A Stata Plugin for Estimating Group-Based Trajectory Models

Bobby L. Jones

University of Pittsburgh Medical Center

Daniel S. Nagin

Carnegie Mellon University

May 21, 2012

This work was generously supported by National Science Foundation Grants SES-102459 and SES-0647576

Abstract

Group-based trajectory models are used to investigate population differences in the developmental courses of behaviors or outcomes . This article demonstrates 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. Censored normal, Poisson, zero-inflated Poisson, and Bernoulli distributions are supported. Applications to psychometric scale data, count data, and a dichotomous prevalence measure are illustrated.

Introduction

A developmental trajectory measures the course of an outcome over age or time. The study of developmental trajectories is a central theme of developmental and abnormal psychology and psychiatry, of life course studies in sociology and criminology, of physical and biological outcomes in medicine and gerontology. A wide variety of statistical methods are used to study these phenomena. This article demonstrates a Stata plugin for estimating group-based trajectory models. The Stata program we demonstrate adapts a well-established SAS-based procedure for estimating group-based trajectory model (Jones, Nagin, and Roeder, 2001; Jones and Nagin, 2007) to the Stata platform.

Group-based trajectory modeling is a specialized form of finite mixture modeling. The method is designed identify groups of individuals following similarly developmental trajectories. 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).

A Brief Overview of Group-Based Trajectory Modeling

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. The maximization is performed using a general quasi-Newton procedure (Dennis, Gay, and Welsch 1981; Dennis and Mei 1979).

The fundamental concept of interest is the distribution of outcomes conditional on age (or time); that is, the distribution of outcome trajectories denoted by $P(Y_i \mid Age_i)$, 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

[1] Trajectories can also be defined by time (e.g., time from treatment).

$$P(Y_i \mid Age_i) = \sum_{j=1}^{J} \pi^j \cdot P(Y_i \mid Age_i, j; \beta^j) \quad (1),$$

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 5th 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 $P(Y_i \mid Age_i, j; \beta^j) = \prod_{t=i}^{T} p(y_{it} \mid age_{it}, j; \beta^j) \quad (2),$

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 logit 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) and binary logit distribution 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 logit 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

[2] See chapter 2 of Nagin (2005) for a discussion of the conditional independence assumption.

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
```

Syntax

traj [if *exp*] , var(*varlist*) indep(*varlist*) model(*string*) order(*numlist*)

[min(*real*) max(*real*) iorder(*numlist*) risk(*varlist*) tcov(*varlist*) plottcov(*matrix*)

start(*matrix*) weight(*varname*) exposure(*varlist*) refgroup(*integer*) dropout(*numlist*)

dcov(*varlist*) obsmar(*varname*) outcome(*varname*) omodel(*string*) detail][3]

Trajectory Variables

var(*varlist*) dependent variables, measured at different times or ages (required).

indep(*varlist*) independent variables i.e. when the dependant variables were measured (required).

Model

model(*string*) probability distribution for the dependent variables (required). Models supported: cnorm, zip, logit.

---

[3] [if exp] is a standard option for Stata commands to allow you to select a data subset for analysis e.g. traj if male == 1, var(opp*) ...

order(*numlist*) polynomial type (0=intercept, 1=linear, 2=quadratic, 3=cubic) for each group trajectory (required).

min(*real*) minimum value for the censored normal model (required for cnorm).

max(*real*) maximum value for the censored normal model (required for cnorm).

iorder(*numlist*) optional polynomial type (0=intercept, 1=linear, 2=quadratic, 3=cubic) for the zero-inflation of each group.

exposure(*varlist*) exposure variables for the zero-inflated Poisson model.

weight(*varname*) a probability weight variable.

Time-Stable Covariates for Group Membership

risk(*varlist*) covariates for the probability of group membership.

refgroup(*integer*) the reference group (default = 1) when the risk option is used.

Time-Varying Covariates for Group Membership

tcov(*varlist*) time-varying covariates for the group trajectories.

plottcov(*matrix*) optional values for plotting trajectories with time-varying covariates.

Dropout Model

dropout(*numlist*) include logistic model of dropout probability per wave. For each group, 0 = constant rate, 1 = depends on the previous response, 2 = depends on the two previous responses.

dcov(*varlist*) optional lagged time-varying covariates for the dropout model.

obsmar(*varname*) a binary variable to mark which observations are to be included in the dropout model and those to be treated as missing at random. This variable = 1 for observations to be treated as data MAR (include completers) and = 0 for observations to be used for the modeled dropout.

Distal Outcome Model

outcome(*varlist*) a distal variable to be regressed on the probability of group membership.

omodel(*string*) the outcome model to be used.

Joint Trajectory Model

The joint model uses the options shown above with a "2" suffix to specify the second model in the joint trajectory model e.g. model2(cnorm) etc.

Miscellaneous

start(*matrix*) parameter start values to override default start values.

The detail option will show the minimization iterations.

Trajplot Syntax

trajplot , [xtitle(*string*) ytitle(*string*) model(*integer*) ci]

xtitle(*string*) x-axis title

ytitle(*string*) y-axis title

xlabel(*string*) passed to twoway scatter xlabel option for x-axis control.

ylabel(*string*) passed to twoway scatter ylabel option for y-axis control.

model(*integer*) indicates which model to graph in the joint trajectory model (1 or 2, default = 1).

The ci option includes 95% confidence intervals on the graph.

Examples

Censored Normal Model

The data consist of annual assessments on 1,037 boys at age 6 (spring 1984) and ages 10 through 15 on an oppositional behavior scale (ranges from 0 to 10) gathered in low socioeconomic areas of Montreal, Canada. See Tremblay et al. (1987) for details. Scores of zero are frequent and the scores decrease in frequency as the score increases. Hence, the censored normal distribution is sensible for modeling the data. The following commands fit a five-group model to the opposition data and provide a graph of the results.

```
. traj , model(cnorm) var(o1-o7) indep(t1-t7) order(1 2 3 2 2) min(0) max(10)

==== traj stata plugin ====  Jones BL  Nagin DS

1037 observations read.
1037 observations used in trajectory model.

                      Maximum Likelihood Estimates
                      Model: Censored Normal (CNORM)

                                          Standard       T for H0:
 Group    Parameter        Estimate          Error     Parameter=0    Prob > |T|

 1        Intercept         0.94328        0.52230           1.806        0.0710
          Linear           -0.25878        0.04670          -5.541        0.0000

 2        Intercept        -6.40139        1.55350          -4.121        0.0000
          Linear            1.90218        0.35576           5.347        0.0000
          Quadratic        -0.08641        0.01734          -4.985        0.0000

 3        Intercept       -14.63833        4.40955          -3.320        0.0009
          Linear            5.09379        1.41972           3.588        0.0003
          Quadratic        -0.48399        0.14225          -3.402        0.0007
          Cubic             0.01394        0.00450           3.098        0.0020

 4        Intercept         3.34775        2.38395           1.404        0.1603
          Linear            1.04058        0.48281           2.155        0.0312
          Quadratic        -0.08839        0.02351          -3.760        0.0002

 5        Intercept         0.44931        3.43962           0.131        0.8961
          Linear            1.14858        0.58290           1.970        0.0488
          Quadratic        -0.05814        0.02651          -2.193        0.0283

          Sigma             2.51297        0.03350          75.007        0.0000

  Group membership

 1        (%)              21.55060        3.00956           7.161        0.0000
 2        (%)              17.33432        3.29936           5.254        0.0000
 3        (%)              41.18017        3.97147          10.369        0.0000
 4        (%)               7.08815        1.93424           3.665        0.0002
 5        (%)              12.84676        6.36291           2.019        0.0435

 BIC=-11675.44 (N=6231)  BIC=-11657.50 (N=1037)  AIC=-11608.06  L=-11588.06

. trajplot, xtitle("Age") ytitle("Opposition")

. list _traj_Group - _traj_ProbG5 if _n < 3, ab(12)
```

```
     +-------------------------------------------------------------------------------------+
     | _traj_Group   _traj_ProbG1   _traj_ProbG2   _traj_ProbG3   _traj_ProbG4   _traj_ProbG5 |
     |-------------------------------------------------------------------------------------|
  1. |           1       .8698951       8.89e-06       .1300557       .0000404       1.82e-10 |
  2. |           3       .0052927       .0225955       .9638394       .0082598       .0000126 |
     +-------------------------------------------------------------------------------------+
```

In Figure 1 we see that there is a group of subjects exhibiting little or no oppositional behavior (group 1, 21.6%); a group showing moderate levels of oppositional behavior (group 2, 17.3%); a group exhibiting low and somewhat decreasing levels of oppositional behavior (group 3, 41.2%); a group that starts out with high levels of oppositional behavior that drops steadily with age (group 4, 7.1%); and a fifth group exhibiting chronic problems with oppositional behavior (group 5, 12.8%). Also shown are the group assignment and group membership probabilities for the first two subjects.

Zero Inflated Poisson (ZIP) Model

The next example is an analysis of Poisson data with extra zeros. The data are the annual number of criminal offense convictions for 411 subjects from a prospective longitudinal survey conducted in a working-class section of London (Farrington and West, 1990). The annual criminal offense convictions were recorded for boys from age 10 through age 30. The Poisson model is appropriate here; however, more zeros are present than would be expected in the purely Poisson model, so we will use the ZIP model. The following commands fit a four-group model to the data and provide a graph of the results.

```
. traj, model(zip) var(y*) indep(t*) order(2 0 2 3) iorder(1)

==== traj stata plugin ====  Jones BL  Nagin DS

403 observations read.
403 observations used in trajectory model.

                        Maximum Likelihood Estimates
                        Model: Zero Inflated Poisson (ZIP)

                                          Standard       T for H0:
 Group    Parameter        Estimate          Error     Parameter=0    Prob > |T|

 1        Intercept        -6.94381        1.27775          -5.434        0.0000
          Linear            5.78661        1.27104           4.553        0.0000
          Quadratic        -1.27200        0.30836          -4.125        0.0000

 2        Intercept        -4.42373        0.32406         -13.651        0.0000

 3        Intercept       -22.06650        5.19544          -4.247        0.0000
          Linear           26.97483        6.83217           3.948        0.0001
          Quadratic        -8.39924        2.22557          -3.774        0.0002

 4        Intercept       -16.44563        2.91087          -5.650        0.0000
```

```
         Linear          25.89964      4.71593          5.492       0.0000
         Quadratic      -12.35311      2.45616         -5.029       0.0000
         Cubic            1.84687      0.40854          4.521       0.0000

         Alpha0          -3.20124      0.95673         -3.346       0.0008
         Alpha1           1.01007      0.42901          2.354       0.0186


  Group membership

 1       (%)             12.86916      2.53525          5.076       0.0000
 2       (%)             67.16193      3.61322         18.588       0.0000
 3       (%)             12.84928      3.19528          4.021       0.0001
 4       (%)              7.11963      1.50838          4.720       0.0000

 BIC= -1516.62 (N=4433)  BIC= -1497.43 (N=403)  AIC= -1465.44  L= -1449.44
```

```
. trajplot, xtitle("Scaled Age") ytitle("Annual Conviction Rate") ci
```

In Figure 2 we see a low chronic offending group (group 1, 12.9%), a negligible-offending group (group 2, 67.2%), an adolescent-limited offending group (group 3, 12.8%) that desists after age 20, and a high offending group (group 4, 7.1%) which has the highest offense rate, occurring during adolescence and early adulthood.

Logistic Model Example

It is common in research on criminal careers to analyze the absence or presence of offenses i.e. a dichotomous prevalence measure. The ZIP analysis is repeated for a derived criminal offense prevalence measure using a logistic model (i.e., periods in which 1 or more convictions are reported are coded as “1” and periods with no convictions are coded as “0”). The following commands fit a three-group model to the prevalence measure data and graph the results.

```
. traj , model(logit) var(p1-p23) indep(tt1-tt23) order(3 3 3)

==== traj stata plugin ====  Jones BL  Nagin DS

403 observations read.
403 observations used in trajectory model.

                        Maximum Likelihood Estimates
                        Model: Logistic (LOGIT)

                                          Standard       T for H0:
 Group   Parameter        Estimate          Error     Parameter=0   Prob > |T|

 1       Intercept       -25.35013      12.32967          -2.056       0.0398
         Linear           29.76964      18.14679           1.640       0.1009
         Quadratic       -13.90460       8.52790          -1.630       0.1030
         Cubic             2.06302       1.28226           1.609       0.1077
```

```
 2       Intercept      -23.57755      3.85244         -6.120       0.0000
         Linear          32.17585      5.86600          5.485       0.0000
         Quadratic      -14.83642      2.87438         -5.162       0.0000
         Cubic            2.13003      0.44848          4.749       0.0000

 3       Intercept      -16.70902      4.06486         -4.111       0.0000
         Linear          24.07013      6.44102          3.737       0.0002
         Quadratic      -10.90374      3.20221         -3.405       0.0007
         Cubic            1.52698      0.50403          3.030       0.0025


  Group membership

 1       (%)             69.28631      4.35489         15.910       0.0000
 2       (%)             24.86304      3.63898          6.832       0.0000
 3       (%)              5.85064      2.04204          2.865       0.0042

 BIC= -1544.21 (N=9269)  BIC= -1522.26 (N=403)  AIC= -1494.27  L= -1480.27
```

```
. trajplot, xtitle("Scaled Age") ytitle("Prevalence")
```

Figure 3 shows a group of subjects, 69.3%, classified as never convicted, 24.9% percent have a low prevalence rate that peaks during adolescence, and the remaining 5.9% percent exhibit a high prevalence rate.

Introducing Time-Stable Covariates

A common modeling objective is to establish whether a trait (e.g., being prone to oppositional behavior) is linked to measured covariates. Suppose we were interested in investigating if high inattention, IQ, and adverse home life are risk factors for elevated levels of opposition. Note that the procedure drops observations with missing risk factor data.

```
. traj, model(cnorm) var(o1-o7) indep(tt1-tt7) order(3 3 3 3 3) min(0) max(10)
risk(qiver91 advers84 h_inatt)


==== traj stata plugin ====  Jones BL  Nagin DS

1037 observations read.
169 had missing values in risk factors/covariates or weights=0.
868 observations used in trajectory model.


                        Maximum Likelihood Estimates
                        Model: Censored Normal (CNORM)

                                          Standard       T for H0:
 Group    Parameter        Estimate          Error     Parameter=0   Prob > |T|

 1       Intercept         -0.10091        8.65067          -0.012       0.9907
         Linear             1.64879       27.92236           0.059       0.9529
         Quadratic         -5.09960       27.90012          -0.183       0.8550
         Cubic              1.93624        8.82656           0.219       0.8264

 2       Intercept        -31.51074        8.16457          -3.859       0.0001
         Linear            97.87435       25.57636           3.827       0.0001
```

```
         Quadratic      -85.43687     25.09830         -3.404       0.0007
         Cubic           23.95584      7.84366          3.054       0.0023

 3       Intercept      -11.57052      4.30136         -2.690       0.0072
         Linear          42.24753     13.75063          3.072       0.0021
         Quadratic      -39.77034     13.67900         -2.907       0.0037
         Cubic           11.14537      4.31101          2.585       0.0098

 4       Intercept        9.06549     10.92393          0.830       0.4066
         Linear         -12.00292     34.77165         -0.345       0.7300
         Quadratic       16.77470     34.55777          0.485       0.6274
         Cubic           -8.31971     10.88237         -0.765       0.4446

 5       Intercept       12.49888     12.79459          0.977       0.3287
         Linear         -20.98325     40.21239         -0.522       0.6018
         Quadratic       21.21777     39.71333          0.534       0.5932
         Cubic           -6.88491     12.49889         -0.551       0.5818

         Sigma            2.51488      0.03483         72.198       0.0000

  Group membership

 1       Constant        (0.00000)           .              .           .

 2       Constant         2.50478      0.75190          3.331       0.0009
         qiver91         -0.43559      0.07907         -5.509       0.0000
         advers84         2.91461      0.61290          4.755       0.0000
         h_inatt          0.38632      0.76662          0.504       0.6143

 3       Constant         1.31502      0.70280          1.871       0.0614
         qiver91         -0.11713      0.06857         -1.708       0.0877
         advers84         1.09161      0.49255          2.216       0.0267
         h_inatt          1.62529      0.57478          2.828       0.0047

 4       Constant        -3.51027      1.48782         -2.359       0.0183
         qiver91          0.09158      0.12837          0.713       0.4756
         advers84         3.26347      0.84741          3.851       0.0001
         h_inatt          3.10326      0.62563          4.960       0.0000

 5       Constant        -1.30740      1.26089         -1.037       0.2998
         qiver91         -0.30800      0.10634         -2.896       0.0038
         advers84         4.66138      1.05203          4.431       0.0000
         h_inatt          3.47356      0.72187          4.812       0.0000

 BIC=-10414.14 (N=5642)  BIC=-10379.51 (N=868)  AIC=-10291.34  L=-10254.34
```

The portion of the output below “Group membership” gives log-odds estimates for the risk factors for each group relative to group 1. As an example, taking the estimates for group 5 the high oppositional behavior group, we see that as adversity in the home and inattention scores increase, so do the likelihood of problems with high oppositional behavior. However, as IQ increases, the likelihood of belonging to the high opposition group decreases.

To aid in illustrating the effect of risk factors on group membership probabilities, dummy observations without trajectory data but with risk factor variables set to desired values can be added to the data. These will not affect the trajectory model, but group membership probabilities based on the risk factor settings are generated.

Time-Varying Covariates / An Example of the Use of Wald Tests

Including time-varying covariates allows a trajectory to depend on additional variables beyond age or time. For example, Laub et al. (1998) examine the impact of marriage on deflecting trajectories of offending from high levels of criminality toward desistance. Life events may also have transitory affects on enduring trajectories of behavior. For instance, spells of mental illness may temporarily alter trajectories of high-level productivity.

Consider an analysis of the effect of gang membership on violent delinquency. This analysis is based on self-reports from the Montreal data on violent delinquency from age 11 to 17 and companion self-reports of whether or not the boy was involved in a delinquent group at that age. In this five-group analysis, the estimate of the effect of gang membership was positive and highly significant for each group implying that gang membership is associated with increased violence.

To aid in the graphical presentation of estimated time-varying covariate effects, the plottcov option will calculate the trajectory for each group using a specified set of values for time-varying covariates. These calculations are done post-model estimation based on the estimated values of the coefficients that define the trajectory over time including the coefficients measuring the effects of all the covariates included in the trajectory. Figure 4 shows the graph of the predicted trajectories for not in a gang from age 11 to 17 compared to those for joining a gang at age 14.

The following commands create the predicted trajectories shown in Figure 4:

```
. matrix tc1 = (0, 0, 0, 0, 0, 0, 0)

. matrix tc2 = (0, 0, 0, 1, 1, 1, 1)

. traj, model(zip) var(bat*) indep(t*) tcov(gang*) order(2 2 2 2 2)
plottcov(tc1)

. trajplot, xtitle("Scaled Age") ytitle("Rate")

. traj, model(zip) var(bat*) indep(t*) tcov(gang*) order(2 2 2 2 2)
plottcov(tc2)

. trajplot, xtitle("Scaled Age") ytitle("Rate")


==== traj stata plugin ====  Jones BL  Nagin DS

909 observations read.
909 observations used in trajectory model.

                        Maximum Likelihood Estimates
                        Model: Zero Inflated Poisson (ZIP)

                                          Standard       T for H0:
 Group    Parameter        Estimate         Error     Parameter=0   Prob > |T|

 1        Intercept        10.11041       2.90206           3.484       0.0005
```

```
          Linear          -14.13295      4.36383          -3.239       0.0012
          Quadratic         4.11709      1.60717           2.562       0.0104
          gang89            1.25526      0.13013           9.646       0.0000

 2        Intercept       -14.85884      2.30554          -6.445       0.0000
          Linear           19.38033      3.17994           6.095       0.0000
          Quadratic        -6.18508      1.08792          -5.685       0.0000
          gang89            1.01374      0.07299          13.888       0.0000

 3        Intercept        -3.70066      1.86139          -1.988       0.0468
          Linear            8.39237      2.76156           3.039       0.0024
          Quadratic        -3.93817      1.00693          -3.911       0.0001
          gang89            0.76535      0.07401          10.342       0.0000

 4        Intercept        -2.35470      1.26687          -1.859       0.0631
          Linear            4.83802      1.81563           2.665       0.0077
          Quadratic        -1.67985      0.63920          -2.628       0.0086
          gang89            0.61311      0.04181          14.663       0.0000

 5        Intercept        -5.83974      1.65795          -3.522       0.0004
          Linear           10.85757      2.38069           4.561       0.0000
          Quadratic        -3.81160      0.84685          -4.501       0.0000
          gang89            0.48966      0.05969           8.204       0.0000


  Group membership

 1        (%)              31.65966      2.18419          14.495       0.0000
 2        (%)              21.45111      2.27352           9.435       0.0000
 3        (%)              23.71713      2.32856          10.185       0.0000
 4        (%)              18.69802      1.66333          11.241       0.0000
 5        (%)               4.47408      0.86109           5.196       0.0000

 BIC= -9645.47 (N=5962)  BIC= -9622.90 (N=909)  AIC= -9565.15  L= -9541.15
```

Wald tests can be performed on model parameter estimates. As an example we investigate differential gang membership effects by trajectory group. In the tests below, we see that the coefficient estimates of gang effect differ for groups 1 and 5 ($p < 0.0001$) but do not differ for groups 4 and 5 ($p = 0.0917$).

```
. testnl _b[gang891] = _b[gang895]

  (1)  _b[gang891] = _b[gang895]

               chi2(1) =       28.45
           Prob > chi2 =        0.0000

. testnl _b[gang894] = _b[gang895]

  (1)  _b[gang894] = _b[gang895]

               chi2(1) =        2.84
           Prob > chi2 =        0.0917
```

Joint Trajectory Model

The joint model was designed to analyze the developmental course of two distinct but related outcomes (Nagin and Tremblay 2001). The model can be used to analyze connections between the developmental trajectories of two outcomes that are evolving contemporaneously (e.g., depression and alcohol use) or that evolve over different time periods (e.g., prosocial behavior in childhood and school achievement in adolescence). The three key outputs of the dual model are as follows: (1) the trajectory groups for both measurement series, (2) the probability of membership in each identified trajectory group, and (3) conditional probabilities linking membership across the trajectory groups of the two respective behaviors. Loeber (1991) has argued that covert behaviors in childhood, such as opposition, are linked to another form of covert behavior in adolescence, property delinquency. We illustrate the joint trajectory model with an analysis of the linkage of opposition from ages 6 to 13 with property delinquency from ages 13 to 17. The model is estimated with data from the Montreal based longitudinal study.

The following commands fit the joint trajectory model:

```
. traj , model(cnorm) var(qcp84op qcp88op qcp89op qcp90op qcp91op) indep(tt1-
tt5) order(2 2 2) max(10) var2(qas91det qas92det qas93det qas94det qas95det)
indep2(tt3 tt4 tt5 tt6 tt7) model2(zip) order2(2 2 2 2)

. trajplot, ytitle("Opposition") xtitle("Scaled Age")

. trajplot, model(2) ytitle("Rate") xtitle("Scaled Age")


==== traj stata plugin ====  Jones BL  Nagin DS

733 observations read.
733 observations used in trajectory model.


                        Maximum Likelihood Estimates
        Model 1: Censored Normal (CNORM)  Model 2: Zero Inflated Poisson (ZIP)
Model 1: Censored Normal (CNORM)

                                          Standard       T for H0:
 Group    Parameter         Estimate        Error     Parameter=0    Prob > |T|

 1        Intercept         -0.03901      1.84391          -0.021       0.9831
          Linear             1.05880      4.33208           0.244       0.8069
          Quadratic         -2.01183      2.36294          -0.851       0.3946

 2        Intercept         -4.21044      1.40907          -2.988       0.0028
          Linear            15.69645      3.23269           4.856       0.0000
          Quadratic         -8.94934      1.72296          -5.194       0.0000

 3        Intercept         -4.73618      2.04613          -2.315       0.0207
          Linear            21.90191      4.66222           4.698       0.0000
          Quadratic        -11.62520      2.48220          -4.683       0.0000

          Sigma              2.56923      0.04439          57.874       0.0000

  Group membership

 1        (%)               35.50349      3.24974          10.925       0.0000
 2        (%)               45.02307      3.05005          14.761       0.0000
```

3	(%)	19.47343	2.78947	6.981	0.0000

Model 2: Zero Inflated Poisson (ZIP)

1	Intercept	-15.51399	8.61636	-1.801	0.0718
	Linear	18.83007	12.85447	1.465	0.1430
	Quadratic	-5.18513	4.75720	-1.090	0.2758
2	Intercept	-3.71824	10.21154	-0.364	0.7158
	Linear	1.89780	15.90764	0.119	0.9050
	Quadratic	-0.56751	6.11637	-0.093	0.9261
3	Intercept	-31.08416	6.27997	-4.950	0.0000
	Linear	51.78083	10.15843	5.097	0.0000
	Quadratic	-21.37932	4.08337	-5.236	0.0000
4	Intercept	-23.51564	3.71308	-6.333	0.0000
	Linear	39.21227	5.76463	6.802	0.0000
	Quadratic	-15.39084	2.21988	-6.933	0.0000

Group membership (model 2 group | model 1 group)

1|1	(%)	5.45137	2.18204	2.498	0.0125
2|1	(%)	71.80609	5.00214	14.355	0.0000
3|1	(%)	22.74144	5.22749	4.350	0.0000
4|1	(%)	0.00111	0.02819	0.039	0.9687
1|2	(%)	10.09008	2.67692	3.769	0.0002
2|2	(%)	50.91071	4.51102	11.286	0.0000
3|2	(%)	28.48404	4.20934	6.767	0.0000
4|2	(%)	10.51516	2.52952	4.157	0.0000
1|3	(%)	14.94874	4.36556	3.424	0.0006
2|3	(%)	33.78339	5.89085	5.735	0.0000
3|3	(%)	31.96875	6.37747	5.013	0.0000
4|3	(%)	19.29912	4.37185	4.414	0.0000

Group membership (model 1 group | model 2 group)

1|1	(20.6%)
2|1	(48.4%)
3|1	(31.0%)
1|2	(46.4%)
2|2	(41.7%)
3|2	(12.0%)
1|3	(29.8%)
2|3	(47.3%)
3|3	(23.0%)
1|4	(0.0%)
2|4	(55.7%)
3|4	(44.3%)

Group membership (model 1 group and model 2 group)

1 1	(1.9%)
2 1	(4.5%)
3 1	(2.9%)
1 2	(25.5%)
2 2	(22.9%)
3 2	(6.6%)
1 3	(8.1%)
2 3	(12.8%)
3 3	(6.2%)
1 4	(0.0%)
2 4	(4.7%)

```
3 4    ( 3.8%)

 Group membership (model 2 group)

1 ( 9.4%)
2 (55.0%)
3 (27.1%)
4 ( 8.5%)

BIC=-10522.45 (N=7174)  BIC=-10484.81 (N=733)  AIC=-10408.96  L=-10375.96
```

Figure 5 displays the form of the trajectories identified for these two behaviors. The top graph shows the trajectories of opposition from ages 6 to 13, which were a product of the censored normal model. One trajectory starts off low at age 6 and declines steadily thereafter. The second trajectory starts off at a modest level of opposition at age 6, rises slightly until age 10, and then begins a gradual decline. The third trajectory starts off high and remains high over the age period. These trajectories of childhood opposition were estimated to account for 35.5%, 45.0%, and 19.5% of the population, respectively.

The bottom graph shows the trajectories for property delinquency using the Poisson model. The largest trajectory group (group 2, 55.0%) exhibits negligible offending. Group 3, 27.1%, shows a low and declining rate of property delinquency. One trajectory group (group 1, 9.4%) follows a pattern of rising property delinquency over the measurement period, whereas the fourth group, 8.5%, remains high over the entire period.

The group membership of the output shows the two conditional probability as well as the joint probability representations of the linkage between opposition and property delinquency.

Modeling a Subsequent Outcome on Trajectory Group Membership

This option links trajectory groups with a cross-sectional outcome measured at or after the termination of the trajectory. As an example we investigate how the number of sexual partners at age 14 might differ by opposition trajectory groups in the childhood opposition model. The following command fits the model with the outcome variable described in this example:

```
. traj , model(cnorm) max(10) var(qcp84op qcp88op qcp89op qcp90op qcp91op)
indep(t1-t5) order(0 2 2) outcome(nbp14) omodel(poisson)

==== traj stata plugin ====  Jones BL  Nagin DS

1037 observations read.
1037 observations used in trajectory model.


                        Maximum Likelihood Estimates
                        Model: Censored Normal (CNORM)
```

Group	Parameter	Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
1	Intercept	-1.30319	0.16275	-8.007	0.0000
2	Intercept	-4.28999	1.09455	-3.919	0.0001
	Linear	1.63419	0.25153	6.497	0.0000
	Quadratic	-0.09415	0.01347	-6.989	0.0000
3	Intercept	-2.04720	1.60319	-1.277	0.2017
	Linear	1.66381	0.36388	4.572	0.0000
	Quadratic	-0.08805	0.01947	-4.523	0.0000
	Sigma	2.59779	0.03814	68.107	0.0000
Group membership					
1	(%)	29.41511	2.49939	11.769	0.0000
2	(%)	47.67685	2.43417	19.587	0.0000
3	(%)	22.90803	2.10466	10.884	0.0000
	Group1	-1.02464	0.11151	-9.189	0.0000
95% CI (Mean): (	0.288,	0.359,	0.447)		
	Group2	-0.40189	0.07394	-5.436	0.0000
95% CI (Mean): (	0.579,	0.669,	0.773)		
	Group3	0.47558	0.07390	6.435	0.0000
95% CI (Mean): (	1.392,	1.609,	1.860)		

BIC=-10370.54 (N=4661) BIC=-10360.77 (N=1037) AIC=-10328.63 L=-10315.63

The numbers of sexual partners differ by childhood opposition trajectory group, with greater levels of oppositional behavior associated with higher numbers of sex partners. The estimated average number of sexual partners at age 14 in the low-opposition group is 0.36 (95 percent CI: 0.29, 0.45). In contrast, the average number of sexual partners in the moderate-opposition group is 0.67 (95 percent CI: 0.58, 0.77), while the number of partners in the high-opposition group is 1.61 (95 percent CI: 1.39, 1.86).

Sample Weights And Exposure Times

Sample weights may be used with any model; however, exposure times are only valid for a ZIP model. When sample weights are included, a robust (sandwich) estimator of the variance-covariance matrix is calculated. This example illustrates both the use of sample weights and an adjustment for exposure times in a ZIP model using data from the Rochester Youth

Development Study. This study tracked a sample of students from the seventh and eighth grades in the Rochester, New York public schools. Male students and students from high-crime areas were oversampled since it was assumed that they were at greater risk for offending. Sample weights were used to account for the oversampling. In addition, assessment intervals were not constant over the course of the study or across study participants. Thus we use exposure times to account for these differences in availability to commit crimes. Early assessments were semiannual (6 months of exposure time) followed by annual assessments. An adjustment for exposure time is also required if individuals are somehow placed in a situation where they are restricted from engaging in the activity of interest. For example, the exposure time adjustment was demonstrated in the Piquero et al. (2001) analysis of the arrest histories of individuals who had been under the supervision of the California Youth Authority. Exposure time adjustment accounted for spells of imprisonment, during which times individuals could not be arrested for crimes. The following command fits a two-group model to the arrest counts using both weight and exposure time variables:

```
. traj , model(zip) var(g2 - g13) indep(t*) order(2 2) iorder(0 2) expos(e*)
weight(wt50)

==== traj stata plugin ====  Jones BL  Nagin DS

247 observations read.
247 observations used in trajectory model.
```

Maximum Likelihood Estimates
Model: Zero Inflated Poisson (ZIP)

Group	Parameter	Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
1	Intercept	14.28603	2.48094	5.758	0.0000
	Linear	-1.63405	0.29613	-5.518	0.0000
	Quadratic	0.04940	0.00859	5.752	0.0000
2	Intercept	5.46429	3.64181	1.500	0.1336
	Linear	-0.62882	0.42522	-1.479	0.1393
	Quadratic	0.02573	0.01211	2.125	0.0337
1	Alpha0	1.30912	0.10574	12.380	0.0000
2	Alpha0	-31.45969	7.07883	-4.444	0.0000
	Alpha1	4.13551	0.87071	4.750	0.0000
	Alpha2	-0.13088	0.02617	-5.002	0.0000
Group membership					
1	(%)	76.97337	3.51061	21.926	0.0000
2	(%)	23.02663	3.51061	6.559	0.0000

```
 BIC= -4167.52 (N=2821)  BIC= -4154.12 (N=247)  AIC= -4134.82  L= -4123.82
```

```
. trajplot, xtitle("Age") ytitle("Annual Arrest Rate")
```

In Figure 6 we see that the first group (77.0%) consists of individuals with a very low rate of offending over the 12-wave period. The other trajectory group (23.0%) shows increasingly high levels of offending during the 9th through 12th waves.

Discussion

We demonstrated the use of a new Stata command, *traj*, to analyze longitudinal data by fitting a mixture model. We illustrated the use of *traj* through various applications including analysis of psychometric scale data (oppositional behavior) using the censored normal mixture, offense counts using the ZIP mixture, and an offense prevalence measure using the logistic mixture. Time-stable covariates (risk factors) were incorporated into the model by assuming that the risk factors are independent of the developmental trajectories, given group membership. A time-dependent covariate can also directly affect the observed behavior trajectory. While we focused on applications from research on antisocial behavior, any application that proposes to differentiate observations by type or category can be analyzed by our method.

References

Dennis, J. E., D. M. Gay, and R. E. Welsch. 1981. An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software* 7:348-383.

Dennis, J. E. and H. W. Mei. 1979. Two new unconstrained optimization algorithms which use function and gradient values. *Journal of Optimization Theory and Applications* 28:453-483.

Farrington, D. P. and D. J. West. 1990. The Cambridge study in delinquent development: a prospective longitudinal study of 411 males. In *Criminality: Personality, Behavior, and Life History*, edited by Hans-Jürgen Kerner and G. Kaiser. New York: Springer-Verlag.

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

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

Loeber, R. 1991. Questions and advances in the study of developmental pathways. In *Models and Integrations*, edited by D. Cicchetti and S. Toth. New York: University of Rochester Press.

Nagin, D. S. and R. E. Tremblay. 2001. Analyzing developmental trajectories of distinct but related behaviors: a group-based method. *Psychological Methods* 6:18-34.

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

Nagin, D. and C. L. Odgers. 2010. Group-based trajectory modeling in clinical research. *Annual Review of Clinical Psychology* 6:109-138.

Piquero, A., A. Blumstein, R. Brame, R. Haapanen, E. Mulvey, and D. S. Nagin. 2001. Assessing the impact of exposure time and incapacitation on longitudinal trajectories of offending. *Journal of Adolescent Research* 16:54-76.

Roeder K., K. G. Lynch, and D. S. Nagin. 1999. Modeling uncertainty in latent class membership: a case study in criminology. *Journal of the American Statistical Association* 94:766-776.

Tremblay, R. E., L. Desmarais-Gervais, C. Gagnon, and P. Charlebois. 1987. The preschool behavior questionnaire: stability of its factor structure between culture, sexes, ages, and socioeconomic classes. *International Journal of Behavioral Development* 10: 467–484.

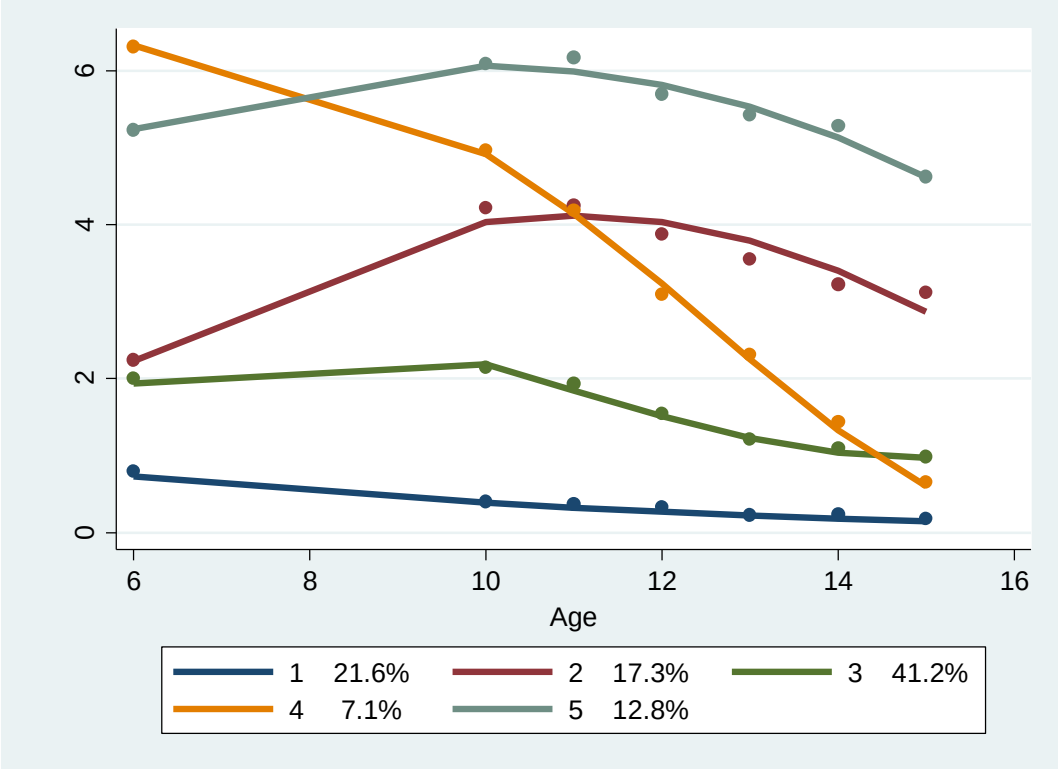


Figure 1: Censored normal model: estimated trajectories (solid lines), observed group means at each age (dot symbols), and estimated group percentages.

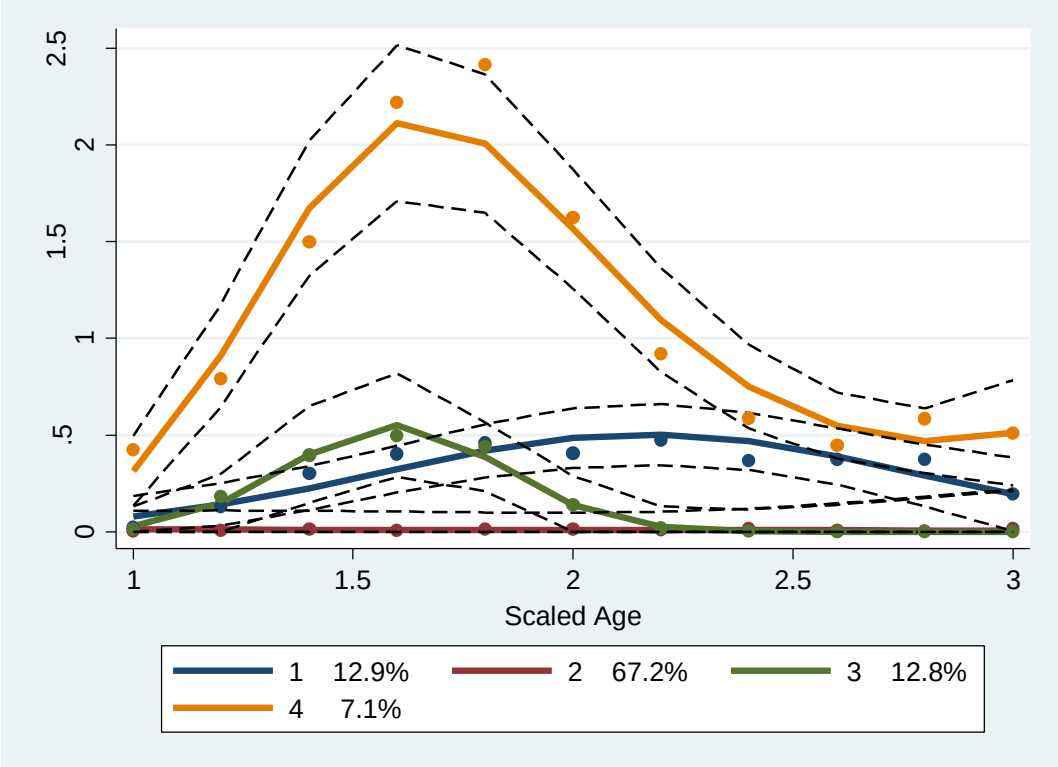


Figure 2: Zero-inflated Poisson model: estimated trajectories (solid lines), observed group means at each age (dot symbols), and estimated group percentages. Dashed lines are 95% pointwise confidence intervals on the estimated trajectories.

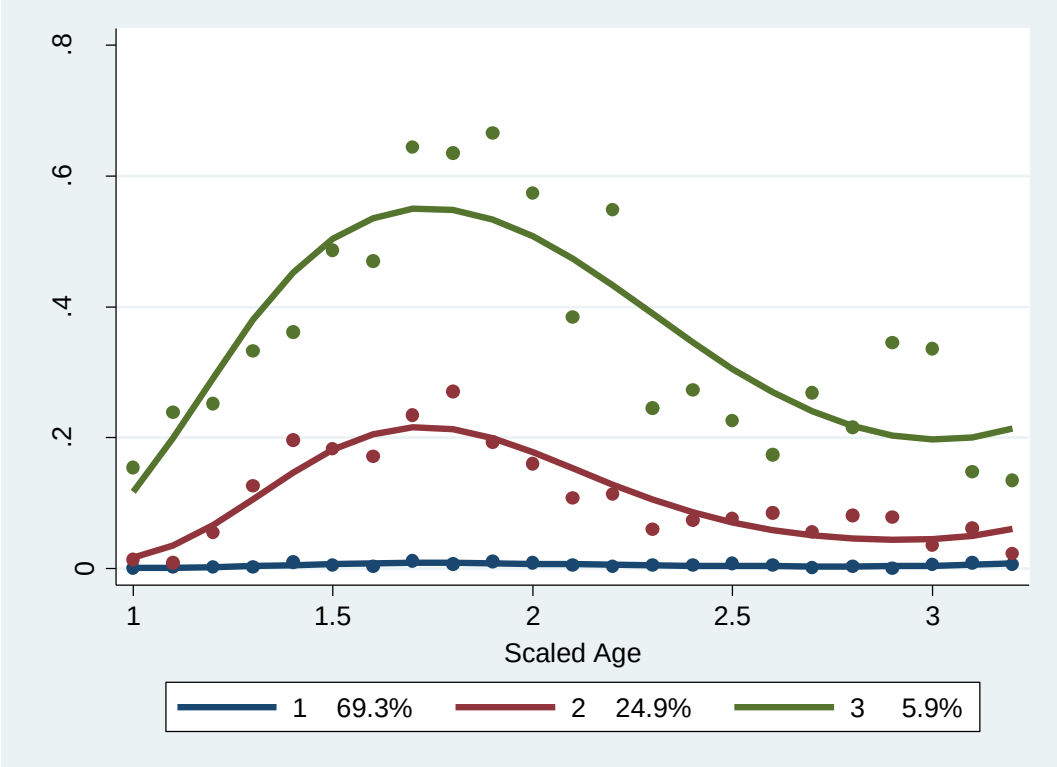


Figure 3: Logit model: estimated trajectories (solid lines), observed group means at each age (dot symbols), and estimated group percentages.

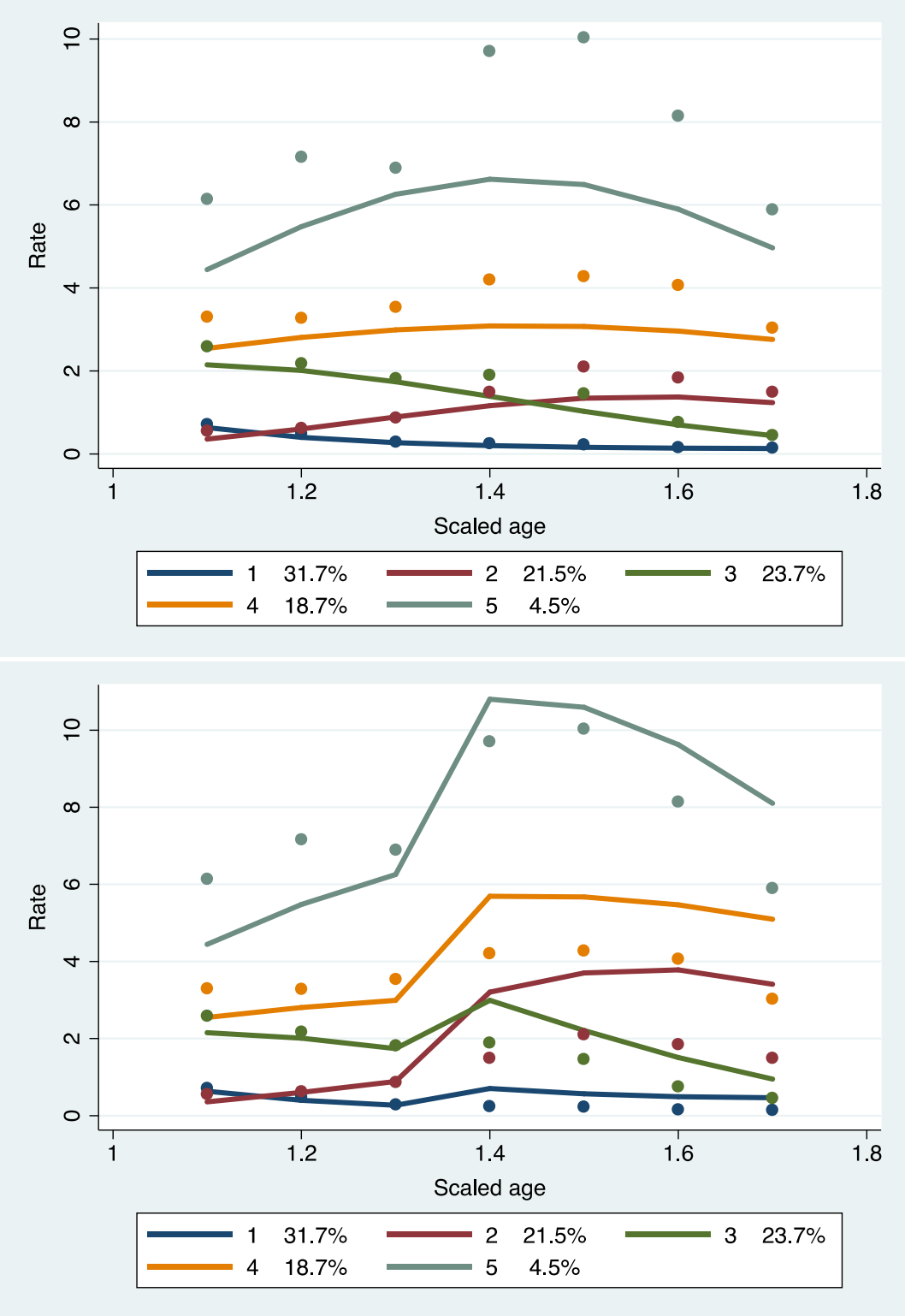
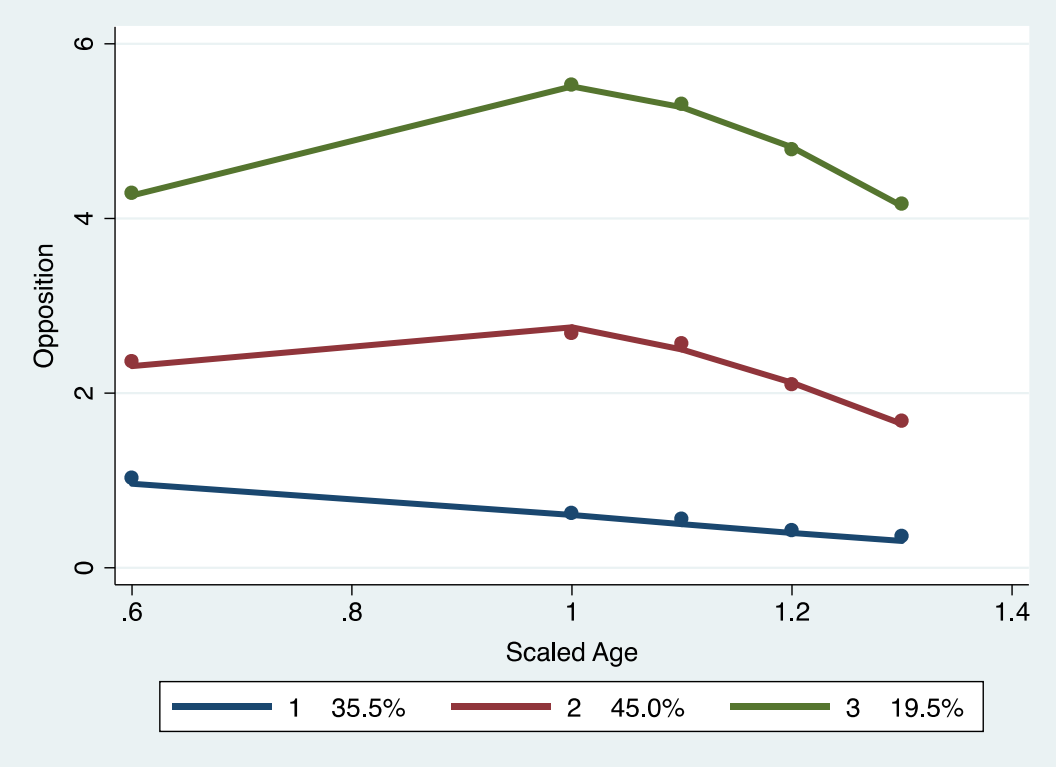


Figure 4: Time-varying covariates: illustrating trajectory paths of violent delinquency for not in a gang (top) and membership in a gang from age 14 to 17 (bottom). Dot symbols are observed data reflecting the average gang membership rates.



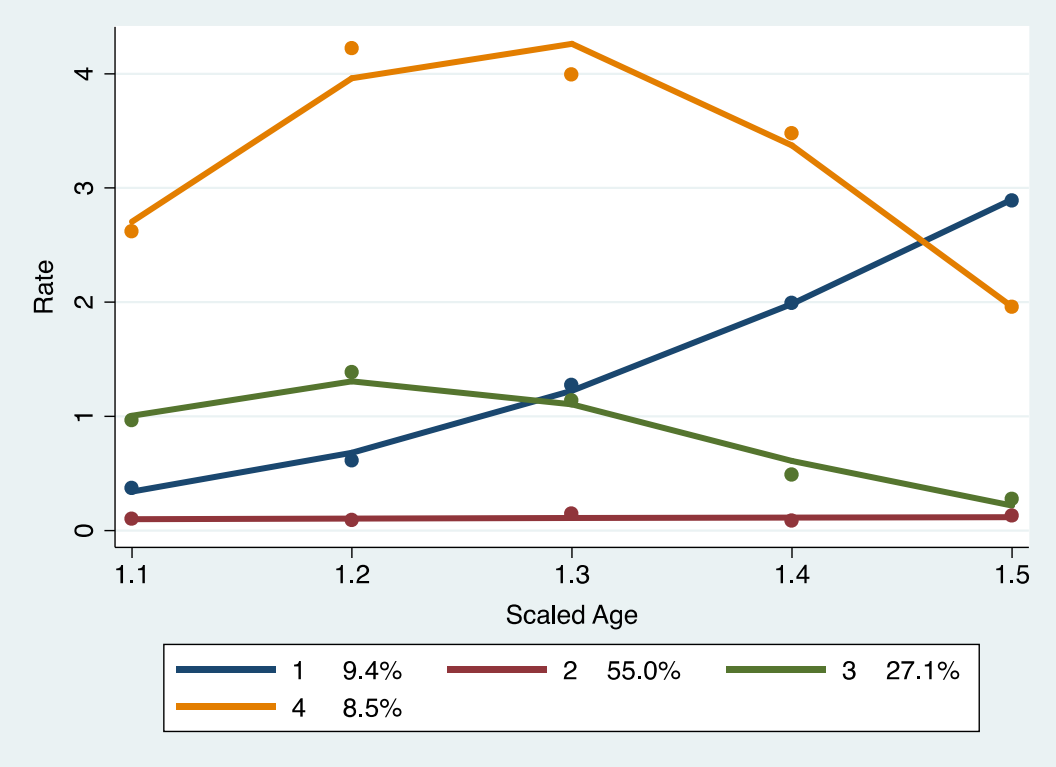


Figure 5: Joint trajectory model: estimated trajectories (solid lines), observed group means at each age (dot symbols), estimated group percentages for the two models.

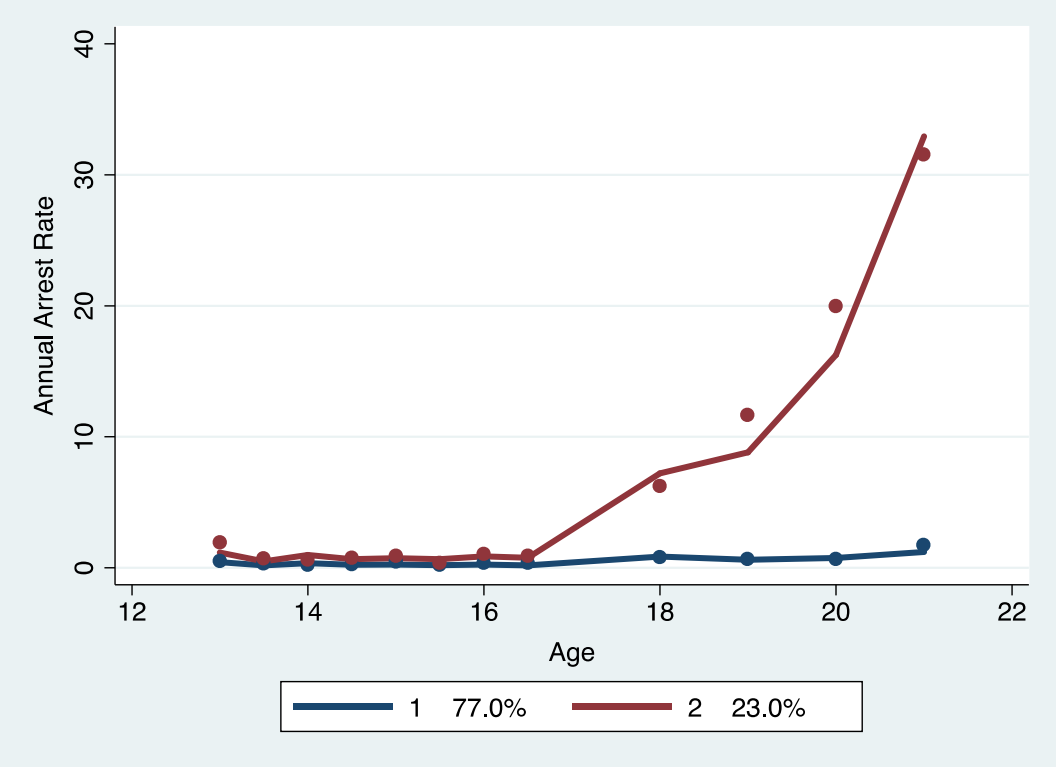


Figure 6: Zip model with exposure variables and sample weights: estimated trajectories (solid lines), observed group means at each age (dot symbols), and estimated group percentages.