Article

Sociological Methods & Research
00(0) 1-6

DOI: 10.1177/0049124113503141
smr.sagepub.com

A Note on a Stata Plugin for Estimating Group-based Trajectory Models

Bobby L. Jones[1] and Daniel S. Nagin[2]

Abstract

Group-based trajectory models are used to investigate population differences in the developmental courses of behaviors or outcomes. This note introduces a new Stata command, *traj*, for fitting to longitudinal data finite (discrete) mixture models designed to identify clusters of individuals following similar progressions of some behavior or outcome over age or time. Normal, Censored normal, Poisson, Zero-inflated Poisson, and Logistic distributions are supported.

Keywords

Stata, plugin, group-based, trajectory, models

Introduction

A developmental trajectory measures the course of an outcome over age or time. This note introduces a Stata plugin for estimating group-based trajectory models that adapts to the Stata platform a well-established Statistical Analysis System (SAS)-based procedure for estimating group-based

[1] University of Pittsburgh Medical Center, Pittsburgh, PA, USA
[2] Carnegie Mellon University, Pittsburgh, PA, USA

Corresponding Author:
Daniel S. Nagin, Carnegie Mellon University, Carnegie Mellon, 5000 Forbes Ave., Pittsburgh, PA 15213, USA.
Email: dn03@andrew.cmu.edu

trajectory model demonstrated in two prior articles in this journal (Jones, Nagin, and Roeder 2001; Jones and Nagin 2007).

Using finite mixtures of suitably defined probability distributions, the group-based approach for modeling developmental trajectories is intended to provide a flexible and easily applied method for identifying distinctive clusters of individual trajectories within the population and for profiling the characteristics of individuals within the clusters. Thus, whereas the hierarchical and latent curve methodologies model population variability in growth with multivariate continuous distribution functions, the group-based approach utilizes a multinomial modeling strategy. Technically, the group-based trajectory model is an example of a finite mixture model. Maximum likelihood is used for the estimation of the model parameters. For a recent review of applications of group-based trajectory modeling, see Nagin and Odgers (2010); and for an extended discussion of the method, including technical details, see Nagin (2005).

The fundamental concept of interest is the distribution of outcomes conditional on age (or time); that is, the distribution of outcome trajectories denoted by $P(Y_i|\text{Age}_i)$, where the random vector Y_i represents individual i's longitudinal sequence of behavioral outcomes and the vector Age_i represents individual i's age when each of those measurements is recorded.[1] The group-based trajectory model assumes that the population distribution of trajectories arises from a finite mixture of unknown order J. The likelihood for each individual i, conditional on the number of groups J, may be written as

$$P(Y_i|\text{Age}_i) = \sum_{j=1}^{J} \pi^j \times P(Y_i|\text{Age}_i, j; \beta^j), \tag{1}$$

where π^j is the probability of membership in group j, and the conditional distribution of Y_i given membership in j is indexed by the unknown parameter vector β^j which among other things determines the shape of the group-specific trajectory. The trajectory is modeled with up to a fifth-order polynomial function of age (or time). For given j, conditional independence is assumed for the sequential realizations of the elements of Y_i, y_{it}, over the T periods of measurement. Thus, we may write

$$P(Y_i|\text{Age}_i, j; \beta^j) = \prod_{t=i}^{T} p(y_{it}|\text{age}_{it}, j; \beta^j), \tag{2}$$

where $p(.)$ is the distribution of y_{it} conditional on membership in group j and the age of individual i at time t.[2]

The software provides three alternative specifications of $p(.)$: *the censored normal distribution* also known as the *Tobit model*, the *zero-inflated Poisson distribution*, and the *binary logistic distribution*. The censored normal distribution is designed for the analysis of repeatedly measured, (approximately) continuous scales which may be censored by either a scale minimum or maximum or both (e.g., longitudinal data on a scale of depression symptoms). A special case is a scale or other outcome variable with no minimum or maximum. The zero-inflated Poisson distribution is designed for the analysis of longitudinal count data (e.g., arrests by age). The Poisson distribution is a special case with no zero inflation. The binary logistic distribution is available for the analysis of longitudinal data on a dichotomous outcome variable (e.g., whether hospitalized in year t or not).

The model also provides capacity for analyzing the effect of time stable covariate effects on probability of group membership and the effect of time dependent covariates on the trajectory itself. Let x_i denote a vector of time stable covariates thought to be associated with probability of trajectory group membership. Effects of time stable covariates are modeled with a generalized logistic function where without loss of generality $\theta_1 = 0$:

$$\pi_j(x_i) = \frac{e^{x_i\theta_j}}{\sum_j e^{x_i\theta_j}}.$$

Effects of time-dependent covariates on the trajectory itself are modeled by generalizing the specification of the polynomial function of age or time that defines the shape of the trajectory in the basic model without other covariates to include such covariate whether time varying (e.g., grade point average) or not (e.g., cohort membership). All parameter effect estimates are trajectory group specific. This allows parameters estimates not only for age or time to vary freely across trajectory group but also the parameter estimates for the other covariates included in the specification of the trajectory.

Installation

Traj can be installed by issuing the following commands within Stata. An additional command, trajplot, supports plotting the results.

```
. net from http://www.andrew.cmu.edu/user/bjones/traj
. net install traj, replace
. help traj
```

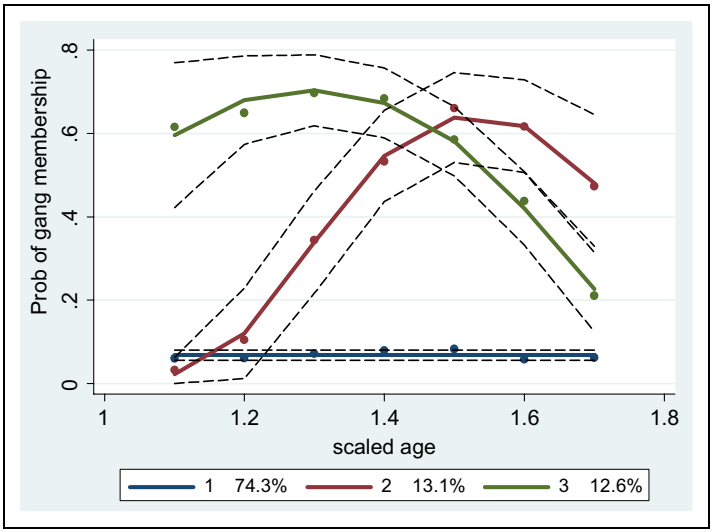

Figure 1. Trajectories of delinquent group membership.

An Example

Figure 1 illustrates an application of the method to data on self-reported delinquent group membership from age 11 to 17 in a large Montreal-based longitudinal study of over 1,000 males. The self-report is in the form of a binary indicator variable ($yes = 1/no = 0$). The model was estimated with the logistic specification of $p(*)$ and therefore the trajectories are defined by the probability of delinquent group membership over age. A three-group model was found to be best based on the Bayesian information criterion. One group, estimated to account for 74.3 percent of the sampled population, followed a trajectory of no involvement in delinquent groups. Another group estimated to account for 13.1 percent of population followed a trajectory of rising delinquent group membership in contrast to that of another group of about equal size which followed a trajectory of declining delinquent group membership. For details of this application, see Lacourse et al. (2002).

The solid lines in Figure 1 are based on the parameter estimates of the model itself. The dashed lines form a 95 percent confidence interval on the estimated probabilities of delinquent group membership. The dots are calculated with the actual data where each individual's responses are weighted based on posterior probabilities of group membership. This figure is the product of plotting software that is installed along with the estimation software.

Documentation

Documentation on the use of both the Stata plugin and the original SAS-based software is available at www.andrew.cmu.edu/user/bjones.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was generously supported by National Science Foundation Grants SES-102459 and SES-0647576.

Notes

1. Trajectories can also be defined by time (e.g., time from treatment).
2. See chapter 2 of Nagin (2005) for a discussion of the conditional independence assumption.

References

Jones, B. L. and D. S. Nagin. 2007. "Advances in Group-based Trajectory Modeling and an SAS Procedure for Estimating Them." *Sociological Methods & Research* 35:542-71.

Jones, B. L., D. S. Nagin, and K. Roeder. 2001. "A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories." *Sociological Methods & Research* 29:374-93.

Lacourse, R., S. Côte, D. S. Nagin, F. Vitaro, M. Brendgen, and R. E. Tremblay. 2002. "A Longitudinal-experimental Approach to Testing Theories of Antisocial Behavior Development." *Development and Psychopathology* 14:909-24.

Nagin, D. 2005. *Group-based Modeling of Development*. Cambridge, MA: Harvard University Press.

Nagin, D. and C. L. Odgers. 2010. "Group-based Trajectory Modeling in Clinical Research." *Annual Review of Clinical Psychology* 6:109-38.

Author Biographies

Bobby L. Jones is a statistician for the Center for Research on Health Care Data Center, University of Pittsburgh. Recent publications include "Advances in Group-based Trajectory Modeling and a SAS Procedure for Estimating Them" in *Sociological Methods and Research* (2007) with Daniel Nagin.

Daniel S. Nagin is the Teresa and H. John Heinz III University Professor of Public Policy and Statistics, Heinz College, Carnegie Mellon University. His research focuses on the evolution of criminal and antisocial behaviors during the life course, the deterrent effect of criminal and noncriminal penalties on illegal behaviors, and the development of statistical methods for analyzing longitudinal data. He is the author of *Group-based Modeling of Development* (2005).