STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. 175–191, DOI 10.21307/stattrans-2021-033
Received – 24.11.2019; accepted – 07.06.2021

Modelling the occupational and educational choices of young people in Poland using Bayesian multinomial logit models

Wioletta Grzenda[1]

ABSTRACT

Binomial logit models are commonly used in the analysis of the situation of respondents on the labour market. Consequently, in most cases researchers consider two states: of being unemployed and employed or economically inactive and active. This paper focuses on the situation of young people aged 18 to 29 on the labour market in Poland. A major part of the people who comprise the studied group are still in education or combine education with work. Therefore, the participants of the research were divided into the following groups: the employed and not learning, those combining education with work, the unemployed, learners/students only, and those economically inactive and not at school. The model allowing an analysis which includes both the most common division into working and non-working persons as well as the division proposed in this study is a nested logit model. This model has a hierarchical structure and is a special case of a multinomial logit model. In this paper, all models were estimated within the Bayesian approach. The findings show that continuing education by young people may result from their problems with finding a job; moreover, combining work with education is not the preferred form of professional activity. In addition, the study examines the inequalities observed on the Polish labour market.

Key words: young people, labour market, education, multinomial logit model, Bayesian approach.

1. Introduction

In socio-economic research, models for the dichotomous dependent variables are very popular (Cramer, 2003; Allison, 2009). Unfortunately, with their use, only two states or two events for a given unit can be analysed. In the case of issues related to the labour market, division into economically active and economically inactive, as well as employed and unemployed persons is usually made. In the case where the examined feature has more than two levels, a better solution than combining selected categories is to use models for discrete outcome variables that can take more than two possible

[1] SGH Warsaw School of Economics, Collegium of Economic Analysis, Institute of Statistics and Demography, Poland. E-mail: wgrzend@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-2226-4563.

values. Among this group of models, two main classes are distinguished: models for ordinal response variables and models for dependent variables with unordered categories. The second group includes: the MultiNomial Logit Model (MNLM), the Conditional Logit Model (CLM), the Mixed Logit Model (MLM) and the Nested Logit Model (NLM) (Cameron and Trivedi, 2005).

The choice of a model depends primarily on whether the independent variables included in the model vary across alternatives or they are the same across alternatives (Cameron and Trivedi, 2005). The standard multinomial logit model can be used when the model takes into account only the features of the individuals studied without taking into account the features of the selected categories. If this assumption is not met, the conditional logit model is used unless both types of features are considered. In the latter case the mixed logit model is used. In addition, according to Stanisz (2016), in order for the standard multinomial model to be used, the categories of the responding variable should be independent and distinguishable for the decision maker. Both the multinomial logit model and the conditional logit model have some limitations regarding the assumption of independence from irrelevant (unrelated) alternatives (IIA). The model in which this assumption can be slightly weakened, and also can take into account the hierarchy of alternatives, is the nested logit model considered in this paper. This model is not widely used due to the problems related to the estimation of its parameters. To avoid these problems, the Bayesian approach and Markov Chain Monte Carlo methods (MCMC) were used in this work (Robert and Casella, 2004).

The purpose of this study is to analyse the occupational and educational choices of young people aged 18 to 29 in Poland. In most studies on this issue, the division of young people into those who have already completed education and those who continue their education, e.g. at a higher level (de Dios Jiménez and Salas-Velasco, 2000), economically active and economically inactive (MRPiPS, 2018) or unemployed and employed (Gallie and Paugam, 2000; Grzenda, 2012; Bieszk-Stolorz and Markowicz, 2013) are considered. The binary divisions presented above can be further detailed. For example, among the economically inactive there are both those who are unwilling to take up employment despite their abilities and young people who remain in the education system and have not started their careers yet. In addition, it is worth considering in the research that young people sometimes combine education with work. Therefore, in this study, the respondents were divided into employed, combining education with work, learners only, and unemployed or persons economically inactive but not being learners. The methodological approach proposed in this work makes it possible to consider in the analysis both a more general division into working and non-working persons, as well as a more detailed division taking into account education of youth. Information on educational and economic activity of young people in Poland was obtained from the Labour Force Survey (LFS).

The subject addressed in this study is very important because according to many reports (CSO, 2016a; CSO, 2016b; MRPiPS, 2018) the situation of young people on the labour market in Poland is the worst compared to other age groups. In addition, economists are concerned about the growing phenomenon of NEET (not in employment, education or training) (Chłoń-Domińczak and Strawiński, 2013), which affects young people who are neither in education nor working. The consequences of this phenomenon apply to the entire economy as well as to individuals who lose their competence over time. Youth unemployment has also a social dimension, lack of employment negatively affects family and fertility decisions, and, as a result, the demographic situation of the country. Therefore, the identification of factors determining the educational and professional decisions of young people may help identify solutions that may improve the situation of these people on the labour market in Poland.

2. Multinomial models

Models for unordered categorical dependent variable are also considered as discrete choice models and are most often used in marketing research (Anderson, De Palma and Thisse, 1992). In the case of the binomial logit model, it can be assumed that a given unit has two variants to choose from. Suppose now that the i-th unit $(i = 1, \dots, n)$ has to select not two but J unordered categories. These categories are mutually exclusive and constitute a whole set of possible selection options for the units under consideration. In the case where the independent variables do not differ for the alternatives considered, a standard multinomial logit model (MNLM) is considered. For this model, the probability of observing the choice by the i-th unit $(i = 1, \dots, n)$ of j-th category $(j = 1, \dots, J)$ is given by the formula:

$$p_{ij} = \frac{exp(\mathbf{x}'_i \boldsymbol{\beta}_j)}{\sum_{k=1}^{J} exp(\mathbf{x}'_i \boldsymbol{\beta}_k)}, i = 1, \dots, n, j = 1, \dots, J,$$

where $\mathbf{x}$ denotes the vector of independent variables and $\boldsymbol{\beta}$ is the vector of parameters. The sum of these probabilities for all categories $j = 1, \dots, J$ is 1.

If the independent variables differ for the alternatives considered, the standard multinomial model cannot be used; the conditional logit model (CLM) is considered then. In the case of this model, the probability of observing the selection of the j-th category $(j = 1, \dots, J)$ by the i-th unit $(i = 1, \dots, n)$ is given by the formula:

$$p_{ij} = \frac{exp(\mathbf{x}'_{ij} \boldsymbol{\beta})}{\sum_{k=1}^{J} exp(\mathbf{x}'_{ik} \boldsymbol{\beta})}, i = 1, \dots, n, j = 1, \dots, J.$$

The combination of both considered models is the mixed logit model (MLM) (Cameron and Trivedi, 2005).

The presented models can also be considered more generally in the context of the additive random utility models (ARUM) and discrete choice theory. In this approach, each unit assigns to each category j certain utility $U_j, j = 1, \dots, J$ and selects the one with the highest utility. Let

$$U_{ij} = \mathbf{x'}_{ij}\boldsymbol{\beta} + \varepsilon_{ij}, i = 1, \dots, n, j = 1, \dots, J,$$

denote the utility function. By making different assumptions about the random component of utility, different multinomial logit models can be obtained.

In the standard multinomial logit model, the random components ε_j $(j = 1, \dots, J)$ are independent and identically Gumbel distributed (have the type I extreme-value distribution), with the density function given by the formula:

$$f(\varepsilon_j) = e^{-\varepsilon_j} exp(-e^{-\varepsilon_j}), j = 1, \dots, J.$$

According to assumptions made in (McFadden, 1974), to be able to use a standard logit multinomial model, the categories analysed must meet the assumption of independence from irrelevant alternatives (IIA). This assumption also applies to the conditional logit model. However, it is often not fulfilled. By eliminating or adding one alternative, the quotient of the probability of the categories considered so far often changes. Unfortunately, there are no tests that conclusively determine whether IIA assumption is met. Cheng and Long (2007) have shown that two existing tests by Hausman and McFadden (1984) and Small and Hsiao (1985) can be unreliable. Then the solution may be to use another model, namely the nested logit model (Train, 2009).

The nested logit model has a hierarchical structure. The set of all possible alternatives is divided into the so-called nests so that the assumption of independence from irrelevant alternatives (IIA) is met only in each nest, but it does not have to be met between the nests. Therefore, in the nested logit model, all random components ε_{ij} $(j = 1, \dots, J)$ do not have to be independent. In addition, instead of the Gumbel distribution, the generalized extreme-value distribution (GEV) is assumed for these components.

Let K denote the number of disjoint subsets (nests) $S_1, S_2, \dots, S_K$, into which the possible alternatives have been divided. Then, the cumulative distribution function for the random components vector $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iJ})$, is given by the formula:

$$F(\boldsymbol{\varepsilon}_i) = exp\left(-\sum_{k=1}^{K}\left(\sum_{j\in S_k} exp(-\varepsilon_{ij}/\lambda_k)\right)^{\lambda_k}\right).$$

Within each of the nests, random components ε_{ij} $(j = 1, \dots, J)$ are correlated. The λ_k parameter is a function of the correlation coefficient between possible alternatives in the k-th nest and is used to measure the correlation between the categories in the nest. The value of 1 for the λ_k parameter means no correlation in the

k-th nest, therefore if the value of this parameter for all nests is 1, then the nested logit model can be replaced with a standard multinomial logit model.

With the previously introduced notation, the choice probability for alternative $j \in S_k$ by i-th ($i = 1, \dots, n$) unit for the nested logit model is given by the formula:

$$P(y_{ij} = 1) = \frac{exp(\mathbf{x'}_{ij}\boldsymbol{\beta}/\lambda_k)\left(\sum_{m \in S_k} exp(\mathbf{x'}_{im}\boldsymbol{\beta}/\lambda_k)\right)^{\lambda_k - 1}}{\sum_{l=1}^{K}\left(\sum_{m \in S_l} exp(\mathbf{x'}_{im}\boldsymbol{\beta}/\lambda_l)\right)^{\lambda_l}}.$$

Then, the likelihood function is in the form:

$$p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\lambda}) = \prod_{i=1}^{N}\prod_{j=1}^{J}\left(P(y_{ij} = 1)\right)^{y_{ij}},$$

where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$.

In this article, the Bayesian approach was used to estimate the parameters of the nested logit model (Lahiri and Gao, 2002; Rossi, Allenby and McCulloch, 2005). This approach requires a prior distribution for the vector of coefficient parameters $\boldsymbol{\beta}$ and the parameter vector $\boldsymbol{\lambda}$. For the parameter vector $\boldsymbol{\beta}$, depending on the prior information, the most common are flat priors or the normal prior distributions. For the components of the $\boldsymbol{\lambda}$ parameter vector and for $a > 0$, the following prior distribution was used in this paper:

$$p(\lambda) = \begin{cases} a\lambda^{a-1}\exp(-\lambda^a) & \text{for } \lambda > 0, \\ 0 & \text{for } \lambda \leq 0. \end{cases}$$

Examples of other prior distributions for the parameter vector $\boldsymbol{\lambda}$ can be found in Lahiri and Gao (2002). This could be, for example, a beta or gamma distribution. Using the notation applied for the nested logit model, the formula for the posterior distribution has the form:

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\lambda})p(\boldsymbol{\beta})p(\boldsymbol{\lambda}).$$

In this paper, Markov Chain Monte Carlo (MCMC) methods were used to determine the marginal posterior distributions, in particular the methods used were the Metropolis algorithm (Gelman, et al., 2000) and the Gamerman algorithm (Gamerman, 1997).

3. Reference data

To analyse the situation of young people on the labour market in Poland, data from the Labour Force Survey (LFS) were used. The LFS is a quarterly panel survey with a rotational sample selection scheme. In this study, the research sample comprised units that were surveyed for two consecutive quarters in 2015. These are people from the samples numbered 63-65 and 67-69. This selection of the sample enabled, *inter alia*, verification of the answers given. In the first stage of the analysis, in accordance with the adopted research objective, people aged 18 to 29 were selected from the entire data set, thus separating a sample of 16,144 respondents. Then, the respondents were divided into five categories due to their situation on the labour market:

1. only learners/students,
2. employed but not being learners,
3. combining education with work,
4. unemployed persons but economically active,
5. economically inactive people but not being learners.

Learners were selected based on question No. 90 (*In the last 4 weeks, including as a last week the week of the survey, were you a student*?). Then, they were divided into people who had a job and those who did not. Having a job as described in this article means doing professional work in accordance with survey question 12 or having a job but temporarily not doing related work, as identified based on the answer to question 13. (Questions: *12. Did you perform work for at least 1 hour, which provided earnings or income in the week under study from Monday to Sunday, or assist in a family business for free?* 13. *Did you have a job in the week under study, but did not perform it temporarily?*). Then, from among persons who did not have a job and did not learn, economically active and economically inactive people were distinguished. Economically active persons mean those who were looking for a job and were ready to take up a job in accordance with survey questions 71 and 79. (*71. In the last 4 weeks including the week of the survey, did you look for a job? 79. Could you take a job in the 2 weeks following the week of the survey?*).

Table 1. A set of potential explanatory variables

Variable	Description	Categories	Percent
age_group	Age group at the time of the survey	1 = from 18 to 19 years old	17.68
		2 = from 20 to 24 years old	40.96
		3 = from 25 to 29 years old	41.35
sex	Sex	0 = woman	49.27
		1 = man	50.73
education	Level of education	1 = higher	22.76
		2 = post-secondary and secondary professional	22.01
		3 = secondary general	23.19
		4 = basic vocational	12.38
		5 = primary school	19.66
marital_status	Marital status	0 = unmarried, a widower, a widow, separated or divorced	78.81
		1 = married	21.19

Table 1. A set of potential explanatory variables (cont.)

Variable	Description	Categories	Percent
child	The presence of a child under 15 years in the household	0 = no 1 = yes	77.87 22.13
place_residence	Class of place of residence during the survey	0 = village 1 = town	47.63 52.37
region	Region of Poland	1 = Central Łódzkie, Mazowieckie) 2 = Southwest (Dolnośląskie, Opolskie) 3 = South (Małopolskie, Śląskie) 4 = Northwest (Wielkopolskie, Zachodniopomorskie, Lubuskie) 5 = North (Kujawsko-Pomorskie, Warmińsko-Mazurskie, Pomorskie) 6 = East (Lubelskie,Podkarpackie, Świętokrzyskie, Podlaskie)	15.03 11.92 14.33 15.04 18.14 25.55

The employed only persons were the largest part of the entire group – 43.99%. Learners constituted 30.22%, among them were both economically inactive and unemployed people. Introducing a more detailed breakdown of learners would mean introducing more values of a dependent variable, and when interpreted against one reference level, it could give hardly clear results. In addition, substantive considerations also had an impact on this division. Namely, this group includes, for example, part-time students who did not enter full-time studies and often have difficulties in determining whether they are not working because they cannot find a job, or because a lot of their time is consumed by studying or it can be an obstacle that they have to attend weekend classes starting on Fridays. The next subgroup includes persons economically inactive but not being learners. The high percentage of economically inactive and not learning persons is worrying as the share of this group is 10.83%. The share of unemployed was 8.46%. Considering the general population in the period under study, it is worth emphasizing that among all the unemployed people aged 18 to 29 accounted for as much as 37.39% (CSO, 2016a). The smallest percentage share was obtained for working and studying people – 6.5%.

Based on the presented breakdown, the dependent variable was constructed. To do this, the last two groups, i.e. groups 4 and 5 were combined into one group: the unemployed and the inactive but not learning. In this way, a group of people unemployed and persons economically inactive but not being learners, which, consider phenomenon of NEET, was then selected as a reference group in the paper. One of the research objectives was to analyse the impact of individual characteristics of the respondents on their situation on the labour market. Therefore, a set of potential explanatory variables included in this study was developed, which is presented in Table 1.

4. The model estimation

In the first stage of the analysis, the nested logit model was estimated in the Bayesian approach. Due to the primary division of the surveyed respondents into working and non-working persons, the two-nest model was chosen. The first nest contains both learning and unemployed, and persons economically inactive but not being learners, while the second one employed and people who combine work and education. Taking into account the large sample size, all considered models were estimated using normal non-informative prior distributions. For the parameter vector $\boldsymbol{\beta}$, the normal prior distributions with mean equal to 0 and variance equal to 100 were adopted in all models. The formula for the prior distribution for the lambda parameter has been presented in Section 2. In this paper, the Metropolis algorithm (Gelman, et al., 2000) or the Gamerman algorithm (Gamerman, 1997) have been used for sampling from multidimensional distributions, depending on the model under consideration.

The results for the nested logit model are presented in Table 2. The assessment of convergence of generated chains was made using the Geweke test. Based on the results obtained for both models at the significance level of $\alpha = 0.05$, the null hypothesis that the obtained chains for the considered parameters of these models are convergent cannot be rejected (Table 2). Two nests were included in the model and none of them was degenerated, therefore posterior values for two lambda parameters were determined. These parameters are used to measure the correlation between alternatives in each nest. The lambda values obtained are less than 1, therefore the nested logit model is a better model for analysing the situation of young people on the labour market in Poland in the examined period, compared to the standard multinomial logit model, because it takes into account the correlation in the considered nests.

Based on the results contained in Table 2, it can be concluded that if the option of non-working and not in education is not considered, then the second option, i.e. employed but not learning is the most important for the respondents, while the second most important one is only learning, in both cases compared to the option of non-working and not in education. On the other hand, the option of combining education with professional work definitely loses its significance, also compared to the reference option.

Table 2. Statistics of the posterior samples and Geweke convergence diagnostics for the nested logit model

Parameter	Posterior expected values	Posterior standard deviation	Highest probability density interval ($\alpha = 0.05$)		Geweke diagnostics	
					z	*p-value*
option 1	0.4025	0.1577	0.0935	0.7055	0.4957	0.6201
option 2	0.7383	0.3372	0.0743	1.3663	0.6567	0.5114
option 3	-1.0159	0.8474	-2.5516	0.7159	1.2513	0.2108
lambda 1	0.8966	0.3483	0.2164	1.5690	0.4646	0.6422
lambda 2	0.9172	0.3664	0.0294	1.5417	-1.0613	0.2886

In the next stage of the study, attempts were made to estimate the nested logit model with variables describing the characteristics of the respondents. Unfortunately, despite various attempts to improve the quality of generated chains, their convergence could not be achieved. Therefore, a standard multinomial model was considered, which is a generalization of the nested logit model (Allison, 2009). This approach was possible because the independent variables included in the model only describe the characteristics of the respondents and not the characteristics of the alternatives. The results of the estimation are presented in Table 3. This model was estimated under the same initial conditions as adopted in the first model. Prior to the interpretation of the results, the convergence of generated chains was also assessed using the Geweke test. Based on the results obtained, it was found that at the significance level $\alpha = 0.01$, the null hypothesis that the obtained chains for the considered parameters of these models are convergent cannot be rejected. Then, the received posterior expected values were interpreted.

In the case of the feature describing the respondent's sex, it was obtained that women, compared to men, were 2.18% less likely to remain in the education system as compared to the option of remaining unemployed or inactive, but not learning. In addition, they had a 58.25% less chance of doing work and a 27.42% less chance of staying in the education system and at the same time working than men, in both cases compared to persons non-working and not in education.

Young people aged 18 and 19 were more than 72 times more likely to remain in the education system, compared to people from the oldest age group 25–29, while people aged 20-24 were about 15 times more likely to remain in the education system compared to the same age group, in both cases compared to the option of persons non-working and not in education. The youngest and those aged 20–24 had 60.60% and 21.46%, respectively, less chance of having a job than people from the oldest age group. In addition, people aged 18 and 19 had more than 7 times more chances to have a job and remain in the education system compared to people from the oldest age group, while people aged 20 to 24 had these chances four times higher.

Of course, all interpretations presented in the paper remain valid under the assumption of *ceteris paribus*. In addition, in further interpretation of the results obtained, the reference level of the target variable is the same, i.e. we assume that for each of the considered options the reference level is a sum of unemployed and persons economically inactive but not being learners.

Single people were more than six times more likely to remain in the education system than married people, and more than twice as likely to combine work and study compared to married people. However, there were no major differences due to marital status in terms of performing only professional work.

Table 3. Statistics of the posterior samples and Geweke convergence diagnostics for the multinomial logit model with variable

Parameter		Posterior expected values	Posterior standard deviation	Highest probability density interval (α = 0.05)		Geweke diagnostics z	Geweke diagnostics *p-value*
sex 0	1	-0.0220	0.0585	-0.1375	0.0918	0.1953	0.8452
sex 0	2	-0.8734	0.0482	-0.9714	-0.7837	-0.8965	0.3700
sex 0	3	-0.3205	0.0784	-0.4709	-0.1647	0.1490	0.8815
age_group 1	1	4.2766	0.1199	4.0418	4.5087	0.2409	0.8096
age_group 2	1	2.6963	0.0922	2.5088	2.8724	0.3924	0.6948
age_group 1	2	-0.9315	0.1214	-1.1637	-0.6913	-0.2102	0.8335
age_group 2	2	-0.2415	0.0524	-0.3445	-0.1386	-0.7510	0.4527
age_group 1	3	2.0762	0.1605	1.7528	2.3803	-0.7171	0.4733
age_group 2	3	1.3900	0.0956	1.2001	1.5712	0.2087	0.8347
marital_status 0	1	1.8844	0.1347	1.6100	2.1376	0.1940	0.8461
marital_status 0	2	-0.0004	0.0606	-0.1139	0.1218	0.5431	0.5871
marital_status 0	3	0.7593	0.1152	0.5345	0.9834	-0.8147	0.4152
education 1	1	1.1135	0.1090	0.8954	1.3210	-0.8506	0.3950
education 2	1	-0.5467	0.0960	-0.7313	-0.3558	0.4423	0.6583
education 3	1	0.8835	0.0891	0.7035	1.0521	-0.6394	0.5226
education 4	1	-2.3453	0.1294	-2.6017	-2.1006	-0.7995	0.4240
education 1	2	2.0446	0.0900	1.8702	2.2228	-0.6612	0.5085
education 2	2	1.4084	0.0839	1.2459	1.5712	-1.3094	0.1904
education 3	2	1.0270	0.0891	0.8618	1.2082	-0.0209	0.9834
education 4	2	0.9030	0.0856	0.7361	1.0753	-0.4503	0.6525
education 1	3	2.4821	0.1550	2.1858	2.7869	-0.8424	0.3996
education 2	3	1.0031	0.1491	0.7126	1.2938	-1.0947	0.2737
education 3	3	1.5160	0.1429	1.2411	1.7987	-0.3626	0.7169
education 4	3	-0.4994	0.1857	-0.8590	-0.1374	0.3333	0.7389
child 0	1	0.0519	0.0967	-0.1373	0.2364	-0.1032	0.9178
child 0	2	-0.6439	0.0614	-0.7656	-0.5236	0.7671	0.4430
child 0	3	-0.8808	0.0988	-1.0697	-0.6886	0.7285	0.4663
place_residence 0	1	-0.3460	0.0584	-0.4582	-0.2324	-0.3128	0.7544
place_residence 0	2	0.0440	0.0473	-0.0522	0.1332	-0.3450	0.7301

Table 3. Statistics of the posterior samples and Geweke convergence diagnostics for the multinomial logit model with variable (cont.)

Parameter		Posterior expected values	Posterior standard deviation	Highest probability density interval (α = 0.05)		Geweke diagnostics	
						z	*p-value*
place_residence 0	3	-0.2625	0.0798	-0.4181	-0.1079	-2.4367	0.0148
region 1	1	0.1133	0.0919	-0.0664	0.2927	1.3343	0.1821
region 2	1	-0.0421	0.0971	-0.2329	0.1488	-0.1950	0.8454
region 3	1	0.2145	0.0938	0.0263	0.3925	0.1968	0.8440
region 4	1	-0.0856	0.0903	-0.2702	0.0819	1.1744	0.2402
region 5	1	-0.2119	0.0847	-0.3766	-0.0457	-0.6986	0.4848
region 1	2	0.5751	0.0751	0.4310	0.7237	-0.4479	0.6542
region 2	2	0.4017	0.0786	0.2486	0.5566	-0.8153	0.4149
region 3	2	0.4922	0.0765	0.3412	0.6410	0.6756	0.4993
region 4	2	0.4283	0.0724	0.2867	0.5681	0.8986	0.3688
region 5	2	0.2772	0.0687	0.1437	0.4110	0.7898	0.4297
region 1	3	0.5633	0.1233	0.3302	0.8084	0.6649	0.5061
region 2	3	0.4660	0.1310	0.2130	0.7237	-1.3811	0.1672
region 3	3	0.4651	0.1291	0.2152	0.7202	0.7235	0.4694
region 4	3	0.4529	0.1211	0.2223	0.6948	-0.4774	0.6330
region 5	3	0.3632	0.1141	0.1431	0.5893	0.7771	0.4371

People with higher education had three times more chances to remain only in the education system than people with basic education, for people with post-secondary and secondary vocational education these chances were 42.11% lower, for people with general secondary education the chances were over twice as large, and 90.42% less for people with basic vocational education. Chances for performing only professional work were more than seven times higher for people with higher education, four times higher for people with post-secondary and secondary vocational education, more than twice higher for people with general secondary education and basic vocational education, in each case compared to persons with basic education. The chances of combining education with work were more than eleven times higher for people with higher education, more than twice as high for people with post-secondary and secondary vocational education, more than four times higher for people with general secondary education and 39.31% lower for people with basic vocational education. In each case, compared to people with basic education.

For the variable describing the presence of a child under the age of 15 in a household in which the respondent or his/her spouse is its head, it was obtained that the lack of a child was only associated with a 5.32% increase in the chances of staying in the education system, compared to persons residing in households, in which a child or children were present. On the other hand, the chances of only working and combining education with work were about 50% lower for these people, also compared to people living in households with children.

People living in the countryside had 29.25% less chance of staying only in the education system compared to urban residents. In addition, they were more than four times more likely to work only and had 23.09% less chance to combine education with work, in both cases compared to young people living in cities.

Comparing the eastern region to other regions of Poland, it was found that the inhabitants of each of them had at least 31% greater chances of only working and combining education with work compared to the inhabitants of the eastern region. In addition, residents of the central (Łódzkie, Mazowieckie) and southern (Małopolskie, Śląskie) regions were more likely to remain in the education system only, by 12% and 23%, respectively, compared to the eastern region. In other regions, these chances were lower compared to the eastern region, with the smallest chance of remaining only in the education system obtained for the northern region.

5. Conclusions

The occupational and educational choices of young people depend on many different factors related to both individual characteristics of these people including their work motivation (Davidescu, Roman, Strat and Mosora, 2019) as well as the socio-economic situation of a country. This work focuses on this first group of factors except for work motivation, due to the lack of relevant data for Poland in this regard. For modelling, the multinomial logit model, and its special case, i.e. the nested logit model (Allison, 2009) were chosen. Using the latter model, it was possible to take into account the hierarchical division of young people due to their status on the labour market as well as their education. However, as far as the analysis of the impact of the individual characteristics of the respondents on their occupational and educational choices is concerned, the standard multinomial model turned out to be the better model.

According to the human capital theory, the wage differences among occupations and the ability to learn during education are the main factors influencing occupational and educational choices of young people (Dale, 2009). What follows from our study is that young people in Poland prefer to focus mainly on working and to a lower extent on education. This may be associated with high opportunity costs of higher education, mainly foregone earnings. Combining education with work is not their preferred form of economic activity too. It can, therefore, be concluded that remaining in the education system is associated with the inability to find a suitable job, or with the prospect of getting a better job after obtaining higher education. Given the current massification of higher education (Jasiński, Bożykowski, Chłoń-Domińczak, Zając and Żółtak, 2017), it is important to look for other solutions to improve the situation of young people on the labour market in Poland. Importantly, the professional situation of this age group is the worst compared to other age groups (CSO, 2016b).

In recent years, in Poland, the approach of young people to the balance between learning and work has been constantly changing. In this paper, we have shown to what extent this approach depends on their age too. It was found that people from the youngest age group from 18 to 19 years old had the best chance of remaining only in the education system, these people also had the least chance of being employed in comparison with people aged 25 to 29. Moreover, the people who most often combined work and study also belonged to the youngest age group, for the next age group, i.e. persons aged from 20 to 24, these chances were almost two times lower. The obtained result may be slightly disturbing in the face of "lifelong learning" widely promoted in the European Union. On the other hand, it is difficult to say unequivocally whether the experience of combining education and employment in youth may facilitate lifelong learning or it is a factor discouraging from pursuing further learning. However, many people from this age group decide to continue to combine both work and study as they are aware of the need of further education (Brooks, 2006).

According to G.S Becker (1991), the main determinant of the economic activity of an individual is education. Our empirical evidence suggests that people with higher education as well as post-secondary and secondary vocational education have greater chance of having a job compared to persons with basic education. On the other hand, people with general secondary education most often combined work and education, such a combination was least often among people with basic vocational education. People who had basic vocational education most often ended their educational activity at this stage. According to Jasiński, Bożykowski, Chłoń-Domińczak, Zając and Żółtak (2017), young people should choose carefully their educational pathway because the employment chances of university graduates in Poland depend on the study area, moreover they also change over time. In the context of combining work and learning they indicated that prior experience in the labour market has an impact on employment chances, but only in the first months after graduation.

The social inequalities in the labour market, including inequalities due to gender (Becker, 2010), have been a challenge for many labour markets in Europe. Our study indicates that during the period considered in Poland women had less chance of employment compared to men, and even less chance to combine education with work. According to Castellano and Rocca (2017,) education is the most important factor determining the gender gap in the labour market. In Poland, women are better educated than men. Therefore, we can agree with Castellano and Rocca that gender inequalities in the labour markets may depend on cultural factors, too. Moreover, as expected, single people were more likely to remain in the education system than married people, and they were also more likely to combine work with education. In the case of respondents working only, there were no major differences due to marital status. In the context of having a family, it was found that having a child is no longer a major problem with continuing education and is also conducive to greater professional activity. According to other studies (Michaud and Tatsiramos, 2011), having a child mainly affects women's employment, but the effects of this influence vary from country to country. Based on one of the latest studies on Polish women (Grzenda, 2019), it was found that the differences in the professional careers of women without children and having children are becoming smaller.

Considering access to education of young people living in cities and in the countryside, the results of other studies have been confirmed (Kołaczek, 2005), according to which at the primary level of compulsory education there are no differences in access to education between cities and villages, these differences are only revealed at a secondary and higher level. In this study, it was found that rural residents had less chance of continuing education, as well as combining work with education, but these differences did not exceed 30% compared to young people living in cities. Given territorial division it was observed that in comparison with the inhabitants of the eastern region of Poland, the inhabitants of all other regions had a greater chance of employment as well as combining work with education. In addition, the inhabitants of the central and southern regions, in which larger scientific centres are concentrated, had a better chance of staying only in the education system compared to the eastern region.

The research methods proposed in this paper made it possible to determine the impact of individual characteristics of young people in Poland on their occupational and educational choices. In addition, our contribution to research in this area consisted of including in the model as many as four different states of their activity: employed but not learning; combining education with work; only learners/students; unemployed and persons economically inactive but not being learners. This provided new insight into how young people enter the labour market in Poland.

Acknowledgements

This paper was presented on the MSA 2019 conference which financed its publication. Organization of the international conference "Multivariate Statistical Analysis 2019" (MSA 2019) was supported from resources for popularization of scientific activities of the Minister of Science and Higher Education in the framework of agreement No. 712/P-DUN/202019.

This study has been prepared as part of the project granted by the National Science Centre, Poland, entitled "The modelling of parallel family and occupational careers with Bayesian methods" (2015/17/B/HS4/02064).

References

Allison, P. D., (2009). Logistic Regression Using the SAS®. Theory and Application. 8th ed. Cary, NC: SAS Institute Inc.

Anderson, S. P., De Palma, A. and Thisse, J. F., (1992). Discrete choice theory of product differentiation. Massachusetts: The MIT Press.

Becker, G. S., (1991). A Treatise on the Family. Cambridge: Harvard University Press.

Becker, G. S., (2010). The Economics of Discrimination. USA: University of Chicago Press.

Bieszk-Stolorz, B., Markowicz, I., (2013). Men's and Women's Economic Activity in Poland. Acta Universitatis Lodziensis. Folia Oeconomica, 285, pp. 221–227.

Brooks, R., (2006). Learning and work in the lives of young adults. International Journal of Lifelong Education, 25(3), pp. 271–289.

Cameron, A. C., Trivedi, P. K., (2005). Microeconometrics: methods and applications. Cambridge: Cambridge University Press.

Castellano, R., Rocca, A., (2017). The dynamic of the gender gap in the European labour market in the years of economic crisis. Quality & Quantity, 51(3), pp. 1337–1357.

Cheng, S., Long, J. S., (2007). Testing for IIA in the multinomial logit model. Sociological Methods & Research, 35(4), pp. 583–600.

Chłoń-Domińczak, A. and Strawiński, P., (2013). Wchodzenie osób młodych na rynek pracy w Polsce. In Proceedings of the 9th Congress of Polish Economists, pp. 28–29.

Cramer, J. S., 2003. Logit Models from Economics and Other Fields. Cambridge: Cambridge University Press.

Central Statistical Office of Poland, (2016a). Aktywność ekonomiczna ludności Polski w latach 2013 – 2015. Warszawa: CSO.

Central Statistical Office of Poland, (2016b). Monitoring Rynku Pracy, Kwartalna informacja o rynku pracy. Warszawa: CSO.

Dale, K., (2009). Household skills and low wages. Journal of Population Economics, 22(4), pp. 1025–1038.

Davidescu, A. A. M., Roman, M., Strat, V. A. and Mosora, M., (2019). Regional sustainability, individual expectations and work motivation: A multilevel analysis. Sustainability, 11(12), 3331.

de Dios Jiménez, J., Salas-Velasco, M., (2000). Modeling educational choices. A binomial logit model applied to the demand for higher education. Higher Education, 40(3), pp. 293–311.

Gallie, D., Paugam, S. eds., (2000). Welfare regimes and the experience of unemployment in Europe. Oxford: OUP Oxford.

Gamerman, D., (1997). Sampling from the posterior distribution in generalized linear mixed models. Statistics and Computing, 7(1), pp. 57–68.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B., (2000). Bayesian Data Analysis. London: Chapman & Hall/CRC.

Grzenda, W., (2012). Badanie determinant pozostawania bez pracy osób młodych z wykorzystaniem semiparametrycznego modelu Coxa. Przegląd Statystyczny, 59(1), pp. 123–139.

Grzenda, W., (2019). Modelowanie karier zawodowej i rodzinnej z wykorzystaniem podejścia bayesowskiego. Warszawa: Wydawnictwo Naukowe PWN.

Hausman, J., McFadden, D., (1984). Specification tests for the multinomial logit model. Econometrica, 52, pp. 1219–1240.

Hsiao, C., Small, K., (1985). Multinomial logit specification tests. International economic review, 26(3), pp. 619–627.

Jasiński, M., Bożykowski, M., Chłoń-Domińczak, A., Zając, T. and Żółtak, M., (2017). Who gets a job after graduation? Factors affecting the early career employment chances of higher education graduates in Poland. Edukacja Quarterly, 143(4).

Kołaczek, B., (2005). Podstawowe uwarunkowania społeczne dostępu młodzieży do kształcenia. Polityka Społeczna, 1.

Lahiri, K., Gao, J., (2002). Bayesian Analysis of Nested Logit Model by Markov Chain Monte Carlo. Journal of Econometrics, 11, pp. 103–133.

McFadden, D., (1978). Modelling the Choice of Residential Location. In: A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, eds. Spatial Interaction Theory and Planning Models. Amsterdam: North-Holland, pp. 75–96.

Michaud, P. C., Tatsiramos, K., (2011). Fertility and female employment dynamics in Europe: the effect of using alternative econometric modeling assumptions. Journal of Applied Econometrics, 26(4), pp. 641–668.

Ministerstwo Rodziny, Pracy i Polityki Społecznej, Departament Rynku Pracy, (2018). Sytuacja na rynku pracy osób młodych w 2017 roku, Warszawa: MRPiPS.

Robert, C. P., Casella, G., (2004). Monte Carlo Statistical Methods. 2nd ed. New York: Springer.

Rossi, P. E., Allenby, G. M. and McCulloch, R., (2005). Bayesian Statistics and Marketing. Chichester. UK: John Wiley & Sons.

Stanisz, A., (2016). Modele regresji logistycznej: zastosowania w medycynie, naukach przyrodniczych i społecznych, Kraków: Wydawnictwo StatSoft Polska.

Train, K. E., (2009). Discrete Choice Methods with Simulation. 2nd ed. Cambridge: Cambridge University Press.