Annals of Nutrition & Metabolism

Ann Nutr Metab 2014;65:205–210
DOI: 10.1159/000360229

Published online: November 18, 2014

Group-Based Trajectory Modeling: An Overview

Daniel S. Nagin

Carnegie Mellon University, Pittsburgh, Pa., USA

Key Words

Trajectory groups · Finite mixture modeling · Group-based trajectory modeling

Abstract

This article provides an overview of a group-based statistical methodology for analyzing developmental trajectories – the evolution of an outcome over age or time. Across all application domains, this group-based statistical method lends itself to the presentation of findings in the form of easily understood graphical and tabular data summaries. In so doing, the method provides statistical researchers with a tool for figuratively painting a statistical portrait of the predictors and consequences of distinct trajectories of development. Data summaries of this form have the great advantage of being accessible to nontechnical audiences and quickly comprehensible to audiences that are technically sophisticated. Examples of the application of the method are provided. A detailed account of the statistical underpinnings of the method and a full range of applications are provided by the author in a previous study.

Introduction

This paper provides an overview of group-based trajectory modeling (GBTM), a statistical methodology for analyzing developmental trajectories – the evolution of an outcome over age or time. A detailed account of the statistical underpinnings of the method and a full range of applications are provided in the literature [1, 2].

In this discussion, the term developmental trajectory is used to describe the progression of any phenomenon, whether behavioral, biological or physical. Charting and understanding developmental trajectories is among the most fundamental and empirically important research topics in the social and behavioral sciences and medicine. Longitudinal data – data with a time-based dimension – provide the empirical foundation for the analysis of developmental trajectories. Most standard statistical approaches for analyzing developmental trajectories, including hierarchical modeling [3, 4] and latent curve analysis [5, 6], are designed to account for individual variability about a mean population trend. However, many of the most interesting and challenging problems in longitudinal analysis have a qualitative dimension that allows for the possibility that there are meaningful subgroups within a population that follow distinctive developmental trajectories that are not identifiable ex ante based on some measured set of individual characteristics (e.g. gender or socioeconomic status). In psychology and medicine, for example, there is a long tradition of taxonomic theorizing about distinctive developmental progressions of these subcategories. For research problems with a taxonomic dimension, the aim is to chart out the distinctive trajectories, to understand what factors account for their distinctiveness and to test whether individuals following the different trajectories also respond differently to a treat-

KARGER

E-Mail karger@karger.com
www.karger.com/anm

0250–6807/14/0653–0205$39.50/0

Daniel S. Nagin
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 (USA)
E-Mail dn03@andrew.cmu.edu

ment, such as a medical intervention, or a major life event, such as the birth of a child. GBTM, which was first advanced by Nagin and Land [7] in 1993, provides the capacity for conducting group-based analysis with time- and age-based data.

Muthén and colleagues [8–10] have since developed an alternative group-based approach called growth mixture modeling [for comparative discussion of these alternative approaches, see 1, 2, 10]. However, for the purpose of this overview, the differences between the methods are secondary to their both being group-based approaches involving application of finite mixture modeling.

Across all application domains, the group-based trajectory statistical method, whether of the GBTM or growth mixture modeling variety, lends itself to the presentation of findings in the form of easily understood graphical and tabular data summaries. In so doing, the method provides statistical researchers with a tool for figuratively painting a statistical portrait of the predictors and consequences of distinct trajectories of development. Data summaries of this form have the great advantage of being accessible to nontechnical audiences and quickly comprehensible to audiences that are technically sophisticated.

An Illustration of GBTM

Figure 1 reports a well-known application of GBTM that was first reported by Nagin and Tremblay [11]. It is based on data assembled as part of the Montreal Longitudinal-Experimental Study of Boys that has tracked 1,037 males from school entry through young adulthood. Assessments were made on a wide range of factors. Among these were teacher reports of each boy's physical aggression at the age of 6 years and again annually from the age of 10 to 15 years. The scale was based on items such as frequency of fighting and physically bullying.

The best model was found to involve four groups. A group called 'lows' is comprised of individuals who display little or no physically aggressive behavior. This group is estimated to comprise about 15% of the sample population. A second group, comprising about 50% of the population, is best labeled 'moderate declining'. At age 6, boys in this group displayed a modest level of physical aggression, but by 10 years they had largely desisted. A third group, comprising about 30% of the population, is labeled 'high declining'. This group starts off scoring high on physical aggression at age 6 but scores far lower by the age of 15 years. Notwithstanding this marked decline, at age

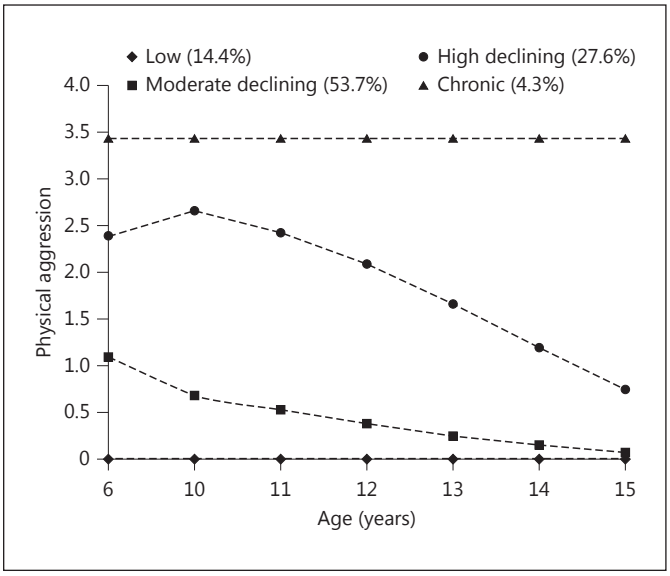


Fig. 1. Trajectories of physical aggression.

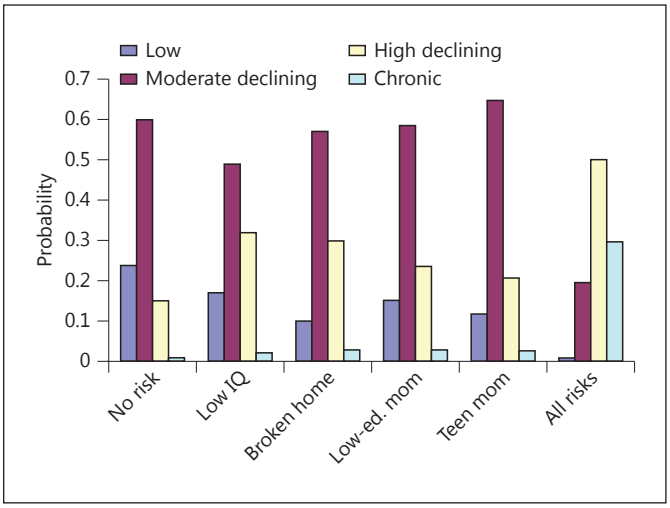
15 they continue to display a modest level of physical aggression. Finally, there is a small group of 'chronics', comprising less than 5% of the population, who display high levels of physical aggression throughout the observation period.

Much could be said about the implications of these trajectories for the development of physical aggression, but for our purposes here, two implications are emphasized. One implication follows from the observation that all the trajectories are either stable or declining from the initial assessment at age 6. This implies that to understand the developmental origins of physical aggression it is important to begin studying physical aggression at an even earlier age. A second and related observation is that the onset of physical aggression is not in adolescence as suggested by many theories of delinquent behavior (see Tremblay and Nagin [12] for a full development of these two observations).

These two points are highlighted because they illustrate the value of conducting longitudinal analysis in terms of groups. Nagin [1] and Nagin and Odgers [2] emphasize that the groups should not be interpreted as literal entities. Instead, they should be thought of as latent longitudinal strata in the data that are composed of individuals following approximately the same development course on the outcome of interest. These strata identify distinctive longitudinal features of the data. In this application, the fact that all of the trajectories are stable or declining is a feature of the data that is of great substantive

Table 1. Physical aggression group profiles

Variable	Group			
	low	moderate declining	high declining	chronic
Years of school – mother	11.1	10.8	9.8	8.4
Years of school – father	11.5	10.7	9.8	9.1
Low IQ, %	21.6	26.8	44.5	46.4
Completed 8th grade on time, %	80.3	64.6	64.6	6.5
Juvenile record, %	0.0	2.0	6.0	13.3
Sexual partners at age 17 (past year), %	1.2	1.7	2.2	3.5



Color version available online

Fig. 2. Impact of risk factors on group membership probabilities. ed. = Educated.

significance. Further, the absence of a feature, namely a trajectory reflecting the adolescent onset of physical aggression, also has important substantive significance.

The group-based methodology is intended to be responsive to calls for the development of 'person-based' approaches to analyzing development [13, 14]. Such appeals are motivated by a desire for methods that can provide a statistical snapshot of the distinguishing characteristics and behaviors of individuals following distinctive developmental pathways. The group-based method lends itself to creating such profiles. Table 1 reports profiles of the characteristics of individuals following the four physical aggression trajectories shown in figure 1. As developed in chapter 5 of Nagin [1], the parameter estimates of the model can be used to calculate the probability of an individual's belonging to each of the trajectory groups. These probabilities are called the posterior probability of group membership. To create the profiles reported in table 1, individuals were assigned to the trajectory group to which they most likely belonged based on their measured history of physical aggression. The summary statistics reported in table 1 are simply the product of a cross-tabulation of group membership with the various individual characteristics and outcomes. (The posterior probabilities can also be used to compute weights that account for uncertainty in individual-level trajectory group membership. However, use of these weights usually does not materially alter the profiles in well-fitting models.)

The profiles conform to long-standing findings on the predictors and consequences of problem behaviors such as physical aggression. Individuals in the chronic aggression group tend to have the least educated parents and most frequently score in the lowest quartile of the IQ distribution of the sample. By contrast, individuals in the low aggression group are least likely to suffer from these risk factors. Further, 90% of the chronic aggression group fail to reach the eighth grade on schedule and 13% have a juvenile record by age 18. By comparison, only 19% of the low aggression group had fallen behind grade level by the eighth grade and none have a juvenile record. In between are the moderate- and high-declining groups.

Table 1 demonstrates that trajectory group membership varies systematically with the individual's psychosocial characteristics. An important generalization of the base model that is laid out in chapter 6 of Nagin [1] allows for joint estimation of both the shapes of the trajectory groups and the impact of psychosocial characteristics on the probability of trajectory group membership. For example, such an analysis shows that the probability of physical aggression trajectory group membership is significantly predicted by low IQ, low paternal education, family breakup prior to age 6 and being born to a mother who began childbearing as a teenager [15]. Figure 2 reports calculations of trajectory group membership for various combinations of these risk factors for physical aggression based on estimated model coefficients.

Trajectories are not immutable. Life events or interventions may alter trajectories for the better or worse. Nagin et al. [16] explored the effect of grade retention from the age of 6 to 15 years on the trajectories of physical aggression shown in figure 1. They found that grade retention seems to exacerbate physical aggression in the low- and high-declining trajectory groups but has no ap-

parent effect on the physical aggression of the extreme groups – the lows and the chronics. The model extension allowing for this sort of analysis is developed in chapter 7 of Nagin [1] (see also Haviland et al. [17, 18] for a discussion of the use of propensity score matching in combination with GBTM in making causal inferences about the effect of life events and interventions on developmental trajectories).

Other important extensions of the basic model include joint/multitrajectory modeling and accounting for nonrandom subject dropout. Joint and multitrajectory modeling are designed to link trajectories of behaviors or outcomes that are thought to be theoretically related, such as trajectories of the level of physical activity in childhood and of body mass index in childhood and beyond, for example. Joint trajectory modeling probabilistically links trajectories of two different outcomes. Multitrajectory modeling is designed to link trajectories for two or more outcomes by defining a trajectory in terms of trajectories for all of the outcomes of interest. This modeling extension is described in chapter 8 of Nagin [1]. The other extension generalizes the basic model laid out in the next section to account for nonrandom subject dropout. In this model extension, each trajectory group is described by a trajectory and the probability of trajectory group membership at baseline and also by the probability of dropout for each period after baseline. This extension is described in Haviland et al. [19].

Likelihood Function

Group-based trajectory models are a specialized application of finite mixture models. While the conceptual aim of the analysis is to identify clusters of individuals with similar trajectories, the estimated parameters of the model are not the result of a cluster analysis. Rather they are the product of maximum likelihood estimation. As such, they share the many desirable characteristics of maximum likelihood parameter estimates – they are consistent and asymptotically normally distributed.

The specific form of the likelihood function to be maximized depends on the type of data being analyzed, but all are a special form of the following underlying likelihood function: let $Y_i = y_{i1}, y_{i2}, ..., y_{it}$ denote a longitudinal sequence of measurements on individual i over T periods. For expositional convenience, y_{it} will generally be described as the behavior of an individual. However, the outcome of interest does not have to pertain to an individual or a behavior – y_{it} can reference an entity such as a community, block face or an organization, or it can measure a quantity such as a poverty rate or a mean salary level.

Let $P(Y_i)$ denote the probability of Y_i. As developed in chapter 2 of Nagin [1], for count data $P(Y_i)$ is specified as the zero-inflated Poisson distribution, for censored data it is specified as the censored normal distribution and for binary data it is specified as the binary logit distribution. Whatever the probability distribution, the ultimate objective is to estimate a set of parameters, Ω, that maximizes the probability of Y_i. The particular form of this parameter set is distribution specific. However, across all distributions, these parameters perform the basic function of defining the shapes of the trajectories and the probability of group membership. As in standard growth curve modeling, the shapes of the trajectories are described by a polynomial function of age or time.

If the parameters of this polynomial function were constant across population members, the expected trajectory of all population members would be identical. Neither standard growth curve methods nor the group-based method assume such homogeneity. Indeed the assumption of homogeneity is antithetical to the objective of either approach because both aim to analyze the reason for individual differences in development. Standard growth curve modeling assumes that the parameters defining the polynomial describe only a population mean and that the trajectories of individual population members vary continuously about this mean, usually according to the multivariate normal distribution. The group-based method assumes that individual differences in trajectories can be summarized by a finite set of different polynomial functions of age or time. Each such set corresponds to a trajectory group which is hereafter indexed by j. Let $P^j(Y_i)$ denote the probability of Y_i given membership in group j, and π_j denote the probability of a randomly chosen population member belonging to group j.

If it were possible to observe group membership, the sampled individuals could be sorted by group membership and their trajectory parameters estimated with readily available Poisson, censored normal (Tobit) and logit regression software packages. However, group membership is not observed. Indeed, the proportion of the population comprising each group j, π_j, is an important parameter of interest in its own right. Thus, construction of the likelihood function requires the aggregation of the J conditional likelihood functions, $P^j(Y_i)$, to form the unconditional probability of the data, Y_i:

$$P(Y_i) = \sum_j^J \pi_j P^j(Y_i), \quad (1)$$

where $P(Y_i)$ is the unconditional probability of observing individual *i*'s longitudinal sequence of behavioral measurements, Y_i. It equals the sum across the *J* groups of the probability of Y_i given *i*'s membership in group *j* weighted by the probability of membership in group *j*. Equation 1 describes what is called a 'finite mixture model' because it sums across a finite number of discrete groups that comprise the population. The term 'mixture' is included in the label because the statistical model specifies that the population is composed of a mixture of unobserved groups.

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,

$$P^j(Y_i) = \prod^{T} p^j(y_{it}), \quad (2)$$

where $p^j(y_{it})$ is the probability distribution function of y_{it} given membership in group *j*.

The rationale for the conditional independence assumption deserves elaboration. This assumption implies that for each individual within a given trajectory group *j*, the distribution of y_{it} for period *T* is independent of the realized level of the outcome in prior periods, y_{it-1}, y_{it-2}, …. Thus, $p^j(y_{it})$ does not include prior values of y_{it} in its specification. This assumption greatly reduces the complexity of an already complex model. Due to this reduction in complexity, most applications of finite mixture modeling with longitudinal data assume conditional independence for the sake of tractability.

On its face, the conditional independence assumption may seem implausible because it would seem to imply that current behavioral outcomes are uncorrelated with past outcomes. At the level of the group which is not observed, this is indeed the case. For individuals within a given group *j*, behavioral outcomes over time are assumed not to be serially correlated in the sense that individual level deviations from the group trend are uncorrelated. However, even with the assumption of conditional independence at the level of the latent group, there will still be serial dependence over time at the level of the population. Specifically, past outcomes will be correlated with current outcomes. Such serial dependence results from the group-specific specification of $p^j(y_{it})$. Differences in this specification across groups allow for persistent differences in the outcome variable across population members.

The conditional independence assumption is also invoked in the standard random effect model that underlies conventional growth curve models. The random effect model assumes that the sequential realizations of y_{it} are independent, conditional upon the individual's random effect. Thus, in the group-based model, the conditional independence assumption is made at the level of the group, whereas in the random effect model it is invoked at the level of the individual. In this sense, the conditional independence assumption is stronger in the group-based model than in the standard random effect model. Balanced against this disadvantage is the advantage that the group-based model does not make the very strong assumption that the random effect is independently and identically distributed according to the normal distribution.

The likelihood for the entire sample of *N* individuals is simply the product of the individual likelihood functions of the *N* individuals comprising the sample:

$$L = \prod^{N} P(Y_i).$$

Intuitively, the estimation procedure for all data types identifies distinctive trajectory groups as follows. Suppose a population is composed of two distinct groups: (i) youth offenders (comprising 50% of the population) who up to age 18 have an expected offending rate, λ, of 5 and who after age 18 have a λ of 1 and (ii) adult offenders (comprising the other 50% of the population) whose offending trajectory is the reverse of that of the youth offenders – through age 18 their $\lambda = 1$ and after age 18 their λ increases to 5. Longitudinal data on the recorded offenses of a sample of individuals from this population would reveal two distinct groups: a clustering of about 50% of the sample who have had many offenses prior to 18 years and relatively few offenses after age 18, and another 50% clustering with the reverse pattern.

If these data were analyzed under the assumption that the relationship between age and λ was identical across all individuals, the estimated value of λ would be a 'compromise' estimate of about 3 for all ages. From this, one might mistakenly conclude that the rate of offending is invariant with age in this population. If the data were instead analyzed using the group-based approach, which specifies the likelihood function as a mixing distribution, no such mathematical 'compromise' would be necessary. The parameters of one component of the mixture would effectively be used to accommodate (i.e. match) the youth offending portion of the data whose offending declines with age and another component of the mixing distribution would be available to accommodate the adult offender data whose offending increases with age.

Concluding Remarks

A hallmark of modern longitudinal studies is the variety and richness of measurements that are made about the study subjects and their circumstances. Less often acknowledged is that this abundance of information is accompanied by a difficult companion – complexity. Commonly, researchers are confronted with the dilemma of how best to explore and communicate the rich set of measurements at their disposal without becoming so bogged down in complexity that the lessons to be learned from the data are lost on them and their audience.

An important motivation for my commitment to developing and promoting the group-based trajectory method is the belief that alternative methods for analyzing development in longitudinal data sets too often leave the researcher with a Hobson's choice of balancing comprehensibility against an adequate exploration of complexity. GBTM does not solve the problem of balancing comprehensibility and complexity. However, it does provide researcher's with a valuable tool for identifying, summarizing and communicating complex patterns in longitudinal data.

Summarizing data necessarily requires reduction. Reduction requires approximation. In the case of group-based models, the approximation involves the grouping of individuals who are not entirely homogenous. Balanced against this reduction error is a greatly expanded capability for creating dense, yet comprehensible, descriptions of groups of people through time.

Acknowledgment

This research has been supported by the National Science Foundation (NSF: SES-99113700 and SES-0647576) and the National Institute of Mental Health (RO1 MH65611-01A2).

Disclosure Statement

The author has nothing to disclose.

References

1 Nagin DS: Group-Based Modeling of Development. Cambridge, Harvard University Press, 2005.

2 Nagin DS, Odgers CL: Group-based trajectory modeling in clinical research. Annu Rev Clin Psychol 2010;6:109–138.

3 Bryk AS, Raudenbush SW: Application of hierarchical linear models to assessing change. Psychol Bull 1987;101:147–158.

4 Goldstein H: Multilevel Statistical Models. London, Arnold, 1995.

5 McArdle JJ, Epstein D: Latent growth curves within developmental structural equation models. Child Dev 1987;58:110–113.

6 Willett JB, Sayer AG: Using covariance structure analysis to detect correlates and predictors of individual change over time. Psychol Bull 1994;116:363–381.

7 Nagin DS, Land KC: Age, criminal careers, and population heterogeneity – specification and estimation of a nonparametric, mixed Poisson model. Criminology 1993;31:327–362.

8 Muthén B, Shedden K: Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics 1999;55:463–469.

9 Muthén B, Brown CH, Masyn K, Jo B, Khoo ST, Yang CC, et al: General growth mixture modeling for randomized preventive interventions. Biostatistics 2002;3:459–475.

10 Muthén B: Latent variable analysis: growth mixture modeling and related techniques for longitudinal data; in Kaplan D (ed): Handbook of Quantitative Methodology for the Social Sciences. Newbury Park, Sage, 2004, pp 345–368.

11 Nagin DS, Tremblay RE: Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. Child Dev 1999;70:1181–1196.

12 Tremblay RE, Nagin DS: Aggression in humans; in Tremblay RE, Hartup WW, Archer J (eds): Developmental Origins of Aggression. New York, Guilford, 2005, pp 83–106.

13 Bergman LR: A pattern-oriented approach to studying individual development: snapshots and processes; in Cairns RB, Bergman LR, Kagan J (eds): Methods and Models for Studying the Individual. Thousand Oaks, Sage, 1998, pp 83–122.

14 Magnusson D: The logic and implications of a person-oriented approach; in Cairns RB, Bergman LR, Kagan J (eds): Methods and Models for Studying the Individual. Thousand Oaks, Sage, 1998, pp 33–64.

15 Nagin D S, Tremblay R E: Parental and early childhood predictors of persistent physical aggression in boys from kindergarten to high school. Arch Gen Psychiatry 2001;58:389–394.

16 Nagin D, Pagani L, Tremblay R, Vitaro F: Life course turning points: a case study of the effect of school failure on interpersonal violence. Dev Psychopathol 2003;15:343–361.

17 Haviland A, Nagin DS, Rosenbaum PR: Combining propensity score matching and group-based trajectory modeling in an observational study. Psychol Methods 2007;12:247–267.

18 Haviland A, Nagin DS, Rosenbaum PR, Tremblay RE: Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data. Dev Psychol 2008;44:422–436.

19 Haviland A, Jones B, Nagin DS: Group-based trajectory modeling extended to account for non-random subject attrition. Sociol Methods Res 2011;41:367–390.