A supervised learning assessment of income mobility in Italy

Statistical Learing & Large Data I & II

Luna Boiago
Chiara Ferrara
Mariachiara Mariani
Anita Sammarini

- **Distribution** and **variation** of **income** as indicators of **quality of life**

- **In this work: estimates** on the **distribution of income** in **Italy from 2015 to 2019** from the **Italian Statistics Income Register (ISTAT)**

- The **first** conceptual definition of **mobility: positional change**

- The first part, investigating:
  - how individuals' **income change over time** in Italy
  - **heterogeneity** of mobility dynamics
  - which **categories** suffer the most from a stagnant **income mobility**

  - The second part: mobility indexes to expand on **drivers of intragenerational mobility (or factors hindering it)** + focus on factors that contribute to **the most stagnant/flexible income dynamics**

- Income inequality in Italy is **comparatively high**
  - Between **1948 and 1968**, no significant changes
  - During **1970s**, income distribution recorded **increasing trends**
  - **From the 1980s to the 2000s**, fluctuations around **flattened trend**
  - Since **2000s, stagnating dynamics**

- **Recent main studies' results** on mobility and inequality

- The main contributions of this analysis:

  - estimating the **impact of education and gender** differentials on an enhanced intragenerational mobility
  - describing differences in mobility dynamics across **age classes**
  - examining the impact of **various sources of income** on mobility (self-employment, employment, and pension income).

- Sample of 337,396 individuals and 40 million taxable persons, for each one a **coefficient** of the **relative weight**
- **From 2015 to 2019** and **longitudinal data** on income records
- Focus on the **employment**, the **self-employment** and the **pension income**
- A set of **individual characteristics** (gender, age class, education, geographical partition)
  - **dummy variables** for studying the **income differences** among individuals **depending on** their **specific characteristics**
- Subdivision of the **distribution into percentiles** and **deciles**
  - whether each individual was **in the highest or in the lowest** decile and percentile and if they **switched** from one to another from 2015 to 2019.

The methodology that has been used in this study is **four parts**:

- The **first** one: **multinomial logistic regression**
  - calculate the proportion of each individual with his/her particular features to be in the four different quartiles/decile

- The **second** step: **logistic regressions**

  - For the **first two,** estimate the probability of passing to **the richest** decile and quartile
  - For the **latter two**, estimate the probability to pass to **the poorest** quartile and decile

- The **third** pace was **delta regressions**, for estimating how many steps each individual has done from 2015 to 2019

- **Finally**, **transition matrix** for describing the probabilities of moving from one state to another in a dynamic system

Multinomial logistic regression:

- Considering each **age class**: the probability for women to move to a different quartile was in general lower than for men. The highest changes to switch into an higher quartile are for women in their twenties or thirties that starts from the second quartile.
- People with a **low degree**: among the four quartiles, the one to increase is the proportion of women that move to the lower quartile which in this case is the first.
- **North-South divide**: in general, a woman from the South with the same characteristics of a woman from the North is less likely to improve her condition

Logistic regressions:

The **estimated regression models**:

$$Quantile_j = \beta_0 + \beta_1 Gender_j + \sum_{j=2}^{9} \beta_j AgeClass_j + \beta_{10} Degree_j + \beta_{11} NoStudy_j +$$
$$\beta_{12} EMP_j + \beta_{13} SEMP_j + \beta_{14} PEN_j + \sum_{j=1}^{14} \gamma_j Region_j +$$
$$+ \sum_{j=1}^{4} \theta_j Quartile_j + \sum_{j=1}^{10} \xi_j Decile_j,$$

Most significant results:

- **Women** → **lowest quartile with higher probability than men**
- Age **15-54** → **lower probability** of moving to the **poorest decile** and **higher probability** of moving to **the richest and to the lowest quartile (than 55-64)**
- **Degree** → **higher bands**
- Share of **self-employment** both highest and lowest quantiles
- For Southern regions, future worsening of economic conditions is more probable compared to Northern regions such as Piemonte and Valle d'Aosta.
- Models (1) and (4): Coefficients ξ signal the relative difficulty of reaching the highest income class and the relative ease of moving down to the lowest income deciles given any starting point

Table 2: Regression analysis output of the four logit models

	Dependent variable:			
	10^{th} Decile	4^{th} Quartile	1^{st} Quartile	1^{st} Decile
	(1)	(2)	(3)	(4)
Gender	−0.006***	−0.022***	0.019***	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Age 15-24	0.028***	0.043***	0.044***	−0.197***
	(0.006)	(0.008)	(0.009)	(0.008)
Age 25-34	0.034***	0.058***	0.037***	−0.189***
	(0.006)	(0.008)	(0.009)	(0.008)
Age 35-44	0.027***	0.043***	0.035***	−0.189***
	(0.006)	(0.008)	(0.009)	(0.008)
Age 45-54	0.021***	0.035***	0.029***	−0.192***
	(0.006)	(0.008)	(0.009)	(0.008)
Study title: Degree	0.024***	0.055***	−0.017***	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Self-employment share	0.026***	0.027***	0.044***	0.026***
	(0.001)	(0.001)	(0.001)	(0.001)
Region Trentino Alto Adige	0.014***	0.020***	−0.004*	−0.008***
	(0.002)	(0.002)	(0.003)	(0.002)
Region Calabria-Sicilia	−0.003**	−0.013***	0.029***	0.010***
	(0.001)	(0.002)	(0.002)	(0.002)
First decile 2015	−0.049***			
	(0.001)			
Second decile 2015	−0.047***			0.221***
	(0.001)			(0.002)
Third decile 2015	−0.047***			0.083***
	(0.001)			(0.002)
Fourth decile 2015	−0.045***			0.061***
	(0.001)			(0.002)

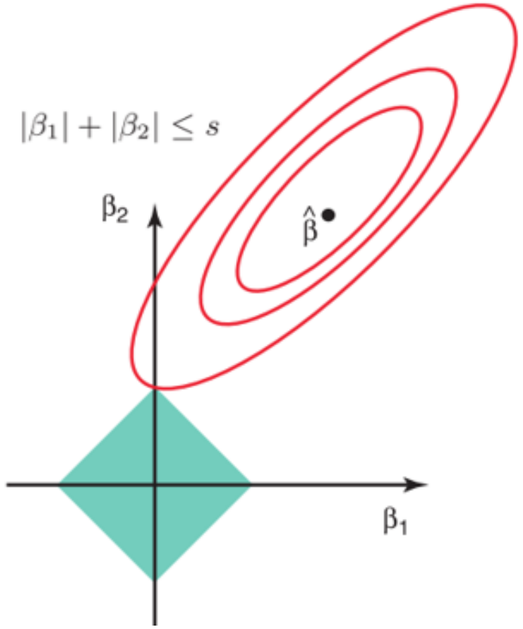
"Delta" regressions:

$$DeltaQuantile_j / RelativeDeltaQuantile_j = \beta_0 + \beta_1 Gender_j + \sum_{j=2}^{9} \beta_j AgeClass_j + \beta_{10} Degree_j +$$
$$\beta_{11} NoStudy_j + \beta_{12} EMP_j + \beta_{13} SEMP_j + \beta_{14} PEN_j + \sum_{j=1}^{14} \gamma_j Region_j +$$
$$+ \sum_{j=1}^{4} \theta_j Quartile_j + \sum_{j=1}^{10} \xi_j Decile_j,$$

- Estimates suggest that:
  - Being a **woman** is associated with a **falling** in the difference between the 2019 and 2015 percentiles **with respect to men**
  - Individuals of **all age classes increase their position** moving towards both richer percentiles and deciles **than those aged between 55 and 64 years**
  - Those with a degree in 2019 rose by about 6 percentiles and 0.6 deciles compared to those without one
  - **Self-employment share** has a **negative** impact, whereas the impact of **employment share** is **positive**

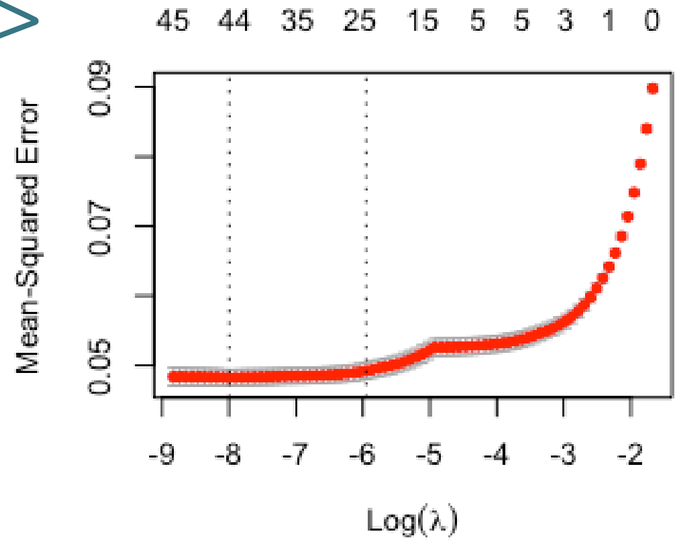
PROCEDURE

- Choice of **4** different **response variables** and LASSO
- Cross-validation to identify the lambda values (optimal and 1se)
- Predictions to assess **variance explained** on train set
- Identification of selected predictors
- Identification of **common set of selected predictors** among the 4 response variables
- Re-running of logit regressions on the dependent variables using the common set of regressors



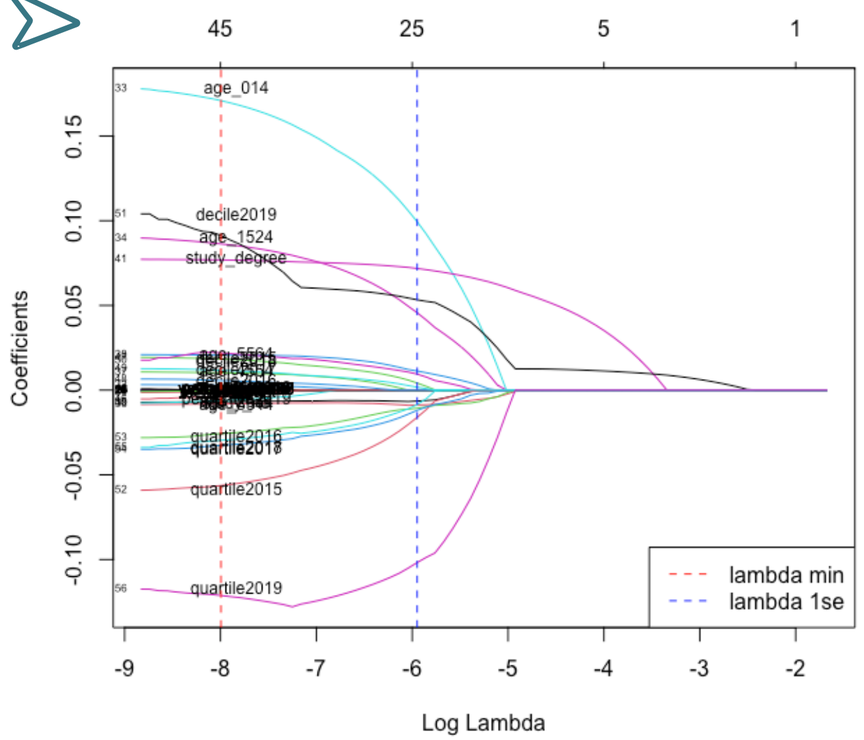
Least Absolute Shrinkage and Selection Operator: consists in a statistical formula for the regularisation of data models and feature selection

INTERPRETATION OF RESULTS



Tuning of lambda
(optimal value vs +1 standard deviation)

Observe how the curve stays flat between 2 thresholds: choice of 1se as lambda value



Selection of predictors according to value of lambda chosen

Selection of **core predictors**, recurrent in all 4 models:

- cittad = citizenship;
- ripart4 = geographic area;
- titostud = study title;
- Y_dispo15 = income 2015;
- ye3_lav_aut19
- north_west
- north_east
- age_014
- age_3544
- percentile2019
- decile2019

Re-running logit regressions using as independent variables these core set of predictors: not all variables result significant!

*the graphs reported relate to the first response variable (dummy richest decile 2019)

- **Introduction** to the topic and **literature/historical review**
- Multinomial logistic regressions: prove that men move from one quartile to another in a larger proportion
- Logit regressions: people between 14 and 54 have larger probability of moving to higher quantiles of income than people older than 54
- LASSO: selection of relevant predictors; results prove that gender, study title, geographical residence and age are main drivers of income mobility

LIMITS

wise to consider a **wider range of time**

would be interesting to assess income mobility **after 2019**, since the pandemic, inflation and general uncertainty might have changed partly the drivers and magnitude of income mobility

Thank you for the attention!