Muthén, B. (2000). Methodological issues in random coefficient growth modeling using a latent variable framework: Applications to the development of heavy drinking. Multivariate Applications in Substance use Research, J. Rose, L. Chassin, C. Presson & J. Sherman (eds.), Hillsdale, N.J.: Erlbaum. pp. 113-140. **(#81)**

4

Methodological Issues in Random Coefficient Growth Modeling Using A Latent Variable Framework: Applications to the Development of Heavy Drinking Ages 18–37

Bengt Muthén
University of California, Los Angeles

This chapter discusses four methodological issues arising in growth modeling of alcohol use development with a national longitudinal data set. First, with development over a long period of time it is reasonable that the different background variables have different importance for the growth process during different segments of time. It is shown how this change in importance can be captured in the modeling using the concept of varying centering points. The variation of centering points is illustrated by both linear and quadratic growth models. Second, in a heterogeneous group of individuals such as obtained by a national sample, it is likely that individual trajectories belong to several different, unknown subpopulations corresponding to normative development and several forms of non-normative development. It will be shown how new modeling techniques can capture this notion. Third, the data are obtained in the form of multiple cohorts where nobody in the sample has observations on the outcome variables for all ages, leading to missing data modeling. Fourth, many of the households in the sample have multiple siblings and given that siblings within a household share the same home environment, this gives rise to intraclass correlation that violates customary analysis assumptions of independent observations. The chapter gives a nontechnical overview of solutions to these four complications, where the solutions reveal interesting information in the data. Emphasis is placed on the first two areas, whereas the last two areas are described briefly.

This chapter discusses methodological issues of interest to researchers who study developmental theories with longitudinal data. The discussion focuses on advanced growth modeling analyses. Although methodologically oriented, the discussion will be as nontechnical as possible, instead focusing on the interpretation of the analyses. The chapter by Curran also discusses growth modeling and is a suitable introduction to this chapter.

To motivate the methodological discussion and have a specific data source and growth analysis as a reference point, data from the National Longitudinal Survey of Youth (NLSY) will be used. This data set was collected by the Bureau of Labor Statistics with alcohol supplements from NIAAA. The NLSY is a nationally representative household survey of young adults living in the United States in 1979. It was obtained as a multistage probability sample with oversampling of Blacks, Hispanics, and economically disadvantaged non-Blacks and non-Hispanics. The NLSY includes eight birth-year cohorts from 1957 to 1964 with approximately 1,000 individuals per cohort. The sample includes every age-eligible household member. Alcohol measures are available from the years 1982, 1983, 1984, 1985, 1988, 1989, and 1994 covering ages 18 to 37.

The focus of this chapter is on the development of heavy drinking and alcohol-related problems. The heavy drinking variable is obtained from the question "How often have you had 6 or more drinks on one occasion during the last 30 days?" The response categories (and their scores) are never (0), once (1), 2 or 3 times (2), 4 or 5 times (3), 6 or 7 times (4), 8 or 9 times (5), 10 or more times (6). Also of interest is an outcome variable representing a severity score reflecting alcohol-related problems. The NLSY contains 22 symptom items designed to capture DSM-IV diagnoses of alcohol dependence and abuse. From a factor analysis of these items, a sum of the 17 items measuring the most severe dimension is created to form the severity score. Background variables, also referred to as time-invariant covariates, are gender, ethnicity (Black, Hispanic, other), family history of alcohol problems, early onset, dropping out of high school, and college education. Family history information is obtained by the question "Have any of your relatives listed on this card been alcoholics or problem drinkers at any time in their lives?", recorded as the three dummy variables FH1 (family history among first-degree relatives only), FH23 (family history among second- or third-degree relatives only), and FH123 (family history

among first-and second- or third-degree relatives). Early onset information is obtained from the question "How old were you when you first started drinking?" with the probe "For example, having two or more drinks a week." This variable is scored 0/1 with 1 for onset at or before age 14. High school dropout is scored 0/1 with 1 for not having completed high school by age 22. College education is scored 0/1 with 1 for some college by age 22.

The sample means for frequency of heavy drinking are shown in Fig. 4.1 for ages 18 to 37. The mean curve shows the typical increase from age 18 to the peak around age 21 with a subsequent decline. The variation in individual trajectories is, however, large and it is of interest to assess the amount of variation, how much of this variation can be explained by the set of time-invariant covariates, and which covariates are particularly important.

The NLSY growth modeling gives rise to several interesting and common methodological issues, four of which are discussed here. The first two will be discussed at some length, and the last two are described more briefly in order to show the range of issues common in longitudinal analysis.

The first issue concerns the influence of background variables on the growth process. In the NLSY, development of problematic alcohol use is studied over an age span of 20 years. With development over such a long period of time, it is reasonable that different background variables have different importance for the growth process during different segments of time. For example, it is of interest to study if dropping out of high school has more or less of an impact on problematic alcohol use when an individual is in his/her 20s or 30s. It is shown how such changes in importance can be studied using modeling that applies the notion of varying centering points.

The second issue concerns the heterogeneity of development across individuals in the sample. In a heterogeneous group of individuals such as obtained by a national sample, it is likely that individual trajectories belong to several different, unknown subpopulations corresponding to normative development and several forms of non-normative development. It is shown how emerging modeling techniques can capture this type of heterogeneity. The new type of approach is very promising for developmental studies in that it provides estimates for all the different growth models for the subpopulations as well as estimates of a person's likelihood of belonging to a certain subpopulation.

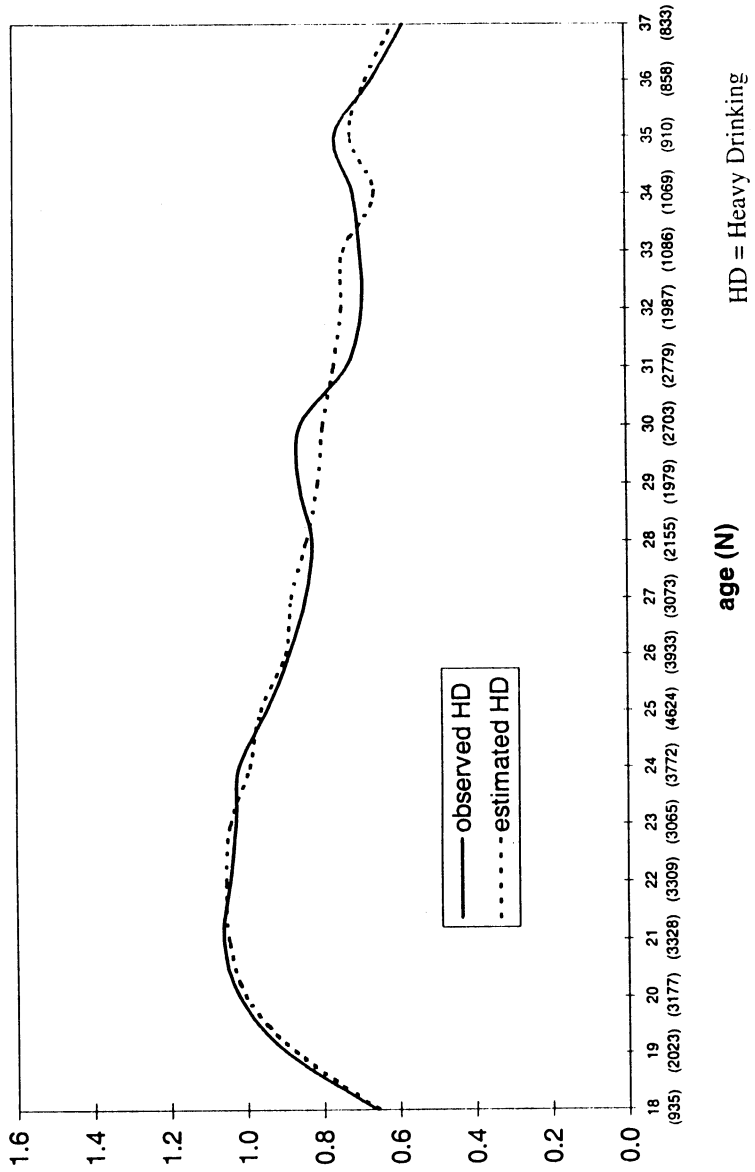

FIG. 4.1. Sample means for frequency of heavy drinking in the NLSY.

The third issue concerns the form of the data for studies covering a long time period. The data are often obtained in the form of multiple cohorts where nobody in the sample has observations on the outcome variables for all ages. This leads to modeling using missing data techniques. An overview is given of some key issues when modeling with missing data.

The fourth issue concerns nonindependence of sample units due to cluster sampling. In the NLSY, many of the households in the sample have multiple siblings and given that siblings within a household share the same home environment, this gives rise to intraclass correlation that violates customary analysis assumptions of independent observations. An overview is given of modeling with cluster samples.

The discussion of these four methodological issues draws on recent work in Muthén and Muthén (in press), Muthén and Shedden (in press), and Muthén and Khoo (1998). All analyses were performed using Mplus, a new comprehensive modeling program for applied researchers (Muthén & Muthén, 1998).

VARYING CENTERING POINTS

This section discusses the use of varying centering points to study changes over time in how a set of background variables influences the growth process. The concept of centering arises in longitudinal analyses carried out using random coefficient growth modeling, also known as multilevel or mixed linear modeling. Muthén and Curran (1997) presented a discussion of such modeling in the latent variable framework used here. An introduction is given in Curran (1998), in this book, and also in Muthén and Khoo (1998).

Briefly stated, growth modeling in the latent variable framework uses latent growth factors to describe the variation across individuals in development over time. For each individual, the growth factor values determine the systematic part of the development over time. This is the person's trajectory over time. Time-specific residuals add to this systematic development to give the observed outcome at each time point.

Conceptually, centering at a certain age can be seen as a particular way of defining the meaning of the growth factors. As an example, in a model of linear growth over time there is an intercept factor and a slope

factor. Our discussion focuses on the intercept factor. The intercept factor determines the level of the systematic part of the development at a particular age. This age is chosen by the analyst as an age that is of particular substantive interest. If the lowest age is chosen, the intercept factor is referred to as initial status. Alternatively, the highest age may be of primary interest in which case the intercept factor concerns ending status. Relating the intercept factor to background variables assesses the influence of the background variables on the intercept defined at the age chosen. As is intuitively clear, however, the influence of the background variables need not be the same at different ages. The estimates will be different as will the standard errors and the t values. In this way, a background variable that is insignificant with a certain centering point may be significant with another. Here, it is proposed that it is valuable to look at the results for all possible centering points in order to study the change over age of the covariate influence.

Methodologically, centering at a certain age can be seen as a particular parameterization of the growth model. Different centering points are statistically equivalent in that they give the same model fit. Different choices correspond to different ways of looking at the same data. This is much like different rotations in exploratory factor analysis being different ways of representing the same sample correlation matrix. The growth analysis need in principle only be done for one centering point, whereupon a transformation of the solution to any other centering point can be carried out. Until this facility is commonly available in software, however, the user can simply reanalyze the model with the different centering points of interest.

Next, the principles of centering are shown for linear and quadratic growth. In the subsequent illustration, these centering types are applied, and are discussed in the context of joint growth modeling of two different outcomes.

Centering With Linear Growth

A simple linear growth model illustrates the formulas clearly. Consider an outcome y_{it} for individual i observed at time point t,

$$y_{it} = \eta_{0i} + \eta_{1i}\, x_t + \epsilon_{it} \qquad (4.1)$$

where η_{0i} is the intercept factor, η_{1i} is the growth rate factor, x_t represents a time-related score such as age, and ϵ_{it} is the time-specific residual. A given individual, i, has the values η_{0i} and η_{1i} on the two growth factors. The growth process for this individual develops over time as x_t changes as $\eta_{0i} + \eta_{1i}\, x_t$. This is individual i's trajectory, describing the systematic part of the variation of the outcome at different time points. The individual's outcome at a certain time point t, y_{it}, is equal to the sum of the systematic part of the variation plus the time-specific residual ϵ_{it}.

Equation 4.1 is often referred to as the Level 1 equation, describing the repeated measures over time. The Level 2 equation describes the variation in the η_{0i} and η_{1i} factors as a function of covariates x,

$$\eta_{0i} = \alpha_0 + \sum_r \beta_{0r}\, x_{ri} + \zeta_{0i}, \qquad (4.2)$$

$$\eta_{1i} = \alpha_1 + \sum_r \beta_{1r}\, x_{ri} + \zeta_{1i}, \qquad (4.3)$$

where α coefficients are intercept parameters, β coefficients are regression weights for the covariates, and the ζ's represent residuals.

Centering relates to the choice of x_t scores in Equation 4.1, so it is important to consider how the x scores are chosen for different t values. For simplicity, say that t represents wave of measurement. With equidistant measurements, t would assume values 1, 2, ... , T, but nonequidistant time points are also possible. Or, t could represent year of measurement, such as 1998, 1999, 2000, 2001. Examples of x_t are age at time point t and school grade at time point t. To a certain degree, however, x_t values can be chosen by the researcher. For example, if age assumes the values 25, 27, 30, the same analysis would be obtained with age rescaled as 0, 1, 1.5. The key issue is that the distance between any two x_t values is comparable in the two scoring schemes. Centering is determined by choosing the time point t for which $x_t = 0$. Next, two alternative centering points are shown, defining the interpretation of the intercept factor as initial status and final status, respectively.

Initial Status Centering. Centering at the first time point ($t = 1$) implies that we set $x_1 = 0$ so that

$$y_{i1} = \eta_{0i} + \eta_{1i}\, 0 + \epsilon_{i1}, \qquad (4.4)$$

which means that growth rate factor η_{1i} is not involved in the outcome y_{i1}, but the outcome is a function of the intercept factor η_{0i} and the time-specific residual ϵ_{i1}. The intercept factor η_{0i} is therefore the systematic part of the outcome y_{i1}, the part determined by the growth process. Because of this, η_{0i} is interpreted as the initial status growth factor, that is, the level of the growth process at the first time point, $t = 1$.

Final Status Centering. Centering at the last time point ($t = T$) implies that we set $x_T = 0$ so that $y_{iT} = \eta_{0i} + \epsilon_{iT}$. Here, the growth factor η_{0i} represents the level of the growth process at the last time point.

Any other centering point in between $t = 0$ and $t = T$ may be chosen. In this way, varying centering expresses the level of the growth process in terms of the η_{0i} factor at any time point t.

It may be noted that analysis of the regression coefficients for the intercept factor defined at different time points is not the same as doing regular regression analysis for the outcome y at these different time points. A set of regression analyses for each time point would give information about how the influence of covariates change over time. Growth modeling, however, considers the intercept factor as the more fundamental dependent variable given that it represents the systematic part of the developmental trajectory. There are two statistical reasons why it is advantageous to use the intercept factor as the dependent variable. First, the intercept factor avoids the time-specific influence of the residual contained in y, which results in a more powerful analysis. Second, in contrast to regular regression, the intercept factor draws on information for y from all time points, which also results in a more powerful analysis. Taken together, this means that a set of regression analyses might overlook important changes in covariate influence over time.

Centering With Quadratic Growth

In growth modeling with a quadratic growth function, three growth factors are involved: an intercept factor, a linear growth rate factor, and a quadratic growth rate factor. The quadratic growth model is written as

$$y_{it} = \eta_{0i} + x_t\eta_{1i} + x^2_t\eta_{2i} + \epsilon_{it}, \qquad (4.5)$$

where η_{0i} is the intercept factor, η_{1i} is the linear rate factor, and η_{2i} is the quadratic rate factor. Consider centering at $t = 1$, using $x_1 = 0$. This means that the intercept factor η_{0i} is interpreted as an initial status factor. As in the linear case, choosing $x_t = 0$ defines the centering point at time point t. Even though the growth function is more complex in the quadratic case, the same simple interpretation as in the linear case is obtained for the intercept factor. With varying centering point corresponding to different time points, the factor is interpreted as the level of the growth process at that time point.

Application of Varying Centering to Heavy Drinking

Given the shape of the mean curve for heavy drinking in the NLSY shown in Fig. 4.1, a quadratic growth model with three random coefficients seems reasonable. Letting the subscript i denote individuals ($i = 1, 2, \ldots, 7{,}933$) and the subscript t time points ($t = 1, 2, \ldots, 20$), the model therefore expresses the development of the heavy drinking outcome variable over time as a quadratic function of three random coefficients η,

$$y_{it} = \eta_{0i} + \lambda_t\eta_{1i} + \lambda^2_t\eta_{2i} + \epsilon_{it}, \qquad (4.6)$$

where η_{0i} is an intercept term, η_{1i} a coefficient for the linear term, and η_{2i} a coefficient for the quadratic term. In Equation 4.6, λ_t replaces x_t to show that these time steps are parameters that can be estimated. The centering point is defined as the time point t where the parameter λ_t is fixed to zero.

Letting the λ_t time steps be estimated gives a more flexible functional form for the growth model. This proved to be very important for the NLSY data where the heavy drinking sample means follow a curve that increases more rapidly than it decreases. Such a nonsymmetric form is not always well-fitted by a conventional quadratic function with fixed λ_t time steps. Corresponding to a centering at age 18, the first three λ_t values are fixed at 0, 1, and 2 for identification purposes, and the next 17 steps are estimated in the analysis.

The variation across individuals in the three growth factors is described by a set of covariates x, adding the model parts where α

$$\eta_{0i} = \alpha_0 + \sum_r \beta_{0r} x_{ri} + \zeta_{0i}, \quad (4.7)$$

$$\eta_{1i} = \alpha_1 + \sum_r \beta_{1r} x_{ri} + \zeta_{1i}, \quad (4.8)$$

$$\eta_{2i} = \alpha_2 + \sum_r \beta_{2r} x_{ri} + \zeta_{2i}, \quad (4.9)$$

and β are regression parameters and ζ are residuals. The model is estimated by maximum likelihood using the Mplus software (Muthén & Muthén, 1998).

Varying the Centering Over Ages 18 to 37: Estimated Effects of Covariates on Heavy Drinking. Figure 4.2 shows the results of varying centering points for the estimated growth model for the frequency of heavy drinking. The figure varies the centering point so that the random intercept is defined as the level of frequency of heavy drinking at ages 18, 21, 25, 27, 32, and 37. At each centering age, the figure shows the partial regression coefficients for the intercept factor regressed on the covariates, denoted β in Equations 4.7 through 4.9. Note that all covariates are 0/1 dummy variables. As usual in regression analysis, the partial regression coefficient for a 0/1 dummy variable represents the effect of going from status 0 to status 1, holding other background constant. The broken horizontal line represents a zero coefficient. Positive coefficients correspond to a higher frequency of heavy drinking and negative coefficients correspond to a lower frequency of heavy drinking. Covariates corresponding to risk factors are expected to have positive coefficients, and covariates corresponding to protective factors are expected to have negative coefficients.

Consider as an example centering at age 18. Here, the largest coefficient is for Onset with a value close to 0.9, saying that individuals who started drinking regularly before age 15 have a higher frequency of heavy drinking at age 18 than those who did not, holding other covariates constant. The legend shows that this coefficient is significantly different from zero (marked s as in significant). The second largest coefficient is for Black with a value close to -0.6, saying that Blacks have a lower frequency of heavy drinking than non-Blacks at age 18, holding other covariates constant. These findings are in line with the alcohol literature. The new findings of Muthén and Muthén (in press) are based on varying

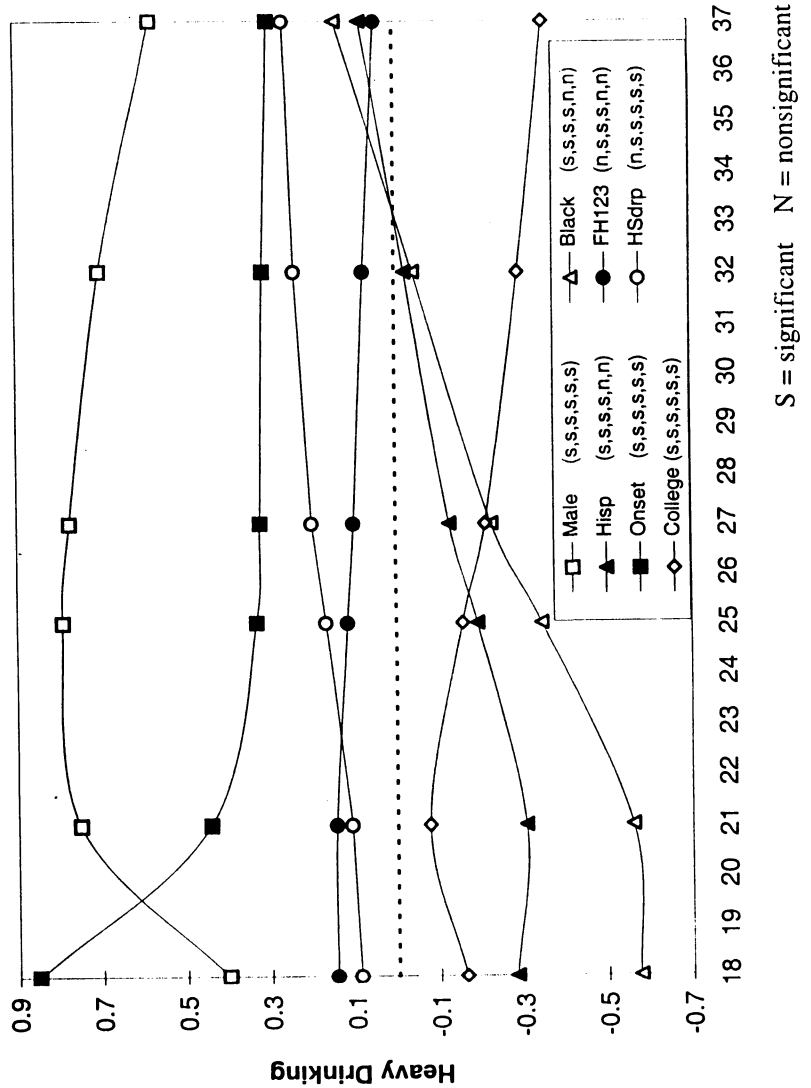

FIG. 4.2. Effects of background variables at different ages: Heavy drinking.

the centering point from 18 to 37. Following are some of the key findings for heavy drinking.

The heavy drinking coefficients in Figure 4.2 show that the largest partial regression coefficient at any age occurs for the early onset variable at age 18. This effect, however, diminishes over age. Figure 4.2 also shows that although Blacks and Hispanics have lower values than Whites at ages 18 and 21, their heavy drinking is not different from Whites at the later ages, 32 and 37. Furthermore, it is interesting to note that there is an increasing protective effect of college education.

Centering and Multiple Outcomes: Joint Analysis of Heavy Drinking and Severity Score

The severity score development in the NLSY is used to illustrate varying centering with a linear growth model. Before these results are presented, however, it is useful to discuss more generally some of the analysis considerations involved when studying two outcome variables. Furthermore, special considerations are needed for modeling the severity score given that, for this variable, only two repeated measures are available per individual.

Figure 4.3 shows the mean values for both of the outcome variables, heavy drinking and the severity score for alcohol-related problems. It is seen that the shape of the mean curve is rather similar for the ages where both are observed, 25 to 37. It is of interest to also study the individual variation in the severity trajectories and compare the relative importance of the time-invariant covariates for severity with that of heavy drinking. Using varying centering points, the comparison of relative importance of covariates can be made over age.

The severity score is based on items asked only in 1989 and 1994. An advantage of the multiple-cohort design is that this still results in coverage of a wide age range, 25 to 37 (this will be discussed later). The problem for growth modeling is, however, that for all individuals in the sample, the score is only available at two time points. With two time points, a general random coefficient growth model cannot be identified even when restricted to a linear model. This is clearly seen when noting that the growth model describes the correlations among the outcomes in terms of the variance and covariance parameters for the growth factors. With a linear model, there are three such parameters but with two measurement occasions per cohort there is only one observed

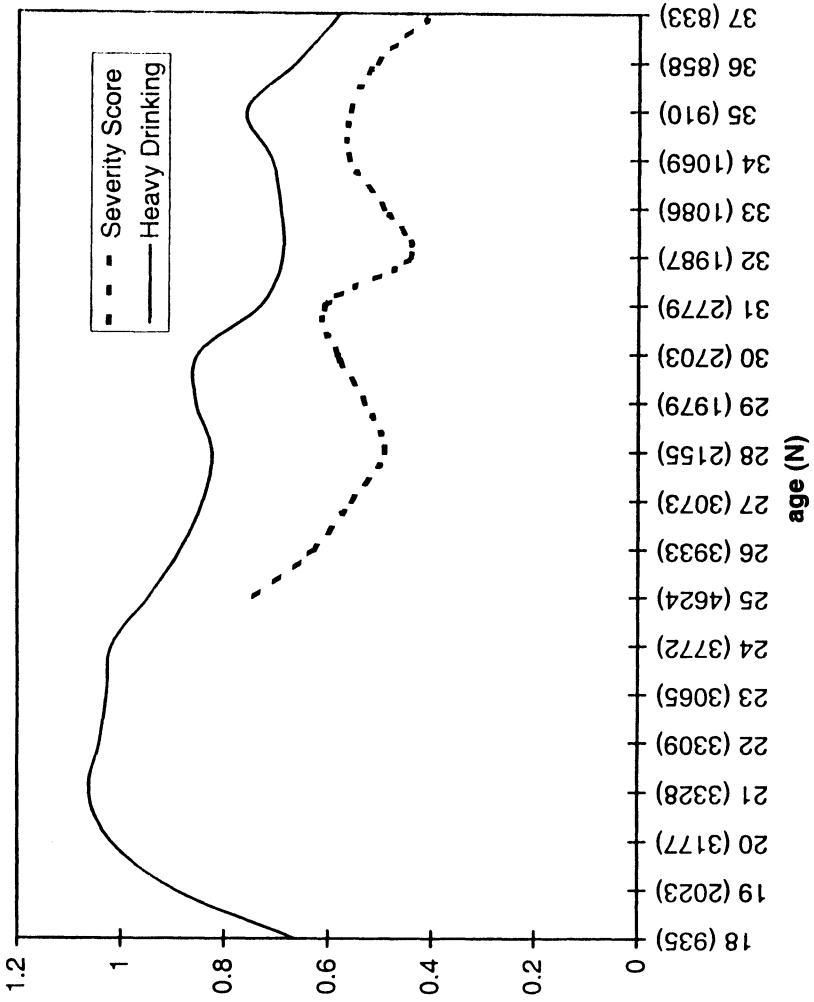

FIG 4.3. Sample means for heavy drinking and severity score in the NLSY.

covariance. This means that none of the cohorts supports the identification of the growth model. There are two possible solutions to this problem. One is to treat the two outcome variables as indicators of a single alcohol problem construct and the other is to specify a simplified linear model for the severity score. These two solutions will be discussed in turn.

Growth Modeling With Multiple Indicators. The severity score and the heavy drinking variable can be viewed as two indicators of the same unobserved construct, say alcohol problems. They reflect different aspects of this construct and it may be assumed that a factor analytic measurement model holds where the two indicators have different loadings and specific (error) variances. This formulation makes it possible to use the model in (Equation 4.1) with y replaced by this construct. The severity score indicator is only available in 1989 and 1994. The growth model can however be formulated for the full age range of 18 to 37, which means that an extrapolation of the severity score growth curve to younger ages than 25 is achieved. This multiple indicator approach does, however, carry with it certain assumptions. The most important assumption given our interest in the influence of covariates is that the model postulates that the regression coefficients for the two indicators on the covariates have the same proportionality factor across all covariates for all time points. This assumption would be violated if different covariates have different influence for the two indicators. Given that one key goal of the study is determining to which extent this holds, presupposing this in the modeling is not desirable.

Linear Growth Modeling With Two Time Points. A more straightforward approach is to specify a simplified linear model that is restricted to have a random coefficient only for the intercept and letting the slope be fixed, that is, not varying across individuals. This growth model is identified using only two time points. This fact was used in the modeling of Muthén and Muthén (in press). There are three reasons that the assumption of a linear growth model with fixed slope is plausible for the severity score in the 25 to 37 age range. First, this age range is past the normative peak age and shows roughly a linear decline. Second, estimates for heavy drinking show that most of the individual variation can be captured by variation in the intercept defined at 25. Third, the slope can still be allowed to vary as a function of the covariates so that

its mean is different for different subgroups. The availability of only two time points does, however, limit the study of other growth forms than the linear. For example, it might have been interesting to study the significance of the bumps seen in Fig. 4.3.

The simplified linear model for the severity score can either be analyzed separately or together with the quadratic growth model for heavy drinking. When estimated together with heavy drinking, it is possible to study contemporaneous correlations between the severity score and heavy drinking. This is of interest to determine to which extent a high severity score occurs with high frequency of heavy drinking. Other types of joint growth modeling of several processes simultaneously were discussed in Muthén (1997) including comparisons with auto-regressive modeling with cross-lagged effects.

Varying the Centering Over Ages 18 to 37: Estimated Effects of Covariates on Severity Score. The results for varying centering points for the severity score are now presented. They are based on the growth model obtained by analyzing the two processes jointly using the simplified linear model. The results are given in Fig. 4.4, where the centering point is varied so that the random intercept is defined at ages 25, 27, 32, and 37. The figure shows the partial regression coefficients for the intercept on each covariate, denoted as β in Equations 4.7 to 4.9.

The severity score coefficients in Fig. 4.4 show an interesting trend of increasing importance of dropping out of high school. At age 37, the strongest partial regression coefficient is the lasting effect of having dropped out of high school. Other interesting trends are the diminishing effects of early onset and family history of alcoholism. The plot of the coefficients for varying centering points is clearly useful in that it indicates trends over age, thereby raising questions concerning future development.

MULTIPLE TRAJECTORY CLASSES

We now focus on modeling of heterogeneity across individuals in their development over time. Individuals may belong to different, unknown subpopulations corresponding to normative development and various forms of non-normative development. The applicability of new methods for studying such heterogeneity is summarized. The discussion indicates

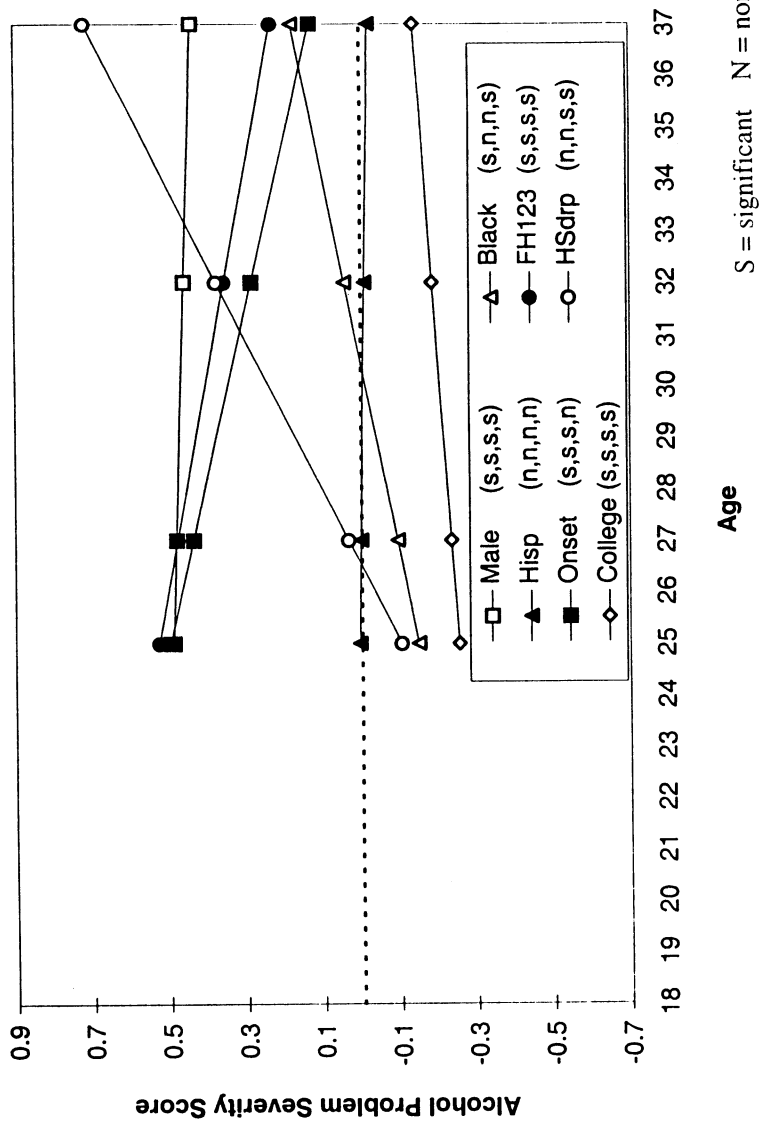

FIG 4.4. Effects of background variables at different ages: Alcohol problem severity score.

that growth mixture modeling is an exciting addition to the developmental analysis tools.

Figure 4.5 shows four different growth trajectories that can be derived from the estimated quadratic growth model for heavy drinking in the NLSY. They are obtained from different individual values on the growth factors for the intercept, linear, and quadratic terms. Given the large differences in the trajectories, one may ask if the individuals following these different trajectories really are from the same population or are better characterized as coming from different subpopulations. The different subpopulations may have different antecedents and different consequences. If this heterogeneity is ignored, incorrect conclusions may be drawn and useful information in the data overlooked.

Alcohol use development in terms of classes of trajectories has been discussed recently in the alcohol literature by Schulenberg, O'Malley, Bachman, Wadsworth, and Johnston (1996) and Bates and Labouvie (1997). Drawing on the Zucker (1994) distinction between different pathways to alcohol problems and alcoholism, one may hypothesize that trajectory Class 4 corresponds to "antisocial alcoholism", and that

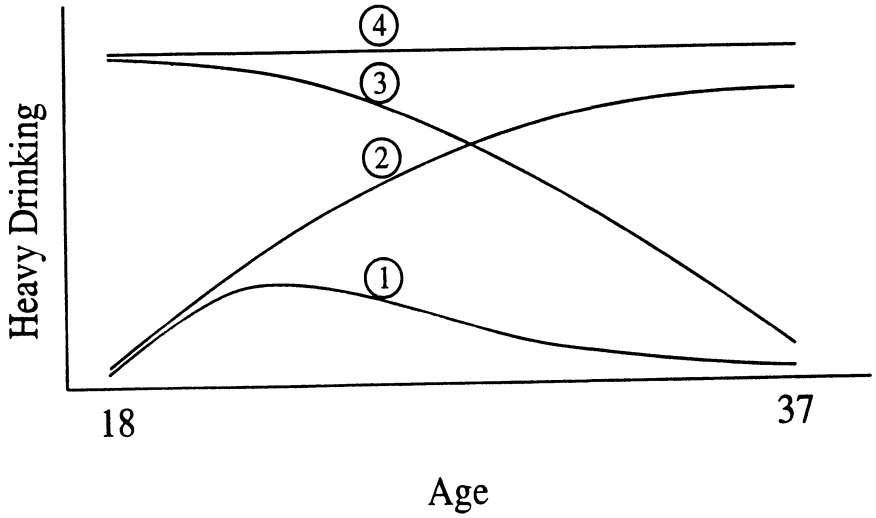

FIG 4.5. Four hypothetical trajectory classes of alcohol development.

trajectory Class 3 corresponds to "developmentally limited alcoholism." Trajectory Class 1 is the normative class of typical development. Class 4 may be more common in males and associated with family history of alcoholism, early onset, and antisocial behavior. Class 4 represents an early start of heavy drinking, which in contrast to Class 3 does not subside in early adulthood. Whereas Class 3 shows a transition from excessive use to more moderate use, Class 2 shows a transition from normative use to problematic use at the typical peak age. An important task of alcohol research is to find predictors of these transitions. The classes also differ in that Class 4 and Class 2 may be associated with alcohol-related problem consequences in young adulthood, whereas Class 1 and Class 3 are not.

Growth Mixture Modeling

The idea of multiple unobserved subpopulations or latent classes can be incorporated into growth modeling using the statistical feature of finite mixture modeling. This is referred to as growth mixture modeling. Muthén and Shedden (in press) used growth mixture modeling to formulate a latent trajectory class model for heavy drinking in the youngest NLSY cohort for ages 18 to 25. Three latent trajectory classes were obtained. The three classes are those of Fig. 4.5 except that Class 3 and Class 4 of Fig. 4.5 are combined.

There are two interesting aspects of growth mixture modeling with latent trajectory classes, studying growth process antecedents and studying growth process consequences.

Growth Mixture Modeling of Antecedents. Growth model covariates play the role of antecedents to the growth process in that they can be used to predict the growth factor values for an individual. With growth mixture modeling, the analysis of antecedents is strengthened by allowing different means and different covariate influence across the different latent trajectory classes.

The quadratic growth model used for NLSY heavy drinking analysis is

$$y_{it} = \eta_{0i} + \lambda_t \eta_{1i} + \lambda^2_t \eta_{2i} + \epsilon_{it}, \tag{4.10}$$

$$\eta_{0i} = \alpha_0 + \sum_r \beta_{0r} x_{ri} + \zeta_{0i}, \tag{4.11}$$

$$\eta_{1i} = \alpha_1 + \sum_r \beta_{1r} x_{ri} + \zeta_{1i}, \tag{4.12}$$

$$\eta_{2i} = \alpha_2 + \sum_r \beta_{2r} x_{ri} + \zeta_{2i}, \tag{4.13}$$

Growth mixture modeling allows the α and β coefficients in these equations to vary across the latent trajectory classes. Consider first the means of the three growth factors, where for simplicity covariates are ignored here so that the means are represented by the α coefficients. The trajectory Classes 2 and 3 in Fig. 4.5 may be taken as an example. The mean of the intercept factor at age 18 (α_0) is lower for Class 2 than for Class 3. The mean of the linear growth rate factor (α_1) is positive for Class 2 and negative for Class 3. The quadratic growth rate factor (α_2) is slightly negative for Class 2 reflecting a downturn at higher ages, whereas it is slightly positive for Class 3. Class-varying β coefficients are also likely across the four classes in Fig. 4.5. Normative development may for example be influenced by different covariate coefficients than non-normative development.

Once trajectory class membership has been found to influence the growth process, the interest turns to finding predictors of trajectory class membership. These might be sought among the growth model covariates but there may also be additional variables. Prediction of class membership using the growth model covariates was studied for the three-class model of Muthén and Shedden (in press). This showed that compared to the normative class, the likelihood of membership in the non-normative class of escalating use was increased by being male, being a high school dropout and having positive family history. The likelihood of membership in the non-normative class of high use already at age 18 was increased by early onset, being male, not being Black, and having positive family history. As discussed in Muthén and Shedden (in press), this analysis can be taken one step further by the inclusion of early latent class indicators, for example, adolescent measures such as early onset and antisocial behavior, in order to provide an early warning system for problematic alcohol development.

Growth Mixture Modeling of Consequences. Conventional growth modeling considers the growth curve as an outcome predicted by the covariates. Growth curves need not, however, be used only as

outcomes but can also be used as predictors. The growth mixture modeling used in Muthén and Shedden (in press) also served a predictive purpose. Muthén and Shedden (in press) were interested in using the heavy drinking growth model to predict alcohol dependence 5 years later. Alcohol dependence was formulated as a binary variable based on the symptom items discussed earlier forming DSM-IV diagnostic criteria to give dependence diagnoses. They argued that it was difficult to use the conventional growth model for this prediction because the growth factors have a nonlinear interplay in their determination of curve shape. For example, curve Classes 4 and 2 in Fig. 4.5 may be hypothesized to be associated with a higher probability of later alcohol dependence than curve Class 1. Using the intercept factor and the linear factor value for this prediction, however, does not fit this hypothesis. An increasing intercept factor value is not always predictive of dependence under this assumption given that a low intercept value is associated with both curve Class 1 and curve Class 2. Likewise, although curve Class 2 has a high linear slope factor value, curve Class 4 has a zero value. This shows that it is hard to use curve shape for prediction in the conventional growth model. The growth mixture model solves this dilemma in a simple way.

The growth mixture model allows for different curve classes such as those in Fig. 4.5 by letting different latent classes of individuals have different growth factor means. In this way, a categorical latent variable, the latent class variable, is introduced in addition to the continuous latent variables corresponding to the growth factors. The categories (classes) of the categorical latent variable correspond directly to the curve shapes and the model lets the dependence probability vary over these categories. The estimated three-class model in Muthén and Shedden (in press) showed that the normative Class 1 had a dependence probability of 0.08. The two non-normative classes, however, had probabilities of 0.23 and 0.39. The corresponding odds ratios for dependence in each of the two non-normative classes relative to the normative class are 3.43 and 7.35. This clearly shows the elevated risk of developing alcohol dependence for individuals who are in the non-normative classes. It is therefore of interest to study the possibility of predicting class membership as early as possible.

MULTIPLE-COHORT ANALYSIS

This section discusses methodological issues related to the use of multiple cohorts in longitudinal studies. The data structure of multiple-cohort analysis is similar to that of missing data and some general missing data issues are also discussed.

Figure 4.6 shows the multiple-cohort design of the NLSY for the 18 to 37 age range when alcohol measurements were included in the survey. Nonshaded areas correspond to years when alcohol measures were obtained. In our analyses, 1985 is excluded given that a question format change was made that may have made across-time comparisons invalid. Figure 4.6 shows that different ages are covered by different cohorts and are represented by different number of observations. For example, age 18 is only represented by the approximately 1,000 individuals in cohort 64, yet age 25 is represented by several cohorts. For each cohort only six ages/occasions are represented (excluding the 1985 measure). The design is sometimes referred to as an *accelerated cohort design* given that there are only 12 years between the first measurement year of 1982 and the last measurement year of 1994, whereas 20 years of development from 18 to 37 is covered.

Multiple Cohorts Viewed as Missing Data

Figure 4.6 may be viewed in terms of missing data. The figure may be seen as a data matrix with rows corresponding to (groups of) people and columns corresponding to variables. For example, the 20 columns may be seen as corresponding to the heavy drinking outcome variable measured at ages 18 to 37. In this way, the shaded areas in the matrix represent missing data. Data are missing by design, which means that if the cohorts can be assumed to be drawn from the same population, we can assume that data are missing completely at random (MCAR) to use a term introduced in Little and Rubin (1987). Simply put, MCAR means that the probabilities of the values being missing are not related to the values that would have been observed. Here, the missing data cannot be handled by the common approach of listwise deletion because nobody in the sample has complete data.

The missing data structure in Fig. 4. 6 can be handled analytically in several ways using a latent variable framework. Using maximum-likelihood estimation with continuous-normal outcome variables, a multiple-group mean and covariance structure analysis of the eight cohorts may be carried out as in Muthén, Kaplan, and Hollis (1987). In this analysis, the groups are assumed to come from one and the same population so that across-group equality constraints are imposed on all parameters

Cohort	Age 18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
57								82	83	84	85	86	87	88	89	90	91	92	93	94
58							82	83	84	85	86	87	88	89	90	91	92	93	94	
59						82	83	84	85	86	87	88	89	90	91	92	93	94		
60					82	83	84	85	86	87	88	89	90	91	92	93	94			
61				82	83	84	85	86	87	88	89	90	91	92	93	94				
62			82	83	84	85	86	87	88	89	90	91	92	93	94					
63		82	83	84	85	86	87	88	89	90	91	92	93	94						
64	82	83	84	85	86	87	88	89	90	91	92	93	94							

FIG. 4.6. NLSY alcohol multiple-cohort data structure.

that are in common. The analysis needs to fill in dummy elements in the sample means, variances and covariances corresponding to variables that are missing for the group (cohort). Newer latent variable software can handle the missing data feature directly without such an artificial setup by allowing for a single-group, single-population analysis where eight missing data patterns are present, one corresponding to each cohort.

Hidden Cohort Design

It should be noted that the missing data structure represented by Fig. 4.6 often arises in circumstances where it is not planned, which implies hidden modeling opportunities. A typical example is where, at a given wave of measurement, all individuals cannot be followed up at the same point in time given practical constraints on interview load and so on. Often these differences in timing can be ignored but in many cases they are important and should be taken into account in the growth modeling. For example, in terms of reading development, early and late grade-one measures do not refer to the same developmental age and college drinking experiences may not be same at the beginning and end of the freshman year. As yet another example, consider a study where drinking is to be examined across a 10-year period assuming a starting age that ranges from 20 to 40 in the sample. For individuals starting at age 20, the 10-year development is in a different developmental phase than for individuals starting at age 40. Ignoring these differences in developmental phases and analyzing the data with time point (1, 2, ..., 10), not age (20, 21, ..., 50), as the time axis would amount to analysis of heterogeneous subpopulations and would give misleading results.

This data situation may be referred to as a *hidden cohort* design. There is a direct analogy between these examples and the missing data structure of Fig. 4.6. For example, to make a distinction between early and late in the year, one may consider the age points as fractions of a year, or use the month of the year, which results in no individual being observed more than once in the year. Letting age define the time axis has the useful side effect of "spreading out the time axis" so that more time points are available for growth modeling. This is highly desirable in order to capture more features in the data and avoid restrictive modeling assumptions. Frequently, it is quite reasonable to assume that MCAR is at hand, given that the time point at which an individual is interviewed is likely to be uncorrelated with his/her observed values.

Attrition

In passing, it should be pointed out that in addition to missingness by design, longitudinal data typically show attrition over time with loss of individuals to follow up. It is often a reasonable approximation that such missingness satisfies the assumption of "missing at random" (MAR) in Little and Rubin (1987) terms. In contrast to MCAR, MAR makes the more realistic assumption that the probabilities of values being missing can be related to variables that are not missing, for instance covariates and variables observed at the first time point of the study. When MAR holds, correct maximum-likelihood estimates are obtained by using all available data, including observations on individuals who have missing data on some of the analysis variables.

Cohort Differences

With multiple cohorts it is important to test for across-cohort parameter invariance. Societal changes may have taken place that influence drinking practices in the 7 years between the births of the youngest and oldest cohort in the NLSY. At the one extreme, one may assume that all cohorts come from the same population as just discussed. At the other extreme, one may assume that each cohort corresponds to its own population. Model tests of fit can be compared for models corresponding to these two extremes. If there is a significant difference in model fit, it is of interest to consider the eight-population analysis with full across-population invariance and study model modification indices that point to invariance assumptions that do not match the data well. In such an eight-population analysis of the Fig. 4.6 data on heavy drinking, the growth modeling of Muthén and Muthén (in press) found that full invariance could not be rejected in important ways except for the mean of one of the time-invariant covariates, early onset. The mean of this variable increased steadily from cohort 57 to cohort 64.

INTRACLASS CORRELATION DUE TO MULTIPLE SIBLINGS PER HOUSEHOLD

This section discusses methodological aspects of analyzing data obtained by cluster sampling. Cluster sampling is a common feature in national samples such as the NLSY. In the NLSY, clustering occurs because for each household all siblings in a given age range are interviewed. Clustered data causes nonindependence among the observations that needs to be taken into account. The methods in this area range from those focusing on nonindependence corrections of standard errors and chi-square tests of model fit to methods that express the nonindependence explicitly in the models.

The left part of Table 4.1 shows the number of interviewed siblings per household in the NLSY sample. It is seen that for almost a quarter of the households more than one sibling was interviewed. The average number of siblings per household is 1.4. The right part of Table 4.1 shows the intraclass correlation, that is, the correlation between siblings within households, for the heavy drinking variable. The intraclass correlation is measured as the ratio of the between-household variance divided by the sum of the within- and between-household variance. If siblings within a household have similar heavy drinking, the within variance is small and the intraclass correlation is large. Table 4.1 shows that the intraclass correlation ranges from 0.19 in 1982 to 0.06 in 1989. It is reasonable that the value decreases by year as the siblings grow older, move out of their original households, and lead more independent lives.

Design Effects

A positive value of the intraclass correlation indicates that the conventional analysis assumption of independence is violated. The seriousness of this violation, however, may not be large. Intraclass correlations are

TABLE 4.1

NLSY Household Clusters

Household Type (Number of respondents)	*Number of Households**	*Intraclass Year*	*Correlations for Siblings Heavy Drinking*
Single	5,944	1982	0.19
Two	1,985	1983	0.18
Three	634	1984	0.12
Four	170	1985	0.09
Five	32	1988	0.04
Six	5	1989	0.06

Total number of households: 8,770

Total number of respondents: 12,686

Average number of respondents per household: 1.4

*Source: NLS User's Guide, 1994, p. 247

discussed in the context of cluster sampling where the ratio of the variance of an estimator computed under cluster sampling (correlated observations) to that computed under simple random sampling (independent observations) is described as the design effect. Cluster sampling gives a larger variance than when using simple random sampling formulas. The degree to which the design effect exceeds 1 reflects the error in the variance assessment when ignoring the non-independence of observations obtained under cluster sampling. In other words, ignoring the intraclass correlation results in standard errors of estimates that are too small by a factor corresponding to the square root of the design effect. For an estimator of the mean and clusters of equal size, the design effect is $1 + \rho(c - 1)$, where ρ is the intraclass correlation and c is the common cluster size. This shows that it is the product of the intraclass correlation and the cluster size that is the important factor.

The NLSY data does not have equal household sizes and the estimates in growth modeling concern more complex parameters than means, but as Muthén and Satorra (1995) pointed out in their simulations' the product formula just shown may give a rough indication of the potential problem at hand. From Table 4.1, one finds that the product is maximally 0.19 (1.4 - 1) = 0.076, which assuming equal household sizes would result in a design effect of 1.076, a very small value that indicates that the lack of independence can probably be ignored.

Latent Variable Analysis With Cluster Samples

Muthén and Satorra (1995) discussed ways to compute correct standard errors and chi-square tests of model fit when the design effects are not ignorable. Even when they are ignorable, however, interesting modeling opportunities exist for describing the correlations among siblings within households as long as these correlations are not zero. As the discussion in Muthén and Satorra (1995) pointed out, sampling statisticians sometimes make a distinction between aggregated and disaggregated analysis.

Aggregated analysis amounts to a conventional analysis that ignores the lack of independence when computing parameter estimates, but uses special formulas to compute standard errors and chi-square tests of fit.

Disaggregated analysis instead deals with the lack of independence directly by modeling it. Multilevel modeling is an example of disaggregated analysis, where in this case one would consider a three-level model for repeated measures over time for each sibling for each household. Here, independence of observations is only assumed for clusters (households), not individuals. Three-level modeling in a latent variable framework was discussed in Muthén (1997). Muthén and Khoo (1998), see also Khoo in this volume, discuss disaggregated analysis alternatives for household (family) data.

CONCLUSIONS

This chapter discussed four analysis complications arising in growth modeling of a national longitudinal sample and analysis approaches were presented that handled these complications. More importantly, each of the four complications can be seen as representing an opportunity for understanding deeper features in the data that are not usually uncovered by conventional techniques. The discussion pointed to solutions that use modeling with a richer set of parameters in order to answer new research questions.

All techniques proposed here are available in the new latent variable software package Mplus (Muthén & Muthén, 1998b). Hopefully, this chapter will stimulate the use of these techniques for new, interesting longitudinal studies in the alcohol area and developmental research in general.

ACKNOWLEDGMENTS

This research was supported by Grant 1 K02 AA 00230-01 from NIAAA, and by Grant 1 R21 AA10948-01A1 from NIAAA. I thank Tom Harford for suggestions regarding the NLSY application.

REFERENCES

Bates, M. E., & Labouvie, E. W. (1997). Adolescent risk factors and the prediction of persistent alcohol and drug use into adulthood. *Alcoholism Clinical and Experimental Research, 21*, 944–950.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Muthén, B.(1997). Latent variable modeling with longitudinal and multilevel data. In A. Raftery (Ed.), *Sociological methodology 1997* (pp. 453–480). Boston: Blackwell.

Muthén, B., & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*, 371–402.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 42*, 431–462.

Muthén, B., & Khoo, S. T. (1998). Longitudinal studies of achievement growth using latent variable modeling. *Learning and individual differences, Special issue: Latent growth curve analysis, 10*, 73–101.

Muthén, B., & Khoo, S. T. (1998). *Growth modeling of family data.* In preparation.

Muthén, B., & Muthén, L. (In press). The development of heavy drinking from age 18 to 37 in a U.S. national sample. Forthcoming in *Journal of Studies on Alcohol.*

Muthén, L., & Muthén, B. (1998b). *Mplus user's guide.* Los Angeles: Muthén & Muthén.

Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267–316). Washington, DC: American Sociological Association.

Muthén, B.,& Shedden, K. (in press). Finite mixture modeling with mixture outcomes using the EM algorithm. Forthcoming in *Biometrics.*

Schulenberg, J., O'Malley, P. M., Bachman, J. G., Wadsworth, K. N., & Johnston, L. D. (1996). Getting drunk and growing up: Trajectories of frequent binge drinking during the transition to young adulthood. *Journal of Studies on Alcohol, 57*, 289–304.

Zucker, R. A. (1994). Pathways to alcohol problems and alcoholism: A developmental account of the evidence for multiple alcoholisms and for contextual contributions to risk. In R. A. Zucker, G. M. Boyd, & J. Howards (Eds.), *The development of alcohol problems: Exploring the biopsychosocial matrix of risk* (pp. 225–289). NIAAA Research Monograph 26.