What does explain the heterogeneity in early family trajectories? A non-parametric approach for sequence analysis

Nicola Barban
Department of Statistical Sciences, University of Padua, Padua, Italy.
Francesco C. Billari
Bocconi University, Milan, Italy.

Abstract. This study examines the impact of demographic and socio-economic variables on the variability of family-life trajectories during young adulthood. We define life-courses as sequences on a monthly time scale and we apply optimal matching (OM) to compute dissimilarities between individuals. We propose a generalization of analysis of variance (ANOVA) to evaluate the link between non-metric measures (sequence dissimilarities) with categorical variables. We conduct hypothesis test using a permutation approach. Data come from the first and third wave of The National Longitudinal Study of Adolescent Health (Add Health). The sample is restricted to women from age 18 to 24.

Extended abstract prepared for PAA 2010 conference

Introduction

Since the profound changes in partnering and childbearing occurred in the United States during the last decades (cohabitation and raise of non-marital fertility) have particularly affected women in their early 20s (Schoen et al., 2007; Amato et al., 2008), it is important to examine divergences in the initial years of early adulthood. In terms of family transitions, those years are very "dense" (Rindfuss, 1991), with more demographic events occurring than during any other part of the life course. The under-25 age group exhibits especially great heterogeneity in family formation behavior, with some women postponing all family-related transitions, others making commitments (e.g., cohabitation), and still others making choices with enduring consequences (e.g., becoming a parent).
These heterogeneous individual trajectories are in part the outcome of individual life planning, sometimes with the participation and help of the parents and/or partner, but they are also influenced by the social origin and the dynamic context around young adults. It is not completely clear thus, how much of the heterogeneity in the life course can be explained by demographic and socioeconomic variables. We expect individuals living in similar context and sharing the same characteristics before starting family transition, to show similar behavior when they start the family transition. On the contrary, we expect to see greater variability between individuals grown in different context.

Sequence analysis gives the representation of the occurrence, the timing and the order of a set of events for individuals observed across time. We compute the monthly life trajectory looking at the simultaneous distribution of cohabitation, marriage and parenthood. In each month individuals can be classified as: Single (S), Single Parent (SP), Cohabiting (C), Cohabiting parent (CP), Married (M) and Married parent (MP). For each individual the resulting sequence is a string composed by 72 values (6 years from age 18 to 24), indicating the evolution of family transition. Using Optimal Matching Analysis (OMA) (Abbott, 1995) we compute a measure of dissimilarity between sequences of family formation. In general, to calculate a pairwise distance between two sequences, the number of minimum transformations (insertion, deletion, and substitution) necessary to transform one sequence into the other is tallied, each transformation is assigned a cost, and these costs are summed. The cost of a single substitution is determined empirically by calculating the substitution matrix for each element. The result of OMA is a matrix of pairwise dissimilarities that usually is the starting point for data reduction techniques (mainly clustering) (Billari and Piccarreta, 2007; Billari et al., 2007). Following the approach of Elzinga and Liefbroer (2006), we model the dyad between element i and j using the OM distance as dependent variables. Independent variables are constructed as

categorical variable with value 1 if both individuals share the same value and 0 otherwise. Analogously to linear regression models we can then evaluate the explicative value of each variable inserted in the model and select the model with best adaptive power. As standard ANOVA for linear models, we decompose the dissimilarity of life trajectories in a component explained by a "model" and in a "residual" component. Then we evaluate a pseudo-R^2 statistic giving a measure of explanatory power of each variable and perform statistical tests using data permutations.
The aim of this paper is therefore both substantive and methodological. We attempt to describe the impact of demographic and socioeconomic variables on the variability of life-course trajectories answering questions like: How similar are the life courses of individuals that share the same characteristics at the beginning of the transition? What are the variables that most influence the divergence of trajectories? On the other side, we propose a semi-parametric to model the variability of sequence analysis without recur to data reduction techniques.

Data and methods

Data

The National Longitudinal Study of Adolescent Health (Add Health) is a school-based, nationally representative sample of U.S. students in grades 7 through 12 in 1994. Nearly all were born in the years 1976 through 1982. The Add Health data include three waves of in-home interviews, which were conducted in 1995 (Wave I), 1996 (Wave II), and 2001-2002 (Wave III). The data for the present study are taken from Waves I and III. Of the 10,480 women interviewed in 1995, 8,015 were also interviewed during Wave III. Since we restrict our analysis to women that we can observe from age 18 to 24, the final sample size become 5,507.

Variables

We propose to evaluate the effect of three categories of variables that may have an effect on the variability of the individuals' life courses (demographic, social class and context).

(a) Demographic
- Race/Ethnicity
- Family type (childhood with both biological parents)
- Immigrant generation (parents born outside US)

(b) Social Class
- Parents' education (parents with college)
- Parents income (total family income above median)

(c) Context
- State, County, neighborhood
- Contextual variable. (median income, employment rate, segregation)

Analysis of Variance

We present a method to evaluate the association between a dissimilarity matrix and a set of categorical variables. This method has been introduced in ecology by Anderson (2001) and McArdle and Anderson (2001) to ecosystem analysis. A permutation approach for distance measure has also been used for evaluate genetic relations by Zapala and Schork (2006). Studer et al. (2009) apply the same method to sequence analysis in social sciences using the OMA distance matrix to evaluate the predictive value of a set of explanatory variables. This method is a generalization of the analysis of variance (ANOVA) in the case of semi-metric and non-metric measures. Let D be the matrix of dissimilarities with elements d_{ij} measure of dissimilarity between individuals i and j. Let 1 be an array of length n with every elements equal to 1

and A matrix such that $a_{ij} = -\frac{1}{2}d_{ij}$. Then the sum of squares SS_{tot} is equal to the trace of G (Gower's centered matrix, Gower and Krzanowski (1999)).

$$G = (I - \frac{1}{n}11^{'})A(I - \frac{1}{n}11^{'}) \tag{1}$$

McArdle and Anderson (2001) show that the sum of squares explained by the model and the sum of squares residual can be written as eq. 2and eq. 3, with $H = X(X^{'}X)^{-1}X^{'}$ known as the hat matrix in the linear regression model.

$$SS_{explained} = tr(HGH) \tag{2}$$

$$SS_{residual} = tr[(I - H)G(I - H)] \tag{3}$$

Analogously to ANOVA we can calculate the F statistic. The index c indicate the complete model with n variables that is compared to the model v composed by $m < n$ variables..

$$F_v = \frac{SS_{exp_c} - SS_{exp_v}/p}{SS_{res_c/(n-m-1)}} \tag{4}$$

Being distance measures far from normality, it is not possible that F follows a Fisher distribution as in the standard linear model. We utilize therefore a permutation approach to evaluate the statistical significance of the model. The distribution of F under the null hypothesis is obtained permuting simultaneously rows and columns of the matrix G, and repeating this process r times. The p-value of F is then obtained evaluating the number of times that F_{perm} is greater or equal to F_{obs}. Generally, 1,000 permutation are sufficient to get a significance level of 5%. Statistical analysis and permutation tests are conducted with the software R using the package TraMineR for sequence analysis, (Gabadinho et al., 2009).

Preliminary results

We present as an example, the preliminary results in table 1. We computed for females in Add-Health sample a simple model to explain the heterogeneity in the family transition sequences. As explanatory variables we consider: race/ethnicity; if respondents live with both biological parents at wave I; if both parents were born in US; if parents are college-educated and if the total family income is above the median value. As we can see from table 1, all the variables except income, contribute significantly to explain the heterogeneity on family trajectories. In particular, the variable with greater explanatory variable is the race of respondent, followed by the parents' birthplace and the typology of parents' family. As in the regression framework one can compare the adaptive power of different linear models, in a similar manner we can evaluate the explicatory power of model for sequence heterogeneity testing the F statistic.
In the complete paper we compare different explanatory models using different set of variables and testing for interactions. The analysis of sequence heterogeneity can contribute to explain divergence of family transition trajectories testing the effect of external variables as in regression models. Therefore, we can exclude variables that do not contribute significantly to the model and evaluate what matters more on explaining the divergence of life course trajectories.

Table 1. Association tests with family formation trajectories. Elaboration from a random subsample of 3,000 woman using 1,000 permutations

Variable	Pseudo$F$	PseudoR^2	p-value
Race/ethnicity	8.878073	0.0086785763	0.000
Family type	12.123745	0.0039504387	0.000
Both parents born in US	20.434555	0.0066584588	0.000
Parents' education	8.174696	0.0026636682	0.002
Total Family Income	2.828974	0.0009218016	0.050
Total	11.141117	0.0254032988	0.000

References

Abbott, A., 1995. Sequence analysis: new methods for old ideas. *Annual Review of Sociology*, 21:93–113.

Amato, P., N. Landale, and T. Havasevich-Brooks, 2008. Precursors of young women's family formation pathways. *Journal of Marriage and Family*, 70:1271–1286.

Anderson, M., 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26:32–46.

Billari, F., A. Aassve, and R. Piccarreta, 2007. Strings of Adulthood: A Sequence Analysis of Young British Women's Work-Family Trajectories. *European Journal of Population*, 23:369–388.

Billari, F. and R. Piccarreta, 2007. Clustering work and family trajectories by using a divisive algorithm. *Journal of Royal Statistical Society: Series A*, 170:1061–1078.

Elzinga, C. and A. Liefbroer, 2006. Intergenerational Transmission of Behavioral Patterns: Similarity of Parents' and Children's Family-Life Trajectories. *in press*.

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller, 2009. Mining Sequence Data in R with TraMineR: A User's Guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva. (TraMineR is on CRAN the Comprehensive R Archive Network).

Gower, J. and W. Krzanowski, 1999. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Journal of the Royal Statistical Society: series C*, 48(4):505–519.

McArdle, B. and M. Anderson, 2001. Fitting Multivariate Models to Community Data: A Comment on Distance-Based Redundancy Analysis. *Ecology*, 82(1):290–297.

Rindfuss, R., 1991. The Young Adult Years: Diversity, Structural Change, and Fertility. *Demography*, (29):493–512.

Schoen, R., N. Landale, and K. Daniels, 2007. Sequence analysis: new methods for old ideas. *Demography*, 44(4):807–820.

Studer, M., G. Ritschard, A. Gabadinho, and N. S. Müller, 2009. Analyse de dissimilaritès par arbre d'induction. *Revue des nouvelles technologies de l'information RNTI*, E-15:7–18.

Zapala, M. and N. Schork, 2006. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the National Academy of Science of the United States of America*, 103(51):19430–19435.