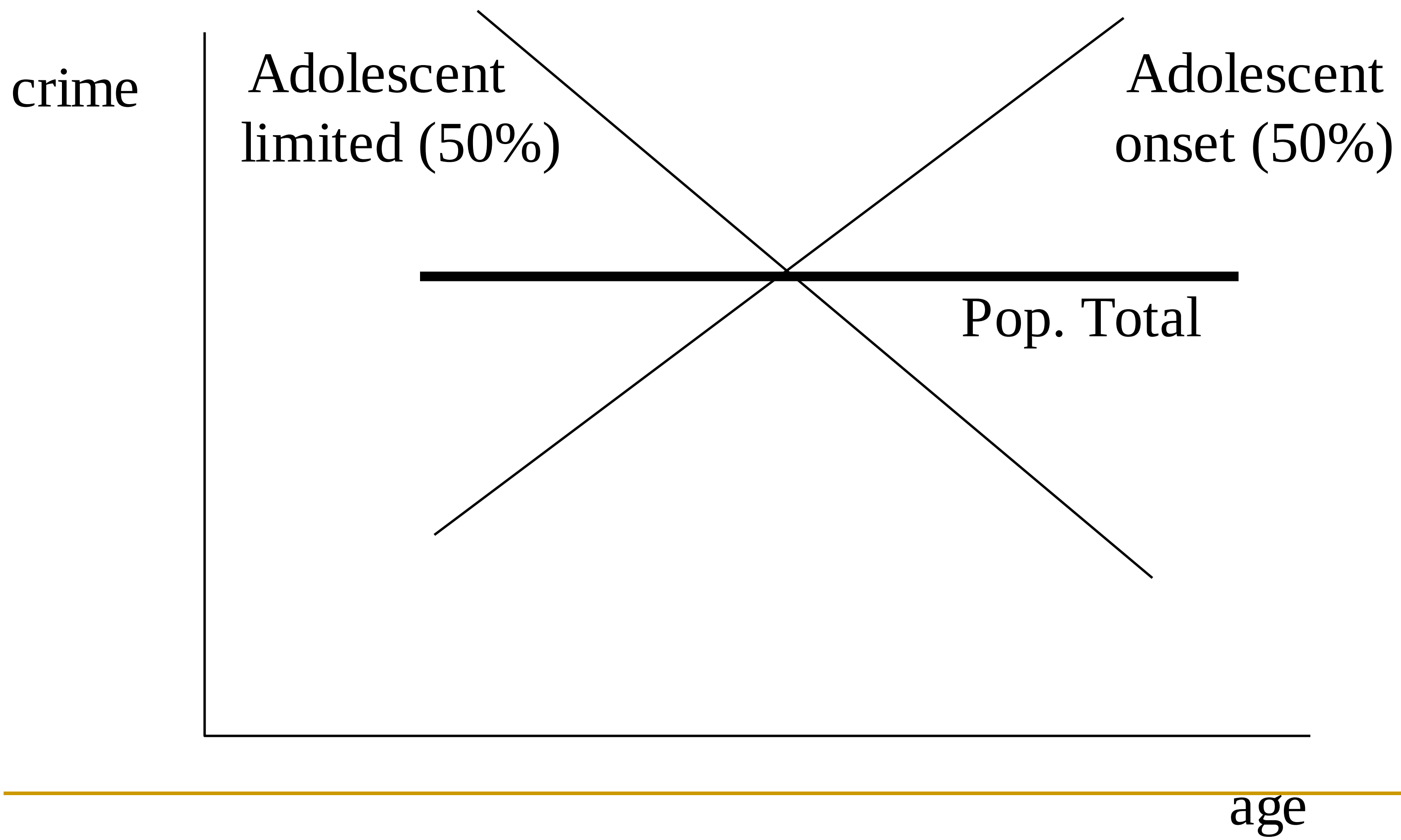
Proc Traj: A SAS Procedure for Group Based Modeling of Longitudinal Data

Bobby L. Jones, PhD
Carnegie Mellon University

An Illustrative Example



Proc Traj

- SAS procedure that identifies clusters of individuals following similar progressions of an outcome over time, age, … , by fitting a group based model
- Proc Traj is used like any SAS proc e.g. Proc Glm, Proc Mixed …
- Traj download, online documentation, examples, and installation information at http://www.andrew.cmu.edu/~bjones

Proc Traj

- Easy-to-use, PC SAS procedure
- Handles missing data
- Handles sample weights
- Allows irregular spacing of measurements
- Accommodates over-lapping cohort designs

Basic Proc Traj Model Estimates

- Number of groups via Bayesian Information Criterion (BIC) model comparisons
- Group mean shapes
- Group sizes
- Individual probabilities of group membership
- Effect sizes for time-independent (risk) and time-dependent variables (tcov)

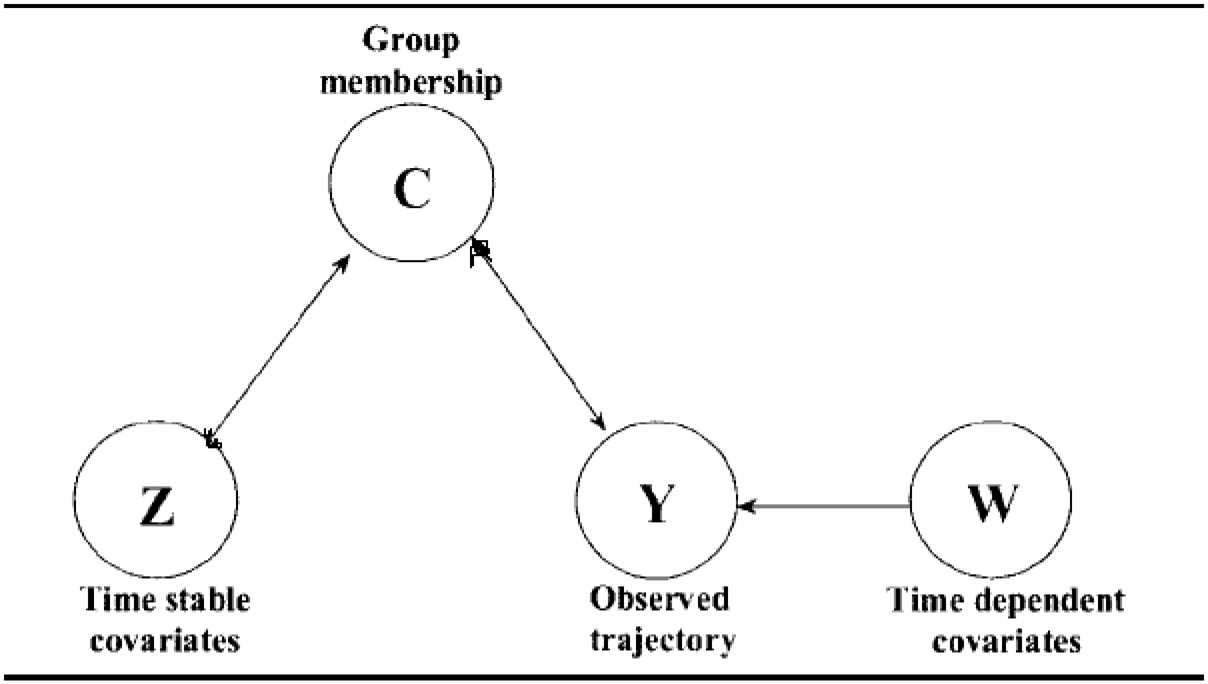
$$BIC = \log(L) - .5 * \log(n) * k$$

$$L = \log likelihood$$

$$n = sample\ size$$

$$k = number\ of\ parameters$$

Model Form and Independence Assumptions



Proc TRAJ Basic Poisson Model Example with Confidence Limits

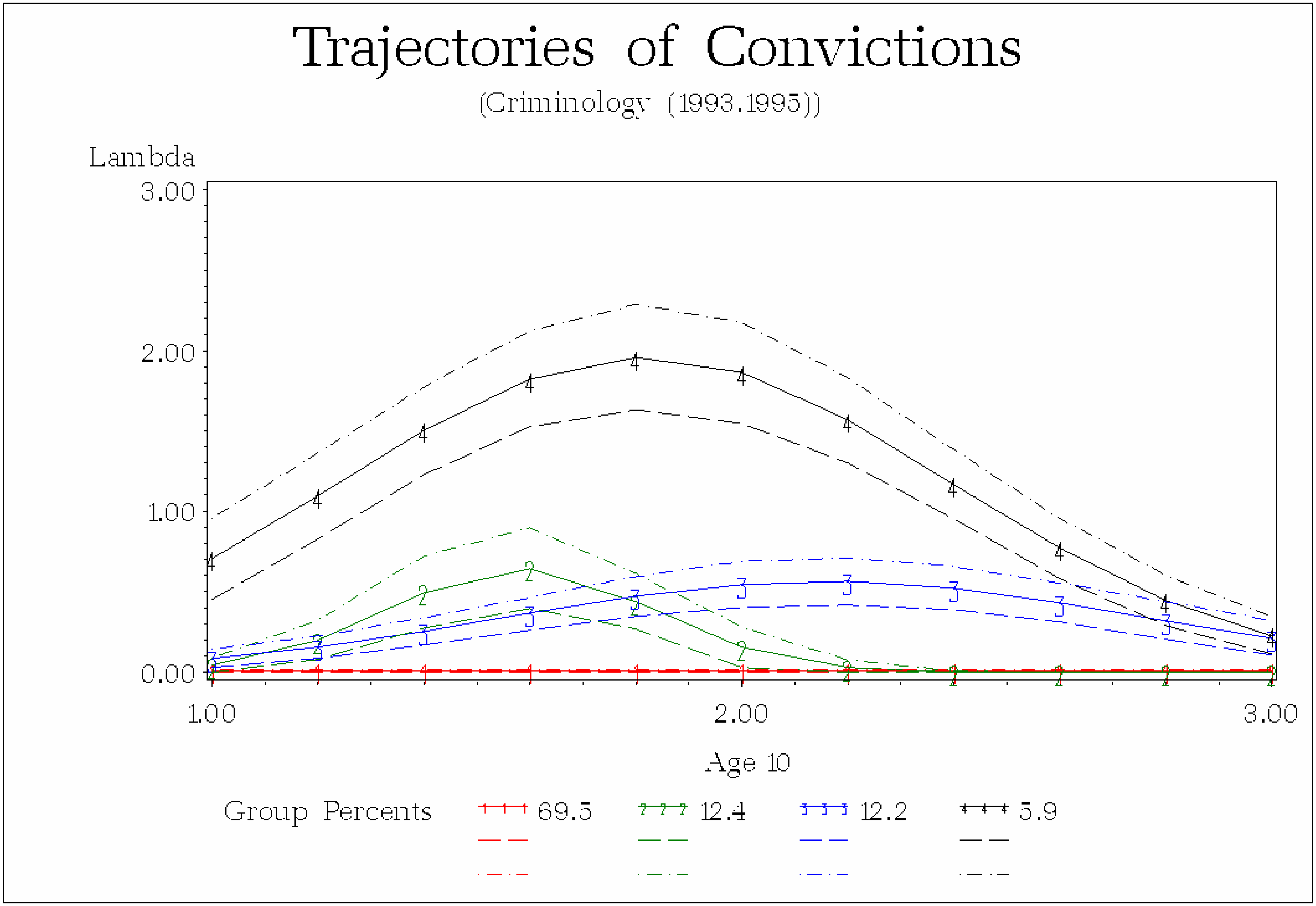
```
PROC TRAJ DATA=A OUT=OF OUTPLOT=OP OUTSTAT=OS CI95M;
        VAR O1-O11;             /* Conviction Variables                    */
        INDEP T1-T11;           /* Age Variables                           */
        MODEL ZIP;              /* Poisson Model                           */
        ORDER 0 2 2 2;          /* Flat and Quadratic for Four Groups      */
RUN;

%TRAJPLOTNEW (OP, OS,,,“Annual Conviction Rate","Age");
```

411 subjects from working-class section of London (Farrington and West, 1990)
Criminal offense convictions recorded annually from age 10 – age 30

Proc Traj Output

1	Intercept	-4.57337	0.27345	-16.724	0.0000
2	Intercept	-20.49791	4.15470	-4.934	0.0000
	Linear	25.38542	5.47675	4.635	0.0000
	Quadratic	-8.02886	1.78662	-4.494	0.0000
3	Intercept	-7.00478	1.16221	-6.027	0.0000
	Linear	5.95561	1.15414	5.160	0.0000
	Quadratic	-1.37783	0.27769	-4.962	0.0000
4	Intercept	-4.36648	0.75693	-5.769	0.0000
	Linear	5.54530	0.81271	6.823	0.0000
	Quadratic	-1.52575	0.21040	-7.252	0.0000
	Group membership				
1	(%)	69.48667	3.17546	21.882	0.0000
2	(%)	12.40384	2.72478	4.552	0.0000
3	(%)	12.19109	2.17877	5.595	0.0000
4	(%)	5.91840	1.29345	4.576	0.0000

BIC= -1516.57 (N=403) BIC= -1532.15 (N=4433) AIC= -1490.57



Trajectories of Convictions
(Criminology (1993,1995))
Lambda
3.00
2.00
1.00
0.00
1.00
2.00
3.00
Age 10
Group Percents
69.5
12.4
12.2
5.9

Short live demo

- Use of Sas Explorer to look at files
- Example 1-Physical Aggression Trajectories
- Montreal Data: 1037 Caucasian, francophone, nonimmigrant males
- First assessment at age 6 in 1984

Types of Data Handled

- Psychometric scale data--censored normal (tobit) model
- Count data--poisson-based model
- Binary data--logit-based model
- Normally distributed data (set min and max outside the range of the data)

The Likelihood Function

Y_i = trajectory data for subject *i*

$$P(Y_i) = \sum_j \pi_j(x_i) P^j(Y_i)$$

$P^j(Y_i)$ = probability of Y_i , if belonging to group *j*

$\pi_j(x_i)$ = probability of belonging to group *j* for covariates (risk) x_i

$$\pi_j(x_i) = e^{x_i \theta_j} \Big/ \sum e^{x_i \theta_j}$$

$$L = \prod_{i=1}^{N} P(Y_i).$$

Zero-inflated Poisson Model for Count Data

IORDER used to specify alpha polynomials

$$p(x) = \begin{cases} 0 \text{ with probability } \rho \\ Poisson(\lambda) \text{ with probability } 1-\rho \end{cases}$$

$$\ln(\lambda) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 age^3$$

$$\rho = \frac{e^{\alpha_0 + \alpha_1 age + \alpha_2 age^2 + \alpha_3 age^3}}{1 + e^{\alpha_0 + \alpha_1 age + \alpha_2 age^2 + \alpha_3 age^3}}$$

Logit Model for Binary Data

$$P(y=1)=\frac{e^{\beta_0+\beta_1 age+\beta_2 age^2+\beta_3 age^3}}{1+e^{\beta_0+\beta_1 age+\beta_2 age^2+\beta_3 age^3}}$$

$$P(y=0)=1-P(y=1)$$

Censored Normal (CNORM) Model

$$\Pr(\mathbf{Y}_i = \mathbf{y}_i \mid C_i = k, \mathbf{W}_i = \mathbf{w}_i) =$$

$$\prod_{y_{ij}=\min} \Phi\left(\frac{Min - \mu_{ijk}}{\sigma}\right) \prod_{Min<y_{ij}<Max} \frac{1}{\sigma}\varphi\left(\frac{y_{ij} - \mu_{ijk}}{\sigma}\right) \prod_{y_{ij}=Max}\left(1 - \Phi\left(\frac{Max - \mu_{ijk}}{\sigma}\right)\right),$$

where

$$\mu_{ijk} = \beta_{0k} + age_{ij}\beta_{1k} + age_{ij}^{2}\,\beta_{2k} + \ldots + w_{ij}\delta_k.$$

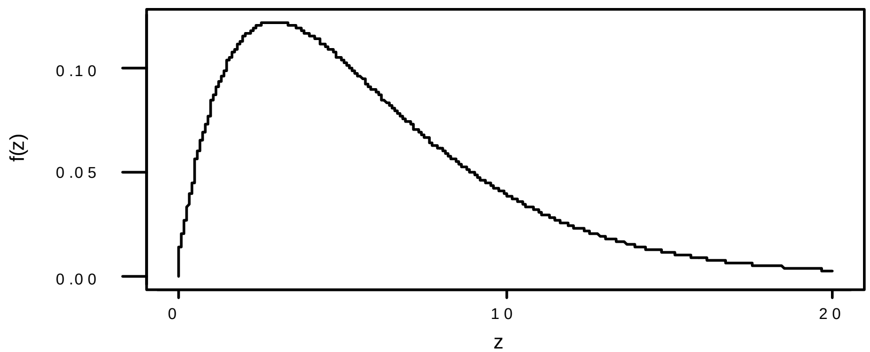
Trajectory mean calculation complicated!
Use plot file for trajectory means (for all models)

Group-Based Trajectory Modeling Compared to Conventional Growth Curve Modeling (HLM)

- Common point of departure: both model individual level trajectories by a polynomial equation in age or time: $\beta_0 + \beta_1 Age + \beta_2 Age^2$
- Point of departure: how to model individual- level differences in developmental trajectories (e.g., population heterogeneity)
  - HLM use normally distributed random effects
  - Group-based trajectory can approximate an unknown distribution of individual differences with groups

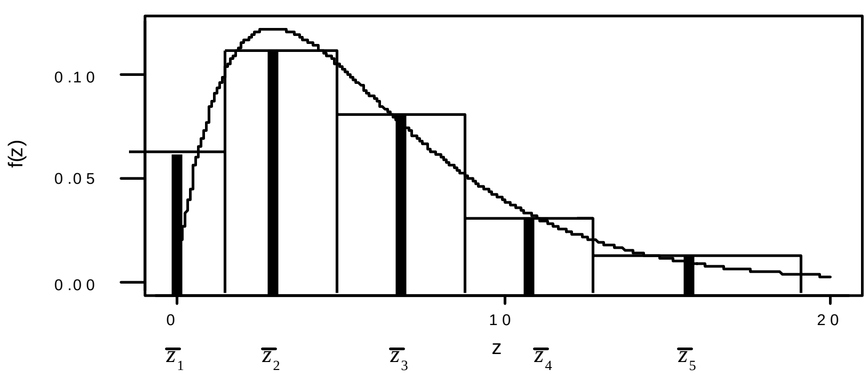
Groups Can Approximate an Unknown Distribution

Panel A



z

Panel B



Advances In Proc Traj

Missing Trajectory Data

- Intermittently Missing
  - Maximum Likelihood Robust Here
- Subject Dropout
  - Model Probability of Dropout on Previous Measurements and Other Variables

Likelihood Function Modified for Dropout

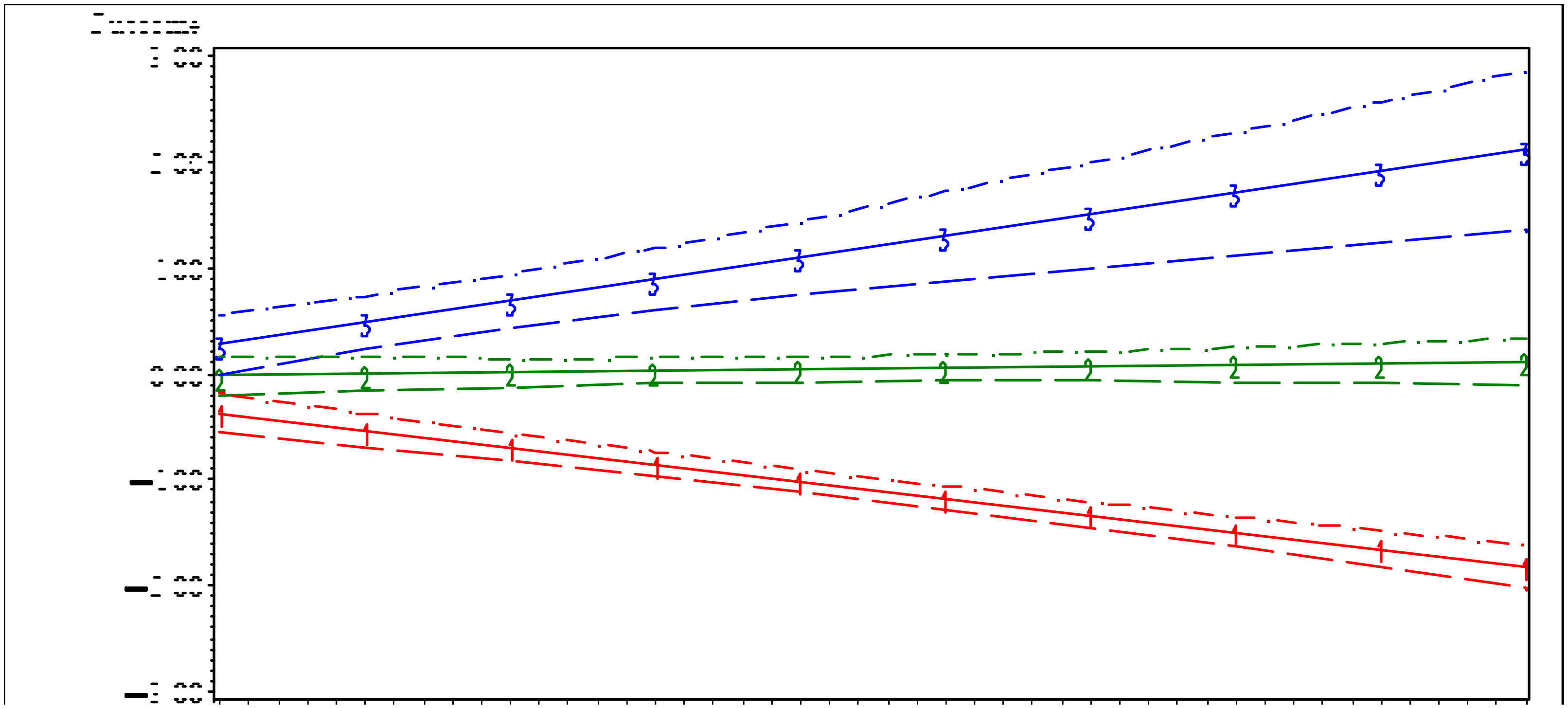
$$P(Y_i) = \sum_j \pi_j(x_i) P^j(Y_i) P^j(D_i \mid Y_i)$$

$P^j(Y_i)$ = probability of Y_i , for group j

$P^j(D_i \mid Y_i)$ = probability of *Dropout*, for group j as a function of measurements before dropout.

$$P(D_i \mid Y_i) = \frac{e^{\beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2}}}{1 + e^{\beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2}}}$$

Data w/Dropout (N=200: 25%, 50%, 25%)



	1	2	3	4	5	6	7	8	9	10	Didn't Dropout
G1	0	0	0	0	1	0	0	1	0	0	48
G2	2	3	7	6	8	5	7	1	3	6	52
G3	1	8	7	8	7	8	1	3	6	1	0

Estimated Group Percentages

- Group size estimates (no dropout modeling)
- Bias in group size

1	(%)	30.9 ± 4.2
2	(%)	53.3 ± 5.4
3	(%)	15.8 ± 4.5

With dropout modeled no bias in group size

1	(%)	27.1 ± 3.7
2	(%)	45.9 ± 7.3
3	(%)	27.0 ± 7.1

- But no bias in trajectory shape estimates with or without dropout modeled

Proc Traj Syntax for Dropout Model

```
proc traj data=a out=b outstat=os outplot=os;
var opp1-opp5; indep t1-t5;
model cnorm; max 10; order 0 2 2;
dropout 2 2 2;
run;
```

Missing Covariate Data

- Default is complete case analysis
- Alternative: multiple imputation
  - PROC MI (produce replicate data sets)
  - PROC MIANALYZE (analyze results)

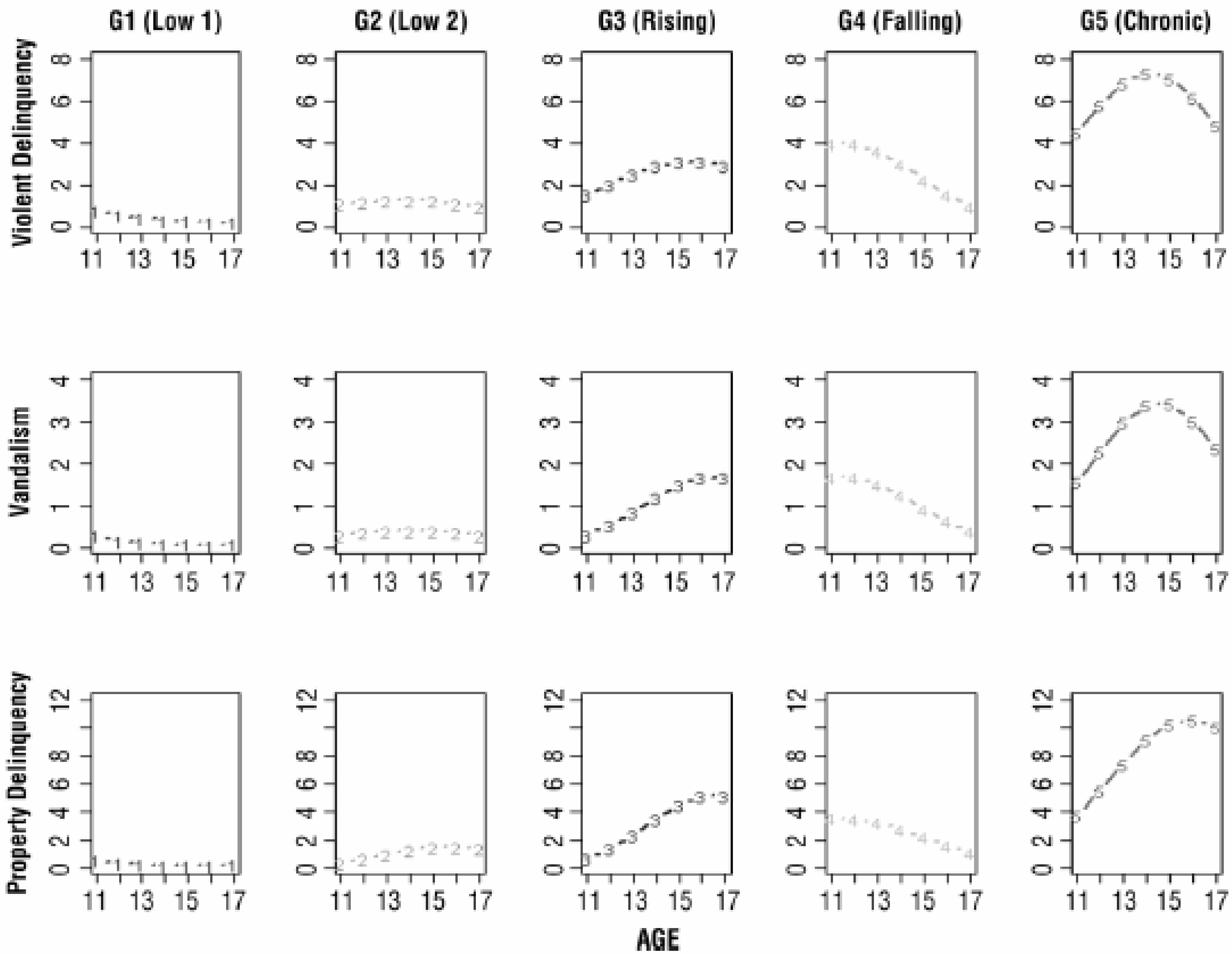
Multitrajectory Modeling

- Extend model to more than two repeated measures.
- Each group is now defined by multiple trajectories: one for each response of interest.

Proc Traj Syntax for Multitrajectory Model

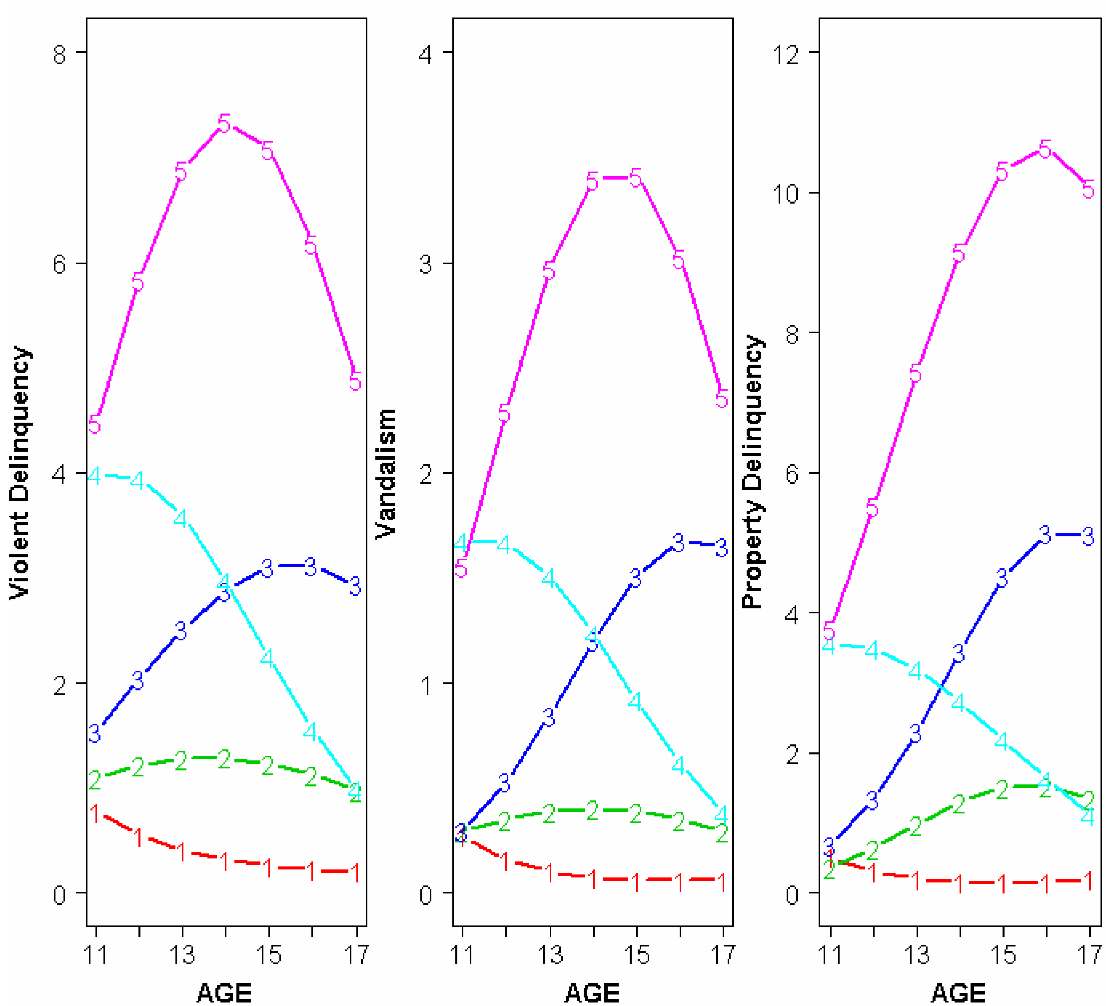
```
proc traj data=montreal outplot=op outstat=os
outplot2=op2 outstat2=os2 outplot3=op3 outstat3=os3;
var bat89-bat95;  indep t1-t7;  model zip;
order 2 2 2 2 2;
var2 det89-det95; indep2 t1-t7; model2 zip;
order2 2 2 2 2 2;
var3 vol89-vol95; indep3 t1-t7; model3 zip;
order3 2 2 2 2 2;
multgroups 5;
run;
```

Five Group Multiple Trajectory Model



Five Group Multiple Trajectory Model

Figure 5B: Trajectory Plots Organized by Behavior



Relating Trajectory Groups to a Subsequent Outcome Variable

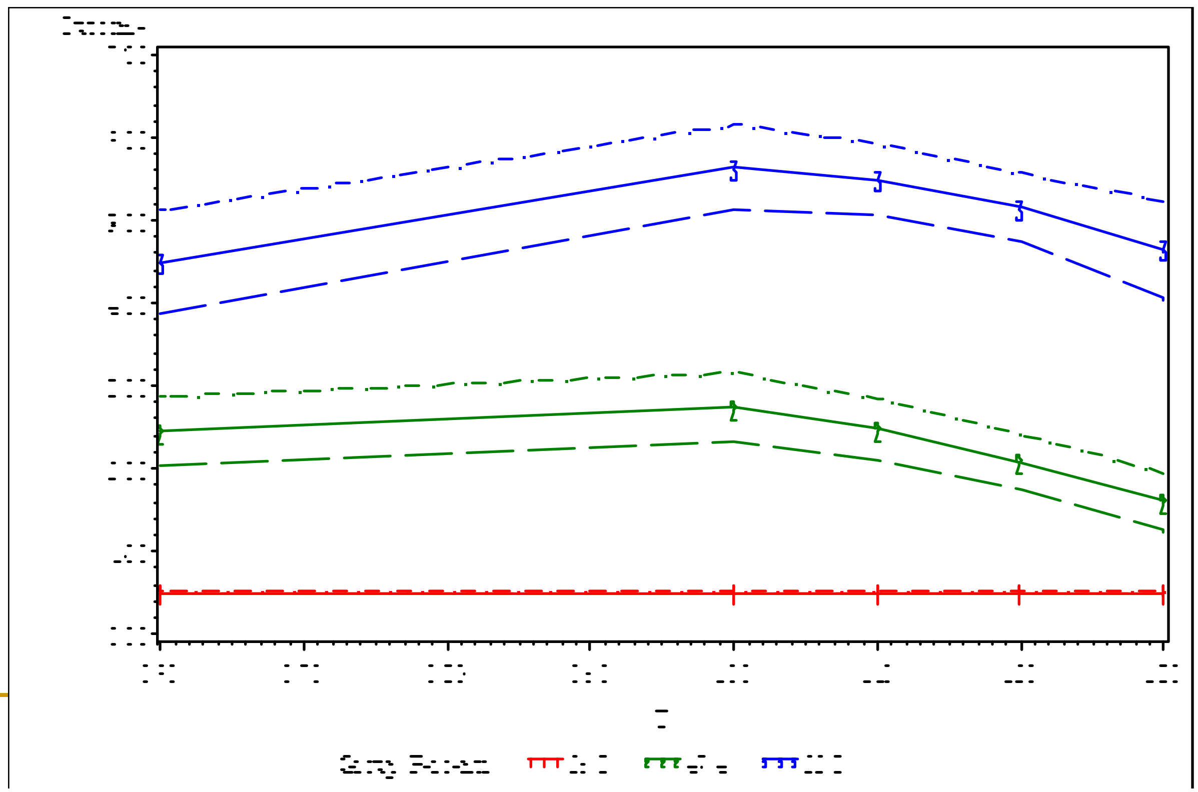
- Link trajectory groups with an outcome variable, measured once on or after the termination of the trajectory.
- Example: Relating trajectories of opposition from age 6 to 13 to average number of sexual partners at age 14.

Proc Traj Syntax

```
proc traj data=a out=b outstat=os outplot=os
  outcomeresults=or;
var opp1-opp5; indep t1-t5;
model cnorm; max 10; order 0 2 2;
outcome nbp14;
omodel poisson;
```

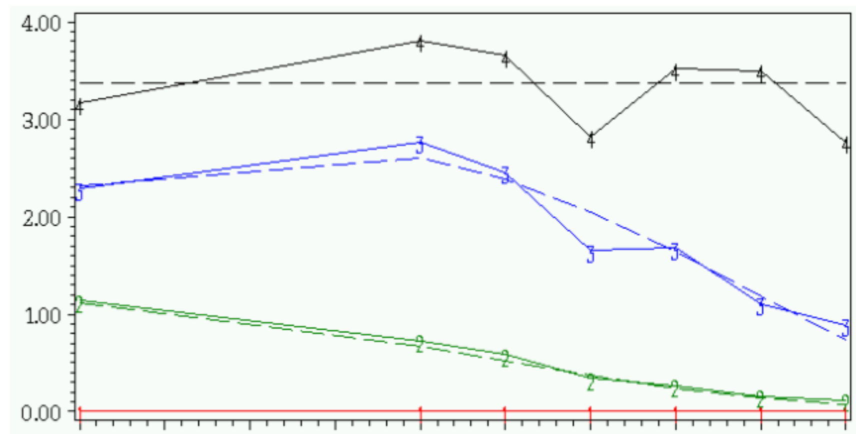
Outcome Analysis (opp 8-13, nbp14)

Gp	L95 Mean	Mean Outcome	U95 Mean
1	0.27	0.35	0.44
2	0.59	0.68	0.78
3	1.48	1.72	2.00



Setting Risk Factors to Calculate Effect on Group Membership

- Physical Aggression Trajectory Model
- Risk Factors
  - Broken Home at Age 5
  - Low IQ
  - Low Maternal Education
  - Mother Began Childbearing as a Teenager



Setting Risk Factors

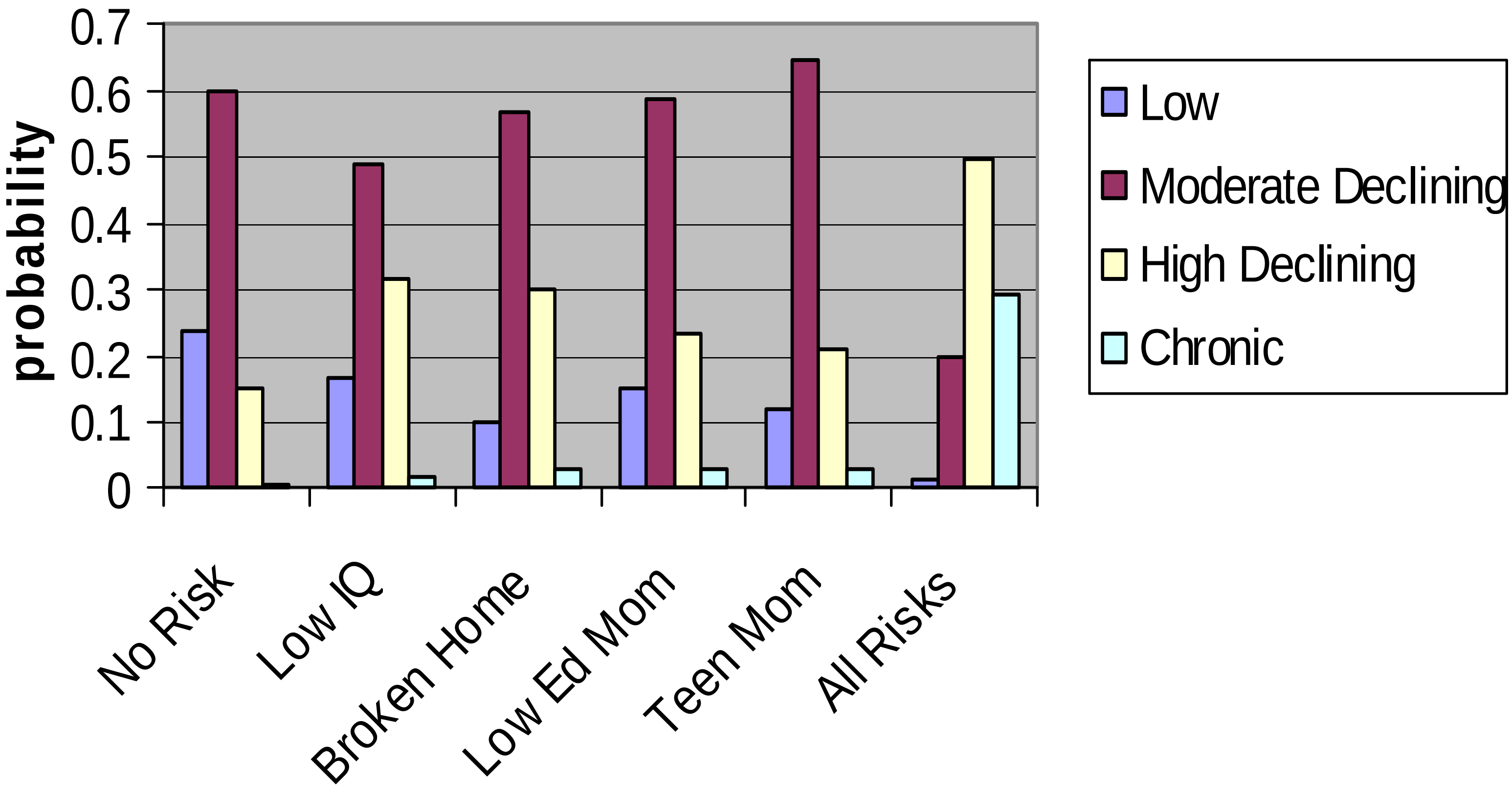
```
data predict;
input nint5 lowiq edmom teenmom;
cards;
0 0 0 0
1 0 0 0
0 1 0 0
1 1 1 1
run;

data combine;
set predict originalDat;
run;
```

Results in Outfile

nint5 lowiq edmom teenmom

RISK				G1Prob	G2Prob	G3Prob	G4Prob
0	0	0	0	0.237612	0.599554	0.154430	0.008404
1	0	0	0	0.102632	0.570928	0.298253	0.028187
0	1	0	0	0.171699	0.490363	0.317862	0.020076
1	1	1	1	0.009400	0.196002	0.498584	0.296015

Impact of Risk Factors on Group Membership Probabilities



%trajtest

- The trajtest macro tests hypotheses about model parameters

- For linear hypothesis, test based on (Lb - c)' $(S')^{-1}$ (Lb - c) (Wald test) $c_1 \times variable_1 \pm c_2 \times variable_2 \pm \cdots \pm c_n \times variable_n$

- Variable names correspond to names in the OUTEST= file.

- %trajtest('linear1=linear2=linear3')
- %trajtest('linear1=linear2,quadra1=quadra2')
- %trajtest(‘risk1=risk2=risk3’)

Refgroup;

- Allows specification of reference group for risk factor analysis.
- Refgroup 2;
- Demo

Readings (Jones articles available from Traj website)

- Jones, B., D.S. Nagin. And K. Roeder. 2001. “A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories.” *Sociological Research and Methods,* 29: 374-393.
- Jones, B. and D.S. Nagin. 2007. “Advances in Group-based Trajectory Modeling and a SAS Procedure for Estimating Them,” *Sociological Research and Methods,* 35: 542-571.
- Nagin, D. S. 2005. *Group-based Modeling of Development*. Cambridge, MA.: Harvard University Press.
- Nagin, D.S. and R. E. Tremblay. 2005. “Developmental Trajectory Groups: Fact or a Useful Statistical Fiction?.” *Criminology,* 43:873-904.
- Nagin, D. S., and R. E. Tremblay. 2001. “Analyzing Developmental Trajectories of Distinct but Related Behaviors: A Group-based Method.” *Psychological Methods*, 6(1): 18-34.
- Nagin, D. S. 1999. “Analyzing Developmental Trajectories: A Semi-parametric, Group-based Approach.” *Psychological Methods*, 4: 139-177.