MODELLING OF THE RISK OF MONETARY POVERTY IN THE CZECH REGIONS

Jitka Bartošová – Marie Forbelská

Abstract

The European Union Statistics on Income and Living Conditions (EU-SILC) is the main source of information about poverty and economic inequality in the member states of the European Union. The sample sizes of its annual national surveys provide reliable information not only at national but at the sub-national (e.g. regional) level too. The article deals with cluster analysis of regional household income dynamics via mixture models. We focus on modelling empirical curves using a set of models based on clustering algorithms known as regression mixtures. We apply generalized linear mixed models on each Czech NUTS3 region. The R environment (R Development Core Team, 2010) is used for the mixture model analysis.

Key words: generalized linear mixed models, linear mixed models, monetary poverty

JEL Code: D31, I32, R10

Introduction

Poverty currently presents serious social and economic problem in both developing and developed countries. When comparing poverty rates in advanced countries, and now in all countries of the EU, risk-of-poverty rate is most frequently used. This is represented by a percentage share equivalent disposable income lower than the poverty line of all the given number of groups of individuals. It acts as a relative measurement which evaluates the financial security of households (individuals) with respect to the national level.

The current financial crises affect the poorest households who have incomes under the line of relative poverty or near it. Poverty in the Czech Republic affects the ‘lower’ strata of our society, those with a worse approach to the labour market. It is understandable that there are regional differences in poverty due to the fact that in regions with a higher concentration of these risk factors it must be expected to find a higher rate of poverty and unemployment (see e.g. Bartošová and Forbelská, 2010, Sipková and Sipko, 2010, Stankovičová, 2010, Želinský, 2010a,b). Regional disparity concerning the financial potential and poverty of its inhabitants is connected to the development of the individual regions, their economic and demographic structure . It is necessary to recognize that though there may be many hidden cause of poverty, which will be shown in the problems in which we will successfully classify the limits of the subgroups (cluster) with similar financial situations. This will enable us to forecast the whole spectrum of factors which affects the unfortunate situation in the regions, and consequently find a way for the leaders to improve or eliminate the problem.

The EU-SILC (European Union Statistics on Income and Living Conditions) is an instrument aiming at collecting timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions. This instrument is anchored in the European Statistical System (ESS). For the first time this investigation was carried out by the Czech Statistical Office in 2005 under the name Living Conditions 2005. Investigation is carried out by the so-called rotating panel, where the same households were re-interviewed in the annual intervals for four years. After this time are replaced by other households living in the newly visited homes that are added to the investigation file continuously by the random selection. Longer monitoring of a household permits building image of their social situation, not only in the year, but also the changes and developments over time (see e.g. Marek, 2010, Bílková, 2012, Bílková and Malá, 2012, Malá, 2012, Řezanková, Loster, 2011).

The analysis has been carried out on household income, adjusted for different household types using an equivalence scale. The equivalized household income is obtained by dividing the available household income by the number of consumption equivalents in the household. It is assumed that, as the size of the household increases and depending on the age of the children, cost savings are achieved in the household through joint budgeting (economies of scale). For weighting purposes, the EU scale (modified OECD scale) is used to calculate a household's resource requirements. An adult living on his or her own is taken as the reference point (consumption equivalent), with an allocated weighting of 1. For each additional adult, the assumed resource requirement increases by 0.5 consumption equivalents. Each child under the age of 14 is weighted with a consumption equivalent of 0.3. So a household comprising a father, mother and child would have a calculated consumption equivalent of 1.8 compared with a single-person household. Formally, suppose a household comprises H_1 adults and H_2 children. Then its equivalized household size (eHs) is $1.0 + 0.5H_1 + 0.3H_2$.

1 Models Specification

Modelling clustered and longitudinal data with and without nested factors has gained importance in recent years. Early exposition is e.g. the book by McCulloch and Searle (2001), which deals primarily with linear mixed models (LMMs). Hierarchical linear model (HLM) or multi-level formulations can be rewritten as LMMs. Extension to generalized LMM (GLMM) is considered in Molenberghs and Verbeke (2005) and an up-to-date mathematical treatment is given by Jiang (2007). A Bayesian perspective of HLMs is taken in Gelman and Hill (2006).

1.1 Linear Mixed Models (LMM)

Linear mixed models extend classical linear models by incorporating random effects in the structure. Assume that the data set at hand consists of N subjects (here households). The general linear mixed model is specified as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i , \qquad (1)$$

where $\mathbf{Y}_i = (Y_{i1},...,Y_{in_i})'$ is the vector of n_i observations for the ith subject (household), $1 \le i \le N$. Vector $\boldsymbol{\beta} = (\beta_1,...,\beta_p)'$ contains the p fixed-effects parameters (β_j, $j = 1,..., p$, are fixed, but unknown regression parameters, common to all subjects). $\mathbf{b}_i = (b_{i1},...,b_{iq})'$ is the vector with the random effects for the ith subject in the data set. The use of random effects reflects the belief that there is heterogeneity among subjects for a subset of the regression coefficients in **β**. $\mathbf{X}_\mathrm{i}$ $(n_i \times p)$ and $\mathbf{Z}_\mathrm{i}$ $(n_i \times q)$ are the design matrices for the p fixed and q random effects, and $\boldsymbol{\varepsilon}_i$ contains the residual components for ith subject. Independence between subjects is assumed. $\mathbf{b}_\mathrm{i}$ and $\boldsymbol{\varepsilon}_\mathrm{i}$ also are assumed to be independent and normally distributed with mean vector **0** and covariance matrices, $\mathbf{D}$ $(q \times q)$ and $\boldsymbol{\Sigma}_\mathrm{i}$ $(n_i \times n_i)$, respectively. Then $\mathbf{Y}_i$ has a marginal normal distribution with mean $\mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $\mathbf{V}_i = Var(\mathbf{Y}_i)$, where $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}_i\mathbf{Z}_i' + \boldsymbol{\Sigma}_i$.

It becomes clear that the fixed effects determine only the mean $E(\mathbf{Y}_i)$, and the inclusion of subject–specific effects mediates structure of the covariance between observations on the same unit. Assuming a normal distribution of $\mathbf{Y}_i|\mathbf{b}_i$ and $\mathbf{b}_i$

($\mathbf{Y}_i | \mathbf{b}_i \sim N(\mathbf{X}_i + \mathbf{Z}_i\mathbf{b}_i, \mathbf{\Sigma}_i)$, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$), it becomes clear that the residual terms model variability within a subject.

If we denote the unknown parameters in the covariance matrix $\mathbf{V}_\text{i}$ as $\boldsymbol{\psi}$, then a closed–form expression for the maximum likelihood estimator of $\boldsymbol{\beta}$ exists, and has the form

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{V}_1^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{V}_1^{-1} \mathbf{Y}_i \,. \tag{2}$$

To predict the random effects, the mean of the posterior distribution of the random effects given the data, $\mathbf{b}_i | \mathbf{Y}_i$, is used. Conditional on $\boldsymbol{\psi}$, we get

$$\widehat{\mathbf{b}} = \mathbf{D}\mathbf{Z}_i' \mathbf{V}_1^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}). \tag{3}$$

Estimation of $\boldsymbol{\psi}$ is mostly performed using of maximum likelihood (ML) or restricted maximum likelihood (REML) methods. The expression maximized by the ML *(l_1)*, or REML *(l_2)* estimates is given by

$$l_1(\boldsymbol{\psi}; \mathbf{y}_1, ..., \mathbf{y}_N) = c_1 - \frac{1}{2}\sum_{i=1}^{N} \log(|\mathbf{V}_i|) - \frac{1}{2}\sum_{i=1}^{N} \mathbf{r}_i' \mathbf{V}_1^{-1} \mathbf{r}_i \,, \tag{4}$$

$$l_2(\boldsymbol{\psi}; \mathbf{y}_1, ..., \mathbf{y}_N) = c_1 - \frac{1}{2}\sum_{i=1}^{N} \log(|\mathbf{V}_i|) - \frac{1}{2}\sum_{i=1}^{N} \mathbf{r}_i' \mathbf{V}_1^{-1} \mathbf{r}_i - \frac{1}{2}\sum_{i=1}^{N} \log(|\mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i|), \tag{5}$$

where $\mathbf{r}_i = \mathbf{y}_\text{i} - \mathbf{X}_i' \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{V}_1^{-1} \mathbf{y}_i\right)^{-1}$ and *c_1, c_2* are appropriate constants. Equations (4) and (5) are maximized using iterative numerical techniques such as Fisher scoring or Newton–Raphson (for details, see Demidenko, 2004). In equations (2) and (3) the unknown $\boldsymbol{\psi}$ is then replaced with $\widehat{\boldsymbol{\psi}}_{ML}$ or $\widehat{\boldsymbol{\psi}}_{REML}$. For inference regarding the fixed and random effects and the variance components, appropriate likelihood ratio and Wald tests are suitable (see Verbeke and Molenberghs (2001)).

The predictor for the conditional expectation $E(\mathbf{Y}_i | \mathbf{b}_i) = \boldsymbol{\mu}_{i|\mathbf{b}_i} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ is obtained from equation (2) and (3). We get a weighted average of $\mathbf{X}_i\widehat{\boldsymbol{\beta}}$ (related to the whole population) and $\mathbf{Y}_\text{i}$ (related to subject *i*) $\widehat{\boldsymbol{\mu}}_{i|\mathbf{b}_i} = \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'\mathbf{V}_1^{-1}(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}) =$
$= (\mathbf{I}_{n_i} - \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'\mathbf{V}_1^{-1}) + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'\mathbf{V}_1^{-1}\mathbf{Y}_i = \mathbf{\Sigma}_i\mathbf{V}_1^{-1}\mathbf{X}_i\widehat{\boldsymbol{\beta}} + (\mathbf{I}_{n_i} - \mathbf{\Sigma}_i\mathbf{V}_1^{-1}\mathbf{X}_i\mathbf{V}_1^{-1}\mathbf{X}_i)\mathbf{Y}_i \,.$

1.2 Generalized Linear Mixed Models (GLMM)

Generalized linear models (GLMs) are, as the name suggests, a generalization or extension of normal linear model. (LMs are a special case of GLMs.) GLMs also allow model other error distributions (binomial, Poisson, negative binomial, or gamma distribution). Nelder and Wedderburn (1972) were the first to propose the generalized linear model to encompass these different models under one unified mathematical framework.

The generalized linear mixed model (GLMM) (see McCullagh and Nelder, 1989) consists of three parts – a link function, a linear predictor, and a distributional model.

Given $\mathbf{b}_i = (b_{i1},...,b_{iq})'$, the variables $\mathbf{Y}_i = (Y_{i1},...,Y_{in_i})'$ are mutually independent with a density function (from the exponential family of distribution) given by

$$f\left(y_{ij}\middle|\mathbf{b}_i,\boldsymbol{\beta}\right) = \exp\left\{\frac{y_{ij}\theta_{ij} - a(\theta_{ij})}{d_{ij}(\phi)} + c(y_{ij},\phi)\right\}, \tag{6}$$

where θ_{ij} is the canonical parameter and ϕ is the scale parameter. The functions d_{ij} and c are specific to each distribution.

The conditional mean and the conditional variance of Y_{ij} are given by

$$E\left(\mathbf{Y}_{ij}\middle|\mathbf{b}_i\right) = \boldsymbol{\mu}_{ij|\mathbf{b}_i} = g^{-1}(\eta_{ij}) = g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i), \tag{7}$$

$$Var\left(\mathbf{Y}_{ij}\middle|\mathbf{b}_i\right) = v(\boldsymbol{\mu}_{ij|\mathbf{b}_i})d_{ij}(\phi), \tag{8}$$

where g and v are the link and the variance function, $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are the j-th row of the matrix $\mathbf{X}_i$ and $\mathbf{Z}_i$.

The random effects $\mathbf{b}_1, \ldots, \mathbf{b}_N$, are mutually independent with a common underlying distribution G which depends on the unknown parameter $\boldsymbol{\psi}$. Next the vector of random effects $\mathbf{b}_\mathrm{i}$ is assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{D}$.

2 Modelling the Impact of Regions on the Risk of Poverty Rate in the Czech Republic between 2005 and 2008

The European Union Survey on Income and Living Conditions (EU-SILC) provides reliable statistics at national level but sample sizes do not allow reliable estimates at sub-national level, despite a rising demand from policy makers and local authorities. The standard poverty rate used in this section (60% of the national median equivalized disposable income) is a relative definition as it depends on the average income of the country. But the national value at-risk-of-poverty rate includes regional differences. In this paper we used generalized linear mixed models for detection of these regional disparities.

To estimate the influence of regions on the risk of monetary poverty in the Czech Republic we used a generalized linear regression model with mixed effects (GLMM), specifically, it was a logistic regression model with mixed effects (Logistic Mixed Effect Model – LMEM).

Tab. 1: Maximum likelihood estimates of the random effects caused by the Czech NUTS3 regions (2005 – 2008)

Regions	2005		2006		2007		2008	
11 Capital Prague	-0.877	**1.**	-0.445	**1.**	-0.802	**1.**	-0.217	5.
21 Central Bohemian	-0.113	5.	-0.168	5.	-0.085	4.	0.117	8.
31 South Bohemian	-0.479	**2.**	-0.037	8.	-0.080	5.	-0.175	7.
32 Pilsen	-0.295	4.	-0.367	**2.**	-0.268	**2.**	-0.539	**1.**
41 Carlsbad	-0.041	6.	0.104	10.	0.035	8.	0.230	11.
42 R. of Ústí nad Labem	0.282	11.	0.384	13.	0.163	10.	0.134	9.
51 Liberec	-0.450	3.	-0.107	6.	0.155	9.	-0.297	3.
52 R. of Hradec Králové	0.169	9.	-0.212	4.	-0.203	3.	-0.283	**2.**
53 Pardubice	0.261	10.	0.003	9.	0.164	11.	-0.264	4.
61 Highlands	0.166	8.	-0.102	7.	-0.070	6.	-0.215	6.
62 South Moravian	0.022	7.	0.236	11.	-0.008	7.	0.391	13.
71 Olomouc	0.644	**14.**	0.680	**14.**	0.560	**14.**	0.586	**14.**
72 Zlín	0.364	12.	-0.225	3.	0.223	12.	0.196	10.
81 Moravian-Silesian	0.374	13.	0.273	12.	0.243	13.	0.361	12.

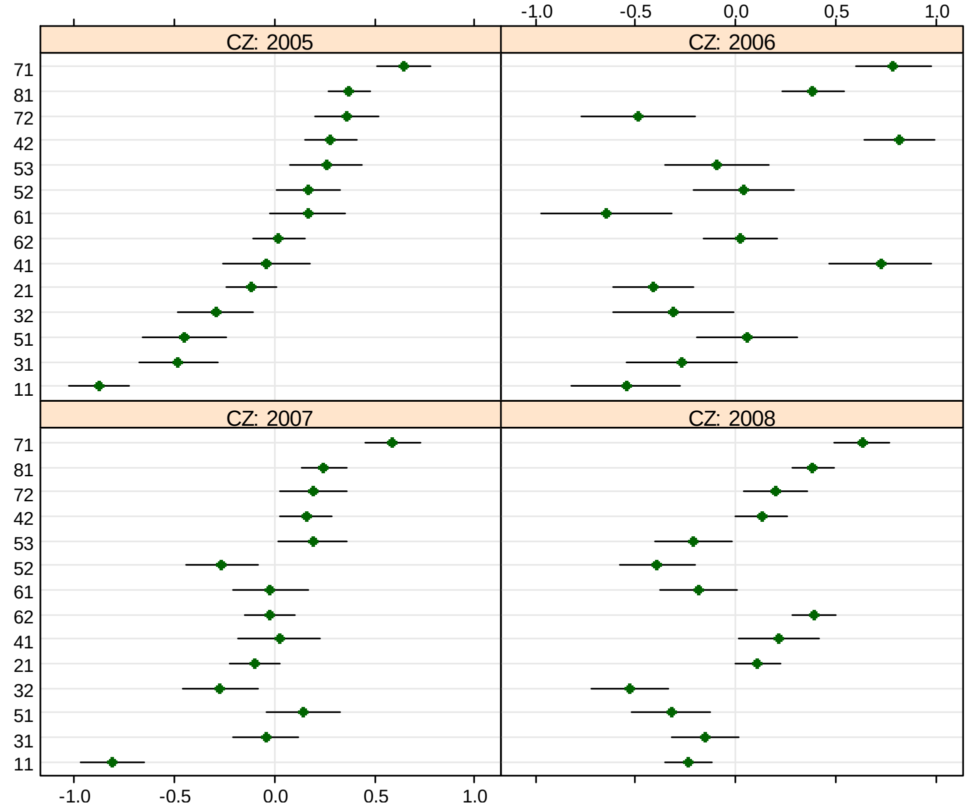
Source: Own calculations; data – EU SILC.

Table 1 shows maximum likelihood estimates of the random effects caused by the Czech NUTS3 regions between 2005 and 2008. The negative sign represents a positive impact on the poverty in region (reduction of the risk of poverty), positive sign represents increasing of the risk of poverty. Maximum likelihood estimates of random effects for each region in Bohemia were obtained using the software package *lme4* in R program.

Time evolution of the regional effects on the risk of monetary poverty during the period 2005 – 2008 is shown in Figure 1. Regions are ranked according to the results from

2005, so you can see the development that occurred during this period. It is changed both – the order of effects (depending on size) and their variability too.

Fig. 1: Interval estimates of random effects caused by the Czech NUTS3 regions (2005 – 2008).



Source: Own calculations; data – EU SILC.

Conclusion

Through the generalized linear regression models with mixed effects (GLMM), specifically using logistic regression with mixed effects (LMEM) was estimated impact of the regions in the Czech Republic at the risk of monetary poverty. The results also were used as a tool for monitoring the development of regional impact on poverty in the Czech Republic in the years 2005 – 2008. As expected – in terms of monetary poverty was best to live in Prague. Very good impact on poverty reduction was also Pilsen Region. In 2008, the estimate of random effect was even better in the case of the Pilsen Region than in the case of Prague.

Acknowledgment

The article was partially supported by project IG F6/3/2012 “Situation of the Juniors and Seniors in the Czech Republic” of the Internal Grant Agency of University of Economics in Prague, and by project VEGA 1/0127/11 “Spatial Distribution of Poverty in the European Union” of the Slovak Scientific Grant Agency.

References

Bartošová, Jitka, and Marie Forbelská. 2010. “ Comparison of Regional Monetary Poverty in the Czech and Slovak Republic.” In *Social Capital, Human Capital and Poverty in the Regions of Slovakia*, ed. Iveta Pauhofová, Oto Hudec, and Tomáš Želinský, 76–84. Košice: Technical University of Košice.

Bílková, Diana. 2012. „Recent Development of the Wage and Income Distribution in the Czech Republic”. *Prague Economic Papers, 2012(2)*: 233-250,

Bílková, Diana, and Ivana Malá. 2012. “Modelling the Income Distributions in the Czech Republic since 1992”. *Austrian Journal of Statistics, 41(2)*: 133–152.

Demidenko, Eugene. 2004. *Mixed models: theory and applications*. NewYork: John Wiley & Sons.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Jiang, Jimming. 2007. *Linear and Generalized Linear Mixed Models and Their Applications*. Berlin: Springer.

Malá, Ivana. 2012. „Použití konečných směsí pro modelování příjmových rozdělení“. *Acta Oeconomica Pragensia, 20(4)*: 26-39.

Marek, Luboš. 2010. „Analýza vývoje mezd v ČR v letech 1995–2008“. *Politická ekonomie* 58(2): 168–206 .

McCullagh, Peter, and John A. Nelder. 1994. *Generalized Linear Models*. London: Chapman and Hall.

McCulloch, Charles E., and Shayle R. Searle. 2001. *Generalized, Linear, and Mixed Models*. New York: Wiley.

Molenberghs, Geert, and Geert Verbeke, G. 2005. *Models for Discrete Longitudinal Data*. Berlin: Springer.

Nelder, John A., and Robert W. M. Wedderburn. 1972. “Generalized linear models”. *Journal of the Royal Statistical Society - Series A 135(3)*: 370–384.

Řezanková, Hana, and Tomáš Loster. 2011. "Analysis of the Dependence of the Housing Characteristics on the Household Type in the Czech Republic". *APLIMAT – Journal of Applied Mathematics, 4(3)*: 351–358.

Sipková, Ľubica, and Jurnaj Sipko. 2010. "Wage Levels in the Regions of the Slovak Republic." In *Social Capital, Human Capital and Poverty in the Regions of Slovakia*, ed. Iveta Pauhofová, Oto Hudec, and Tomáš Želinský, 51–66. Košice: Technical University of Košice.

Stankovičová, Iveta. 2010. "Regional Aspects of Monetary Poverty in Slovakia." In *Social Capital, Human Capital and Poverty in the Regions of Slovakia*, ed. Iveta Pauhofová, Oto Hudec, and Tomáš Želinský, 67–75. Košice: Technical University of Košice.

Verbeke, Geert, and Geert Molenberghs. 2001. *Linear Mixed Models for Longitudinal Data*. Berlin: Springer.

Želinský, Tomáš. 2010a. "Analysis of Poverty in Slovakia Based on the Concept of Relative Deprivation." *Politická ekonomie, 58(4)*: 542–565 .

Želinský, Tomáš. 2010b. "Regions of Slovakia from the View of Poverty." In *Social Capital, Human Capital and Poverty in the Regions of Slovakia*, ed. Iveta Pauhofová, Oto Hudec, and Tomáš Želinský, 37–50. Košice: Technical University of Košice.

Contact

Jitka Bartošová

University of Economics in Prague, Faculty of Management, Department of Information Management, Jarošovská 1117/II, 37701 Jindřichův Hradec, Czech Republic

bartosov@fm.vse.cz

Marie Forbelská

Masaryk University in Brno, Faculty of Science, Department of Mathematics and Statistics, Kotlářská 2, 611 37 Brno, Czech Republic

forbel@math.muni.cz