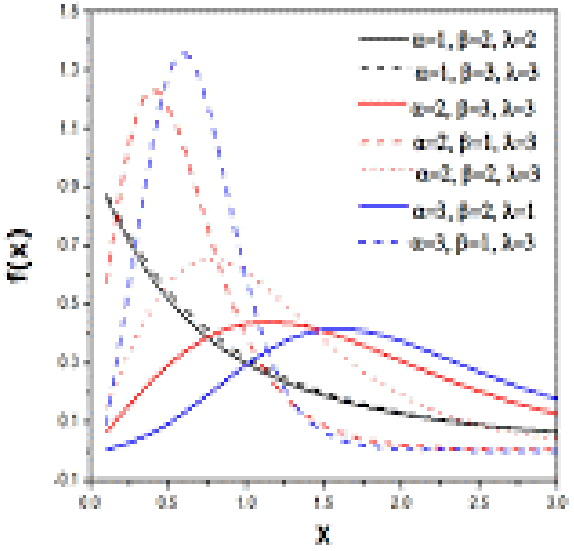
IGM	Transition Matrix/Monte Carlo
Labour market component - Hierarchical system assures the sequential calibration of in-work, employment, occupational, industrial structure) - We allow transitions in-work and out-of-work	**Labour market component** - Applies employment probabilities to transition people out-of-work
Income component - Simulate all income components, taking into account the changes in the labour market structure (e.g. those that are classified in a different occupation in the calibration, have their wage updated)	**Income component** - Simulate incomes of those out of work

Modelling Issue – Earnings Distributional Assumptions

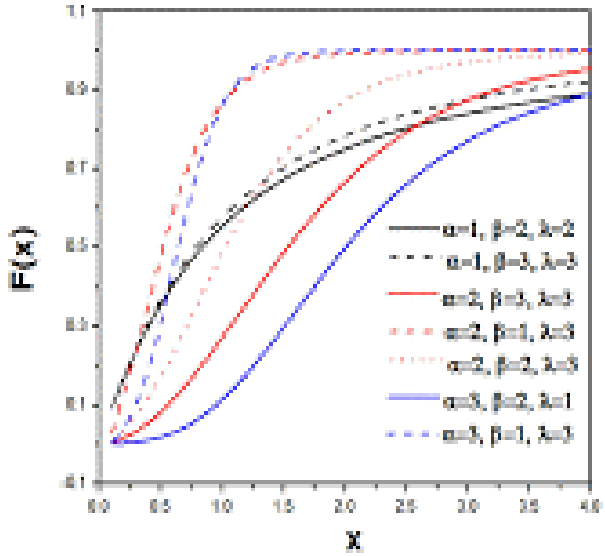


Singh Maddala Distribution – more control in parameterisation of curve

pdf



cdf



Data Issues

- Time lag between collection and release for research and analysis.
  - The most recent analysis of SILC undertaken is for 2007 – compared with 2008 and 2009.
  - In normal times a lot happens in a two year period,
  - In a crisis the changes are so significant that such a lag can mean the data is relatively meaningless.
- There are more recent datasets available
  - LFS → available on a quarterly basis at a six week lag
  - the Live Register data and Price data → monthly basis on a short lag.
  - Pandemic Benefit Impact → weekly
- However these datasets do not contain income information.

ALIGNMENT IN IGM

- Simulate GZ + v → Logit Model
- Calibration Total by Sex and Gender
- Rank within Group
  - Select the highest GZ + v according to the external calibration totals

MODELLING CHOICES TO BE TESTED

- Do Models change over time?
- Is it important to model in more specific detail than in-work (ie structural changes)?
- How should we model income equations (Log-norm vs Singh Maddala Distributional Regression)?
- Comparing IGM with transition matrix/Monte Carlo
- Looking at historical crisis 2007-2009
- Use Simple Model then Full Model

Do Models Change over Time – CHOW Test

Chow Test – Models Different over time

Sex	Male		Male	
	Beta	S.E.	Beta	S.E.
University	1.2856***	0.1051	1.2856***	0.1051
Upper Secondary Education	1.1107***	0.0878	1.1107***	0.0878
Numer of Children Aged 0-3	0.1086	0.1235	0.1086	0.1235
Number of Children aged 4-11	-0.1594**	0.0626	-0.1594**	0.0626
Number of Children aged 12-15	-0.2285***	0.0779	-0.2285***	0.0779
Married	0.653***	0.0959	0.653***	0.0959
Age	0.3168***	0.0103	0.3168***	0.0103
Age Squared	-0.0038***	0.0001	-0.0038***	0.0001
Per-Urban	-0.0884	0.0918	-0.0884	0.0918
Rural	0.4898***	0.0890	0.4898***	0.0890
Year == 2009			-0.5203	0.3162
Year == 2009 x University			0.0949	0.1430
Year == 2009 x Upper Secondary Education			-0.2146*	0.1241
Year == 2009 x Numer of Children Aged 0-3			0.1195	0.1728
Year == 2009 x Number of Children aged 4-11			0.0092	0.0859
Year == 2009 x Number of Children aged 12-15			0.0013	0.1127
Year == 2009 x Married			-0.052	0.1342
Year == 2009 x Age			0.0107	0.0150
Year == 2009 x Age Squared			-0.0001	0.0002
Year == 2009 x Per-Urban			-0.1125	0.1292
Year == 2009 x Rural			-0.1392	0.1247
Constant	-5.7354***	0.2159	-5.7354***	0.2159

Is the Industry Distribution Different?

- No difference in industry distribution when not simulated for Males, but difference for females
- IGM captures changes

Chi Square Test of Industry Distribution		
	Monte Carlo	IGM
Male	0.014	0.000
Female	0.116	0.000

Are Simulated Mean Earnings Different? Ratios

- Log Normal creates means that are different for both Transition Matrix/Monte Carlo and IGM
- Singh Maddalla very close

Ratio of Means		
	Male	Female
Monte Carlo	1.06	1.19
IGM (OLS)	1.07	1.19
IGM (SMD)	1.02	1.02

Are Simulated Mean Earnings Different? Statistical Difference

- Means different for both Transition Matrix/Monte Carlo
- No difference for SMD

Statistically Significant Difference in Means		
	Male	Female
Monte Carlo	1	1
IGM (OLS)	1	1
IGM (SMD)	0	0

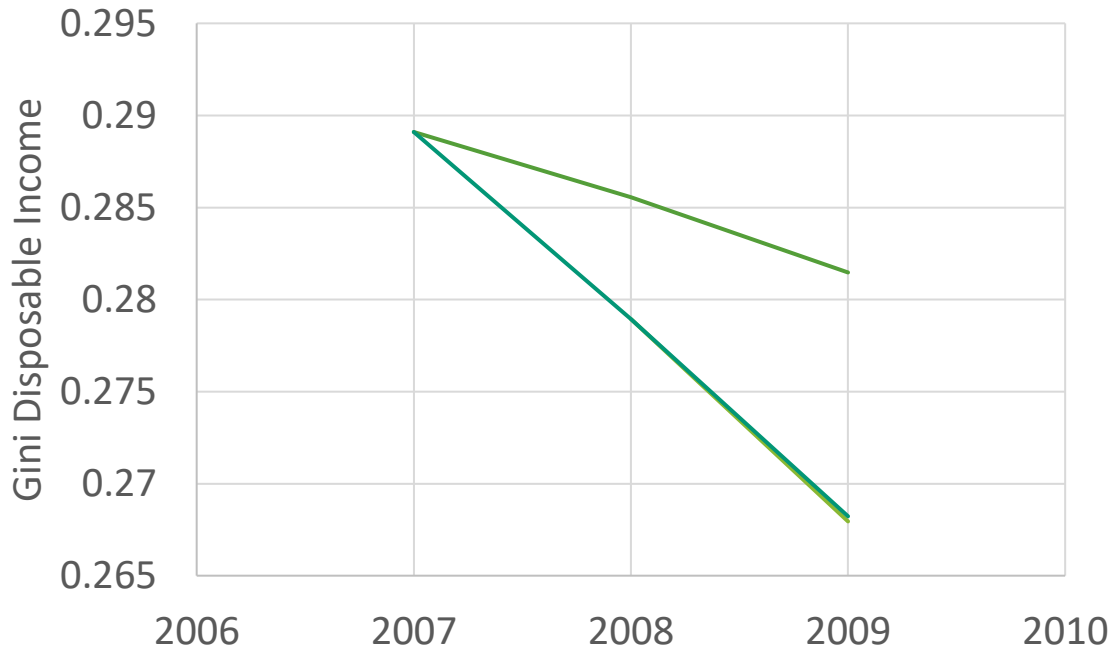
Are Simulated Mean Earnings Different? Kolgomorov-Smirnov Test of Distributions (p. value of difference)

- Transition probability/Monte Carol Different for all
- Distributional Issues using log-normal
- No difference with actual for SMD for Males
- Difference of females due to changes in residuals not

Kolgomorov-Smirnov Test of Distributions (p. value of difference)		
	Male	Female
Relative to Actual 2009		
Monte Carlo	0.006	0.000
IGM (OLS)	0.006	0.000
IGM (SMD)	0.133	0.001
IGM (SMD - No Industry)	0.185	0.001

NOWCASTING – EVALUATION

- Evaluation
  - Looking back at simulation properties of a now cast during the financial crisis
  - Comparing simulated with actual, we find a good fit.
  - However performance weakens the further the now cast
  - Over longer term, method picks up on turning points and trend, but weaker on level



SUMMARY

- Nowcasting a useful methodology
- Not much difference between Transition Matrix and IGM for In-work
- Significant difference if other structural changes
- Problems in using Log-Normal distribution
- Model replicates changes in Income Distribution over 1 year
- Over longer term captures the trend, but weakens the further we get from estimation year
- What is the success bar?
- Future work – trends in B and Sigma's

Thank you

Cathal.odonoghue@nuigalway.ie