Assessing individual poverty status using repeated cross-sectional surveys

M. Grazia Pittau[♡], Roberto Zelli[♡] and Saida Ismailakhunova[♣]

[♡]Sapienza University of Rome; [♣]World Bank

IARIW and World Bank Conference
Washington DC, November 8, 2019

Motivation and background

Motivation

- Growing interest in **poverty dynamics** to shape the design and implementation of social development strategies:
  - mobility into and out of poverty status;
  - transient and chronic poverty (current and persistent poverty);
  - vulnerability to poverty.
- Data demand: reliable longitudinal data, seldom available especially in developing countries.
- What can we say when only (repeated) independent cross-sectional data are available?

Cohort approach

Cohort approach (Deaton, JEc 1985; Deaton and Paxon, JPE 1994;...)

- Follows cohort ("fixed membership group") of individuals or households over repeated cross-sectional surveys.
- Moffit (JEc 1993), McKenzie (JEc 1994), Veerbek (book 2008) discuss conditions to obtain consistent mobility estimates from pseudo-panel models.
- Trade-off between number of cohorts and number of observations in each cohort.

Synthetic panel

Synthetic panel (Dang and Lanjouw, WBwp 2013; Dang et al., JDE 2014; Hérault and Jenkins, JoEI 2019)

- Synthetic panels generally use two cross-sectional data.
- Observed incomes of individuals interviewed in the second round are compared to "artificial" incomes as if the same individuals (with the same time-invariant characteristics) had been surveyed in the first round.
- Artificial incomes are predicted based on parameters of an income model estimated in the first round of cross-sectional data that includes only time-invariant covariates.
- Assumptions on the (joint distribution) of the error terms are crucial for estimation of mobility.

Dynamic modelling

Dynamic modelling using individual data (Chaudhuri, 2003; Bourguignon, Goh and Kim, 2004; Gunther and Harttgen, WD 2009)

- This approach uses the parameters of an individual income/consumption dynamic model to estimate vulnerability to poverty.
- Vulnerability as expected poverty (VEP): probability of an individual to be poor in the next periods.
- Assumptions on the structure of residuals and on the expected shape of consumption/income distribution (log-normal).

A closer look at the dynamic modelling approach

- According to the VEP concept, the **vulnerability to poverty** of individual i in period t is defined as the probability of being poor at time $(t+1)$ given the individual information set $\Im_t$ at time t (i.e. income or consumption before the poverty line):

$$v_{it} = \text{prob}\{\text{poor}_{i,t+1} = 1 | \Im_t\} = \text{prob}\{c_{i,t+1} < z | \Im_t\}$$

- The expected value $\mathbb{E}(c_{i,t+1})$ and the variance are based on a set of observable individual characteristics:

$$\mathbb{E}(c_{i,t+1}) = c(X_i, \beta_{t+1})$$

- where the idiosyncratic shocks on consumption are i.i.d. over time for each individual (even if they are not i.i.d. across individuals)
- uncertainty about future consumption stems only from the idiosyncratic shocks.

Simplified assumptions (Gallardo, JES 2019)

General **assumptions** of the dynamic modelling approach:

- The model assumes a probability distribution for $c_{i,t+1}$, which is usually a log-normal distribution;
- it rules out the possibility of contextual effects and/or aggregate shocks (improvement by Gunther and Harttgen, 2009);
- strong time stationarity $\beta_{t+1} = \beta \rightarrow$ the parameters of the distribution remain invariant over time (structure of the economy is relatively stable over time).

The proposed methodology

Our methodology tries to overcome the previous assumptions. It allows:

1. **to directly estimate the probability** of being poor in the next period;
2. **to include contextual variables** as a second hierarchical level to separate idiosyncratic shocks from aggregate shocks (macro-covariates distinct from individual covariates);
3. **to model** β_t, allowing the effect of predictors to vary over time;
4. to include **macroeconomic forecasts** that can potentially influence, directly and indirectly, individual poverty status.

How can we do that?

A sketch of the model

The dynamic multilevel model I

1. The **direct estimation of the probability** of being poor.
   - Let $\pi_{i[jt]} = P(Y_{i[jt]=1})$ be the probability that members of household i resident in region j fall into poverty at time t, where Y is a binary variable equal to 1 if household is poor.
   - The probability of being poor is directly estimated by the following *varying-intercepts* and *varying-slopes* multilevel logistic model:

$$\pi_{ijt} = logit^{-1}(\underbrace{\alpha_{jt} + \beta_{jt} x_{ijt}}_{\text{time-space varying}} + \underbrace{\beta z_{ijt}}_{\text{time-space invariant}} + error) \qquad (1)$$

   where
   - $logit^{-1}$ is the inverse-logistic function, jt indexes the area j where household i resides at time t,
   - x are individual-level predictors with time-space varying coefficients,
   - z are individual-level predictors with time-space invariant coefficients,
   - α_{jt} (intercept), β_{jt} (slope) are the varying-parameters of the model.

The dynamic multilevel model II

2. The **inclusion of contextual variables** and the **modelling of α_t and β_t**.
   - We include contextual predictors at both regional level and time level;
   - We model simultaneously α_{jt} and β_{jt} allowing them to vary across regions and across time:

$$\alpha_{jt} \sim \mathrm{N}(\alpha_j + \alpha_t + \gamma^{\alpha} U_{jt}, \sigma^2_{\alpha}), \tag{2}$$

$$\beta_{jt} \sim \mathrm{N}(\beta_j + \beta_t + \gamma^{\beta} U_{jt}, \sigma^2_{\beta}), \tag{3}$$

   where
   - U_{jt} are contextual predictors at regional and time level,
   - α_j and β_j can be further modelled to capture cross-sectional contextual effects,
   - α_t and β_t can be further modelled to capture the dynamics of national macro-economic effects.
   - This modelling makes (1) a **dynamic multilevel model**

The dynamic multilevel model III

3. Once, the model has been estimated, **the probability of being poor at time (t+1), vulnerability to poverty**, can be estimated as:

$$v_{it} = \widehat{\pi}_{i,t+1} = f(\underbrace{x_{i,t}}_{\substack{\text{individual} \\ \text{predictors}}} , \underbrace{\widehat{\alpha_{j,t+1}}}_{\substack{\text{endogenous} \\ \text{time-forecast}}} , \underbrace{\widehat{\beta_{j,t+1}}}_{\substack{\text{endogenous} \\ \text{time-forecast}}} , \underbrace{U_{j,t+1}}_{\substack{\text{exogenous} \\ \text{forecast}}}) \quad (4)$$

- The dynamic multilevel model treats individual poverty status as a function of individuals' characteristics and circumstances, in interaction with time-varying features of their economic contexts;
- **Repeated cross-sectional data**– consisting of independent observations drawn from the same context (e.g. the same region)– allows the specification of such a dynamic multilevel model.

Estimation I

- The complexity of these dynamic hierarchical models has prevented their use so far because of their well-known problems of convergence.
- Casting them in a Bayesian statistical framework offers a reasonable solution. Recent developments of simulation techniques such as Markov chain Monte Carlo (MCMC), a useful class of algorithms, help to fit complex models and provide Bayesian inference to statistical uncertainty.
- **Stan** is a programming language designed to make statistical modeling easier and faster, especially for Bayesian estimation problems. Stan can help estimate complex models with large numbers of parameters as the dynamic multilevel models.
- The library `rstanarm` in R which is based on the `lme4` syntax (Carpenter et al. 2017; Stan Development Team 2017) allowed us to use Stan into R.

Estimation II

- Once the the dynamic poverty model has been specified, it can be estimated by a fully Bayesian model using the function `stan_glmer` and through the Hamiltonian Monte Carlo sampling we can finally draw posterior distributions of the estimated parameters assessing the uncertainties in a Bayesian analysis described by a numerically calculated posterior distribution.

A dynamic multilevel model for poverty in Kyrgyzstan

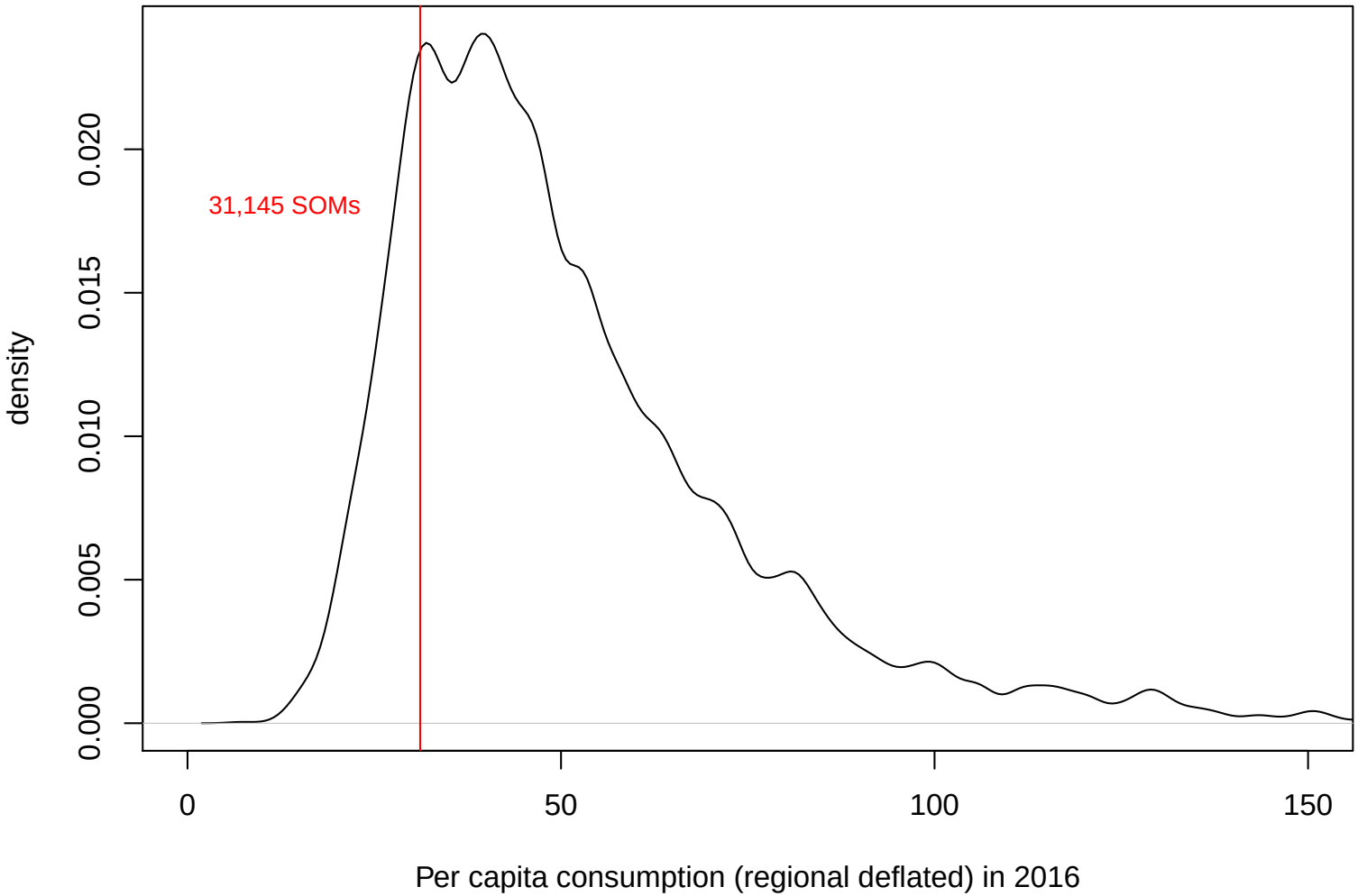
A poverty model for Kyrgyzstan

- We illustrate our approach estimating a dynamic hierarchical model for Kyrgyz Republic that combines independent cross-sectional household surveys and macroeconomic data and forecast;
- Repeated cross-sectional annual data from the Kyrgyz Integrated households budget and labor force surveys (KIHS) conducted by the National Statistical Committee (NSC), available over the period 2013–2017;
- For each year the KIHS covers around 5,000 households and around 20,000 individuals;
- The sample design is a two-stage random sampling, **stratified into 16 strata**, representing urban and rural dimensions of the seven country regions (oblasts) (Batken, Jalal-Abad, Issyk-Kul, Naryn, Osh, Talas and Chui), the city of Bishkek and Osh city;
- Households in the sample are associated with **sampling weights**.

Definition of poverty in Kyrgyzstan

- The poverty measurement applied by the National Statistical Committee of the Kyrgyz Republic follows an **absolute approach**: poor households are those whose **per capita consumption** falls below the national poverty line;
- The national absolute poverty line is calculated to allow a sufficient calorie requirement (2,100 calories per day per person) plus a basket of non-food goods and services;
- All members of a poor household are identified as poor.

Distribution of per capita consumption and the poverty line in 2016



Building the dynamic logistic model involves the following steps:

1. Selection of the **main significant individual predictors** that can explain poverty;
2. Identification of predictors whose effect is **time(-space) varying** and predictors whose effect is **stationary**;
3. Selection of **macro-economic** predictors (variables) at oblast and national level;
4. Modelling dynamics of intercepts and slopes;
5. Estimation. diagnostics and validation;
6. Forecast of the probability to be poor for each household in the sample;
7. (Adjust the sample weights by post-stratification).

Core model I

- **hh-level individual predictors included in the core model**
  - **Time-varying effects (β_t)**
    - educational level,
    - percentage of high educated members of the households,
    - percentage of professional educated members of the households,
    - household size
  - **Time invariant effect (β)**
    - age of the household head,
    - percentage of employed members of the household,
    - residence (urban or rural),
    - ownership of selected durable goods (car, washing machine, electric stove, satellite antenna, refrigerator),
    - access to services and housing (landline, sewer services, number of rooms)
- **Oblast and country variables included in the model**:
  - Unemployment rate by oblast: structural ($\bar{U}_j$) and cyclical ($U_{jt} - \bar{U}_j$);
  - Per capita national GDP (growth).

The dynamic logistic model for Kyrgyzstan

$$\pi_{i[jt]} = \text{logit}^{-1}(\alpha_{jt}^{oblast-year} + \beta_{jt}^{ED,oblast-year} \cdot x_{i,EDUC} +$$

$$+\beta_{jt}^{S,oblast-year} \cdot x_{i,SIZE} + \sum_{k=1}^{K} \beta_k^{fixed} \cdot x_{i,k})$$

$$\alpha_{jt}^{oblast-year} \sim \text{N}(\alpha_j^{oblast} + \alpha_t^{year} + \alpha \cdot (U_{jt} - \bar{U}_j);\ \sigma^2_{\alpha,oblast-year})$$

$$\alpha_j^{oblast} \sim \text{N}(\gamma_0 + \gamma_1 \cdot \bar{U}_j + \gamma_2 \cdot \bar{U}_j \text{Area};\ \sigma^2_{\alpha,oblast})$$

$$\alpha_t^{year} \sim \text{N}(\delta_0 + \delta_1 \cdot GDP_t^{country};\ \sigma^2_{\alpha,year})$$

$$\beta_{jt}^{ED,oblast-year} \sim \text{N}(\beta_j^{ED,oblast} + \beta_t^{ED,year};\ \sigma^2_{\beta,ED,oblast-year})$$

$$\beta_j^{ED,oblast} \sim \text{N}(\zeta_0 + \zeta_1 \cdot \bar{U}_j;\ \sigma^2_{\beta,ED,oblast})$$

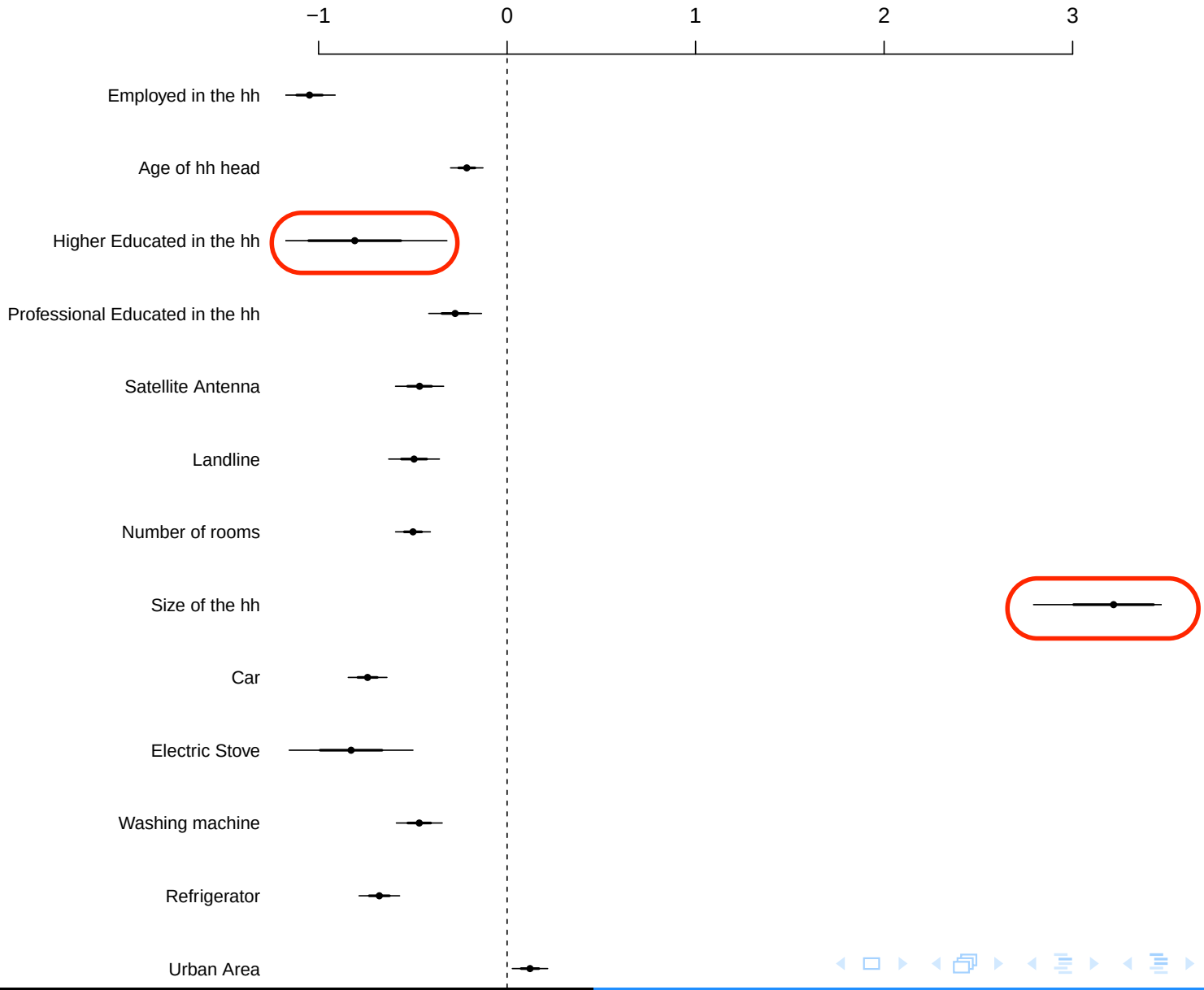
$$\beta_t^{ED,year} \sim \text{N}(\xi_0 + \xi_1 \cdot \text{time};\ \sigma^2_{\beta,ED,year})$$

$$\beta_{jt}^{S,oblast-year} \sim \text{N}(\beta_j^{S,oblast} + \beta_t^{S,year};\ \sigma^2_{\beta,S,oblast-year})$$

$$\beta_j^{S,oblast} \sim \text{N}(\phi_0;\ \sigma^2_{\beta,S,oblast})$$

$$\beta_t^{S,year} \sim \text{N}(\psi_0;\ \sigma^2_{\beta,S,year})$$

Estimated individual coefficients



Main results of the estimated model I

- **Individual factors that negatively affect the probability of being poor** (roughly in order of 'importance'):
  - Share of employed members in the household
  - Share of highly educated members in the household
  - Availability of an electric stove (stationary)
  - Availability of refrigerator
  - Number of rooms
- **Individual factors that positively affect the probability of being poor** (roughly in order of 'importance'):
  - Household size
  - Age (over 50) of the household head
  - Gender (female) of the household head

Time-space varying intercepts

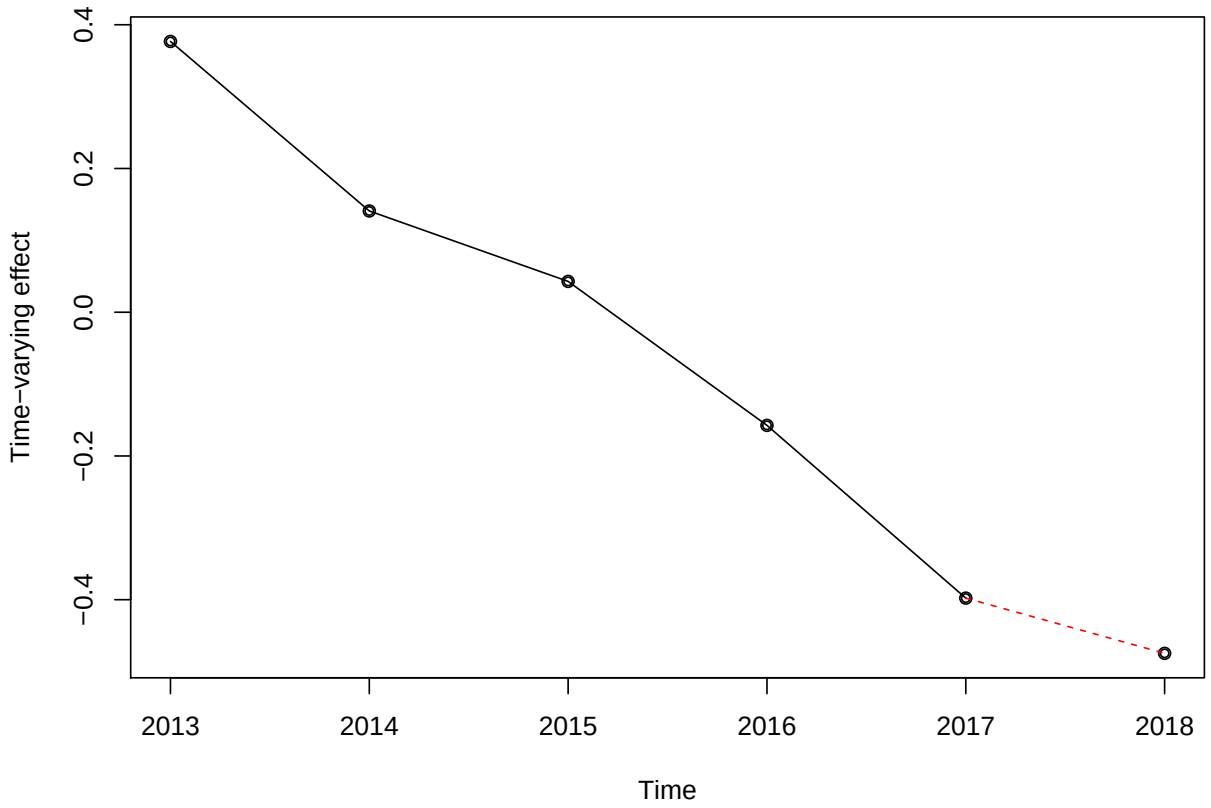
Time-space varying intercepts:

$$\alpha_{jt}^{oblast-year} = -1.71 + 0.75 * \bar{U}_j - 0.31 * \bar{U}_j * \text{Urban} +$$

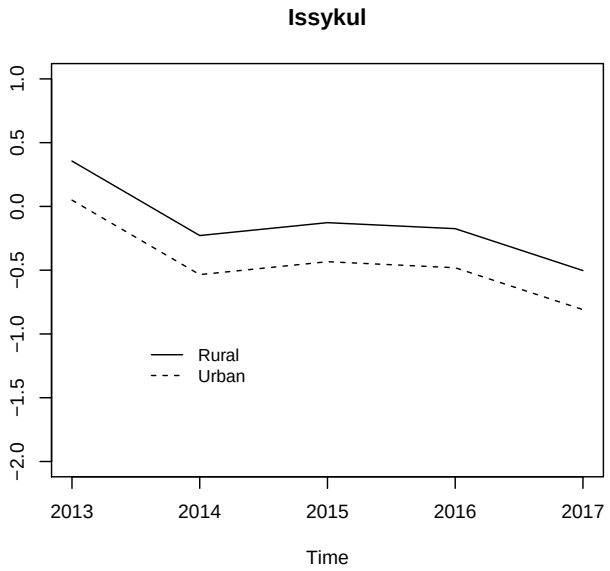
$$+0.08 * (U_{jt} - \bar{U}_j) - 0.058 * GDP_t + \text{errors}$$

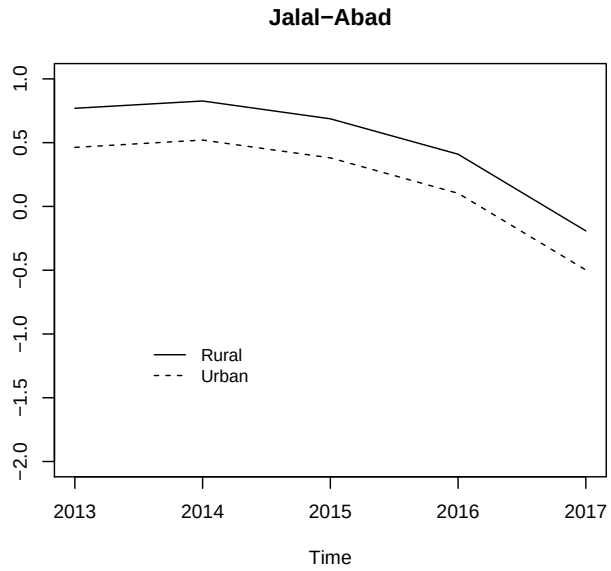
- Unemployment rate matters substantially (0.75) for poverty and this association holds cross-sectionally and less (0.08) longitudinally;
- That is, there is a significant effect of enduring differences in oblast's level of unemployment, but the longitudinal variation in the level of unemployment (measured as deviation from the regional mean) over the period is weakly associated with variation in poverty risk;
- Interaction between unemployment and area of residence: if the area of residence is **urban** the impact of the regional unemployment rate in the probability to be poor is lower (0.75-0.31).
- GDP: higher per capita GDP decreases the individual probability to be poor (trickle-down effect);

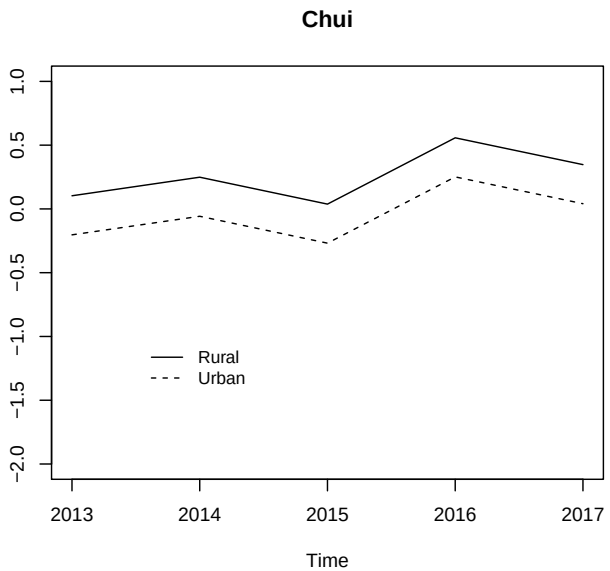
Time-varying effect on π_i due to GDP growth

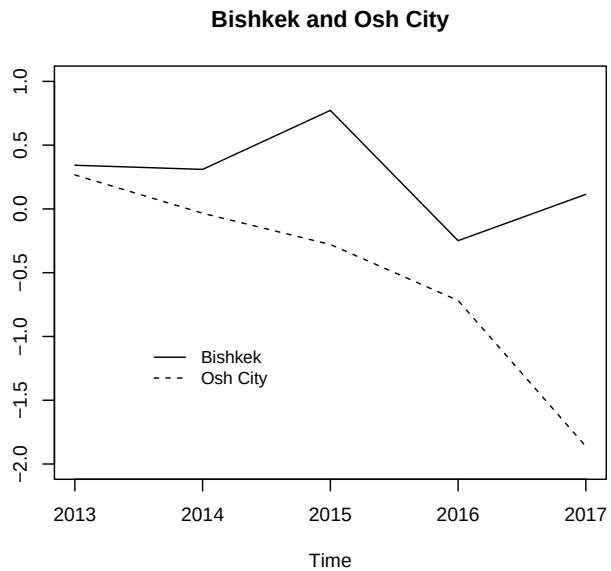


Time-varying overall effect on π_i by oblast









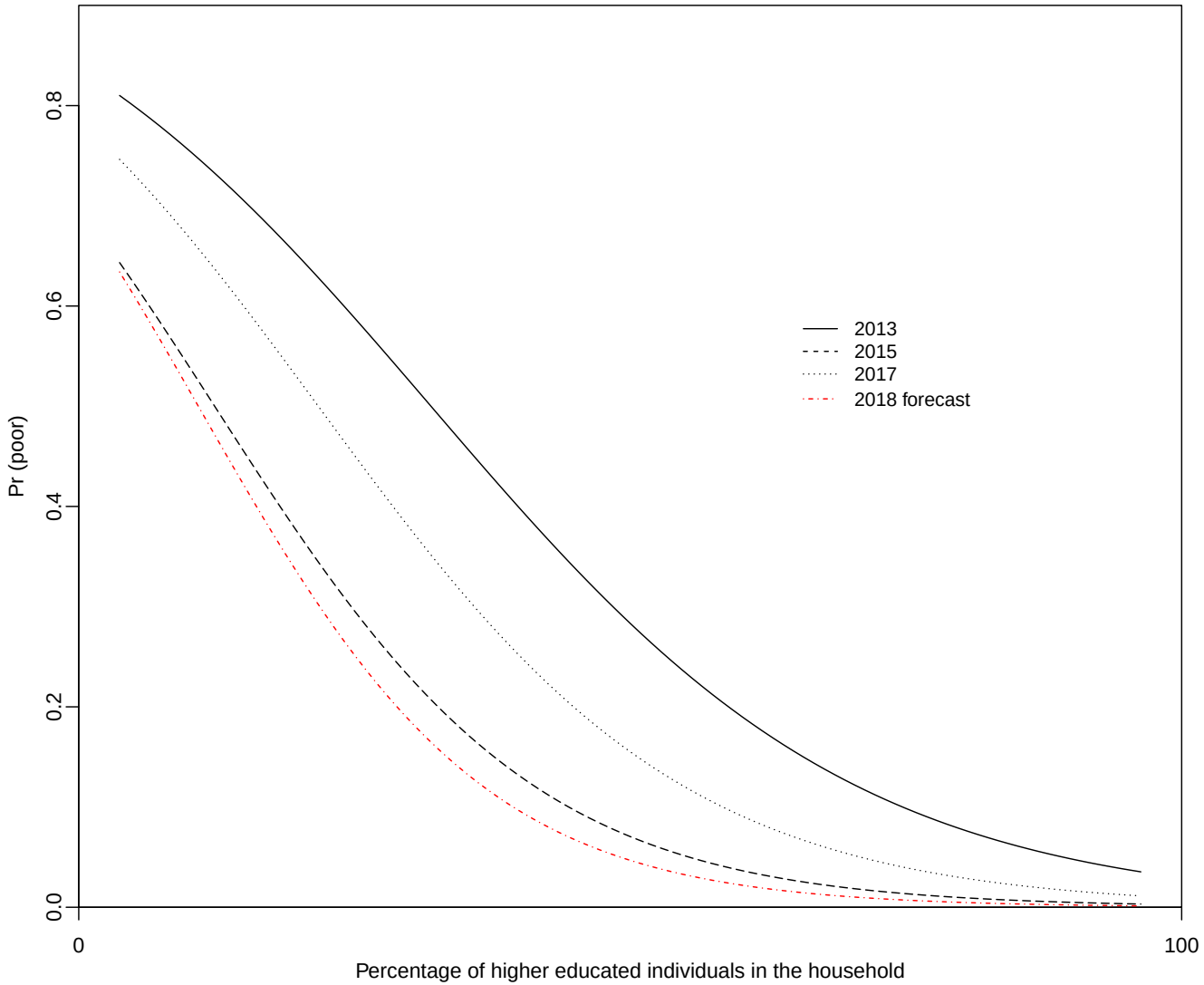
Time-space varying slopes

Time-space varying slopes (education):

$$\beta_{jt}^{ED,oblast-year} = -0.81 + 0.21 * \bar{U}_j - 0.18 * time + \text{errors}$$

- the average effect of education is negative (-0.81): the increase of the number if high educated people in the hh decreases the probability of been poor;
- Positive effect (+0.21) of the oblast unemployment rate: where the unemployment rate is high, the impact of having educated members in the hh is less important;
- The negative effect (-0.18) on the probability to be poor of having educated members in the household **grows over time**: that is education matters more and more;

Probabilities to be poor of an average 5-members hh by level of education over time

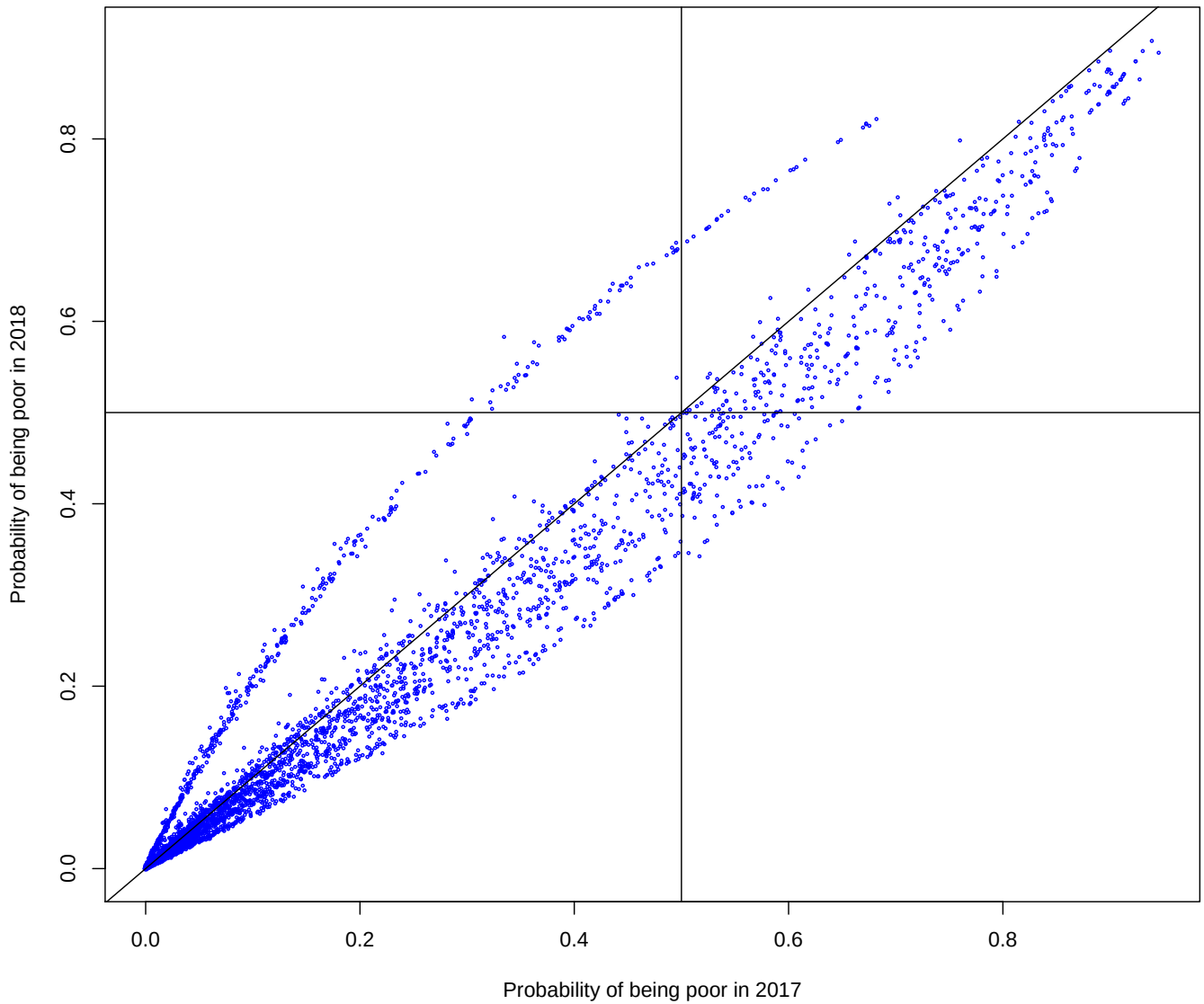


Vulnerability and Mobility Analysis

Vulnerability and mobility and analysis

- The dynamic model estimates for each household *i* of the sample the probability of being poor: **in-sample predictions**;
- **out-of-sample predictions** refer instead to the year 2018 and are the probabilities of been poor in 2018 conditioned on:
  - invariant characteristics of the households in 2017;
  - time-varying effects of some individual predictors;
  - macroeconomic forecast of unemployment rate and GDP.

Vulnerability to poverty: estimated probabilities in 2017 and forecast probabilities in 2018



Mobility

- Based on these predicted probability, we can estimate the number of people that transit **in and out** poverty status;
- Cut-off value equal to 0.5: when the predicted probability of being poor is above 50% the household is predicted to be poor;
- The 2017–2018 estimated **household mobility matrix**:

	2018		
2017	Not poor	Poor	
Not poor	98.7%	1.3%	100%
Poor	**25.2%**	74.8%	100%

- The 2017–2018 estimated **population mobility matrix**:

	2018		
2017	Not poor	Poor	
Not poor	97.0%	3.0%	100%
Poor	**19.3%**	80.7%	100%

Conclusions

Remarks

- The proposed methodology overcomes some strong assumptions that underlie the existing methodologies to evaluate poverty dynamics in absence of panel data;
- Multi-level models are flexible models that allow one to introduce micro and macro predictors in an appropriate statistical framework along with time-varying effects of micro predictors;
- Estimation of these dynamics models within a Bayesian statistical framework offers a reasonable solution for the convergence problems.

Further analysis

- Decomposition of the error term: how much of the total error is due to idiosyncratic shocks and how much is due instead to aggregate shocks;
- Vulnerability analysis under different scenarios: what happens to the vulnerability to poverty if for example the unemployment rate increases over time and/or in some regions?
- Introduction of lagged variables at macro level.