Statistical Analysis of Labor Market Integration: A Mixture Regression Approach

Tapio Nummi, Janne Salonen, and Timothy E. O'Brien

1 Introduction

Integration of young people into the labor market is a complex and socially important issue which must be understood as a process in time. The majority of young people attach to the labor market quite quickly, some after their studies although some do remain unemployed (e.g. Pareliussen 2016). Moreover, different stages may not necessarily follow a straightforward progression. For example, it is quite common in the Finnish system for students to work during their studies. Meaning, one person can have several different statuses at the same time (throughout 1 year). Obtaining an overall picture of such complex and heterogeneous longitudinal data is a challenging task.

In this paper we present one possible approach to this complex data analysis problem. The approach is based on the mixture regression applied to multivariate longitudinal binary data. In applied statistics, these methods are often referred to as latent class regression models, or trajectory analysis (Nagin 2005). The idea is that data consists of unknown sub-populations with some common properties that can be revealed through longitudinal data. Recently these methods have been very popular in many fields of science, including psychology, education, sociology, marketing

T. Nummi (✉)
Faculty of Natural Sciences, University of Tampere, Tampere, Finland
e-mail: tan@uta.fi

J. Salonen
Research Department, The Finnish Centre for Pensions, Helsinki, Finland
e-mail: Janne.Salonen@etk.fi

T.E. O'Brien
Department of Mathematics and Statistics and Institute of Environmental Sustainability, Loyola University of Chicago, Chicago, IL, USA
e-mail: teobrien@gmail.com

D.-G. Chen et al. (eds.), *New Advances in Statistics and Data Science*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-319-69416-0_18

and health sciences (Korpela et al. 2017; Kokko et al. 2008; Jolkkonen et al. 2017; Mani and Nandkumar 2016; Nummi et al. 2014, 2017).

In this paper we present a 4-dimensional binary mixture regression model that is used to identify the sub-groups in the data gathered. We find that there are ten main groups that lead to different development paths of young men. Most people integrated into the labor market quite quickly after various intermediate stages, but in a few groups the integration is weaker or slower. In terms of society, the groups of weak attachment are of central interest, because they may later require special support or action from society. For instance, in some countries a special youth guarantee policy has been promoted (Keränen 2012; Escudero and Mourelo 2015).

2 Methods

2.1 Data

The data comes from the administrative registers of the Finnish Centre for Pensions and Statistics Finland. In the administrative registers there is a range of information pertaining to all of the pension insured (total population) people in Finland. For this study we choose the male cohort born in 1987. For other studies of the same cohort we can refer to Paananen and Gissler (2013). We follow all individuals between 2005 to 2013, when the cohort is 18–26 years of age. We take a sub-population of those who are Finnish citizens and who have lived in Finland during the specified period. The research population is 29,383 males.

Labor market attachment is measured using days when working, in education, in unemployment and on various social benefits per year. This yields the 4-dimensional response vector as follows:

- Variable 1 (Employed): Individual employed for days/year in private or public sector or self-employed.
- Variable 2 (Education): Individual in education leading to a degree and/or is receiving student financial aid.
- Variable 3 (Unemployed): Individual receives unemployment benefits, either earnings related or paid by the state.
- Variable 4 (Leave): Individual receives sickness benefits or is on vocational rehabilitation. Parental leave and the permanently disabled are included here as well.

The original data is measured as days/year. For our analysis, the data was dichotomized (Yes/No) because in this type of longitudinal data, the most important factor in the formation of an individual's career trajectory is the several different statuses of the individuals. This makes the analysis of data much simpler and more uniform.

2.2 Multivariate Binary Mixture

Our aim is to identify clusters of individuals with the same kind of mean developmental profiles (trajectories). Let $\mathbf{y}_i = (y_{ij1}, y_{ij2}, \ldots, y_{ijT})'$ represent the sequence of measurements on individual i for the variable j over T periods and let $f_i(\mathbf{y}_i|\mathbf{X}_i)$ denote the marginal probability distribution of $\mathbf{y}_i$ with possible time dependent covariates $\mathbf{X}_i$ that are same to all $j = 1, \ldots, s$ variables. It is assumed that $f_i(\mathbf{y}_i|\mathbf{X}_i)$ follows a mixture of K densities

$$f_i(\mathbf{y}_i|\mathbf{X}_i) = \sum_{k=1}^{K} \pi_k f_{ik}(\mathbf{y}_i|\mathbf{X}_i), \quad \sum_{k=1}^{K} \pi_k = 1 \text{ with } \pi_k > 0, \tag{1}$$

where π_k is the probability of belonging to the cluster k and $f_{ik}(\mathbf{y}_i|\mathbf{X}_i)$ is the density for the kth cluster (see e.g. McLachlan and Peel 2000). The natural choice is to use the Bernoulli distribution for the mixture components. It is assumed that s variables in $\mathbf{y}_i$, $i = 1, \ldots, N$, are independent. Also measurements given the kth cluster are assumed to be independent. This yields the density

$$f_{ik}(\mathbf{y}_i|\mathbf{X}_i) = \prod_{j=1}^{s} \prod_{t=1}^{T} p_{ijtk}^{y_{ijt}} (1 - p_{ijtk}^{1-y_{ijt}}), \tag{2}$$

where p_{ijtk} is a function of covariates $\mathbf{X}_i$. For modeling the conditional distribution of p_{ijtk} we use the logistic regression model. For the ith individual we can then write

$$p_{ijtk} = \frac{\exp(\mathbf{x}_i' \beta_{jk})}{1 + \exp(\mathbf{x}_i' \boldsymbol{\beta}_{jk})}, \tag{3}$$

where $\mathbf{x}_i'$ is the tth row of $\mathbf{X}_i$, $\boldsymbol{\beta}_{jk}$ is the parameter vector of the jth variable in the kth cluster. For our analysis we took the second degree model

$$\mathbf{x}_i' \boldsymbol{\beta}_{jk} = \beta_{0jk} + \beta_{1jk} t + \beta_{2jk} t^2 \tag{4}$$

for modeling the probabilities within the variable j and cluster k in time t. Maximum likelihood estimates can then be obtained by maximizing the log likelihood $\log \sum_{i=1}^{N} f_i$ over unknown parameters β_{jk}, $j = 1, \ldots s; k = 1, \ldots K$ (Nagin 1999; Jones et al. 2001; Nagin and Tremblay 2001; Jones and Nagin 2007, Nagin and Odgers 2010a,b). When the model parameters are estimated the posterior probability estimate provides a tool for assigning individuals to specific clusters. Individuals can then be assigned to specific clusters to which their posterior probability is the largest.

3 Analysis

Choosing the number of trajectories (clusters) K is an important issue when applying mixture modeling. The selection of K can be based on technical criteria, substantive examination, or both. We used the information criteria BIC, which is perhaps the most widely used in this context. Here we present the values of BIC for $k = 5, \ldots, 10$: -397509.4, -394554.5, -392966.5, -389504.3, -388897.6 and -383596.6. The maximum (note: SAS implementation) is obtained for $k = 10$ and is therefore our choice for the number of clusters K. Increasing K could yield to more difficult to interpret and insignificantly small clusters. The final estimated model is summarized in Tables 1 and 2 (in Appendix). The trajectory plots (Figs. 1, 2, 3, and 4) present conditional point means calculated for each of the four variables.

Table 1 Clusters k, their estimated absolute N_k and relative sizes $\hat{\pi}_k$

Group (k)	N_k	$\hat{\pi}_k$ (%)
1:	545	1.9
2:	3009	10.2
3:	1971	6.7
4:	1975	6.7
5:	5942	20.2
6:	5933	20.2
7:	1510	5.1
8:	3141	10.7
9:	4093	13.9
10:	938	3.2

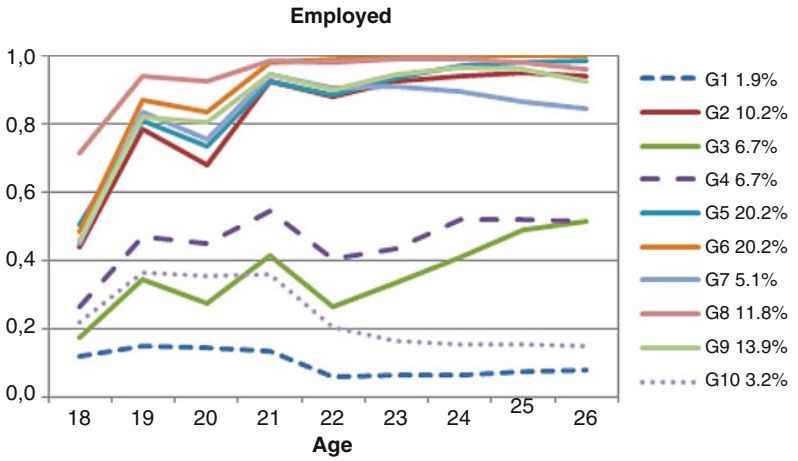

Fig. 1 Time-point means (proportions) over trajectories for the variable Employed

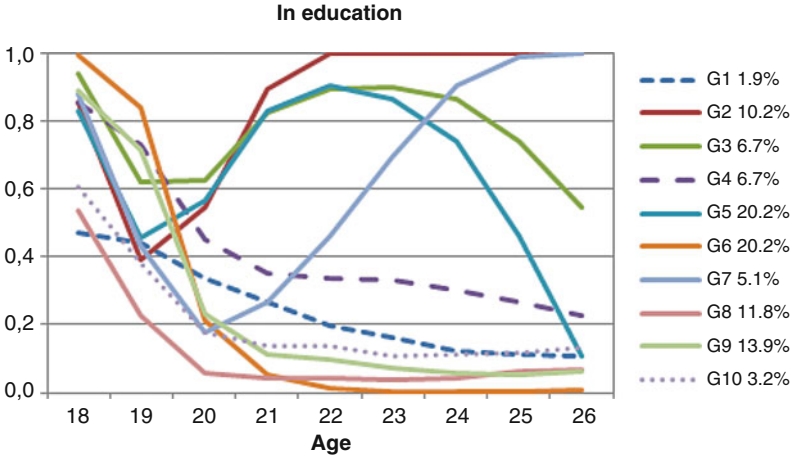

Fig. 2 Time-point means (proportions) over trajectories for the variable Education

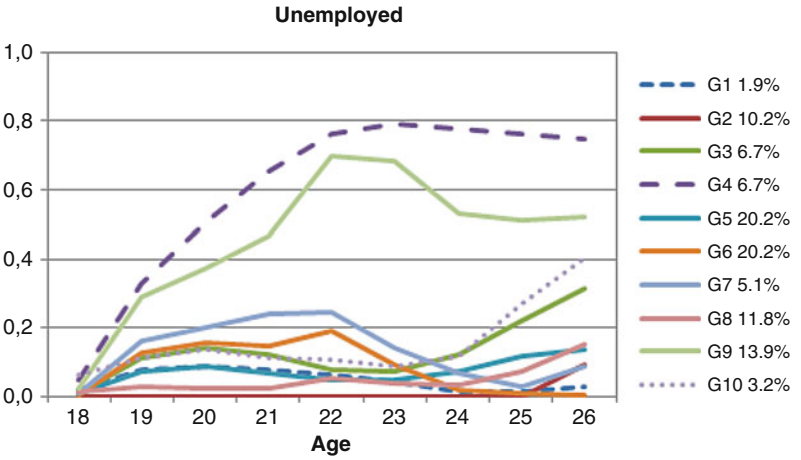

Fig. 3 Time-point means (proportions) over trajectories for the variable Unemployed

These plots are used as the main tool for the interpretation of the results obtained from the mixture regression fit. The computations were carried out by SAS proc traj procedure.

3.1 Normal Life-Course

In the trajectory plots the solid lines indicate groups (total of 88.1%) where labor market integration is good (Fig. 1). It is quite common for young people ages 18–21 to be in post-secondary or vocational education (Fig. 2). From Fig. 1 we note that

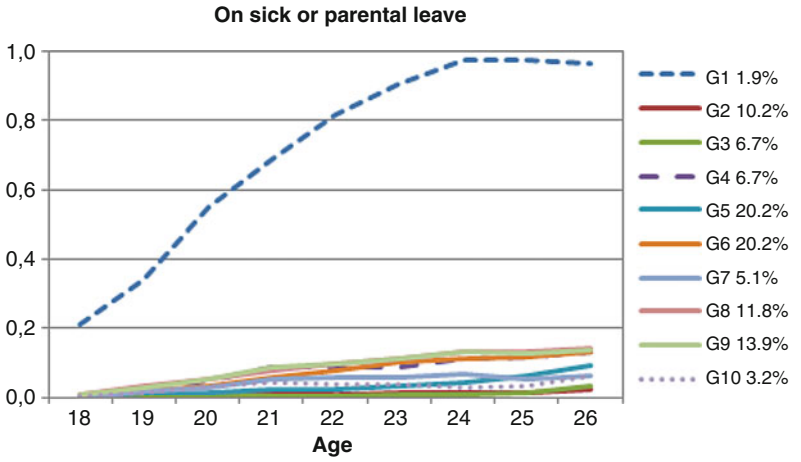

Fig. 4 Time-point means (proportions) over trajectories for the variable Leave

about 74.6% of young males are in the trajectories where the percentage of people who have employment status after 21 years of age is surprisingly high (>80%). We refer to this group of trajectories (groups 2, 5, 6, 7, 8 and 9) as the HES group (high employment status group). Trajectories 5 and 2 (30.4%) from the HES group both have high percentage of people (>50%) in education after 21 years of age (Fig. 2). This reflects the well-known fact that the majority of students in Finland work during their studies. Group 2 (10.2%) contains typical university students. Note that this group also has a high percentage of people in employment.

Group 9 (13.9%) is interesting as we can see from Fig. 3 that this group has a low percentage of people in education after the age of 21 (Fig. 2). However, the percentage of people in employment is still high and increases with age (Fig. 1). This group contains low-educated young men who work in various part-time or temporary jobs and experience unemployment periods.

In general most of the trajectory groups (Fig. 3) have low percentage of people who are unemployed, with many young men experiencing short unemployment periods. These unemployment spells are usually short throughout their normal life-course. This is also true for sickness and disability periods (Fig. 4). As the trajectories indicate, young Finnish men participate in family life (take parental leave), and this participation increases with age. However, due to short follow-up time, a more thorough analysis of parental leave is not possible as the first child is usually born after this time.

From employment and education trajectories we can see a change at age 20 (Figs. 1 and 2). This is due to the fact that the Finnish male cohort (80% of males) enters military service at this age. Service lasts less than a year, and therefore these men in the military have no status in the labor market or on education during this time.

3.2 Weak Labor Market Integration

In the figures the dotted lines indicate groups (total of 11.8%) where labor market integration is poor for longer time periods. We refer to this group of trajectories (groups 1, 3 and 10) as the LES group (low employment status group) which is our main interest and focus. These young men are experiencing difficulties in the labor market and have therefore been the target government support programs.

From Fig. 2 we can identify four groups (1, 3, 4 and 10) with low percentages in terms of employment. There is one positive indication, however, as group 3 contains higher percentage in education (Fig. 2). The LES group is not particularly active in education, as they perhaps receive some vocational education in their late teens.

Unemployment and leave trajectories (Figs. 3 and 4) explain the LES group in more detail. Group 4 (6.7%) is clearly unemployed after secondary and vocational education. The percentage of unemployment in this group is nearly 80% at the end of the follow up period.

The sickness and parental leave trajectory plot indicates that Group 1 (1.9%) has the highest percentage (Fig. 4). In fact this group is not on parental leave, but receives occupational rehabilitation or a disability pension instead. This is clearly the group with the most difficulties in the labor market. In Finland occupational rehabilitation is rather effective, so these young men may have a chance to attend school or work later in life.

As an ex-post validation of these trajectory groups we can measure or sum up the working days over the follow-up period. The length of working life is on average 5.1 years for the HES group and only 1.3 years for the LES group. The results confirm that this analysis has found clusters that also have practical importance.

4 Concluding Remarks

It is clear that our mixture regression analysis is an effective tool for the identification of different clusters of register-based data. Naturally, the central interest is on those who have difficulties with labor market integration. We think that our analysis provides new insight into this important social issue. We mainly concentrated on a descriptive analysis of results. However, it would also be interesting to analyze the identified clusters further using covariates like social class, living area, parents' education, or parents' income. The best way to proceed may be the joint modeling of clusters and mixing percentages using multinomial regression. This more subject-oriented analysis remains a topic of further research.

Acknowledgements The authors wish to thank the Finnish Centre for Pensions and Statistics Finland for providing the research data for this study. We also like to thank the referees for the comments that led to improvements of the paper.

Appendix

Table 2 Summary of the estimated model: variables, clusters, parameter estimates and their standard errors

Variable	Group	$\hat{\beta}_0$	SE($\hat{\beta}_0$)	$\hat{\beta}_1$	SE($\hat{\beta}_1$)	$\hat{\beta}_2$	SE($\hat{\beta}_2$)
1	1	2.398	4.0937	−0.3086	0.3809	0.0043	0.0088
1	2	−33.1751	1.6662	2.8697	0.1567	−0.0572	0.0036
1	3	−1.3307	1.4851	−0.082	0.1356	0.0052	0.0031
1	4	−10.4101	1.4206	0.8602	0.1308	−0.0177	0.003
1	5	−18.5665	1.4307	1.4459	0.1371	−0.0223	0.0033
1	6	−30.6046	3.7969	2.2395	0.3836	−0.0297	0.0097
1	7	−49.6895	2.1645	4.5499	0.201	−0.0993	0.0046
1	8	−64.2231	2.9188	5.9415	0.2719	−0.1288	0.0062
1	9	−48.3674	1.6773	4.3293	0.1584	−0.0912	0.0037
1	10	−10.4439	2.556	1.0078	0.2392	−0.0264	0.0055
2	1	11.1319	2.9527	−0.8535	0.2754	0.013	0.0064
2	2	399.1368	10.4145	−41.2003	1.0772	1.0621	0.0278
2	3	−20.6646	1.6869	2.0785	0.156	−0.0485	0.0036
2	4	36.9462	1.8208	−3.1096	0.1655	0.0635	0.0037
2	5	−54.0624	0.9532	5.2318	0.0885	−0.1231	0.002
2	6	165.358	4.3881	−14.2717	0.3962	0.2974	0.0089
2	7	151.7117	4.8462	−14.7589	0.4754	0.3563	0.0116
2	8	72.251	2.3522	−6.5804	0.2146	0.1428	0.0048
2	9	83.5415	2.0619	−7.2462	0.1871	0.1515	0.0042
2	10	45.0733	2.7618	−4.0274	0.2565	0.0857	0.0059
3	1	−37.1495	7.5126	3.338	0.7054	−0.0805	0.0164
3	2	−3.5307	99.786	−15.2704	54.119	0.5892	2.0791
3	3	7.1229	2.1854	−1.0465	0.1995	0.0284	0.0045
3	4	−61.0014	2.0817	5.2682	0.1901	−0.1111	0.0043
3	5	−1.4631	1.5541	−0.2562	0.1412	0.0092	0.0032
3	6	−110.2661	3.8561	10.3492	0.3723	−0.246	0.009
3	7	−54.1435	3.2046	4.9123	0.2998	−0.1141	0.007
3	8	8.3929	3.0624	−1.3156	0.2777	0.0354	0.0062
3	9	−58.2828	1.2639	5.0717	0.1153	−0.1092	0.0026
3	10	19.9764	2.9505	−2.2136	0.2699	0.0548	0.0061
4	1	−22.1081	4.2901	1.4843	0.4084	−0.0185	0.0097
4	2	−16.5855	5.8831	0.9372	0.5318	−0.0173	0.0119
4	3	6.3987	8.9441	−1.345	0.7973	0.0372	0.0176
4	4	−33.001	3.2404	2.5058	0.2873	−0.0505	0.0063
4	5	−11.5583	2.6232	0.3925	0.2318	−0.0015	0.0051

(continued)

Table 2 (continued)

Variable	Group	$\hat{\beta}_0$	SE($\hat{\beta}_0$)	$\hat{\beta}_1$	SE($\hat{\beta}_1$)	$\hat{\beta}_2$	SE($\hat{\beta}_2$)
4	6	−36.7212	2.2235	2.7618	0.1953	−0.0548	0.0043
4	7	−38.7528	4.6765	3.0266	0.4167	−0.0633	0.0092
4	8	−30.3185	2.4405	2.2794	0.2174	−0.0456	0.0048
4	9	−33.8236	2.1217	2.6022	0.1885	−0.0529	0.0042
4	10	−20.8289	6.5442	1.4166	0.591	−0.0281	0.0132

References

Escudero,V., & Mourelo, M. L. (2015). *The Youth Guarantee programme in Europe: Features, implementation and challenges*. Geneva: ILO.

Jolkkonen, A., Kurvinen, A., Virtanen, P., Lipiäinen, L., Nummi, T., Koistinen, P. (2017). Labour market attachment following major workforce downsizings: A comparison of displaced and retained workers. *Work, Employment and Society*. https://doi.org/10.1177/0950017017706305.

Jones, B. L., & Nagin, D. S. (2007). Advances in group-based trajectory modeling and an SAS procedure for estimating them. *Sociological Methods & Research, 35*(4), 542–571.

Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research, 29*(3), 374–393.

Keränen, K. (2012). Young people within services — best practices for the promotion of the youth guarantee. The Ministry of Employment and the Economy in Finland.

Kokko, K., Pulkkinen, L., Mesiäinen, P., Lyyra, A.-L. (2008). Trajectories based on postcomprehensive and higher education: Their correlates and antecedents. *Journal of Social Issues, 64*(1), 59–76.

Korpela, K., Nummi, T., Lipiäinen, L., De Bloom, J., Sianoja, M., Pasanen, T., & Kinnunen, U. (2017). Nature exposure predicts well-being trajectory groups among employees across two years. *Journal of Environmental Psychology*. https://doi.org/doi:10.1016/j.jenvp.2017.06.002.

Mani, D., & Nandkumar, A. (2016). The differential impacts of markets for technology on the value of technological resources: An application of group-based trajectory models. *Strategic Management Journal, 37*(1), 192–205.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Hoboken, NJ: Wiley.

Nagin, O. S. (1999). Analyzing developmental trajectories: Semi-parametric group-based approach. *Psychological Methods, 4*, 39–177.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nagin, D. S., & Odgers, C. L. (2010a). Group-based trajectory modeling (nearly) two decades later. *Journal of Quantitative Criminology, 26*(4), 445–453.

Nagin, D. S., & Odgers, C. L. (2010b). Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology, 6*, 109–138.

Nagin, D. S., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods, 6*(1), 18–34.

Nummi, T., Hakanen, T., Lipiäinen, L., Harjunmaa, U., Salo, M., Saha, M. T., & Vuovela, N. (2014). A trajectory analysis of body mass index for Finnish children. *Journal of Applied Statistics, 41*(7), 1422–1435.

Nummi, T., Virtanen, P., Leino-Arjas, P., & Hammarström, A. (2017). Trajectories of a set of ten functional somatic symptoms from adolescence to middle age. *Archives of Public Health, 75*, 11.

Paananen, R., & Gissler, M. (2013). *International Journal of Epidemiology, 41*, 941–945.

Pareliussen, J. K. (2016). Age, skills and labour market outcomes in Finland. OECD Economics Department Working Papers, No. 1321. Paris: OECD Publishing.