Sociological Methods & Research
http://smr.sagepub.com/

Robustness of Group-Based Models for Longitudinal Count Data

David L. Weakliem and Bradley R. Entner Wright

Sociological Methods & Research 2009 38: 147
DOI: 10.1177/0049124109339368

Sociological Methods & Research
Volume 38 Number 1
August 2009 147-170

10.1177/0049124109339368
http://smr.sagepub.com
hosted at
http://online.sagepub.com

Robustness of Group-Based Models for Longitudinal Count Data

David L. Weakliem
Bradley R. Entner Wright
University of Connecticut, Storrs

In recent years, there have been efforts to develop latent class models for trajectories. The semiparametric mixed Poisson regression (SMPR) model has been used in many empirical studies, but there have been few attempts to evaluate the robustness of the estimates from this model. The authors use simulated data to evaluate the performance of the SMPR model under a variety of assumptions. They find that estimates are sensitive to the conditional distribution of counts and misspecification of the shape of the trajectory. When there is only one underlying trajectory and overdispersion is present, the SMPR model frequently finds multiple groups, which often appear to differ in shape as well as level. The tendency can be substantially reduced by use of the zero-inflated Poisson distribution in conjunction with top-coding of large counts. The article concludes with a discussion of other extensions and alternatives to the standard SMPR model that might provide more robust estimates.

Keywords: *semiparametric mixed Poisson regression; overdispersion; robustness; latent class models*

Individual and institutional histories often seem to show a pattern in which crucial events or decisions have a lasting impact on later developments. Familiar expressions such as being "at a crossroads," "taking a wrong turn," and "getting back on track" reflect this way of looking at change. Until recently, the great majority of statistical models for change were based on the very different idea of a process in which random shocks produced

Authors' Note: We thank the reviewers and editor for valuable comments and suggestions.
Please address correspondence to David Weakliem, Department of Sociology, Unit-2068, University of Connecticut, Storrs, CT 06269-2068; e-mail: David.Weakliem@uconn.edu.
Data and programs for replication are available at http://web.uconn.edu/weakliem.

temporary deviations from an equilibrium level. In recent decades, however, quantitative researchers have shown increasing interest in the idea of trajectories. Abbott (1997:87), for example, argues that social scientists should turn from "variable-based" to "pattern-based" approaches. A number of different models have been proposed, but the semiparametric mixed Poisson regression (SMPR) model for count data is one of the best articulated and most widely used: Nagin and Tremblay (2005) provide an exposition of the model and a review of the results of empirical studies. The estimates are based on a likelihood model, so it is possible to test hypotheses about the number and nature of the trajectories rather than to offer simply descriptive characterizations.

The approach has been particularly popular in the study of crime and deviance. In aggregate statistics, crime rates show a distinct peak in late adolescence or early adulthood. Some theorists, notably Hirschi and Gottfredson (1983), argue that this is a universal pattern, but others hold that significant groups deviate from it. For example, Moffitt and colleagues (Moffitt 1993; Moffitt et al. 1996) suggest that some people have a "life-course persistent" pattern in which the propensity to commit crimes does not decline in adulthood. Other theoretical accounts that suggest distinct patterns of development include Patterson, DeBaryshe, and Ramsey (1989); Loeber and LeBlanc (1990); Farrington (1986); and Elliott, Huizinga, and Menard (1989). Consequently, this article takes criminal offenses as template-simulating data, but the general idea of trajectories is relevant to many substantive areas and the SMPR model can be used with any time series of count data.

Robustness

The estimation and interpretation of any statistical model requires some auxiliary assumptions. In regression, for example, ordinary least squares estimates will have various desirable properties if errors are normally distributed with equal variance and uncorrelated across cases. In models for count data, the distribution of counts around the expected value plays a role similar to the error term in a regression. It is generally impossible to be sure that the assumptions are met in any given case, and there is often no reason to believe that they are exactly true. Hence, it is important to evaluate the *robustness* of any technique—its ability to provide approximately accurate answers even when the assumptions are not exactly true.[1]

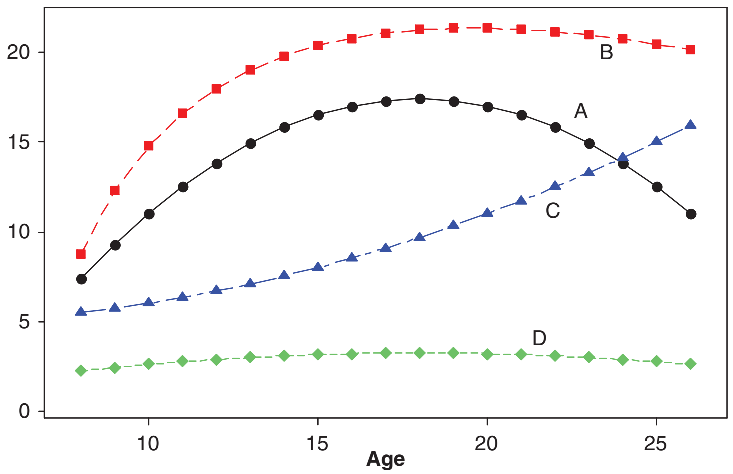
Robustness is a matter of degree: The results produced by any method will be affected if the assumptions are violated to a sufficient degree, but some methods are more robust than others. If a technique is more sensitive to deviations from the assumptions than one would like, it is often possible to modify it in ways that increase robustness. It may also be possible to develop specification tests in order to determine whether a particular assumption can be maintained, such as the familiar Durbin-Watson test for serially correlated residuals.

This article provides a study of the robustness of estimates from the SMPR model as implemented in "Proc Traj," a procedure in the SAS statistical software program (Jones, Nagin, and Roeder 2001; Jones and Nagin 2007). We focus on the number of groups and the nature of differences in their trajectories, and we give some attention to the size of the groups. The first issue has been the focus of many studies: The well-known article of D'Unger et al. (1998), for example, is titled "How Many Latent Classes of Delinquent/Criminal Careers?"

The nature of the differences in trajectories has not received as much systematic attention, but it is of considerable theoretical interest. One possibility is that the difference involves only the levels: that one trajectory differs from another by the same multiplicative factor at all ages. The alternative possibility is that the ratio of propensities changes in magnitude or even direction. Figure 1 illustrates some differences of this kind with hypothetical trajectories based on the findings of previous studies. Trajectory A represents the typical age-crime curve, which peaks in adolescence or early adulthood. Trajectory B represents a life-course persistent or "chronic" group, in which criminal propensity is not only relatively high in youth but also declines relatively slowly with age. This combination implies that the gap between the chronic group and the typical group becomes larger with age. Trajectory C represents a "late-onset chronic" group like that found by D'Unger et al. (1998), in which criminal propensity actually rises after adolescence. Finally, many studies find a group with a very low propensity at all ages, represented by Trajectory D.

We will refer to this question as *parallelism* because trajectories that differ by a constant multiple will be parallel when plotted on a logarithmic scale. Testing hypotheses about parallelism requires manipulation of the covariance matrix of parameter estimates, so discussion of the issue in empirical work has usually been based on informal examination of the estimated trajectories. However, one widely used program for SMPR models, Proc Traj, now includes an option that makes it easy to test hypotheses about parallelism (Jones and Nagin 2007).

Figure 1
Hypothetical Age Trajectories



There are often no clear theoretical expectations about the size of groups, but the issue is still of considerable practical importance. For example, from a theoretical point of view, the existence of even a small late-onset chronic group would refute the idea of a universal age-crime curve. The contribution of this group to the total crime rate, however, would depend on its size.

Models of Trajectories

Models for count data can be thought of as a combination of two parts, the first representing propensity as a function of independent variables and the second representing the relationship between the propensity and the observed counts. These could be described as the systematic and random parts of the model. The systematic part is usually assumed to have a log-linear form, which prevents propensity from falling below zero and means that differences in propensity can be understood in multiplicative terms. The systematic part of the model can be written

$$\log(\lambda_t) = f(t), \tag{1}$$

where λ is the expected number of events per unit of time and $f(t)$ is a function of age. If desired, one can add other independent variables to the right-hand side.

Polynomial functions in t can be used to approximate the unknown function $f(t)$. We will use a second-degree polynomial as an illustration, but the general points apply to polynomials of any degree:

$$\log(\lambda_t) = \beta_0 + \beta_1 t + \beta_2 t^2. \quad (2)$$

Equation (2) implies that all people of the same age have the same criminal propensity. Given this model, individual differences in offending represent merely random fluctuation around the common propensity λ_t. Lasting differences in propensity can be represented by adding an individual-specific term that remains constant over time:

$$\log(\lambda_{ti}) = \beta_0 + \beta_1 t + \beta_2 t^2 + u_i. \quad (3)$$

The index i represents the individual, so the u_i term means that there will be a number of parallel propensity curves at different levels but sharing the same shape. If u_i has a discrete distribution, there will be a finite number of trajectories; if it has a continuous distribution, there is no definite number of trajectories because each individual can have a unique value. In principle, a continuous distribution seems more likely: As Nagin and Land (1993:330) observe, "most modern criminological theories are ultimately theories of degree." However, discrete distributions may have a practical advantage in terms of flexibility. To estimate (3) with a continuous distribution, it is necessary to specify the form of the distribution, and existing programs are limited to a small number of distributions, sometimes only the normal. Hence, discrete distributions may provide a better approximation for distributions that have shapes very different from the available ones.

There may also be differences in the coefficients of the polynomial terms:

$$\log(\lambda_{ti}) = \beta_0 + (\beta_1 + \zeta_i)t + (\beta_2 + \eta_i)t^2 + u_i. \quad (4)$$

Model (4) permits differences in the shape as well as the level of trajectories. Like u, ζ and η may have either discrete or continuous distributions. The model can be estimated using programs for multilevel models assuming continuous distributions for ζ and η, but only a small number of distributions are currently available. Once again, the flexibility of discrete distributions means that they can be useful as approximations even when they are not literally true. Moreover, many theoretical discussions of trajectories involve a small number of ideal types rather than continuous variation, so the use of discrete distributions produces a closer match between theory and analysis.

If the distributions of ζ and η are discrete, it is convenient to write equation (4) as the sum over k latent classes:

$$\log(\lambda_{it}) = \Sigma_k \delta_{ik}(\beta_{0k} + \beta_{1k}t + \beta_{2k}t^2). \quad (4a)$$

The variable δ_{ik} has the value of 1 if case i is a member of class k and 0 otherwise. Each individual is a member of one and only one class. The age-crime curve for a class is defined by the values of β_0, β_1, and β_2, so the ζ, η, and u terms in (4) are absorbed in the different values of the βs. For example, β_{0k} in (4a) is equivalent to $\beta_0 + u_k$ in (4), where u_k is the deviation of group k from the baseline value. Equation (4a) is the form in which the SMPR model is usually presented in the literature, so it will be used from this point onward.

Saying that the curves in two groups are parallel is equivalent to saying that all of the β terms except β_0 are equal for those groups. For example, with a second-degree polynomial, the hypothesis of parallel curves in Groups 1 and 2 implies that both $\beta_{11} = \beta_{12}$ and $\beta_{21} = \beta_{22}$ are true. The covariance matrix of the parameter estimates can be used to construct a Wald test of the hypothesis (Harvey 1990:166-69).[2] Like standard likelihood ratio tests, Wald tests produce a statistic that has a χ^2 distribution when the null hypothesis is true.

Determining the Number of Groups

Because class membership cannot be observed directly, model (4a) is a form of latent class model (Nagin 1999:146). In addition to the parameters shown in the equation, the model includes parameters representing the proportion of cases in each latent class, which we will refer to as π_k. The π_k parameters must sum to 1, so there are only $k-1$ nonredundant π parameters. For example, a two-group model in which the trajectory for each class followed a second-order polynomial would include seven free parameters: three β parameters for each latent class and one nonredundant π parameter.

The number of latent classes itself is not a parameter of the model and cannot be estimated directly from the data. Rather, it must be specified in advance in order to estimate a model. The optimal number of groups can be determined by estimating models with different numbers of latent classes and choosing the one that yields the best fit. There are several possible indexes of fit, but the most popular one is the Bayesian information criterion (BIC). The BIC can be expressed in a number of ways: We will use $-\log(L) + 0.5{*}k \log(N)$, where L is the likelihood, k is the number of parameters estimated in the model, and N is the number of individuals in the data set (Raftery 1995). In this form, the BIC has a positive sign, and smaller values indicate better fits.

Differences in BIC statistics can be interpreted as a measure of the weight of evidence provided by the data in favor of one model against the other. Specifically, if D is the difference between the BIC statistics of two models, the weight of evidence is given by $\exp(D)$. For example, a difference of 5 in the BIC translates to odds of about 150 to 1 in favor of the model with the smaller BIC. The usual procedure is to begin with one latent class and then fit models with increasing numbers of latent classes until the value of the BIC begins to increase.

Conditional Distribution of Counts

The preceding model specifies the chance of an event occurring—whether the event actually occurs is inherently unpredictable. If the hazard rate for some event is constant during a period of time and each occurrence of the event is independent, the total number of events in the period will follow a Poisson distribution conditional on the level of λ_{it} (Johnson, Kotz, and Kemp 1992). One feature of the Poisson distribution is that the mean and variance are equal.

Several conditions could lead to departures from a Poisson distribution. One is short-term individual variation in propensity, which can be represented by adding an error term e_{it} to equation (4a). That is, in a given period, individual propensity could be above or below its normal level as a result of various idiosyncratic factors. The additional error term will produce *overdispersion* in the counts; that is, the variance will be larger than the expected value. Overdispersion will also result if the occurrence of an event increases the hazard rate for subsequent events (Barron 1992). For example, it is possible that people who commit one offense will feel less inhibition against committing another. A final possible source of overdispersion is variation in the hazard rate within the time period.

In the framework used here, overdispersion can be represented by including an error term in the model for propensity:

$$\log(\lambda_{it}) = \Sigma\delta_{ik}(\beta_{0k} + \beta_{1k}t + \beta_{2k}t^2) + e_{it}. \quad (5)$$

The conditional distribution of the counts will depend on the distribution of e_{it}. The case that has been analyzed most extensively is the negative binomial distribution, which occurs when e_{it} is the logarithm of a random variable drawn from a gamma distribution (Johnson et al. 1992:223-24). The negative binomial includes the Poisson distribution as a limiting case when the variance of e_{it} equals zero, and experience has shown that it often provides a good fit to count data.

Although the negative binomial distribution is theoretically appealing, it is computationally burdensome when combined with a latent class model and is not available in standard software for trajectory models. An alternative model for overdispersion is the zero-inflated Poisson (ZIP) distribution, which is a Poisson distribution with some excess number of zero counts (Johnson et al. 1992:186-87). The ZIP distribution is included as an option in Proc Traj. A substantive rationale for the ZIP model is that people "may go through periods of active offending ... interspersed with periods of dormancy" (Nagin and Land 1993:334; see also Barnett, Blumstein, and Farrington 1989). The periods of dormancy, in which by definition no offenses are committed, result in an excess of zeros.

The ZIP distribution can be represented by an "intermittency" variable, which in Proc Traj is modeled as a polynomial function of age. For example, if a second-degree polynomial is used,

$$\log(p_t/(1-p_t)) = \gamma_0 + \gamma_1 t + \gamma_2 t_2, \tag{6}$$

p_t is the probability that a person will be in an active state at time t.

The value of λ from equation (4a) is multiplied by the intermittency variable, so the offending rate falls to zero when a person is not in an active state. Although both equations (4a) and (6) involve powers of age, the γ parameters are entirely distinct from the β parameters. Most researchers devote little attention to the parameter estimates for the intermittency equation, using it simply as a way to allow for the possibility of departures from the Poisson distribution.[3]

Simulated Data

We investigate the robustness of the SMPR model using simulated data. All of the simulated data sets contained 1,000 cases. Loughran and Nagin (2006) find that in samples of this size the parameter estimates have little bias and are approximately normally distributed given correct specification of the model. General considerations of statistical power suggest the chance of underestimating the number of latent classes should decline as the number of cases increases. That is, additional cases provide more information to distinguish different trajectories. When the BIC is used to select models, the probability of overestimating the number of classes declines to zero as sample size increases, and with samples of the size used here, the BIC favors models with fewer parameters than the major alternatives.[4]

The simulated data are based on the 1958 Philadelphia birth cohort study reported in Tracy, Wolfgang, and Figlio (1990), which Loughran and Nagin (2006) also used as a model. This study followed 13,000 members of the cohort from age 8 to age 26 and measured their number of police contacts in each year. The annual rate was about 0.01 at age 8, rose to a peak of 0.28 by age 16, and then declined to about 0.02 at age 26. The general pattern of a sharp rise in adolescence, a peak in the late teens, and then a decline is typical of most data on criminal behavior. Like these data, our simulated data sets contained observations for 19 periods, which we refer to as representing "ages" 8 to 26, with the peak occurring about age 16 or 17.

The simulated data sets vary along four dimensions: the number of trajectories, the shape of the trajectories, the mean number of events, and the degree of overdispersion. We consider three possibilities on the first dimension: one group, two groups, and continuous variation in u_i. On the second dimension, we consider two possibilities: a quadratic polynomial and a curve based on the values observed in the Philadelphia data.[5] On the third dimension, we consider two possibilities: means approximately equal to the values in the Philadelphia data and larger means. On the final dimension, we consider several degrees of overdispersion. The model for overdispersion is a negative binomial distribution: that is, equation (5) plus an error term distributed as the logarithm of a random variable from a gamma distribution. The gamma distribution has two parameters, known as β and γ, which affect the shape and scale, respectively. We specify $\gamma = 1/\beta$, so the mean is always equal to 1 and the value of β determines the shape and variance of the error distribution. With a smaller β, most values are near 0, but a few are large, producing more overdispersion in the counts. The Poisson is the limiting case of the negative binomial distribution as β increases. The conditions generating the data are summarized in Table 1.

We used the same procedure to search for the best fitting model in all data sets. The starting point was a single group with a constant rate and no intermittency parameter. The first stage was to increase the order of the trajectory by adding polynomial functions of time until the best fit was obtained. We then applied the same approach to the intermittency function, starting with a Poisson model, then a ZIP model with a constant rate of intermittency, then ZIP models involving polynomial functions of time.

Table 1
Summary of Simulated Data Conditions

Condition	Shape of Trajectory	Error Distribution	Number of Trajectories	e_{it}	u_i
1	Quadratic function	Poisson	1	No	No
2	Quadratic function	Negative binomial	1	Yes	No
3	Philadelphia data	Poisson	1	No	No
4	Philadelphia data	Negative binomial	1	Yes	No
5	Philadelphia data × 8	Negative binomial	1	Yes	No
6	Philadelphia data	Poisson	Undefined	No	Yes
7	Philadelphia data	Poisson	2	No	No
8	Philadelphia data	Negative binomial	2	Yes	No

The next stage was to successively increase the number of groups given the order of the trajectory and intermittency polynomials. Finally, we checked the order of the trajectory and intermittency polynomials given the number of groups, and if any changes improved the fit we reestimated the number of groups given the order.

The analysis of each data set took a good deal of time because of the number of trial models that had to be estimated. Moreover, it was often necessary to experiment with different starting values in order to achieve convergence.[6] As a result, it was not practical to perform a large number of replications, and we considered only 50 for each condition. Our goal was to identify the general contours of performance under various conditions rather than to estimate precise error rates, and the number of replications was sufficient for this purpose.

Results

Condition 1

In Condition 1, there is a single underlying trajectory following a quadratic curve, and the counts follow a Poisson distribution. Thus, this condition represents the ideal situation in which the assumptions used to estimate the model are identical to those used to generate the data. The results can be described briefly: They were correct in all 50 replications. The best fitting models all had a single group in which the relationship between propensity and age followed a quadratic curve. In most cases, the BIC strongly favored the one-group model, giving odds of 1,000 to 1 or more against a two-group model. The estimated relationship between age

Table 2
Condition 2: Data With One Latent Class, Age-Specific Means Following Quadratic Polynomial, and Varying Distributions

Distribution	One-Group Solutions	Two-Group Solutions	Three-Group Solutions
Poisson	50 (100%)	0	0
Negative binomial ($\beta = .5$)	50 (100%)	0	0
Negative binomial ($\beta = .25$)	19 (38%)	30 (60%)	1 (2%)
Negative binomial ($\beta = .14$)	2 (4%)	42 (84%)	6 (12%)

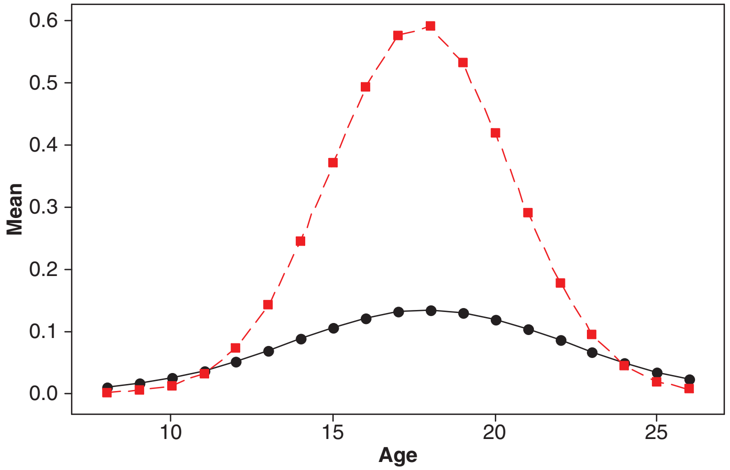
and propensity closely resembled the curve used to generate the data, and the Poisson model provided a better fit than the ZIP.

Condition 2

Condition 2 is identical to Condition 1 except that the counts follow a negative binomial rather than a Poisson distribution. The Philadelphia data reported by Tracy et al. (1990) show a good deal of overdispersion. Given the observed mean at age 16, the maximum value of a random variable following a Poisson distribution would rarely be more than 5; in the actual data, there were 6 cases with 10 police contacts, 1 with 11, and 1 with 24. Results for data generated with differing amounts of overdispersion are summarized in Table 2. The inclusion of an error term generated by a gamma distribution with $\beta = .14$ produces a comparable degree of overdispersion in the simulated data. We also consider two more moderate degrees of overdispersion, $\beta = .5$ and $\beta = .25$. A one-group model was selected in all 50 cases when $\beta = .5$, in 19 when $\beta = .25$, and in only 2 when $\beta = .14$. That is, there is a tendency to overestimate the number of groups that increases with the degree of overdispersion.

Figure 2 provides more detail about the nature of the best fitting solution found for one reasonably typical case from this condition. The lower curve, containing about 90 percent of the cases, peaks at about 0.13 convictions per year. The upper curve, including about 10 percent of the cases, peaks at almost 0.60. The smaller group is estimated to have a much more rapid rate of growth and decline, so the gap between the two groups is largest in the late teens, resembling the "adolescence-peaked" trajectory found by D'Unger et al. (1998:1613) in their analysis of the Philadelphia data. A Wald test rejects the hypothesis that the curves are parallel,

Figure 2
Estimated Trajectories in Example Data Set From Condition 2



$X^2(2) = 11.8$, $p < .001$; $N = 1{,}000$. That is, there is spurious evidence of distinct trajectories that differ in shape as well as level.

The results can be understood by considering the influence of the cases with very large counts. Such counts would be very rare under a Poisson distribution with the given mean, and the ZIP distribution, although it allows for more dispersion, does not dramatically increase the probability of very large counts. Hence, these cases can be accommodated only by the creation of the group with the higher estimated peak propensity.

Condition 3

In this condition, the data are generated by a single underlying trajectory, and variation around the curve follows a Poisson distribution. Rather than the quadratic function in Condition 1, the means are modeled on those in the Philadelphia data. The empirical curve cannot be exactly reproduced by any polynomial of reasonable degree, so any estimated model is necessarily an approximation. Consequently, in this condition we can assess the robustness of the SMPR results against misspecification of trajectory shape.

Table 3
Condition 3: Data With One Latent Class, Poisson Distribution, and Age-Specific Means Following Philadelphia Data

Identifier	Groups in Best Fitting Model	Group Sizes	BIC		Log Likelihood	
			One Group	Two Groups	One Group	Two Groups
3-A	1	100%	5,705	5,727	5,674	5,672
3-B	1	100%	5,588	5,605	5,557	5,550
3-C	1	100%	5,546	5,561	5,515	5,506
3-D	1	100%	5,596	5,616	5,565	5,561
3-E	1	100%	5,683	5,703	5,652	5,647
3-F	1	100%	5,554	5,574	5,523	5,519
3-G	1	100%	5,660	5,675	5,629	5,620
3-H	1	100%	5,481	5,499	5,450	5,444
3-I	1	100%	5,535	5,561	5,504	5,505
3-J	1	100%	5,508	5,525	5,477	5,469

Note: BIC = Bayesian information criterion.

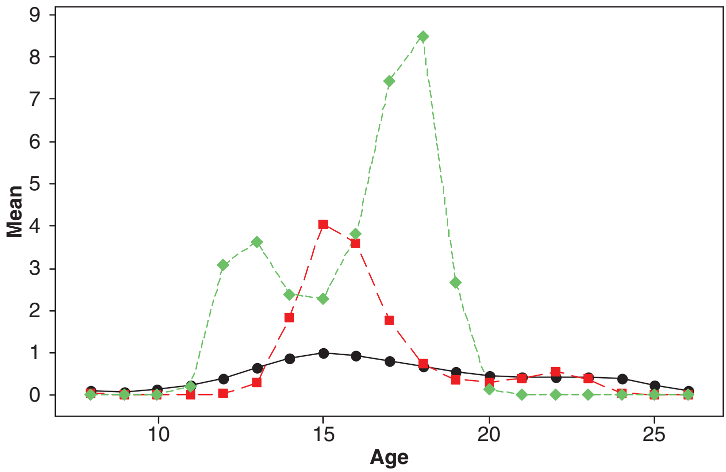
The search procedure resulted in a fifth-order polynomial for the trajectory. Given this specification, the one-group model provided the best fit in all 50 data sets. Table 3 presents the results from the first 10 data sets. The differences in the BIC ranged from about 15 to 25, indicating very strong evidence in favor of the one-group models. In analyses not reported in the table, we used a second-order polynomial to model the trajectories and found that models of multiple groups frequently fit better than single-group models. Thus, estimates of the number of groups appear to be robust against the slight misspecification involved in using a polynomial to approximate the trajectory. However, they are not robust against the more serious misspecification of using a polynomial of insufficiently high degree.

Condition 4

Condition 4 was identical to the previous condition except that the counts followed negative binomial rather than Poisson distributions about the conditional means. As in Condition 2, data were generated using β values of .5, .25, and .14.

Once again, the search procedure resulted in the use of a fifth-degree polynomial. With $\beta = .5$, the one-class model provided the best fit in 48 of 50 cases; with $\beta = .25$, in 36 cases; and with $\beta = .14$, in only 4. At the highest level of overdispersion, a three-group solution was chosen in five

Figure 3
Estimated Trajectories in Example Data Set From Condition 4



data sets. In Figure 3, we graph the results of the estimated trajectories from one of the three-group solutions. The largest group has a gradual rise and decline and a low peak. A second group has a rapid rise and fall: At the beginning and end of the period, propensity is about the same as in the first group, but at its peak it is much higher. Finally, there is a small group with much higher rates and a somewhat later peak. Using a Wald test, the hypothesis of parallelism is strongly rejected for each pair of curves.

These results indicate that estimates of the number of groups are sensitive to overdispersion, even when the underlying trajectory is correctly specified and an intermittency parameter is included. This issue has not received much attention in the literature, but one simple approach has occasionally been used: top-coding high counts. Loughran and Nagin (2006:258) analyzed the Philadelphia data and reported that the "small contingent of outlying data points was the source of considerable model instability." After top-coding at a value of 10 for 2-year blocks, the SMPR produced stable and plausible results, but because they were using real data, the true model was unknown. With the present data, it is possible to see how top-coding affects the chance of selecting the true one-group model.

Table 4
Condition 4: Data With One Latent Class, Age-Specific Means Following Philadelphia Data, and Varying Distributions

Distribution	One-Group Solutions	Two-Group Solutions	Three-Group Solutions
Negative binomial, $\beta = 0.5$	48	2	0
Top coded at 5	50	0	0
Poisson model	0	6	44
Negative binomial, $\beta = .14$	4	41	5
Top coded at 7	14	34	2
Top coded at 5	46	4	0
Poisson model	0	0	50

Table 4 summarizes the results obtained for $\beta = .25$ and $\beta = .14$ under four different approaches: a Poisson model (no intermittency parameter), the ZIP model discussed above, and a ZIP model applied with two levels of top-coding. One set of analyses used top codes of 5 for $\beta = .25$ and 7 for $\beta = .14$. These values were selected to include about the highest 0.1 percent of cases in the peak years, so only a handful of cases in each data set were affected by the top-coding. For $\beta = .14$, we also considered a value of 5 for top-coding, which affected about 1 percent of the cases in the peak years.

With the more moderate degree of underdispersion, the Poisson model performs poorly, never selecting a one-group model. Using a ZIP model produces substantial improvement, and a ZIP model with top-coding at 5 (0.1 percent of the cases) results in a correct solution in all 50 cases. With the higher degree of overdispersion, a combination of the ZIP and top-coding at 7 improved the performance, but multiple-group solutions were still chosen in more than 70 percent of the cases. Reducing the top code to 5 improved the performance considerably, as the rate of multiple-group solutions fell to 8 percent, although this error rate is still large enough to be cause for some concern.

Condition 5

This condition is identical to Condition 4 except that the mean at each age is multiplied by a factor of 8. With a Poisson distribution, the one-group solution provided the best fit in all 50 cases. Overdispersion makes multiple-group solutions more likely, and the deterioration in performance is much more rapid than in Condition 3. Even at the most moderate degree of overdispersion, $\beta = .5$, the one-group solution was never chosen, and usually the three-group solution was preferred.

Condition 6

In this condition, there is a quadratic relationship between age and propensity, with continuous variation in the overall level. The differences in level are modeled by an individual-specific term with a gamma distribution that has a multiplicative effect on the standard trajectory. Each person has a unique trajectory that is parallel to all others but at a different level, so a model with a small number of trajectories can be only an approximation. In a good approximation, the estimated groups would differ only in level, not in shape. In effect, the latent classes would divide the individual variation in propensity into several categories, such as low, medium, and high.

The best fit was obtained with either three or four groups. Although the estimated size of the groups varies substantially, there was generally a group with a low propensity including a majority of cases, a smaller group with a somewhat higher propensity, and one or two small groups with even higher propensities. The differences between the estimated trajectories involved only levels, not shapes. That is, under these conditions the SMPR provides a good approximation to the underlying continuous distribution.

Condition 7

In this condition, there are two latent classes with trajectories that are not parallel. In terms of Moffitt's (1993) typology, the first could be described as representing the "adolescence-limited" pattern and the second as representing a life-course persistent pattern. The second group, including 10 percent of the cases, has a higher rate of offending in childhood and adolescence and a slower decline in the late teens and 20s. The share of all offenses committed by this group rises from 14 percent at age 16 to 34 percent at age 20 and passes 50 percent at age 25. Variation around each trajectory follows a Poisson distribution.

A one-group model was chosen in 3 cases, a two-group model in 44, and a three-group model in 3. The estimated shapes of the trajectories in the two-group solutions were close to the actual curves, and the hypothesis of parallelism could be strongly rejected. In the three-group models, the

Table 5
Condition 7: Two Latent Classes With Poisson Distribution

					BIC			Log Likelihood	
Identifier	Groups in Best Fitting Model	Group Sizes (%)			One Group	Two Groups	Three Groups	One Group	Two Groups	Three Groups
7-A	1	100			5,577	5,582	NC	5,557	5,537	NC
7-B	2	93	7		5,528	5,512	5,531	5,507	5,467	5,462
7-C	2	92	8		5,746	5,722	5,737	5,725	5,677	5,668
7-D	2	90	10		5,713	5,665	5,684	5,692	5,620	5,615
7-E	2	89	11		5,462	5,445	5,469	5,441	5,400	5,400
7-F	2	64	36		5,480	5,477	5,489	5,459	5,432	5,420
7-G	2	63	37		5,635	5,624	5,646	5,614	5,580	5,577
7-H	2	58	42		5,742	5,736	5,737	5,721	5,691	5,667
7-I	2	54	46		5,443	5,429	5,438	5,422	5,384	5,369
7-J	3	74	14	12	5,840	5,837	**5,835**	5,819	5,792	5,766

Note: BIC = Bayesian information criterion; NC = no convergence.

trajectory of the third group tended to be a hybrid of the two actual trajectories. Table 5 gives more detail on the results for the first 10 data sets. These results show that the estimates of the size of the groups were not very accurate; there were 4 cases in which the smaller life-course persistent group was estimated to contain more than 30 percent of the cases. Overall, the procedure was fairly successful in identifying the number of groups and the shape of the trajectories. Estimates of the size of the groups, however, were less reliable.

Condition 8

In this condition, the underlying model is like that of Condition 7, but the conditional counts follow a negative binomial distribution with a β of .14. The best model, as selected by the BIC, contained two groups in 30 of the data sets and three groups in 10 data sets. The three-group models did not converge in the remaining 10 cases. Nonconvergence can be a sign of an attempt to overfit the model, so one might argue that these cases should be counted as two-group models. In either case, comparing the results to those from Condition 7, overdispersion results in a tendency to overestimate the number of latent classes. This tendency is much weaker, however, than that found when there was only one group in the actual data (Conditions 3 and 4). Even with no top-coding, the correct solution was

Table 6
Condition 8: Two Latent Classes With Negative Binomial Distribution ($\beta = .14$)

Identifier	Groups in Best Fitting Model	Group Sizes (%)			BIC One Group	BIC Two Groups	BIC Three Groups	Log Likelihood One Group	Log Likelihood Two Groups	Log Likelihood Three Groups
8-A	2	93	7		5,381	5,347	5,351	5,350	5,292	5,271
8-B	2	86	14		5,094	5,064	5,072	5,063	5,009	4,993
8-C	2	92	8		5,185	5,151	5,156	5,154	5,095	5,077
8-D	2	97	3		5,083	5,050	5,052	5,052	4,995	4,972
8-E	2	89	11		5,323	5,292	5,300	5,291	5,236	5,221
8-F	2	89	11		5,065	5,024	5,037	5,034	4,969	4,957
8-G	2	94	6		5,206	5,178	5,184	5,175	5,123	5,105
8-H	3	81	10	9	5,246	5,207	5,198	5,215	5,152	5,118
8-I	3	87	10	3	5,291	5,210	5,207	5,260	5,152	5,127
8-J	3	80	16	4	5,301	5,258	5,253	5,270	5,152	5,173

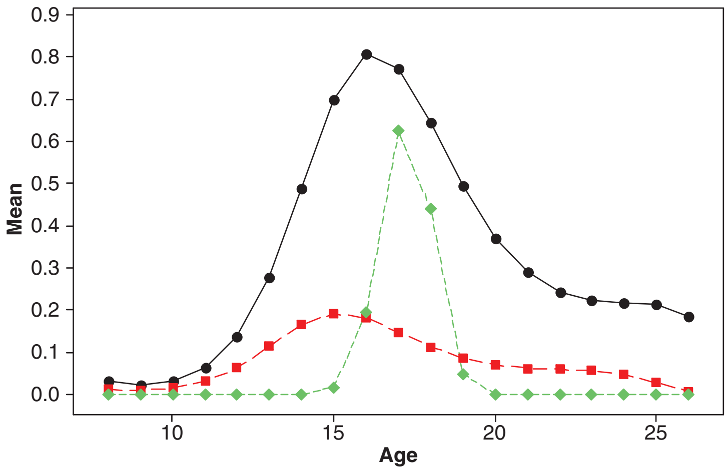
Note: BIC = Bayesian information criterion.

chosen in 60 to 80 percent of the cases in Condition 8, depending on how the failures to converge are counted.

Table 6 gives more detail on results from the first 10 data sets generated under this condition. In the two-group solutions, the estimates of the group sizes are fairly close to the true values of 90 and 10 percent. A comparison between Tables 5 and 6. Suggests that overdispersion results in more accurate estimates of group sizes. It is not clear why this should be the case, but the issue deserves further investigation.

Figure 4 displays the estimated trajectories from the best fitting model for one of the three-group solutions. One group, containing about 2 percent of the sample, seems to follow a life-course persistent pattern—high and relatively persistent rates of offending. The largest group, including about 86 percent of the sample, has a consistently low level, while a third, containing about 11 percent of the sample, begins at a low level but rises to a higher peak. This group, unlike the others, also shows some increase at the very end of the period. The hypothesis that the trajectory in the third group is parallel to the trajectories in either the first or second group can be rejected. Thus, the solution accurately identifies a minority with a relatively high and persistent propensity but also adds a spurious group with a distinctive pattern. The estimated trajectory in this group is similar to that found for the spurious minority group in Figure 3.

Figure 4
Estimated Trajectories in Example Data Set From Condition 8



Discussion

Table 7 summarizes the results of the various analyses. The major finding of this study is that when there is only one underlying trajectory and overdispersion is present, the SMPR model frequently finds multiple groups. In effect, it has difficulty distinguishing between overdispersion and the presence of multiple groups. The resulting groups may differ in shape as well as level, resulting in spurious inferences about the life course. Sensitivity to overdispersion seems to be greater when the mean number of events is larger. A second important point arose in the course of the analysis of Condition 3: Serious misspecification of the underlying trajectory increases the chance of spurious findings of multiple groups. Consequently, researchers should search for the correct specification rather than simply choose a convenient low-order polynomial and vary the number of groups.

When there was continuous variation in the level of propensity (Condition 6), the SMPR model gave a good approximation to the actual pattern. The results indicated differences only in levels, with no spurious evidence

Table 7
Summary of Results Under Different Conditions

Condition	Data	Number of Groups	Correct Results (%)
1	Quadratic function, Poisson distribution	1	100
2	Quadratic function, negative binomial distribution	1	4
3	Philadelphia means, Poisson distribution	1	100
4	Philadelphia means, negative binomial distribution	1	8 (original data) 28 (top-coded at 7) 92 (top-coded at 5)
5	Philadelphia means × 8, negative binomial distribution	1	0 (original data)
6	Philadelphia means, Poisson distribution	Undefined	100 qualitatively correct
7	Philadelphia means, Poisson distribution	2	88
8	Philadelphia means, negative binomial distribution	2	60-80

of differences in shape. These results support Nagin's (2005:45-55) observations about the potential value of the model even when the assumption of a finite number of groups is not literally true.

Finally, when there actually were two groups with distinctly different shapes, the procedure produced fairly accurate estimates of the number of groups and their general shapes. Although overdispersion produced some tendency to overestimate the number of groups, it appeared to be weaker than in the other conditions. These results suggest that the effect of overdispersion on estimates depends on the nature of the true model, although more work is needed to discover the nature of the relationship.

The conclusions about the effect of overdispersion add to those of a number of previous studies. R. Matsueda (personal communication, 2004) reports that the presence of outliers tends to lead to overestimation of the number of groups in the SMPR model. Bauer and Curran (2003) find that estimates in latent class models that assume a normal distribution are sensitive to departures from that distribution. These findings suggest that it is important to reduce the dependence of the estimates on assumptions about the conditional distribution of the counts. When the counts follow a negative binomial distribution around the conditional mean, the use of the ZIP distribution produces some improvement but does not eliminate the tendency to overestimate the number of groups.

The most direct way to improve robustness would be to include other options for the distribution of counts. Until this is done, the combination of the ZIP distribution and top-coding of high counts provides substantial improvement at the degrees of overdispersion considered in this study. Although the ZIP distribution can include a larger number of zeros, it does not substantially increase the probability of very large counts, so top-coding is required to reduce their influence. On the other hand, if multiple trajectories are actually present in the data, top-coding reduces the power to distinguish between them. More work is necessary in order to discover how to strike the best balance between maintaining power to distinguish real group differences and minimizing the risk of finding spurious ones. In this case, top-coding about 1 percent of the peak values produced fairly good results, but further analysis is needed to see whether this is an effective rule in general.

Following the recommendations about top-coding and specification would mean that the application of SMPR models would become more labor intensive. Consequently, it would also be useful to develop specification tests for the existence of nonparallel trajectories. From a practical point of view, models with continuously distributed parallel trajectories are inherently easier to fit. Such models can be implemented with at least two publicly available programs, the SAS procedure GLIMMIX and the stand-alone program SABRE (Wolfinger and O'Connell 1993; Centre for Applied Statistics 1996). However, if there are differences in the shapes as well as the levels of trajectories, more complex models such as the SMPR or growth curve models are necessary. As a result, specification tests for the existence of nonparallel trajectories would be useful in deciding on the appropriate model for the analysis of a given data set.

Godfrey (1988) outlines general principles for specification tests, and Lindsay and Roeder (1992) discuss diagnostics for mixture models. The observation that could be applied to develop a specification test in this case is that parallel and nonparallel trajectories would result in different patterns of association between the counts at various ages. Enduring differences in propensity would mean that the pattern of associations could be reduced to a single component with equal loadings at all ages; that is, a person who had a relatively high count at one age would tend to have relatively high counts at all other ages. Nonparallel trajectories would mean that the association between counts at different ages would differ, producing a more complex structure. Although many details would need to be worked out to develop such a test, experimentation with the simulated data for this project suggests that this approach has promise.

It is also useful to consider the relationship between the estimated class membership and any other variables that are available. Because overdispersion involves the error term, it will not be related to any other variables. As a result, membership in any spurious classes produced by overdispersion will be unpredictable. Conversely, if membership can be predicted from other variables, that is a sign that the class does exist.

In conclusion, the idea of multiple trajectories is an important and theoretically appealing one. Empirical conclusions derived from previous applications of the models should be regarded with caution because findings of multiple groups can merely reflect overdispersion in the conditional counts or the use of a polynomial of insufficiently high degree to model the trajectories. A combination of efforts to extend the model to allow for alternative distributions and the development of specification tests can help to improve robustness. Until that time, investigators can obtain better performance by two simple practices: correctly specifying the shape of the trajectory and top-coding large counts.

Notes

1. Some discussions of robustness focus on what Andersen (2008:3) terms "robustness of efficiency," or the efficiency of an estimator under a range of distributional assumptions, but the general term encompasses a wider range. See Portnoy and He (2000) for a review of work in the area.

2. Wald tests require manipulation of the vector of parameter estimates, the covariance matrix of the estimates, and a matrix representing the constraints on the parameters. The tests were implemented using the SAS matrix language Proc Iml.

3. It is possible to have a separate intermittency model for each latent class. However, we use the intermittency model only as a control for overdispersion, we use a single equation.

4. The major alternatives are the Akaike information criterion and standard hypothesis tests, although the use of hypothesis tests is problematic with latent class models (Nagin 2005:62-63).

5. Means for the simulated data were obtained by smoothing the observed means using running medians (Tukey 1977).

6. Proc Traj uses the expectation-maximization (EM) algorithm, which is stable but can have a very slow rate of convergence. We found that experimentation with starting values was often needed to achieve convergence within a reasonable amount of time, especially when polynomials of a high order were used. See Karlis and Xekalaki (2003) for discussion of convergence of the EM algorithm for mixture models.

References

Abbott, Andrew. 1997. "On the Concept of Turning Point." *Comparative Social Research* 16:85-105.

Andersen, Robert. 2008. *Modern Methods for Robust Regression*. Thousand Oaks, CA: Sage.

Barnett, Arnold, Alfred Blumstein, and David P. Farrington. 1989. "A Prospective Test of a Criminal Career Model." *Criminology* 27:373-85.

Barron, David N. 1992. "The Analysis of Count Data: Overdispersion and Autocorrelation." *Sociological Methodology* 22:179-220.

Bauer, Daniel J. and Patrick J. Curran. 2003. "Distributional Assumptions of Growth Mixture Models: Implications for Overextraction of Latent Trajectory Classes." *Psychological Methods* 8:338-63.

Centre for Applied Statistics, University of Lancaster. 1996. *Sabre: Software for the Analysis of Binary Recurrent Events.* Lancaster, UK: Author.

D'Unger, Amy V., Kenneth Land, Patricia L. McCall, and Daniel S. Nagin. 1998. "How Many Latent Classes of Delinquent/Criminal Careers? Results From Mixed Poisson Regression Analysis." *American Journal of Sociology* 103:1593-1630.

Elliott, Delbert S., David Huizinga, and Scott Menard. 1989. *Multiple Problem Youth.* New York: Springer-Verlag.

Farrington, David P. 1986. "Age and Crime." Pp. 189-250 in *Crime and Justice: A Review of Research,* vol. 7, edited by Michael Tonry and Norval Morris. Chicago: University of Chicago Press.

Godfrey, L. G. 1988. *Misspecification Tests in Econometrics.* Cambridge, UK: Cambridge University Press.

Harvey, A. C. 1990. *The Econometric Analysis of Time Series.* Cambridge, MA: MIT Press.

Hirschi, Travis and Michael Gottfredson. 1983. "Age and the Explanation of Crime." *American Journal of Sociology* 89:552-84.

Johnson, Normal L., Samuel Kotz, and Adrienne W. Kemp. 1992. *Univariate Discrete Distributions*. 2d ed. New York: John Wiley.

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

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

Karlis, Dimitris and Evdokia Xekalaki. 2003. "Choosing Initial Values for the EM Algorithm for Finite Mixtures." *Computational Statistics and Data Analysis* 41:577-90.

Lindsay, Bruce G. and Kathryn Roeder. 1992. "Residual Diagnostics for Mixture Models." *Journal of the American Statistical Association* 87:785-94.

Loeber, Rolf and Marc Le Blanc. 1990. "Toward a Developmental Criminology." Pp. 375-473 in *Crime and Justice: A Review of Research*, vol. 12, edited by Michael Tonry and Norval Morris. Chicago: University of Chicago Press.

Loughran, Tom and Daniel S. Nagin. 2006. "Finite Sample Effects in Group-Based Trajectory Models." *Sociological Methods & Research* 35:250-78.

Moffitt, Terrie E. 1993. "Adolescence-Limited and Life-Course-Persistent Antisocial Behavior: A Developmental Taxonomy." *Psychological Review* 100:674-701.

Moffitt, Terrie E., Avshalom Caspi, Nigel Dickson, Phil Silva, and Warren Stanton. 1996. "Childhood-Onset Versus Adolescent-Onset Antisocial Conduct Problems in Males: Natural History From Ages 3 to 18 Years." *Development and Psychopathology* 8:399-424.

Nagin, Daniel S. 1999. "Analyzing Developmental Trajectories: A Semiparametric, Group-Based Approach." *Psychological Methods* 4:139-57.

———. 2005. *Group-Based Modeling of Development.* Cambridge, MA: Harvard University Press.

Nagin, Daniel S. and Kenneth C. Land. 1993. "Age, Criminal Careers, and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model." *Criminology* 31:327-62.

Nagin, Daniel S. and Richard E. Tremblay. 2005. "What Has Been Learned From Group-Based Trajectory Modeling? Examples From Physical Aggression and Other Problem Behaviors." *Annals of the American Academy of Political and Social Science* 602:82-117.

Patterson, Gerald R., Barbara D. DeBaryshe, and Elizabeth Ramsey. 1989. "A Developmental Perspective on Antisocial Behavior." *American Psychologist* 44:329-35.

Portnoy, Stephen and Xuming He. 2000. "A Robust Journey in the New Millennium." *Journal of the American Statistical Association* 95:1331-35.

Raftery, Adrian. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111-63.

Tracy, Paul E., Marvin E. Wolfgang, and Robert M. Figlio. 1990. *Delinquency Careers in Two Birth Cohorts.* New York: Plenum.

Tukey, John Wilder. 1977. *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Wolfinger, Russ and Michael O'Connell. 1993. "Generalized Linear Mixed Models: A Pseudo-Likelihood Approach." *Journal of Statistical and Computer Simulation* 48:233-43.

David L. Weakliem is a professor of sociology at the University of Connecticut. He is currently writing a book on model selection and hypothesis testing. His recent publications have appeared in *Comparative Sociology*, the *Sociological Forum*, and *Work & Occupations*.

Bradley R. Entner Wright is an associate professor of sociology at the University of Connecticut, where he studies the sociology of religion.