Applied Longitudinal Data Analysis

Modeling Change and Event Occurrence

Judith D. Singer
John B. Willett

OXFORD
UNIVERSITY PRESS
2003

1

A Framework for Investigating Change over Time

Change is inevitable. Change is constant.
—Benjamin Disraeli

Change is pervasive in everyday life. Infants crawl and walk, children learn to read and write, the elderly become frail and forgetful. Beyond these natural changes, targeted interventions can also cause change: cholesterol levels may decline with new medication; test scores might rise after coaching. By measuring and charting changes like these—both naturalistic and experimentally induced—we uncover the temporal nature of development.

The investigation of change has fascinated empirical researchers for generations. Yet it is only since the 1980s, when methodologists developed a class of appropriate statistical models—known variously as *individual growth models, random coefficient models, multilevel models, mixed models,* and *hierarchical linear models*—that researchers have been able to study change well. Until then, the technical literature on the measurement of change was awash with broken promises, erroneous half-truths, and name-calling. The 1960s and 1970s were especially rancorous, with most methodologists offering little hope, insisting that researchers should not even attempt to measure change because it could not be done well (Bereiter, 1963; Linn & Slinde, 1977). For instance, in their paper, "How should we measure change? Or should we?," Cronbach and Furby (1970) tried to end the debate forever, advising researchers interested in the study of change to "frame their questions in other ways."

Today we know that it is possible to measure change, and to do it well, *if you have longitudinal data* (Rogosa, Brandt, & Zimowski, 1982; Willett, 1989). Cross-sectional data—so easy to collect and so widely available—will not suffice. In this chapter, we describe why longitudinal data are necessary for studying change. We begin, in section 1.1, by introducing three

longitudinal studies of change. In section 1.2, we distinguish between the two types of question these examples address, questions about: (1) *within-individual change*—How does *each person* change over time?—and (2) *interindividual differences in change*—What predicts differences among people in their changes? This distinction provides an appealing heuristic for framing research questions and underpins the statistical models we ultimately present. We conclude, in section 1.3, by identifying three requisite *methodological* features of any study of change: the availability of (1) multiple waves of data; (2) a substantively meaningful metric for time; and (3) an outcome that changes systematically.

1.1 When Might You Study Change over Time?

Many studies lend themselves to the measurement of change. The research design can be experimental or observational. Data can be collected prospectively or retrospectively. Time can be measured in a variety of units—months, years, semesters, sessions, and so on. The data collection schedule can be fixed (everyone has the same periodicity) or flexible (each person has a unique schedule). Because the phrases "growth models" and "growth curve analysis" have become synonymous with the measurement of change, many people assume that outcomes must "grow" or *increase* over time. Yet the statistical models that we will specify care little about the direction (or even the functional form) of change. They lend themselves equally well to outcomes that *decrease* over time (e.g., weight loss among dieters) or exhibit complex trajectories (including plateaus and reversals), as we illustrate in the following three examples.

1.1.1 Changes in Antisocial Behavior during Adolescence

Adolescence is a period of great experimentation when youngsters try out new identities and explore new behaviors. Although most teenagers remain psychologically healthy, some experience difficulty and manifest antisocial behaviors, including aggressive *externalizing behaviors* and depressive *internalizing behaviors*. For decades, psychologists have postulated a variety of theories about why some adolescents develop problems and others do not, but lacking appropriate statistical methods, these suppositions went untested. Recent advances in statistical methods have allowed empirical exploration of developmental trajectories and assessment of

Coie, Terry, Lenox, Lochman, and Hyman (1995) designed an ingenious study to investigate longitudinal patterns by capitalizing on data gathered routinely by the Durham, North Carolina, public schools. As part of a systemwide screening program, every third grader completes a battery of sociometric instruments designed to identify classmates who are overly aggressive (who start fights, hit children, or say mean things) or extremely rejected (who are liked by few peers and disliked by many). To investigate the link between these early assessments and later antisocial behavioral trajectories, the researchers tracked a random sample of 407 children, stratified by their third-grade peer ratings. When they were in sixth, eighth, and tenth grade, these children completed a battery of instruments, including the Child Assessment Schedule (CAS), a semi-structured interview that assesses levels of antisocial behavior. Combining data sets allowed the researchers to examine these children's patterns of change between sixth and tenth grade and the predictability of these patterns on the basis of the earlier peer ratings.

Because of well-known gender differences in antisocial behavior, the researchers conducted separate but parallel analyses by gender. For simplicity here, we focus on boys. Nonaggressive boys—regardless of their peer rejection ratings—consistently displayed few antisocial behaviors between sixth and tenth grades. For them, the researchers were unable to reject the null hypothesis of no systematic change over time. Aggressive nonrejected boys were indistinguishable from this group with respect to patterns of externalizing behavior, but their sixth-grade levels of internalizing behavior were temporarily elevated (declining linearly to the nonaggressive boys' level by tenth grade). Boys who were both aggressive *and* rejected in third grade followed a very different trajectory. Although they were indistinguishable from the nonaggressive boys in their sixth-grade levels of either outcome, over time they experienced significant linear increases in both. The researchers concluded that adolescent boys who will ultimately manifest increasing levels of antisocial behavior can be identified as early as third grade on the basis of peer aggression and rejection ratings.

1.1.2 Individual Differences in Reading Trajectories

Some children learn to read more rapidly than others. Yet despite decades of research, specialists still do not fully understand why. Educators and pediatricians offer two major competing theories for these interindividual differences: (1) the *lag* hypothesis, which assumes that every child can become a proficient reader—children differ only in the *rate* at which they acquire skills; and (2) the *deficit* hypothesis, which

assumes that some children will never read well because they lack a crucial skill. If the lag hypothesis were true, all children would eventually become proficient; we need only follow them for sufficient time to see their mastery. If the deficit hypothesis were true, some children would never become proficient no matter how long they were followed—they simply lack the skills to do so.

Francis, Shaywitz, Stuebing, Shaywitz, and Fletcher (1996) evaluated the evidence for and against these competing hypotheses by following 363 six-year-olds until age 16. Each year, children completed the Woodcock-Johnson Psycho-educational Test Battery, a well-established measure of reading ability; every other year, they also completed the Wechsler Intelligence Scale for Children (WISC). By comparing third-grade reading scores to expectations based upon concomitant WISC scores, the researchers identified three distinct groups of children: 301 "normal readers"; 28 "discrepant readers," whose reading scores were much different than their WISC scores would suggest; and 34 "low achievers," whose reading scores, while not discrepant from their WISC scores, were far below normal.

Drawing from a rich theoretical tradition that anticipates complex trajectories of development, the researchers examined the tenability of several alternative nonlinear growth models. Based upon a combination of graphical exploration and statistical testing, they selected a model in which reading ability increases nonlinearly over time, eventually reaching an asymptote—the maximum reading level the child could be expected to attain (if testing continued indefinitely). Examining the fitted trajectories, the researchers found that the two groups of disabled readers were indistinguishable statistically, but that both differed significantly from the normal readers in their eventual plateau. They estimated that the average child in the normal group would attain a reading level 30 points higher than that of the average child in either the discrepant or low-achieving group (a large difference given the standard deviation of 12). The researchers concluded that their data were more consistent with the deficit hypothesis—that some children will *never* attain mastery—than with the lag hypothesis.

1.1.3 Efficacy of Short-Term Anxiety-Provoking Psychotherapy

Many psychiatrists find that short-term anxiety-provoking psychotherapy (STAPP) can ameliorate psychological distress. A methodological strength of the associated literature is its consistent use of a well-developed instrument: the Symptom Check List (SCL-90), developed by

Derogatis (1994). A methodological weakness is its reliance on two-wave designs: one wave of data pretreatment and a second wave posttreatment. Researchers conclude that the treatment is effective when the decrease in SCL-90 scores among STAPP patients is lower than the decrease among individuals in a comparison group.

Svartberg, Seltzer, Stiles, and Khoo (1995) adopted a different approach to studying STAPP's efficacy. Instead of collecting just two waves of data, the researchers examined "the course, rate and correlates of symptom improvement as measured with the SCL-90 during and after STAPP" (p. 242). A sample of 15 patients received approximately 20 weekly STAPP sessions. During the study, each patient completed the SCL-90 up to seven times: once or twice at referral (before therapy began), once at mid-therapy, once at termination, and three times after therapy ended (after 6, 12, and 24 months). Suspecting that STAPP's effectiveness would vary with the patients' abilities to control their emotional and motivational impulses (known as *ego rigidity*), two independent psychiatrists reviewed the patients' intake files and assigned ego rigidity ratings.

Plotting each patient's SCL-90 data over time, the researchers identified two distinct temporal patterns, one during treatment and another after treatment. Between intake and treatment termination (an average of 8.5 months later), most patients experienced relatively steep linear declines in SCL-90 scores—an average decrease of 0.060 symptoms per month (from an initial mean of 0.93). During the two years after treatment, the rate of linear decline in symptoms was far lower—only 0.005 per month—although still distinguishable from 0. In addition to significant differences among individuals in their rates of decline before and after treatment termination, ego rigidity was associated with rates of symptom decline during therapy (but not after). The researchers concluded that: (1) STAPP can decrease symptoms of distress *during* therapy; (2) gains achieved during STAPP therapy *can* be maintained; but (3) major gains *after* STAPP therapy ends are rare.

1.2 Distinguishing Between Two Types of Questions about Change

From a substantive point of view, each of these studies poses a unique set of research questions about its own specific outcomes (antisocial behavior, reading levels, and SCL-90 scores) and its own specific predictors (peer ratings, disability group, and ego rigidity ratings). From a statistical point of view, however, each poses an identical pair of questions: (1)

How does the outcome change over time? and (2) Can we predict differences in these changes? From this perspective, Coie and colleagues (1995) are asking: (1) How does each adolescent's level of antisocial behavior change from sixth through tenth grade?; and (2) Can we predict differences in these changes according to third grade peer ratings? Similarly, Francis and colleagues (1996) are asking: (1) How does reading ability change between ages 6 and 16?; and (2) Can we predict differences in these changes according to the presence or absence of a reading disability?

These two kinds of question form the core of every study about change. The first question is descriptive and asks us to characterize each person's pattern of change over time. Is individual change linear? Nonlinear? Is it consistent over time or does it fluctuate? The second question is relational and asks us to examine the association between predictors and the patterns of change. Do different types of people experience different patterns of change? Which predictors are associated with which patterns? In subsequent chapters, we use these two questions to provide the conceptual foundation for our analysis of change, leading naturally to the specification of a pair of statistical models—one per question. To develop your intuition about the questions and how they map onto subsequent studies of change, here we simply emphasize their sequential and hierarchical nature.

In the first stage of an analysis of change, known as *level-1*, we ask about *within-individual change* over time. Here, we characterize the individual pattern of change so that we can describe each person's *individual growth trajectory*—the way his or her outcome values rise and fall over time. Does this child's reading skill grow rapidly, so that she begins to understand complex text by fourth or fifth grade? Does another child's reading skill start out lower and grow more slowly? The goal of a level-1 analysis is to describe the *shape* of each person's individual growth trajectory.

In the second stage of an analysis of change, known as *level-2*, we ask about *interindividual differences in change*. Here, we assess whether different people manifest different patterns of withinindividual change and ask what predicts these differences. We ask whether it is possible to predict, on the basis of third-grade peer ratings, which boys will remain psychologically healthy during adolescence and which will become increasingly antisocial? Can ego rigidity ratings predict which patients will respond most rapidly to psychotherapy? The goal of a level-2 analysis is to detect heterogeneity in change across individuals and to determine the *relationship* between predictors and the *shape* of each person's individual growth trajectory.

In subsequent chapters, we map these two research questions onto a pair of statistical models: (1) a level-1 model, describing within-individual change over time; and (2) a level-2 model, relating predictors to any interindividual differences in change. Ultimately, we consider these two models to be a "linked pair" and refer to them jointly as the *multilevel model for change*. But for now, we ask only that you learn to distinguish the two types of questions. Doing so helps clarify why research studies of change must possess certain methodological features, a topic to which we now turn.

1.3 Three Important Features of a Study of Change

Not every longitudinal study is amenable to the analysis of change. The studies introduced in section 1.1 share three methodological features that make them particularly well suited to this task. They each have:

- Three or more waves of data
- An outcome whose values change systematically over time
- A sensible metric for clocking time

We comment on each of these features of research design below.

1.3.1 Multiple Waves of Data

To model change, you need longitudinal data that describe how each person in the sample changes over time. We begin with this apparent tautology because too many empirical researchers seem willing to leap from cross-sectional data that describe differences among individuals of different ages to making generalizations about change over time. Many developmental psychologists, for example, analyze cross-sectional data sets composed of children of differing ages, concluding that outcome differences between age groups—in measures such as antisocial behavior—reflect real change over time. Although change is a compelling explanation of this situation—it might even be the *true* explanation—cross-sectional data can never confirm this possibility because equally valid competing explanations abound. Even in a sample drawn from a single school, a random sample of older children may differ from a random sample of younger children in important ways: the groups began school in different years, they experienced different curricula and life events, and if data collection continues for a sufficient period of time, the older sample omits age-mates who dropped out of school. Any observed differences in outcomes between grade-separated cohorts may be due to these explanations and not to systematic individual change. In

statistical terms, cross-sectional studies confound age and cohort effects (and age and history effects) and are prone to selection bias.

Studies that collect two waves of data are only marginally better. For decades, researchers erroneously believed that two-wave studies were sufficient for studying change because they narrowly conceptualized change as an *increment*: the simple difference between scores assessed on two measurement occasions (see Willett, 1989). This limited perspective views change as the acquisition (or loss) of the focal increment: a "chunk" of achievement, attitude, symptoms, skill, or whatever. But there are two reasons an increment's size cannot describe the *process of change*. First, it cannot tell us about the *shape* of each person's individual growth trajectory, the focus of our level-1 question. Did all the change occur immediately after the first assessment? Was progress steady or delayed? Second, it cannot distinguish true change from measurement error. If measurement error renders pretest scores too low and posttest scores too high, you might conclude erroneously that scores increase over time when a longer temporal view would suggest the opposite. In statistical terms, two-waves studies cannot describe individual trajectories of change and they confound true change with measurement error (see Rogosa, Brandt, & Zimowski, 1982).

Once you recognize the need for multiple waves of data, the obvious question is, How many waves are enough? Are three sufficient? Four? Should you gather more? Notice that Coie's study of antisocial behavior included just three waves, while Svartberg's STAPP study included at least six and Francis's reading study included up to ten. In general, more waves are always better, within cost and logistical constraints. Detailed discussion of this design issue requires clear understanding of the statistical models presented in this book. So for now, we simply note that more waves allow you to posit more elaborate statistical models. If your data set has only three waves, you must fit simpler models with stricter assumptions—usually assuming that individual growth is *linear* over time (as Coie and colleagues did in their study of antisocial behavior). Additional waves allow you to posit more flexible models with less restrictive assumptions; you can assume that individual growth is nonlinear (as in the reading study) or linear in chunks (as in the STAPP study). In chapters 2–5, we assume that individual growth is linear over time. In chapter 6, we extend these basic ideas to situations in which level-1 growth is discontinuous or nonlinear.

1.3.2 A Sensible Metric for Time

Time is the fundamental predictor in every study of change; it must be measured reliably and validly in a sensible metric. In our examples,

reading scores are associated with particular *ages*, antisocial behavior is associated with particular *grades*, and SCL-90 scores are associated with particular *months since intake*. Choice of a time metric affects several interrelated decisions about the number and spacing of data collection waves. Each of these, in turn, involves consideration of costs, substantive needs, and statistical benefits. Once again, because discussion of these issues requires the statistical models that we have yet to develop, we do not delve into specifics here. Instead we discuss general principles.

Our overarching point is that there is no single answer to the seemingly simple question about the most sensible metric for time. You should adopt whatever scale makes most sense for your outcomes and your research question. Coie and colleagues used *grade* because they expected antisocial behavior to depend more on this "social" measure of time than on chronological age. In contrast, Francis and colleagues used age because each reading score was based on the child's age at testing. Of course, these researchers also had the option of analyzing their data using grade as the time metric; indeed, they present tables in this metric. Yet when it came to data analysis, they used the child's age at testing so as to increase the precision with which they measured each child's growth trajectory.

Many studies possess several plausible metrics for time. Suppose, for example, your interest focuses on the longevity of automobiles. Most of us would initially assess time using the vehicle's *age*—the number of weeks (or months) since purchase (or manufacture). And for many automotive outcomes—particularly those that assess appearance qualities like rust and seat wear—this choice seems appropriate. But for other outcomes, other metrics may be better. When modeling the depth of tire treads, you might measure time in *miles*, reasoning that tire wear depends more on actual use, not years on the road. The tires of a one-year-old car that has been driven 50,000 miles will likely be more worn than those of a two-year-old car that has been driven only 20,000 miles. Similarly, when modeling the health of the starter/igniter, you might measure time in *trips*, reasoning that the starter is used only once each drive. The condition of the starters in two cars of identical age and mileage may differ if one car is driven infrequently for long distances and the other is driven several times daily for short hops. So, too, when modeling the life of the engine, you might measure time in *oil changes*, reasoning that lubrication is most important in determining engine wear.

Our point is simple: choose a metric for time that reflects the cadence you expect to be most useful for your outcome. Psychotherapy studies can clock time in *weeks* or *number of sessions*. Classroom studies can clock time in *grade* or *age*. Studies of parenting behavior can clock time using *parental age* or *child age*. The only constraint is that, like time itself, the

temporal variable can change only monotonically—in other words, it cannot reverse direction. This means, for example, that when studying child outcomes, you could use height, but not weight, as a gauge of time.

Having chosen a metric for time, you have great flexibility concerning the *spacing* of the waves of data collection. The goal is to collect sufficient data to provide a reasonable view of each individual's growth trajectory. *Equally spaced waves* have a certain appeal, in that they offer balance and symmetry. But there is nothing sacrosanct about equal spacing. If you expect rapid nonlinear change during some time periods, you should collect more data at those times. If you expect little change during other periods, space those measurements further apart. So in their STAPP study, Svartberg and colleagues (1995) spaced their early waves more closely together—at approximately 0, 4, 8, and 12 months—because they expected greater change during therapy. Their later waves were further apart—at 18 and 30 months—because they expected fewer changes.

A related issue is whether everyone should share the same data collection schedule—in other words, whether everyone needs an identical distribution of waves. If everyone is assessed on an identical schedule—whether the waves are equally or unequally spaced—we say that the data set is *time-structured*. If data collection schedules vary across individuals, we say the data set is *time-unstructured*. Individual growth modeling is flexible enough to handle both possibilities. For simplicity, we begin with time-structured data sets (in chapters 2, 3, and 4). In chapter 5, we show how the same multilevel model for change can be used to analyze time-unstructured data sets.

Finally, the resultant data set need not be *balanced*; in other words, each person need not have the same number of waves. Most longitudinal studies experience some attrition. In Coie and colleagues' (1995) study of antisocial behavior, 219 children had three waves, 118 had two, and 70 had one. In Francis and colleagues' (1996) reading study, the total number of assessments per child varied between six and nine. While nonrandom attrition can be problematic for drawing inferences, individual growth modeling does not require balanced data. Each individual's empirical growth record can contain a unique number of waves collected at unique occasions of measurement—indeed, as we will see in chapter 5, some individuals can even contribute fewer than three waves!

1.3.3 A Continuous Outcome That Changes Systematically Over Time

Statistical models care little about the substantive meaning of the individual outcomes. The same models can chart changes in standardized test scores, self-assessments, physiological measurements, or observer ratings. This flexibility allows individual growth models to be used across diverse disciplines, from the social and behavioral sciences to the physical and natural sciences. The *content* of measurement is a substantive, not statistical, decision.

How to measure a given construct, however, *is* a statistical decision, and not all variables are equally suitable. Individual growth models are designed for continuous outcomes whose values change systematically over time.[1] This focus allows us to represent individual growth trajectories using meaningful parametric forms (an idea we introduce in chapter 2). Of course, it must make conceptual and theoretical sense for the outcome to follow such a trajectory. Francis and colleagues (1996) invoke developmental theory to argue that reading ability will follow a logistic trajectory as more complex skills are layered upon basic building blocks and children head toward an upper asymptote. Svartberg and colleagues (1995) invoke psychiatric theory to argue that patients' trajectories of symptomatology will differ when they are in therapy and after therapy ends.

Continuous outcomes support all the usual manipulations of arithmetic: addition, subtraction, multiplication, and division. Differences between pairs of scores, equidistantly spaced along the scale, have identical meanings. Scores derived from standardized instruments developed by testing companies—including the Woodcock Johnson Psychoeducational Test Battery—usually display these properties. So, too, do arithmetic scores derived from most public-domain instruments, like Hodges's Child Assessment Schedule and Derogatis's SCL-90. Even homegrown instruments can produce scores with the requisite measurement properties as long as they include a large enough number of items, each scored using a large enough number of response categories.

Of course, your outcomes must also possess decent psychometric properties. Using well-known or carefully piloted instruments can ensure acceptable standards of validity and precision. But longitudinal research imposes three additional requirements because the metric, validity, and precision of the outcome must also be preserved across time.

When we say that the metric in which the outcome is measured must be preserved across time, we mean that the outcome scores must be equatable over time—a given value of the outcome on any occasion must represent the same "amount" of the outcome on every occasion. Outcome equatability is easiest to ensure when you use the identical instrument for measurement repeatedly over time, as did Coie and colleagues (1995) in their study of antisocial behavior and Svartverg and colleagues (1995) in their study of STAPP. Establishing outcome equatability when

the measures differ over time—like the Woodcock Johnson test battery used by Francis and colleagues (1996)—requires more effort. If the instrument has been developed by a testing organization, you can usually find support for equatability over time in the testing manuals. Francis and colleagues (1996) note that:

> The Rasch-scaled score reported for the reading-cluster score is a transformation of the number correct for each subtest that yields a score with interval scale properties and a constant metric. The transformation is such that a score of 500 corresponds to the average performance level of fifth graders. Its interval scale and constant metric properties make the Rasch-scaled score ideal for longitudinal studies of individual growth. (p. 6)

If outcome measures are not equatable over time, the longitudinal equivalence of the score meanings cannot be assumed, rendering the scores useless for measuring change.

Note that measures cannot be made equatable simply by standardizing their scores on each occasion to a common standard deviation. Although occasion-by-occasion standardization appears persuasive—it seems to let you talk about children who are "1 (standard deviation) unit" above the mean at age 10 and "1.2 units" above the mean at age 11, say—the "units" from which these scores are derived (i.e., the underlying age-specific standard deviations used in the standardization process) are themselves unlikely to have had either the same size or the same meaning.

Second, your outcomes must be equally valid across all measurement occasions. If you suspect that cross-wave validity might be compromised, you should replace the measure *before* data collection begins. Sometimes, as in the psychotherapy study, it is easy to argue that validity is maintained over time because the respondents have good reason to answer honestly on successive occasions. But in other studies, such as Coie and colleagues' (1996) antisocial behavior study, instrument validity over time may be more difficult to assert because young children may not understand all the questions about antisocial behavior included in the measure and older children may be less likely to answer honestly. Take the time to be cautious even when using instruments that appear valid on the surface. In his landmark paper on dilemmas in the measurement of change, Lord (1963) argued that, just because a measurement was valid on one occasion, it would not necessarily remain so on all subsequent occasions even when administered to the same individuals under the same conditions. He argued that a multiplication test may be a valid measure of mathematical skill among young children, but becomes a measure of memory among teenagers.

Third, you should try to preserve your outcome's precision over time,

although precision need not be identical on every occasion. Within the logistical constraints imposed by data collection, the goal is to minimize errors introduced by instrument administration. An instrument that is "reliable enough" in a cross-sectional study—perhaps with a reliability of .8 or .9—will no doubt be sufficient for a study of change. So, too, the measurement error variance can vary across occasions because the methods we introduce can easily accommodate heteroscedastic error variation. Although the reliability of change measurement depends directly on outcome reliability, the precision with which you estimate individual change depends more on the number and spacing of the waves of data collection. In fact, by carefully choosing and placing the occasions of measurement, you can usually offset the deleterious effects of measurement error in the outcome.

2

Exploring Longitudinal Data on Change

Change is the nursery of music, joy, life, and Eternity.
—John Donne

Wise researchers conduct descriptive exploratory analyses of their data before fitting statistical models. As when working with cross-sectional data, exploratory analyses of longitudinal data can reveal general patterns, provide insight into functional form, and identify individuals whose data do not conform to the general pattern. The exploratory analyses presented in this chapter are based on numerical and graphical strategies already familiar from cross-sectional work. Owing to the nature of longitudinal data, however, they are inevitably more complex in this new setting. For example, before you conduct even a single analysis of longitudinal data, you must confront a seemingly innocuous decision that has serious ramifications: how to store your longitudinal data efficiently. In section 2.1, we introduce two different data organizations for longitudinal data—the "person-level" format and the "person-period" format—and argue in favor of the latter.

We devote the rest of this chapter to describing exploratory analyses that can help you learn how different individuals in your sample change over time. These analyses serve two purposes: to identify important features of your data and to prepare you for subsequent model-based analyses. In section 2.2, we address the *within-person* question—How does each person change over time?—by exploring and summarizing *empirical growth records*, which list each individual's outcome values over time. In section 2.3, we address the *between-person* question—How does individual change differ across people?—by exploring whether different people change in similar or different ways. In section 2.4, we show how to ascertain descriptively whether observed differences in change across people (*interindividual differences in change*) are associated with individual

characteristics. These between-person explorations can help identify variables that may ultimately prove to be important predictors of change. We conclude, in section 2.5, by examining the reliability and precision of exploratory estimates of change and commenting on their implications for the design of longitudinal studies.

2.1 Creating a Longitudinal Data Set

Your first step is to organize your longitudinal data in a format suitable for analysis. In cross-sectional work, data-set organization is so straightforward as to not warrant explicit attention—all you need is a "standard" data set in which each individual has his or her own record. In longitudinal work, data-set organization is less straightforward because you can use two very different arrangements:

- A *person-level data set*, in which each person has one record and multiple variables contain the data from each measurement occasion
- A *person-period data set*, in which each person has multiple records—one for each measurement occasion

A person-level data set has as many records as there are people in the sample. As you collect additional waves, the file gains new variables, not new cases. A person-period data set has many more records—one for each person-period combination. As you collect additional waves of data, the file gains new records, but no new variables.

All statistical software packages can easily convert a longitudinal data set from one format to the other. The website associated with our book presents illustrative code for implementing the conversion in a variety of statistical packages. If you are using SAS, for example, Singer (1998, 2001) provides simple code for the conversion. In STATA, the "reshape" command can be used. The ability to move from one format to the other means that you can enter, and clean, your data using whichever format is most convenient. But as we show below, when it comes to data analysis—either exploratory or inferential—you need to have your data in a person-period format because this most naturally supports meaningful analyses of change over time.

We illustrate the difference between the two formats in figure 2.1, which presents five waves of data from the *National Youth Survey* (NYS; Raudenbush & Chan, 1992). Each year, when participants were ages 11, 12, 13, 14, and 15, they filled out a nine-item instrument designed to assess their tolerance of deviant behavior. Using a four-point scale

"Person-Level" data set

ID	*TOL11*	*TOL12*	*TOL13*	*TOL14*	*TOL15*	*MALE*	*EXPOSURE*
9	2.23	1.79	1.9	2.12	2.66	0	1.54
45	1.12	1.45	1.45	1.45	1.99	1	1.16
268	1.45	1.34	1.99	1.79	1.34	1	0.9
314	1.22	1.22	1.55	1.12	1.12	0	0.81
442	1.45	1.99	1.45	1.67	1.9	0	1.13
514	1.34	1.67	2.23	2.12	2.44	1	0.9
569	1.79	1.9	1.9	1.99	1.99	0	1.99
624	1.12	1.12	1.22	1.12	1.22	1	0.98
723	1.22	1.34	1.12	1	1.12	0	0.81
918	1	1	1.22	1.99	1.22	0	1.21
949	1.99	1.55	1.12	1.45	1.55	1	0.93
978	1.22	1.34	2.12	3.46	3.32	1	1.59
1105	1.34	1.9	1.99	1.9	2.12	1	1.38
1542	1.22	1.22	1.99	1.79	2.12	0	1.44
1552	1	1.12	2.23	1.55	1.55	0	1.04
1653	1.11	1.11	1.34	1.55	2.12	0	1.25

"Person-Period" data set

ID	*AGE*	*TOL*	*MALE*	*EXPOSURE*
9	11	2.23	0	1.54
9	12	1.79	0	1.54
9	13	1.9	0	1.54
9	14	2.12	0	1.54
9	15	2.66	0	1.54
45	11	1.12	1	1.16
45	12	1.45	1	1.16
45	13	1.45	1	1.16
45	14	1.45	1	1.16
45	15	1.99	1	1.16
.	.	.	.	.
.	.	.	.	.
1653	11	1.11	0	1.25
1653	12	1.11	0	1.25
1653	13	1.34	0	1.25
1653	14	1.55	0	1.25
1653	15	2.12	0	1.25

Figure 2.1. Conversion of a person-level data set into a person-period data set for selected participants in the tolerance study.

(1 = very wrong, 2 = wrong, 3 = a little bit wrong, 4 = not wrong at all), they indicated whether it was wrong for someone their age to: (a) cheat on tests, (b) purposely destroy property of others, (c) use marijuana, (d) steal something worth less than five dollars, (e) hit or threaten someone without reason, (f) use alcohol, (g) break into a building or vehicle to steal, (h) sell hard drugs, or (i) steal something worth more than fifty dollars. At each occasion, the outcome, *TOL*, is computed as the respondent's average across the nine responses. Figure 2.1 also includes two potential predictors of change in tolerance: *MALE*, representing respondent gender, and *EXPOSURE*, assessing the respondent's self-reported exposure to deviant behavior at age 11. To obtain values of this latter predictor, participants estimated the proportion of their close friends who were involved in each of the same nine activities on another four-point scale (ranging from 0 = none, to 4 = all). Like *TOL*, each respondent's value of *EXPOSURE* is the average of his or her nine responses. Figure 2.1 presents data for a random sample of 16 participants from the larger NYS data set. Although the exploratory methods of this chapter apply in data sets of all sizes, we have kept this example purposefully small to enhance manageability and clarity. In later chapters, we apply the same methods to larger data sets.

2.1.1 The Person-Level Data Set

Many people initially store longitudinal data as a *person-level* data set (also known as the *multivariate format*), probably because it most resembles the familiar cross-sectional data-set format. The top panel of figure 2.1 displays the NYS data using this arrangement. The hallmark feature of a person-level data set is that each person has only one row (or "record") of data, regardless of the number of waves of data collection. A 16-person data set has 16 records; a 20,000-person data set has 20,000. Repeated measurements of each outcome appear as additional variables (hence the alternate "multivariate" label for the format). In the person-level data set of figure 2.1, the five values of tolerance appear in columns 2 through 6 (*TOL11*, *TOL12*, . . . *TOL15*). Suffixes attached to column headings identify the measurement occasion (here, respondent's age) and additional variables—here, *MALE* and *EXPOSURE*—appear in additional columns.

The primary advantage of a person-level data set is the ease with which you can examine visually each person's *empirical growth record*, his or her temporally sequenced outcome values. Each person's empirical growth record appears compactly in a single row making it is easy to assess quickly the way he or she is changing over time. In examining the top panel of figure 2.1, for example, notice that change differs considerably across

Table 2.1: Estimated bivariate correlations among tolerance scores assessed on five measurement occasions ($n = 16$)

	TOL11	*TOL12*	*TOL13*	*TOL14*	*TOL15*
TOL11	1.00				
TOL12	0.66	1.00			
TOL13	0.06	0.25	1.00		
TOL14	0.14	0.21	0.59	1.00	
TOL15	0.26	0.39	0.57	0.83	1.00

adolescents. Although most become more tolerant of deviant behavior over time (e.g., subjects 514 and 1653), many remain relatively stable (e.g., subjects 569 and 624), none of the 16 becomes much less tolerant (although subject 949 declines for a while before increasing).

Despite the ease with which you can examine each person's empirical growth record visually, the person-level data set has four disadvantages that render it a poor choice for most longitudinal analyses: (1) it leads naturally to noninformative summaries; (2) it omits an explicit "time" variable; (3) it is inefficient, or useless, when the number and spacing of waves varies across individuals; and (4) it cannot easily handle the presence of time-varying predictors. Below, we explain these difficulties; in section 2.1.2, we demonstrate how each is addressed by a conversion to a person-period data set.

First, let us begin by examining the five separate tolerance variables in the person-level data set of figure 2.1 and asking how you might analyze these longitudinal data. For most researchers, the instinctive response is to examine wave-to-wave relationships among *TOL11* through *TOL15* using bivariate correlation analyses (as shown in table 2.1) or companion bivariate plots. Unfortunately, summarizing the bivariate relationships between waves tells us little about change over time, for either individuals or groups. What, for example, does the weak but generally positive correlation between successive assessments of *TOLERANCE* tell us? For any pair of measures, say *TOL11* and *TOL12*, we know that adolescents who were more tolerant of deviant behavior at one wave tend to be more tolerant at the next. This indicates that the *rank order* of adolescents remains relatively stable across occasions. But it does not tell us *how* each person changes over time; it does not even tell us about the *direction* of change. If everyone's score declined by one point between age 11 and age 12, but the rank ordering was preserved, the correlation between waves would be positive (at +1)! Tempting though it is to infer a direct link between the wave-to-wave correlations and change, it is a

futile exercise. Even with a small data set—here just five waves of data for 16 people—wave-to-wave correlations and plots tell us nothing about change over time.

Second, the person-level data set has no explicit numeric variable identifying the occasions of measurement. Information about "time" appears in the variable names, not in the data, and is therefore unavailable for statistical analysis. Within the actual person-level data set of figure 2.1, for example, information on *when* these *TOLERANCE* measures were assessed—the numeric values 11, 12, 13, 14, and 15—appears nowhere. Without including these values in the dataset, we cannot address within-person questions about the relationship between the outcome and "time."

Third, the person-level format is inefficient if either the number, or spacing, of waves varies across individuals. The person-level format is best suited to research designs with *fixed* occasions of measurement—each person has the same number of waves collected on the same exact schedule. The person-level data set of figure 2.1 is compact because the NYS used such a design—each adolescent was assessed on the same five annual measurement occasions (at ages 11, 12, 13, 14, and 15). Many longitudinal data sets do not share this structure. For example, if we reconceptualized "time" as the adolescent's *specific* age (say, in months) at each measurement occasion, we would need to expand the person-level data set in some way. We would need either five additional columns to record the respondent's precise age on each measurement occasion (e.g., variables with names like *AGE11*, *AGE12*, *AGE13*, *AGE14*, and *AGE15*) or even more additional columns to record the respondent's tolerance of deviant behavior on each of the many *unique* measurement occasions (e.g., variables with names like *TOL11.1*, *TOL11.2*, ... *TOL15.11*). This latter approach is particularly impractical. Not only would we add 55 variables to the data set, we would have missing values in the cells corresponding to each month not used by a particular individual. In the extreme, if each person in the data set has his or her own unique data collection schedule—as would be the case were *AGE* recorded in days—the person-level format becomes completely unworkable. Hundreds of columns would be needed and most of the data entries would be missing!

Finally, person-level data sets become unwieldy when the values of *predictors* can vary over time. The two predictors in this data set are *time-invariant*—the values of *MALE* and *EXPOSURE* remain the same on every occasion. This allows us to use a single variable to record the values of each. If the data set contained *time-varying predictors*—predictors whose values vary over time—we would need an additional *set* of columns for each—one per measurement occasion. If, for example, exposure to

deviant behavior were measured each year, we would need four additional columns. While the data could certainly be recorded in this way, this leads to the same disadvantages for time-varying predictors as we have just described for time-varying outcomes.

Taken together, these disadvantages render the person-level format, so familiar in cross-sectional research, ill suited to longitudinal work. Although we will return to the multivariate format in chapter 8, when we introduce a covariance structure analysis approach to modeling change (known as *latent growth modeling*), for now we suggest that longitudinal data analysis is facilitated—and made more meaningful—if you use the "person-period" format for your data.

2.1.2 The Person-Period Data Set

In a person-period data set, also known as *univariate format*, each individual has multiple records, one for each period in which he or she was observed. The bottom panel of figure 2.1 presents illustrative entries for the NYS data. Both panels present identical information; they differ only in *structure*. The person-period data set arrays each person's empirical growth record vertically, not horizontally. Person-period data sets therefore have fewer columns than person-level data sets (here, five instead of eight), but many more rows (here, 80 instead of 16). Even for this small example, the person-period data set has so many rows that figure 2.1 displays only a small subset.

All person-period data sets contain four types of variables: (1) a subject identifier; (2) a time indicator; (3) outcome variable(s); and (4) predictor variable(s). The *ID* number, which identifies the participant that each record describes, typically appears in the first column. Time-invariant by definition, *ID*s are identical across each person's multiple records. Including an *ID* number is more than good record keeping; it is an integral part of the analysis. Without an *ID*, you cannot sort the data set into person-specific subsets (a first step in examining individual change trajectories in section 2.2).

The second column in the person-period data set typically displays a *time indicator*—usually labeled *AGE*, *WAVE*, or *TIME*—which identifies the specific occasion of measurement that the record describes. For the NYS data, the second column of the person-period data set identifies the respondent's *AGE* (in years) on each measurement occasion. A dedicated time variable is a fundamental feature of every person-period data set; it is what renders the format amenable to recording longitudinal data from a wide range of research designs. You can easily construct a person-period data set even if each participant has a unique data collection schedule

(as would be the case if we clocked time using each adolescent's precise age on the date of interview). The new *AGE* variable would simply record each adolescent's age on that particular date (e.g., 11.24, 12.32, 13.73, 14.11, 15.40 for one case; 11.10, 12.32, 13.59, 14.21, 15.69 for the next, etc.). A dedicated *TIME* variable also allows person-period data sets to accommodate research designs in which the number of measurement occasions differs across people. Each person simply has as many records as he or she has waves of data in the design. Someone with three waves will have three records; someone with 20 will have 20.

Each outcome in a person-period data set—here, just *TOL*—is represented by a single variable (hence the alternate "univariate" label for the format) whose values represent that person's score on each occasion. In figure 2.1, every adolescent has five records, one per occasion, each containing his or her tolerance of deviant behavior at the age indicated.

Every predictor—whether time-varying or time-invariant—is also represented by a single variable. A person-period data set can include as many predictors of either type as you would like. The person-period data set in figure 2.1 includes two time-invariant predictors, *MALE* and *EXPOSURE*. The former is time-invariant; the latter is time-invariant only because of the way it was constructed (as exposure to deviant behavior at one point in time, age 11). Time-invariant predictors have identical values across each person's multiple records; time-varying predictors have potentially differing values. We defer discussion of time-varying predictors to section 5.3. For now, we simply note how easy it is to include them in a person-period data set.

We hope that this discussion convinces you of the utility of storing longitudinal data in a person-period format. Although person-period data sets are typically longer than their person-level cousins, the ease with which they can accommodate any data collection schedule, any number of outcomes, and any combination of time-invariant and time-varying predictors outweigh the cost of increased size.

2.2 Descriptive Analysis of Individual Change over Time

Having created a person-period data set, you are now poised to conduct exploratory analyses that describe how individuals in the data set change over time. Descriptive analyses can reveal the nature and idiosyncrasies of each person's temporal pattern of growth, addressing the question: How does each person change over time? In section 2.2.1, we present a simple graphical strategy; in section 2.2.2, we summarize the observed trends by superimposing rudimentary fitted trajectories.

2.2.1 Empirical Growth Plots

The simplest way of visualizing how a person changes over time is to examine an *empirical growth plot*, a temporally sequenced graph of his or her empirical growth record. You can easily obtain empirical growth plots from any major statistical package: sort the person-period data set by subject identifier (*ID*) and separately plot each person's outcome vs. time (e.g., *TOL* vs. *AGE*). Because it is difficult to discern similarities and differences among individuals if each page contains only a single plot, we recommend that you cluster sets of plots in smaller numbers of panels.

Figure 2.2 presents empirical growth plots for the 16 adolescents in the NYS study. To facilitate comparison and interpretation, we use identical axes across panels. We emphasize this seemingly minor point because many statistical packages have the annoying habit of automatically expanding (or contracting) scales to fill out a page or plot area. When this happens, individuals who change only modestly acquire seemingly steep trajectories because the vertical axis expands to cover their limited outcome range; individuals who change dramatically acquire seemingly shallow trajectories because the vertical axis shrinks to accommodate their wide outcome range. If your axes vary inadvertently, you may draw erroneous conclusions about any similarities and differences in individual change.

Empirical growth plots can reveal a great deal about how each person changes over time. You can evaluate change in both absolute terms (against the outcome's overall scale) and in relative terms (in comparison to other sample members). Who is increasing? Who is decreasing? Who is increasing the most? The least? Does anyone increase and then decrease (or vice versa)? Inspection of figure 2.2 suggests that tolerance of deviant behavior generally increases with age (only subjects 314, 624, 723, and 949 do not fit this trend). But we also see that most adolescents remain in the lower portion of the outcome scale—here shown in its full extension from 1 to 4—suggesting that tolerance for deviant behavior never reaches alarming proportions (except, perhaps, for subject 978).

Should you examine every possible empirical growth plot if your data set is large, including perhaps thousands of cases? We do not suggest that you sacrifice a ream of paper in the name of data analysis. Instead, you can randomly select a subsample of individuals (perhaps stratified into groups defined by the values of important predictors) to conduct these exploratory analyses. All statistical packages can generate the random numbers necessary for such subsample selection; in fact, this is how we selected these 16 individuals from the NYS sample.

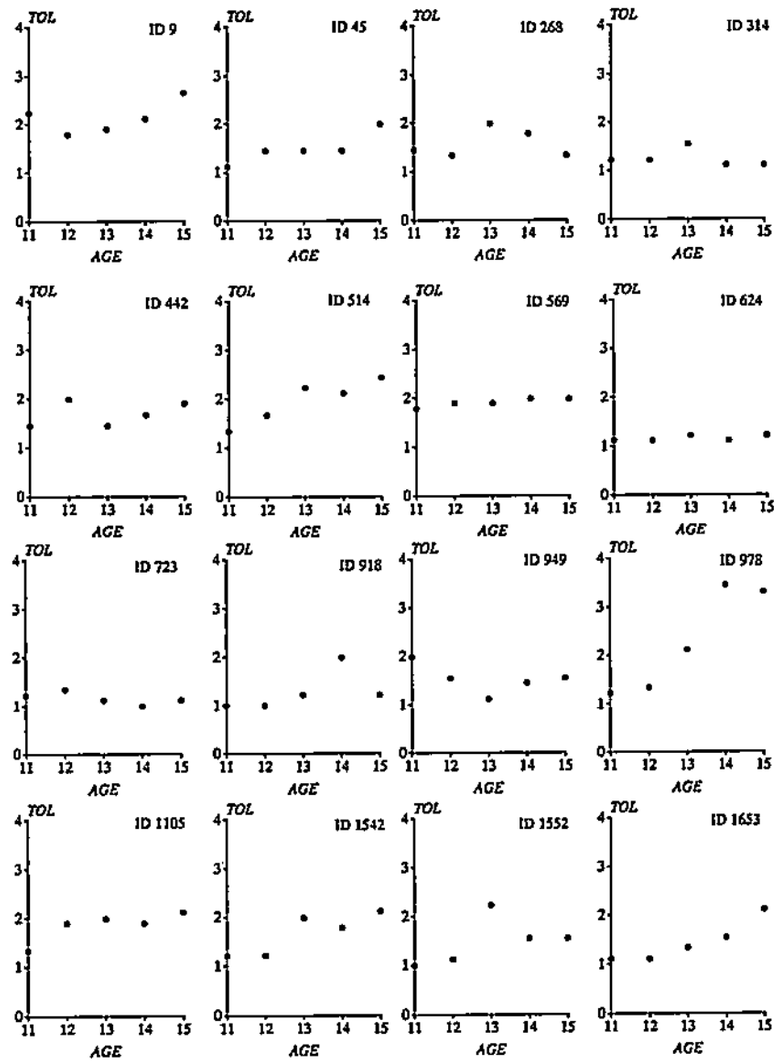


Figure 2.2. Exploring how individuals change over time. Empirical growth plots for 16 participants in the tolerance study.

2.2.2 Using a Trajectory to Summarize Each Person's Empirical Growth Record

It is easy to imagine summarizing the plot of each person's empirical growth record using some type of smooth trajectory. Although we often

begin by drawing freehand trajectories, we strongly recommend that you also apply two standardized approaches. With the *nonparametric* approach, you let the "data speak for themselves" by smoothing across temporal idiosyncrasies without imposing a specific functional form. With the *parametric* approach, you select a common functional form for the trajectories—a straight line, a quadratic or some other curve—and then fit a separate regression model to each person's data, yielding a fitted trajectory.

The fundamental advantage of the nonparametric approach is that it requires no assumptions. The parametric approach requires assumptions but, in return, provides numeric summaries of the trajectories (e.g., estimated intercepts and slopes) suitable for further exploration. We find it helpful to begin nonparametrically—as these summaries often inform the parametric analysis.

Smoothing the Empirical Growth Trajectory Nonparametrically

Nonparametric trajectories summarize each person's pattern of change over time graphically without committing to a specific functional form. All major statistical packages provide several options for assumption-free smoothing, including the use of splines, loess smoothers, kernel smoothers, and moving averages. Choice of a particular smoothing algorithm is primarily a matter of convenience; all are adequate for the exploratory purposes we intend here.

Figure 2.3 plots the NYS empirical growth records and superimposes a smooth nonparametric trajectory (obtained using the "curve" option in *Harvard Graphics*). When examining smoothed trajectories like these, focus on their elevation, shape, and tilt. Where do the scores hover—at the low, medium, or high end of the scale? Does everyone change over time or do some people remain the same? What is the overall pattern of change? Is it linear or curvilinear; smooth or steplike? Do the trajectories have an inflection point or plateau? Is the rate of change steep or shallow? Is this rate of change similar or different across people? The trajectories in figure 2.3 reinforce our preliminary conclusions about the nature of individual change in the tolerance of deviant behavior. Most adolescents experience a gentle increase between ages 11 and 15, except for subject 978, who registers a dramatic leap after age 13.

After examining the nonparametric trajectories individually, stare at the entire set together as a group. Group-level analysis can help inform decisions that you will soon need to make about a functional form for the trajectory. In our example, several adolescents appear to have linear trajectories (subjects 514, 569, 624, and 723) while others have

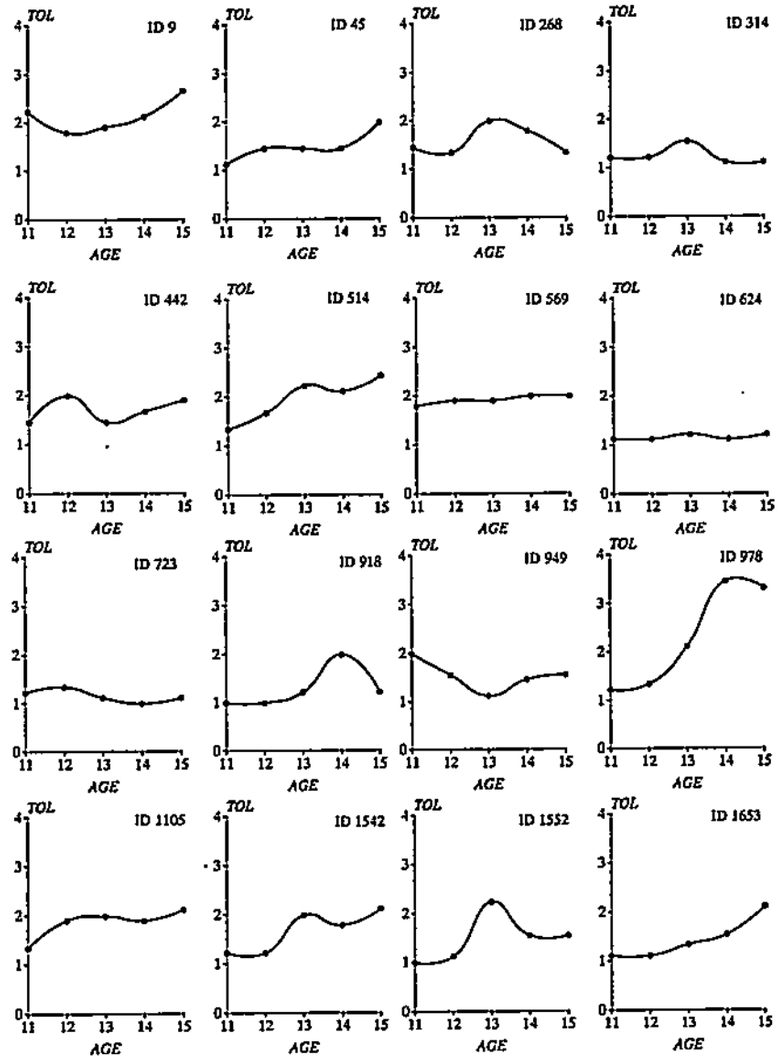


Figure 2.3. Smooth nonparametric summaries of how individuals change over time. Smooth nonparametric trajectories superimposed on empirical growth plots for participants in the tolerance study.

curvilinear ones that either accelerate (9, 45, 978, and 1653) or rise and fall around a central peak or trough (268, 314, 918, 949, 1552).

Smoothing the Empirical Growth Trajectory Using OLS Regression

We can also summarize each person's growth trajectory by fitting a separate parametric model to each person's data. Although many methods of model fitting are possible, we find that ordinary least squares (OLS) regression is usually adequate for exploratory purposes. Of course, fitting person-specific regression models, one individual at a time, is hardly the most efficient use of longitudinal data; that's why we need the multilevel model for change that we will soon introduce. But because the "fitting of little OLS regression models" approach is intuitive and easy to implement in a person-period data set, we find that it connects empirical researchers with their data in a direct and intimate way.

To fit an exploratory OLS regression model to each person's data, you must first select a specific functional form for that model. Not only is this decision crucial during exploratory analysis, it becomes even more important during formal model fitting. Ideally, substantive theory and past research will guide your choice. But when you observe only a restricted portion of the life span—as we do here—or when you have only three or four waves of data, model selection can be difficult.

Two factors further complicate the choice of a functional form. First, exploratory analyses often suggest that different people require different functions—change might appear linear for some, curvilinear for others. We observe this pattern, to some extent, in figure 2.3. Yet the simplification that comes from adopting a common functional form across everyone in the data set is so compelling that its advantages totally outweigh its disadvantages. Adopting a common functional form across everyone in the sample allows you to distinguish people easily using the same set of numerical summaries derived from their fitted trajectories. This process is especially simple if you adopt a linear change model, as we do here; you can then compare individuals using just the estimated intercepts and slopes of their fitted trajectories. Second, measurement error makes it difficult to discern whether compelling patterns in the empirical growth record really reflect true change or are simply due to random fluctuation. Remember, each *observed score* is just a fallible operationalization of an underlying *true score*—depending upon the sign of the error, the observed score can be inappropriately high or low. The empirical growth records do not present a person's true pattern of change over time; they present the fallible observed reflection of that change. Some of what we see in the empirical

These complications argue for parsimony when selecting a functional form for exploratory analysis, driving you to adopt the simplest trajectory that can do the job. Often the best choice is simply a straight line. In this example, we adopted a linear individual change trend because it provides a decent description of the trajectories for these 16 adolescents. In making this decision, of course, we assume implicitly that any deviations from linearity in figure 2.3 result from either the presence of outliers or measurement error. Use of an individual linear change model simplifies our discussion enormously and has pedagogic advantages as well. We devote chapter 6 to a discussion of models for discontinuous and nonlinear change.

Having selected an appropriate parametric form for summarizing the empirical growth records, you obtain fitted trajectories using a three-step process:

1. Estimate a within-person regression model for each person in the data set. With a linear change model, simply regress the outcome (here *TOL*) on some representation of time (here, *AGE*) in the person-period data set. Be sure to conduct a separate analysis for each person (i.e., conduct the regression analyses "by *ID*").
2. Collect summary statistics from all the within-person regression models into a separate data set. For a linear-change model, each person's estimated intercept and slope summarize their growth trajectory; the R^2 and residual variance statistics summarize their goodness of fit.
3. Superimpose each person's fitted regression line on a plot of his or her empirical growth record. For each person, plot selected predicted values and join them together smoothly.

We now apply this three-step process to the NYS data.

We begin by fitting a separate linear change model to each person's empirical growth record. Although we can regress *TOL* on *AGE* directly, we instead regress *TOL* on (*AGE* – *11*) years, providing a *centered* version of *AGE*. Centering the temporal predictor is optional, but doing so improves the interpretability of the intercept. Had we not centered *AGE*, the fitted intercept would estimate the adolescent's tolerance of deviant behavior at age 0—an age beyond the range of these data and hardly one at which a child can report an attitude. Subtracting 11 years from each value of *AGE* moves the origin of the plot so that each intercept now estimates the adolescent's tolerance of deviant behavior at the more reasonable age of 11 years.

Centering *AGE* has no effect on the interpretation of each person's slope: it still estimates his or her annual rate of change. Adolescents with positive slopes grow more tolerant of deviant behavior as they age; those

Table 2.2: Results of fitting separate within-person exploratory OLS regression models for *TOLERANCE* as a function of linear time

	Initial status		Rate of change		Residual			
ID	Estimate	se	Estimate	se	variance	R^2	*MALE*	*EXPOSURE*
0009	1.90	0.25	0.12	0.10	0.11	0.31	0	1.54
0045	1.14	0.13	0.17	0.05	0.03	0.77	1	1.16
0268	1.54	0.26	0.02	0.11	0.11	0.02	1	0.90
0314	1.31	0.15	−0.03	0.06	0.04	0.07	0	0.81
0442	1.58	0.21	0.06	0.09	0.07	0.14	0	1.13
0514	1.43	0.14	0.27	0.06	0.03	0.88	1	0.90
0569	1.82	0.03	0.05	0.01	0.00	0.88	0	1.99
0624	1.12	0.04	0.02	0.02	0.00	0.33	1	0.98
0723	1.27	0.08	−0.05	0.04	0.01	0.45	0	0.81
0918	1.00	0.30	0.14	0.13	0.15	0.31	0	1.21
0949	1.73	0.24	−0.10	0.10	0.10	0.25	1	0.93
0978	1.03	0.32	0.63	0.13	0.17	0.89	1	1.59
1105	1.54	0.15	0.16	0.06	0.04	0.68	1	1.38
1542	1.19	0.18	0.24	0.07	0.05	0.78	0	1.44
1552	1.18	0.37	0.15	0.15	0.23	0.25	0	1.04
1653	0.95	0.14	0.25	0.06	0.03	0.86	0	1.25

cents with negative slopes grow less tolerant of deviant behavior over time; those with the most negative slopes become less tolerant the most rapidly. Because the fitted slopes estimate the annual rate of change in the outcome, they are the parameter of central interest in an exploratory analysis of change.

Table 2.2 presents the results of fitting 16 linear-change OLS regression models to the NYS data. The table displays OLS-estimated intercepts and slopes for each person along with associated standard errors, residual variance, and R^2 statistics. Figure 2.4 presents a stem-and-leaf display of each summary statistic. Notice that both the fitted intercepts and slopes vary considerably, reflecting the heterogeneity in trajectories observed in figure 2.3. Although most adolescents have little tolerance for deviant behavior at age 11, some—like subjects 9 and 569—are more tolerant. Notice, too, that many adolescents register little change over time. Comparing the estimated slopes to their associated standard errors, we find that the slopes for nine people (subjects 9, 268, 314, 442, 624, 723, 918, 949, and 1552) are indistinguishable from 0. Three have moderate increases (514, 1542, and 1653) and one extreme case (978) increases three times faster than his closest peer.

Figure 2.5 superimposes each adolescent's fitted OLS trajectory on his or her empirical growth plot. All major statistical packages can generate

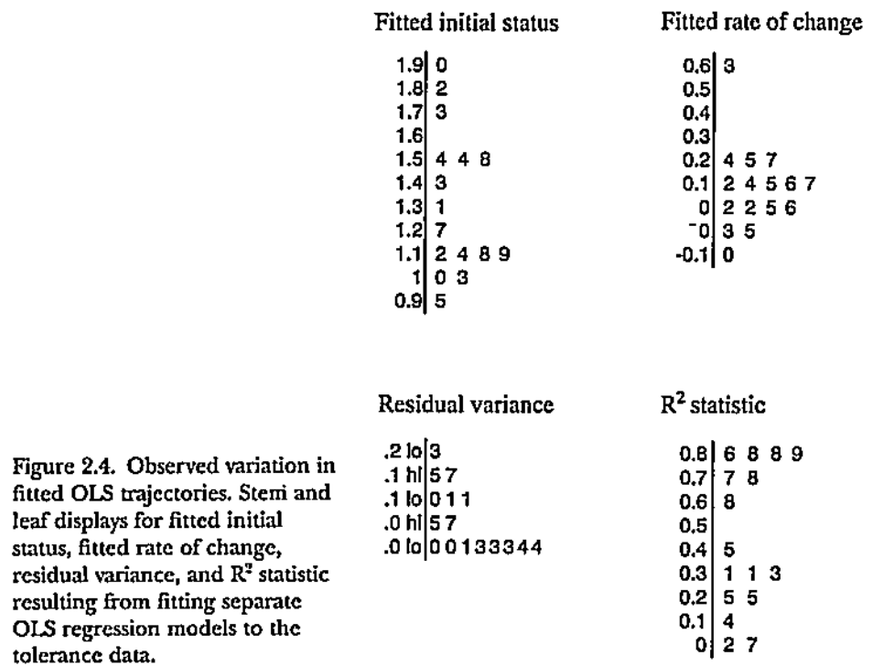


Figure 2.4. Observed variation in fitted OLS trajectories. Stem and leaf displays for fitted initial status, fitted rate of change, residual variance, and R^2 statistic resulting from fitting separate OLS regression models to the tolerance data.

such plots. For example, because the estimated intercept and slope for subject 514 are 1.43 and 0.27, the fitted values at ages 11 and 15 are: 1.43 (computed as $1.43 + 0.27(11 - 11)$) and 2.51 (computed as $1.43 + 0.27(15 - 11)$). To prevent extrapolation beyond the temporal limits of the data, we plot this trajectory only between ages 11 and 15.

Comparing the exploratory OLS-fitted trajectories with the observed data points allows us to evaluate how well the chosen linear change model fits each person's growth record. For some adolescents (such as 569 and 624), the linear change model fits well—their observed and fitted values nearly coincide. A linear change trajectory may also be reasonable for many other sample members (including subjects 45, 314, 442, 514, 723, 949, 1105, and 1542) if we are correct in regarding the observed deviations from the fitted trajectory as random error. For five adolescents (subjects 9, 268, 918, 978, and 1552), observed and fitted values are more disparate. Inspection of their empirical growth records suggests that their change may warrant a curvilinear model.

Table 2.2 presents two simple ways of quantifying the quality of fit for each person: an individual R^2 statistic and an individual estimated residual variance. Even in this small sample, notice the striking variability in

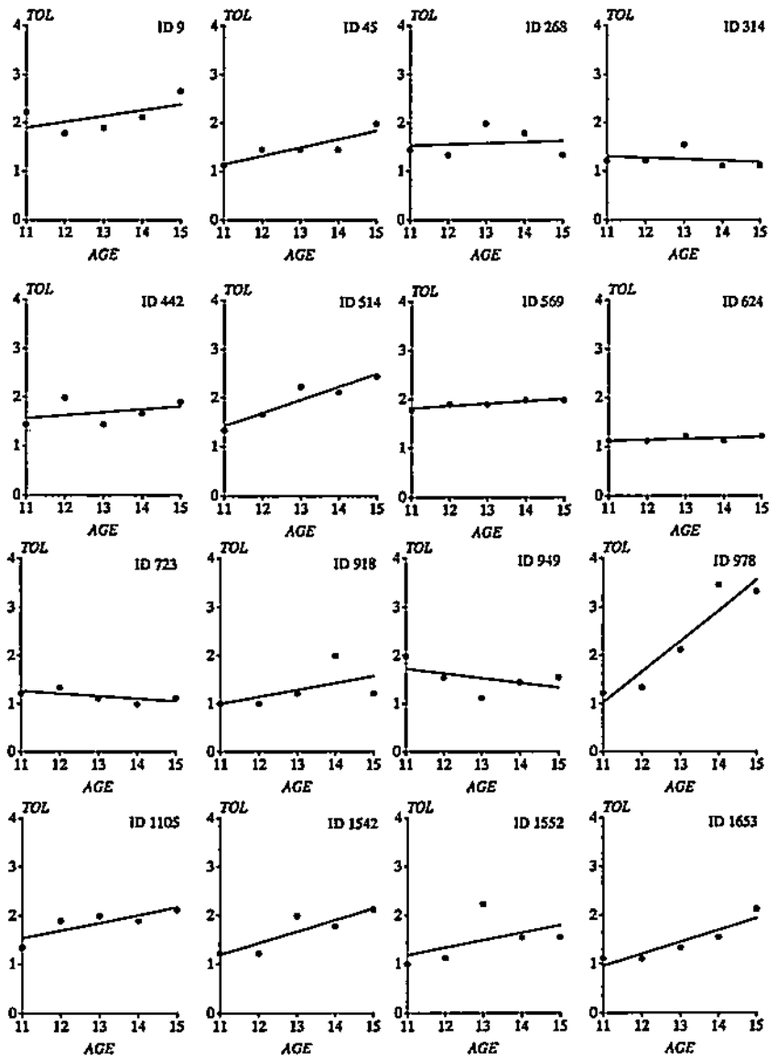


Figure 2.5. OLS summaries of how individuals change over time. Fitted OLS trajectories superimposed on empirical growth plots for participants in the tolerance study.

the individual R^2 statistics. They range from a low of 2% for subject 268 (whose trajectory is essentially flat and whose data are widely scattered) to highs of 88% for subjects 514 and 569 (whose empirical growth records show remarkable linearity in change) and 89% for subject 978 (who has

the most rapid rate of growth). The individual estimated residual variances mirror this variability (as you might expect, given that they are an element in the computation of the R^2 statistic). Skewed by definition (as apparent in figure 2.4), they range from a low near 0 for subjects 569 and 624 (whose data are predicted nearly perfectly) to highs of 0.17 and 0.23 for subjects 978 and 1552 (who each have an extreme observation). We conclude that the quality of exploratory model fit varies substantially from person to person; the linear change trajectory works well for some sample members and poorly for others.

By now you may be questioning the wisdom of using OLS regression methods to conduct even exploratory analyses of these data. OLS regression methods assume independence and homoscedasticity of residuals. Yet these assumptions are unlikely to hold in longitudinal data where residuals tend to be autocorrelated and heteroscedastic over time within person. Despite this concern, OLS estimates can be very useful for exploratory purposes. Although they are less efficient when the assumption of residual independence is violated (i.e., their sampling variance is too high), they still provide unbiased estimates of the intercept and slope of the individual change (Willett, 1989). In other words, these exploratory estimates of the key features of the individual change trajectory—each person's intercept and slope—will be on target, if a little noisy.

2.3 Exploring Differences in Change across People

Having summarized how each individual changes over time, we now examine similarities and differences in these changes across people. Does everyone change in the same way? Or do the trajectories of change differ substantially across people? Questions like these focus on the assessment of *interindividual differences* in change.

2.3.1 Examining the Entire Set of Smooth Trajectories

The simplest way of exploring interindividual differences in change is to plot, on a single graph, the entire set of smoothed individual trajectories. The left panel of figure 2.6 presents such a display for the NYS data using the nonparametric smoother; the right panel presents a similar display using OLS regression methods. In both, we omit the observed data to decrease clutter.

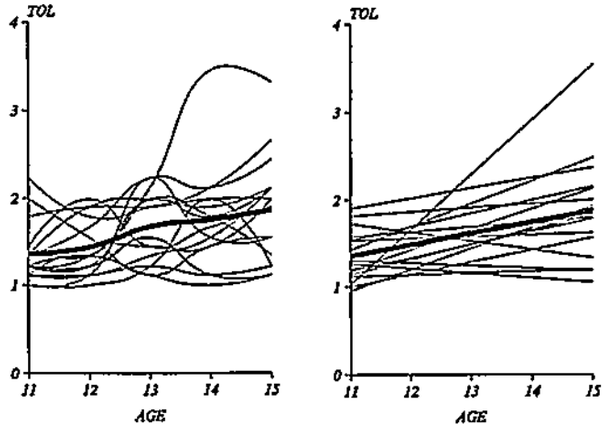


Figure 2.6. Examining the collection of smooth nonparametric and OLS trajectories across participants in the tolerance study. Panel A presents the collection of smooth nonparametric trajectories; Panel B presents the collection of fitted OLS trajectories. Both panels also present an *average change trajectory* for the entire group.

Each panel in figure 2.6 also includes a new summary: an *average change trajectory* for the entire group. Depicted in bold, this summary helps us compare individual change with group change. Computing an average change trajectory is a simple two-step process. First, sort the person-period data set by time (here, *AGE*), and separately estimate the mean outcome (here, *TOLERANCE*) for each occasion of measurement. Second, plot these time-specific means and apply the same smoothing algorithm, nonparametric or parametric, used to obtain the individual trajectories.

Both panels in figure 2.6 suggest that, on average, the change in tolerance of deviant behavior between ages 11 and 15 is positive but modest, rising by one to two-tenths of a point per year (on this 1 to 4 scale). This suggests that as adolescents mature, they gradually tolerate more deviant behavior. Note that even the nonparametrically smoothed average trajectory seems approximately linear. (The slight curvature or discontinuity between ages 12 and 13 disappears if we set aside the extreme case, subject 978.) Both panels also suggest substantial interindividual heterogeneity in change. For some adolescents, tolerance increases moderately with age; for others, it remains stable; for some, it declines. This heterogeneity creates a "fanning out" of trajectories as increasing age engenders greater diversity in tolerance. Notice that the OLS regression panel is somewhat easier to interpret because of its greater structure.

Although the average change trajectory is a valuable summary, we inject a note of caution: the shape of the average change trajectory may not mimic the shape of the individual trajectories from which it derives. We see this disconcerting behavior in figure 2.6, where the nonparametrically smoothed trajectories manifest various curvilinear shapes but the average trajectory is nearly linear. This means that you should never infer the shape of the individual change trajectories from the shape of their average. As we explain in section 6.4, the only kind of trajectory for which the "average of the curves" is identical to the "curve of the averages" is one whose mathematical representation is *linear in the parameters* (Keats, 1983). All polynomials—including linear, quadratic, and cubic trajectories—are linear in the parameters; their average trajectory will always be a polynomial of the same order as the individual trajectories. The average of a set of straight lines will be a straight line; the average of a set of quadratics will be a quadratic. But many other common curves do not share this property. The average of a set of logistic curves, for example, is usually a smoothed-out step function. This means that you must exercise extreme caution when examining an average growth trajectory. We display the average simply for comparison, not to learn anything about underlying shapes of the individual trajectories.

2.3.2 Using the Results of Model Fitting to Frame Questions about Change

Adopting a parametric model for individual change allows us to re-express *generic* questions about interindividual differences in "change" as *specific* questions about the behavior of parameters in the individual models. If we have selected our parametric model wisely, little information is lost and great simplification is achieved. If you adopt a linear individual change model, for instance, you are implicitly agreeing to summarize each person's growth using just two parameter estimates: (1) the fitted intercept; and (2) the fitted slope. For the NYS data, variation in fitted intercepts across adolescents summarizes observed interindividual differences in tolerance at age 11. If these intercepts describe fitted values at the first wave of data collection, as they do here, we say that they estimate someone's "initial status." Variation in the fitted slopes describes observed interindividual differences in the rates at which tolerance for deviant behavior changes over time.

Greater specificity and simplification accrues if we reframe general questions about interindividual heterogeneity in change in terms of key parameters of the individual change trajectory. Rather than asking "Do individuals differ in their changes, and if so, how?" we can now ask "Do

individuals differ in their intercepts? In their slopes?" To learn about the observed *average* pattern of change, we examine the sample averages of the fitted intercepts and slopes; these tell us about the average initial status and the average annual rate of change in the sample as a whole. To learn about the observed *individual differences* in change, we examine the sample *variances* and *standard deviations* of the intercepts and slopes; these tell us about the observed variability in initial status and rates of change in the sample. And to learn about the observed relationship between initial status and the rate of change, we can examine the sample *covariance* or *correlation* between intercepts and slopes.

Formal answers to these questions require the multilevel model for change of chapter 3. But we can presage this work by conducting simple descriptive analyses of the estimated intercepts and slopes. In addition to plotting their distribution (as in figure 2.4), we can examine standard descriptive statistics (means and standard deviations) and bivariate summaries (correlation coefficients) obtained using the data set that describes the separate fitted regression results in table 2.2.

We find it helpful to examine three specific quantities, the:

- *Sample means of the estimated intercepts and slopes.* The level-1 OLS-estimated intercepts and slopes are unbiased estimates of initial status and rate of change for each person. Their sample means are therefore unbiased estimates of the key features of the average observed change trajectory.
- *Sample variances (or standard deviations) of the estimated intercepts and slopes.* These measures quantify the amount of observed interindividual heterogeneity in change.
- *Sample correlation between the estimated intercepts and slopes.* This correlation summarizes the association between fitted initial status and fitted rate of change and answers the question: Are observed initial status and rate of change related?

Results of these analyses for the NYS data appear in table 2.3.

Across this sample, we find an average estimated intercept of 1.36 and an average estimated slope of 0.13. We therefore conclude that the average adolescent in this sample has an observed tolerance level of 1.36 at age 11 and that this increases by an estimated 0.13 points per year. The magnitude of the sample standard deviations (in comparison to their means) suggests that adolescents are scattered widely around both these averages. This tells us that the adolescents differ considerably in their fitted initial status and fitted rates of change. Finally, the correlation coefficient of −0.45 indicates a negative relationship between fitted initial status and fitted rate of change, suggesting that adolescents with greater

Table 2.3: Descriptive statistics for the individual growth parameters obtained by fitting separate within-person OLS regression models for *TOLERANCE* as a function of linear time ($n = 16$)

	Initial status (intercept)	Rate of change (slope)
Mean	1.36	0.13
Standard deviation	0.30	0.17
Bivariate correlation		−0.45

initial tolerance tend to become more tolerant less rapidly over time (although we must be cautious in our interpretation because of negative bias introduced by the presence of measurement error).

2.3.3 Exploring the Relationship between Change and Time-Invariant Predictors

Evaluating the impact of predictors helps you uncover systematic patterns in the individual change trajectories corresponding to interindividual variation in personal characteristics. For the NYS data, we consider two time-invariant predictors: *MALE* and *EXPOSURE*. Asking whether the observed tolerance trajectories differ by gender allows us to explore whether boys (or girls) are initially more tolerant of deviant behavior and whether they tend to have different annual rates of change. Asking whether the observed tolerance trajectories differ by early exposure to deviant behavior (at age 11) allows us to explore whether a child's fitted initial level of tolerance is associated with early exposure and whether the fitted rate of change in tolerance is related as well. All of these questions focus on *systematic interindividual differences in change.*

Graphically Examining Groups of Smoothed Individual Growth Trajectories

Plots of smoothed individual growth trajectories, displayed separately for groups distinguished by important predictor values, are valuable exploratory tools. If a predictor is categorical, display construction is straightforward. If a predictor is continuous, you can temporarily categorize its values. For example, we split *EXPOSURE* at its median (1.145) for the purposes of display. For numeric analysis, of course, we continue to use its continuous representation.

Figure 2.7 presents smoothed OLS individual growth trajectories separately by gender (upper pair of panels) and exposure (lower pair of

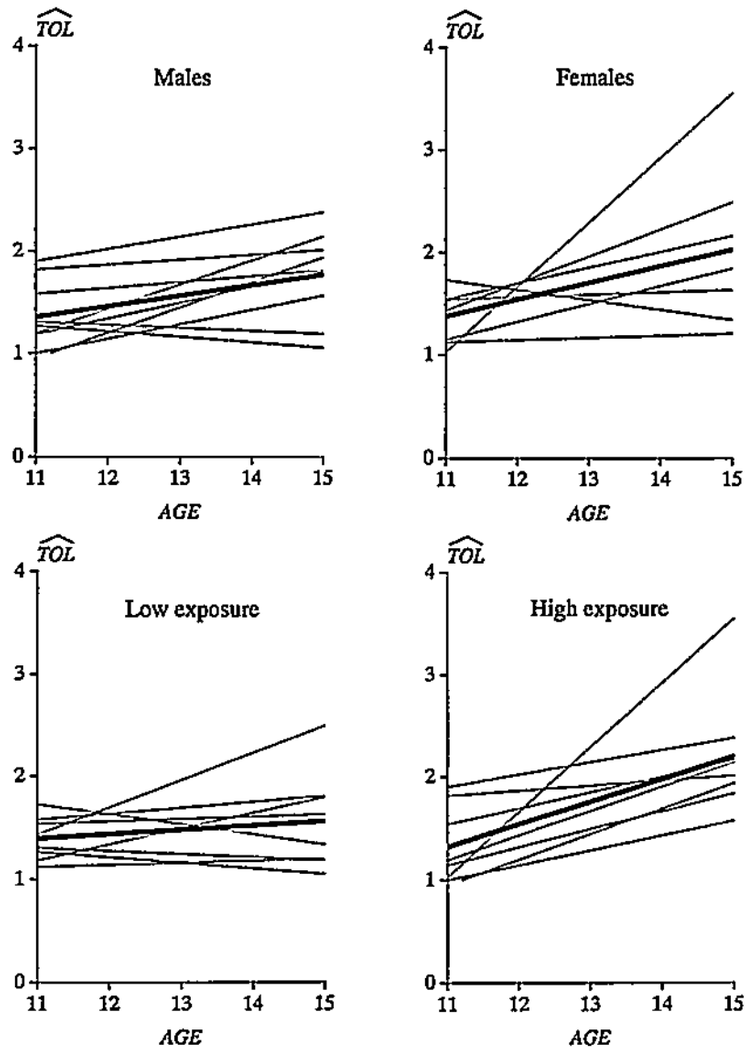


Figure 2.7. Identifying potential predictors of change by examining OLS fitted trajectories separately by levels of selected predictors. Fitted OLS trajectories for the tolerance data displayed separately by gender (upper panel) and exposure (lower panel).

panels). The bold trajectory in each panel depicts the average trajectory for the subgroup. When you examine plots like these, look for systematic patterns: Do the observed trajectories differ across groups? Do observed differences appear more in the intercepts or in the slopes? Are some groups' observed trajectories more heterogeneous than others'? Setting aside subject 978, who had extremely rapid growth, we find little difference in the distribution of fitted trajectories by gender. Each group's average observed trajectory is similar in intercept, slope, and scatter. We also find little difference in fitted initial status by exposure, but we do discern a difference in the fitted rate of change. Even discounting subject 978, those with greater initial exposure to deviant behavior seem to become tolerant more rapidly as they age.

The Relationship between OLS-Estimated Trajectories and Substantive Predictors

Just as we described the distribution of fitted intercepts and slopes in section 2.3, we can also use them as objects of further exploratory analysis. To investigate whether fitted trajectories vary systematically with predictors, we can treat the estimated intercepts and slopes as outcomes and explore the relationship between them and predictors. For the NYS data, these analyses explore whether the initial tolerance of deviant behavior or the annual rate of change in tolerance is observed to differ by: (1) gender or (2) early exposure to deviant behavior.

Because these analyses are exploratory—soon to be replaced in chapter 3 by the fitting of a multilevel model for change—we restrict ourselves to the simplest of approaches: the use of bivariate plots and sample correlations. Figure 2.8 plots the fitted intercepts and slopes versus the two predictors: *MALE* and *EXPOSURE*. Accompanying each plot is a sample correlation coefficient. All signs point to little or no gender differential in either fitted initial status or rate of change. But with respect to *EXPOSURE*, it does appear that adolescents with greater early exposure to deviant behavior become more tolerant at a faster rate than peers who were less exposed.

Despite their utility for descriptive and exploratory analyses, OLS estimated intercepts and slopes are hardly the final word in the analysis of change. Estimates are not true values—they are imperfect measures of each person's true initial status and true rate of change. They have biases that operate in known directions; for example, their sample variances are inflated by the presence of measurement error in the outcome. This means that the variance in the true rate of change will necessarily be smaller than the variance of the fitted slope because part of the latter's

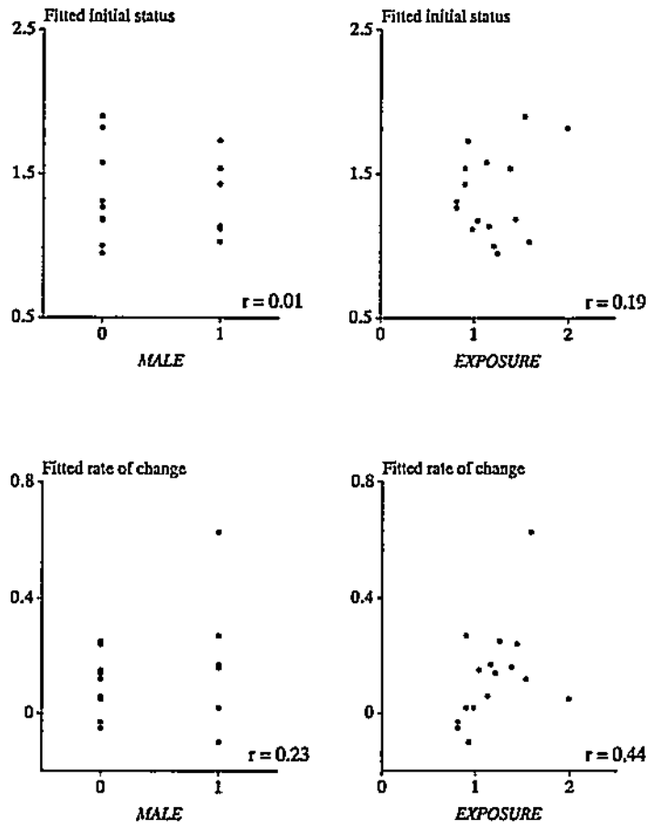


Figure 2.8. Examining the relationship between OLS parameter estimates (for initial status and rates of change) and potential predictors. Fitted OLS intercepts and slopes for the tolerance data plotted vs. two predictors: *MALE* and *EXPOSURE*.

variability is error variation. So, too, the sample correlation between the fitted intercept and slope is negatively biased (it underestimates the population correlation) because the measurement error in fitted initial status is embedded, with opposite sign, in the fitted rate of change.

These biases suggest that you should use the descriptive analyses of this chapter for exploratory purposes only. They can help you get your feet wet and in touch with your data. Although it is technically possible to improve these estimates—for example, we can deflate the sample variances of OLS estimates and we can correct the correlation coefficient for measurement error (Willett, 1989)—we do not recommend expending this extra effort. The need for ad hoc corrections has been effectively

replaced by the widespread availability of computer software for fitting the multilevel model for change directly.

2.4 Improving the Precision and Reliability of OLS-Estimated Rates of Change: Lessons for Research Design

Before introducing the multilevel model for change, let us examine another feature of the within-person exploratory OLS trajectories introduced in this chapter: the precision and reliability of the estimated rates of change. We do so not because we will be using these estimates for further analysis, but because it allows us to comment on—in a particularly simple arena—some fundamental principles of longitudinal design. As you would hope, these same basic principles also apply directly to the more complex models we will soon introduce.

Statisticians assess the precision of a parameter estimate in terms of its *sampling variation*, a measure of the variability that would be found across infinite resamplings from the same population. The most common measure of sampling variability is an estimate's *standard error*, the square root of its estimated sampling variance. Precision and standard error have an inverse relationship; the smaller the standard error, the more precise the estimate. Table 2.2 reveals great variability in the standard errors of the individual slope estimates for the NYS data. For some, the estimated rate of change is very precise (e.g., subjects 569 and 624); for others, it is not (e.g., subject 1552).

Understanding why the individual slope estimates vary in precision provides important insights into how you can improve longitudinal studies of change. Standard results from mathematical statistics tell us that the precision of an OLS-estimated rate of change depends upon an individual's: (1) residual variance, the vertical deviations of observed values around the fitted line; and (2) number and spacing of the waves of longitudinal data. If individual i has T waves of data, gathered at times $t_{i1}, t_{i2}, \ldots, t_{iT}$, the sampling variance of the OLS-estimated rate of change is[1]:

$$\left(\begin{array}{c}\text{Sampling variance}\\ \text{of the OLS rate of change}\\ \text{for individual } i\end{array}\right) = \frac{\sigma_{\varepsilon_i}^2}{\sum_{j=1}^{T}(t_{ij}-\bar{t}_i)^2} = \frac{\sigma_{\varepsilon_i}^2}{CSST_i}, \qquad (2.1)$$

where $\sigma_{\varepsilon_i}^2$ represents the residual variance for the *ith* individual and $CSST_i$ represents his or her corrected sum of squares for *TIME*, the sum of squared deviations of the time values around the average time, $\bar{t}_i$.

Equation 2.1 suggests two ways of increasing the precision of OLS estimated rates of change: (1) decrease the residual variance (because it appears in the numerator); or (2) increase variability in measurement times (because the corrected sums of squares for time appears in the denominator). Of course, the magnitude of the residual variance is largely outside your control; strictly speaking, you cannot directly modify its value. But because at least some of the residual variance is nothing more than measurement error, you can improve precision by using outcome measures with better psychometric properties.

Greater improvements in precision accrue if you work to increase the corrected sum of squares for time by modifying your research design. Inspection of equation 2.1 indicates that the greater the variability in the timing of measurement, the more precise the assessment of change. There are two simple ways of achieving increased variability in the timing of measurement: (1) redistribute the timing of the planned measurement occasions to be further away from their average; and (2) increase the number of waves. Both strategies yield substantial payoffs because it is the *squared* deviations of the measurement times about their average in the denominator of equation 2.1. A change as simple as adding another wave of data to your research design, far afield from the central set of observations, can reap dramatic improvements in the precision with which change can be measured.

We can reach similar conclusions by examining the reliability of the OLS estimated rates of change. Even though we believe that precision is a better criterion for judging measurement quality, we have three reasons for also examining reliability. First, the issue of reliability so dominates the literature on the measurement of change that it may be unwise to avoid all discussion. Second, it is useful to define reliability explicitly so as to distinguish it mathematically from precision. Third, even though reliability and precision are different criteria for evaluating measurement quality, they do, in this case, lead to similar recommendations about research design.

Unlike precision, which describes how well an individual slope estimate measures that person's true rate of change, reliability describes how much the rate of change varies across people. Precision has meaning for the individual; reliability has meaning for the group. Reliability is defined in terms of interindividual variation: it is the proportion of a measure's observed variance that is true variance. When test developers claim that a test has a reliability of .90 in a population, they mean that 90% of the person-to-person variation in observed scores across the population is variability in true scores.

Reliability of change is defined similarly. The population reliability of

the OLS slope is the proportion of population variance in observed rate of change that is variance in true rate of change (see Rogosa et al., 1982; Willett, 1988, 1989). If reliability is high, a large portion of the interindividual differences in observed rate of change will be differences in true rate of change. Were we to rank everyone in the population on their observed changes, we would then be pretty confident that the rankings reflect the rank order of the true changes. If reliability is low, the rankings on observed change might not reflect the true underlying rankings at all.

Improvements in precision generally lead to improvements in reliability —when you measure individual change more accurately, you can better distinguish individuals on the basis of these changes. But as a group-level parameter, reliability's magnitude is also affected by the amount of variability in true change in the population. If everyone has an identical value of true rate of change, you will be unable to effectively distinguish among people even if their observed rates of change are precise, so reliability will be zero. This means that you can simultaneously enjoy excellent individual precision for the rate of change and poor reliability for detecting interindividual differences in change; you can measure everyone's change well, but be unable to distinguish people because everyone's changes are identical. For a constant level of measurement precision, as population heterogeneity in true change increases, so does reliability.

The disadvantage of reliability as a gauge of measurement quality is that it confounds the effect of within-person precision with the effect of between-person heterogeneity in true change. When individual precision is poor or when interindividual heterogeneity in true change is small, reliability tends to 0. When precision is high or when heterogeneity in true change is large, reliability tends to 1. This means that reliability does not tell you uniquely about either precision or heterogeneity in true change; instead, it tells you about both simultaneously, impairing its value as an indicator of measurement quality.

We can confirm these inadequacies algebraically, albeit under a pair of limiting assumptions: (1) that the longitudinal data are fully balanced—everyone in the population is observed on the same set of occasions, $t_1, t_2, \ldots, t_T$; and (2) that each person's residuals are drawn identically and independently from a common distribution with variance σ_ε^2. The population reliability of the OLS estimate of individual rate of change is then:

$$\text{Reliability of the OLS rate of change} = \frac{\sigma^2_{TrueSlope}}{\sigma^2_{TrueSlope} + \dfrac{\sigma^2_\varepsilon}{CSST}}, \qquad (2.2)$$

where $\sigma^2_{True\,Slope}$ is the population variance of the true rate of change and *CSST* is the corrected sum-of-squares-time, now common across individuals (Willett, 1988). Because $\sigma^2_{True\,Slope}$ appears in both the numerator and denominator, it plays a central role in determining reliability. If everyone is growing at the same true rate, all true growth trajectories will be parallel and there will be no variability in the true rate of change across people. When this happens, both $\sigma^2_{True\,Slope}$ and the reliability of change will be 0, no matter how precisely the individual change is measured. Ironically, this means that the OLS slope can be a very precise yet completely unreliable measure of change. If there are large differences in the true rate of change across people, the true growth trajectories will crisscross considerably. When this happens, $\sigma^2_{True\,Slope}$ will be large, dominating both numerator and denominator, and the reliability of the OLS slope will tend to 1, regardless of its precision. This means that the OLS slope can be an imprecise yet reliable measure of change. The conclusion: you can be fooled about the quality of your change measurement if you use reliability as your sole criterion.

We can also use equation 2.2 to reinforce our earlier conclusions about longitudinal research design. First, for a given level of interindividual difference in true change in the population, the reliability of the OLS slope depends solely on the residual variance. Once again, the better the quality of your outcome measurement, the better the reliability with which change can be measured because at least part of the residual variance is simply measurement error. Second, reliability can be improved through design, by manipulating the number and spacing of the measurement occasions. Anything that you can do to increase corrected sum-of-squares time, *CSST*, will help. As you add waves of data or move the existing waves further away from the center of the data collection period, the reliability with which change can be measured will improve.

3

Introducing the Multilevel Model for Change

When you're finished changing, you're finished
—Benjamin Franklin

In this chapter, we introduce the multilevel model for change, demonstrating how it allows us to address within-person and between-person questions about change simultaneously. Although there are several ways of writing the statistical model, here we adopt a simple and common approach that has much substantive appeal. We specify the multilevel model for change by simultaneously postulating a pair of subsidiary models—a level-1 submodel that describes how each person changes over time, and a level-2 model that describes how these changes differ across people (Bryk & Raudenbush, 1987; Rogosa & Willett, 1985).

We begin, in section 3.1, by briefly reviewing the rationale and purpose of statistical models in general and the multilevel model for change in particular. We then introduce the level-1 model for individual change (section 3.2) and the level-2 model for interindividual heterogeneity in change (section 3.3). In section 3.4, we provide an initial foray into the world of estimation, introducing the method of maximum likelihood. (We discuss other methods of estimation in subsequent chapters.) We close, in sections 3.5 and 3.6, by illustrating how the resultant parameter estimates can be interpreted and how key hypotheses can be tested.

We do not intend this chapter to present a complete and general account of the multilevel model for change. Our goal is to provide a single "worked" example—from beginning to end—that illustrates all the steps you must go through when specifying the model, fitting it to data, and interpreting its results. We proceed in this way because we believe it is easier to learn about the model by first walking through a simple, but complete, analysis in a constrained, yet realistic, context. This minimizes notational and analytic complexity and lets us focus on interpretation and

understanding. As a result, this chapter is limited to: (1) a linear change model for individual growth; (2) a time-structured data set in which everyone shares an identical data collection schedule; (3) an evaluation of the impact of a single dichotomous time-invariant predictor; and (4) the use of one piece of dedicated statistical software, HLM. In subsequent chapters, we extend this basic model in many ways, generalizing it to situations in which growth is curvilinear or discontinuous; the timing, spacing, and number of waves of data differ across individuals; interest centers on the effects of many predictors, both discrete and continuous, time-invariant and time-varying; distributional assumptions differ; and other methods of estimation and statistical software are used.

3.1 What Is the Purpose of the Multilevel Model for Change?

Even though you have surely fit many types of statistical models in your data analytic career, experience tells us that when researchers get caught up in a novel and complex analysis, they often need to be reminded just what a statistical model is and what it is not. So before presenting the multilevel model for change itself, we briefly review the purpose of statistical models.

Statistical models are mathematical representations of population behavior; they describe salient features of the hypothesized process of interest among individuals in the target population. When you use a particular statistical model to analyze a particular set of data, you implicitly declare that *this* population model gave rise to *these* sample data. Statistical models are not statements about sample behavior; they are statements about the *population process* that generated the data.

To provide explicit statements about population processes, statistical models are expressed using parameters—intercepts, slopes, variances, and so on—that represent specific population quantities of interest. Were you to use the following simple linear regression model to represent the relationship between infant birth weight (in pounds) and neurological functioning on a single occasion in a cross-sectional data set (with the usual notation) $NEURO_i = \beta_0 + \beta_1 (BWGT_i - 3) + \varepsilon_i$, you would be declaring implicitly that, in the population from which your sample was drawn: (1) β_0 is an unknown intercept parameter that represents the expected level of neurological functioning for a three-pound newborn; and (2) β_1 is an unknown slope parameter that represents the expected difference in functioning between newborns whose birth weights differ by one pound. Even an analysis as simple as a one-sample t-test invokes a statistical model expressed in terms of an unknown population parameter: the population mean, μ. In conducting this test, you use sample data to evaluate the evidence concerning μ's value: Is μ equal to zero (or some other prespecified value)? Analyses may differ in form and function, but a statistical model underpins every inference.

In whatever context, having postulated a statistical model, you then fit the model to sample data and estimate the population parameters' unknown values. Most methods of estimation provide a measure of "goodness-of-fit"—such as an R^2 statistic or a residual variance—that quantifies the correspondence between the fitted model and sample data. If the model fits well, you can use the estimated parameter values to draw conclusions about the direction and magnitude of hypothesized effects in the population. Were you to fit the simple linear regression model just specified above, and find that $\widehat{NEURO}_i = 80 + 5(BWGT_i - 3)$, you would be able to predict that an average three-pound newborn has a functional level of 80 and that functional levels are five points higher for each extra pound at birth. Hypothesis tests and confidence intervals could then be used to make inferences from the sample back to the population.

The simple regression model above is designed for cross-sectional data. What kind of statistical model is needed to represent change processes in longitudinal data? Clearly, we seek a model that embodies two types of research questions: level-1 questions about *within-person change* and level-2 questions about *between-person differences in change*. If the hypothetical study of neurological functioning just described were longitudinal, we might ask: (1) How does each child's neurological functioning change over time? and (2) Do children's trajectories of change vary by birth weight? The distinction between the within-person and the between-person questions is more than cosmetic—it provides the core rationale for specifying a statistical model for change. It suggests that a model for change must include components at two levels: (1) a level-1 submodel that describes how individuals change over time; and (2) a level-2 submodel that describes how these changes vary across individuals. Taken together, these two components form what is known as a multilevel statistical model (Bryk & Raudenbush, 1987; Rogosa & Willett, 1985).

In this chapter, we develop and explain the multilevel model for change using an example of three waves of data collected by Burchinal and colleagues (1997). As part of a larger study of the effects of early intervention on child development, these researchers tracked the cognitive performance of 103 African-American infants born into low-income families. When the children were 6 months old, approximately half ($n = 58$) were randomly assigned to participate in an intensive early intervention program designed to enhance their cognitive functioning; the

Table 3.1: Excerpts from the person-period data set for the early intervention study

ID	*AGE*	*COG*	*PROGRAM*
68	1.0	103	1
68	1.5	119	1
68	2.0	96	1
70	1.0	106	1
70	1.5	107	1
70	2.0	96	1
71	1.0	112	1
71	1.5	86	1
71	2.0	73	1
72	1.0	100	1
72	1.5	93	1
72	2.0	87	1
...	...	...	...
902	1.0	119	0
902	1.5	93	0
902	2.0	99	0
904	1.0	112	0
904	1.5	98	0
904	2.0	79	0
906	1.0	89	0
906	1.5	66	0
906	2.0	81	0
908	1.0	117	0
908	1.5	90	0
908	2.0	76	0
...	...	...	...

other half ($n = 45$) received no intervention and constituted a control group. Each child was assessed 12 times between ages 6 and 96 months. Here, we examine the effects of program participation on changes in cognitive performance as measured by a nationally normed test administered three times, at ages 12, 18, and 24 months.

Table 3.1 presents illustrative entries from the person-period data set for this example. Each child has three records, one per wave of data collection. Each record contains four variables: (1) *ID*; (2) *AGE*, the child's age (in years) at each assessment (1.0, 1.5, or 2.0); (3) *COG*, the child's cognitive performance score at that age; and (4) *PROGRAM*, a dichotomy that describes whether the child participated in the early intervention program. Because children remained in their group for the duration of data collection, this predictor is time-invariant. Notice that all eight empirical growth records in table 3.1 suggest a decline in cognitive performance over time. As a result, although we might wish that we would be determining whether program participants experience a faster rate of *growth*, it appears that we will actually be determining whether they experience a slower rate of *decline*.

3.2 The Level-1 Submodel for Individual Change

The *level-1* component of the multilevel model, also known as the *individual growth model*, represents the change we expect each member of the population to experience during the time period under study. In the current example, the level-1 submodel represents the individual change in cognitive performance that we hypothesize will occur during each child's second year of life.

Whatever level-1 submodel we specify, we must believe that the observed data could reasonably have come from a population in which the model is functioning. To align expectations with reality, we usually precede level-1 submodel specification with visual inspection of the empirical growth plots (although purists might question the wisdom of "peeking"). Figure 3.1 presents empirical growth plots of *COG vs AGE* for the 8 children whose data appear in table 3.1. We also examined plots for the 95 other children in the sample but we do not present them here, to conserve space. The plots reinforce our perception of declining cognitive performance over time. For some, the decline appears smooth and systematic (subjects 71, 72, 904, 908); for others, it appears scattered and irregular (subjects 68, 70, 902, 906).

When examining empirical growth plots like these, with an eye toward ultimate model specification, we ask global questions such as: What type of population individual growth model might have generated these sample data? Should it be linear or curvilinear with age? Smooth or jagged? Continuous or disjoint? As discussed in chapter 2, try and look beyond inevitable sample zigs and zags because plots of observed data confound information on true change with the effects of random error. In these plots, for example, the slight nonlinearity with age for subjects 68, 70, 902, 906, and 908 might be due to the imprecision of the cognitive assessment. Often, and especially when you have few waves of data, it is difficult to argue for anything except a linear-change individual-growth model. So when we determine which trajectory to select for modeling change, we often err on the side of parsimony and postulate a simple linear model.[1]

Adopting an individual growth model in which change is a linear function of *AGE*, we write the level-1 submodel as:

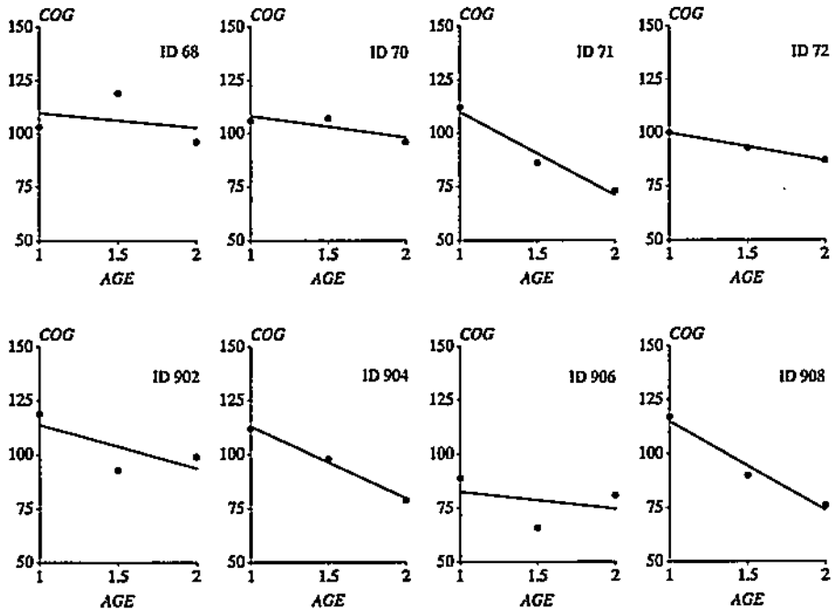


Figure 3.1. Identifying a suitable functional form for the level-1 submodel. Empirical growth plots with superimposed OLS trajectories for 8 participants in the early intervention study.

$$Y_{ij} = [\pi_{0i} + \pi_{1i}(AGE_{ij} - 1)] + [\varepsilon_{ij}]. \quad (3.1)$$

In postulating this submodel, we assert that, in the population from which this sample was drawn, Y_{ij}, the value of *COG* for child i at time j, is a linear function of his or her age on that occasion (AGE_{ij}). This model assumes that a straight line adequately represents each person's true change over time and that any deviations from linearity observed in sample data result from random measurement error (ε_{ij}).

Equation 3.1 uses two subscripts, i and j, to identify individuals and occasions, respectively. For these data, i runs from 1 through 103 (for the 103 children) and j runs from 1 through 3 (for the three waves of data). Although everyone in this data set was assessed on the same three occasions (ages 1.0, 1.5, and 2.0), the level-1 submodel in equation 3.1 is not limited in application to *time-structured* designs. The identical submodel could be used for data sets in which the timing and spacing of waves differs across people.[2] For now, we work with this time-structured

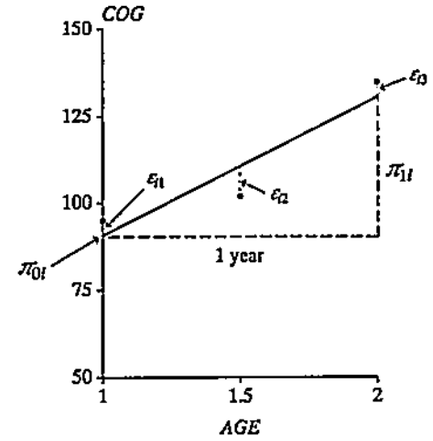


Figure 3.2. Understanding the structural and stochastic features of the level-1 individual growth model. Mapping the model in equation 3.1 onto imaginary data for child i, an arbitrarily selected member of the population.

example; in chapter 5, we extend our presentation to data sets in which data collection schedules vary across people.

In writing equation 3.1, we use brackets to distinguish two parts of the submodel: the *structural* part (in the first set of brackets) and the *stochastic* part (in the second). This distinction parallels the classical psychometric distinction between "true scores" and "measurement error," but as we discuss below, its implications are much broader.

3.2.1 The Structural Part of the Level-1 Submodel

The structural part of the level-1 submodel embodies our hypotheses about the shape of each person's *true trajectory of change* over time. Equation 3.1 stipulates that this trajectory is linear with age and has *individual growth parameters* π_{0i} and π_{1i} that characterize its shape for the ith child in the population. Harkening back to section 2.2.2, these individual growth parameters are the population parameters that lie beneath the individual intercepts and slopes obtained when we fit OLS-estimated individual change trajectories in our exploratory analyses.

To clarify what the individual growth model says about the population, examine figure 3.2, which maps the model onto imaginary data for an arbitrarily selected member of the population, child i. First notice the intercept. Because we specify the level-1 submodel using the predictor (*AGE-1*), the intercept, π_{0i}, represents child i's true cognitive performance at age 1. We concretize this interpretation in figure 3.2 by showing that the child's hypothesized trajectory intersects the Y axis at π_{0i}. Because we hypothesize that each child in the population has his or her own

intercept, this growth parameter includes the subscript i. Child 1's intercept is π_{01}, child 2's intercept is π_{02}, and so on.

Notice that equation 3.1 uses a special representation for the predictor, *AGE*. We used a similar approach in chapter 2, when we subtracted 11 from each adolescent's age before fitting exploratory OLS change trajectories to the tolerance data. This practice, known as *centering*, facilitates parameter interpretation. By using (*AGE-1*) as a level-1 predictor, instead of *AGE*, the intercept in equation 3.1 represents child i's true value of Y at age 1. Had we simply used *AGE* as a level-1 predictor, with no centering, π_{0i} would represent child i's true value of Y at age 0, an age that precedes the onset of data collection. This representation is less attractive because: (1) we would be predicting beyond the data's temporal limits; and (2) we don't know whether the trajectory extends back to birth linearly with age.

As you become adept at positing level-1 submodels, you will find that it is wise to consider empirical and interpretive issues like these when chosing the scale of your temporal predictor. In section 5.4, we explore other temporal representations, including those in which we center time on its *middle* and *final* values. The approach we adopt here—centering time on the first wave of data collection—is usually a good way to start. Aligning π_{0i} with the first wave of data collection allows us to interpret its value using simple nomenclature: it is child i's true *initial status*. If π_{0i} is large, child i has a high true initial status; if π_{0i} is small, child i has low true initial status. We summarize this interpretation in the first row of the top panel of table 3.2, which defines all parameters in equation 3.1.

The second parameter in equation 3.1, π_{1i}, represents the *slope* of the postulated individual change trajectory. The slope is the most important parameter in a level-1 linear change submodel because it represents the rate at which individual i changes over time. Because *AGE* is clocked in years, π_{1i} represents child i's true annual rate of change. We represent this parameter in figure 3.2 using the right triangle whose hypotenuse is the child's hypothesized trajectory. During the single year under study in our example—as child i goes from age 1 to 2—the trajectory rises by π_{1i}. Because we hypothesize that each individual in the population has his (or her) own rate of change, this growth parameter is subscripted by i. Child 1's rate of change is π_{11}, child 2's rate of change is π_{12}, and so on. If π_{1i} is positive, child i's true outcome increases over time; if π_{1i} is negative, child i's true outcome decreases over time (this latter case prevails in our example).

In specifying a level-1 submodel that attempts to describe everyone (all the i's) in the population, we implicitly assume that all the true individual change trajectories have a common algebraic form. But we do not assume that everyone has the same exact trajectory. Because each person

Table 3.2: Definition and interpretation of parameters in the multilevel model for change

	Symbol	Definition	Illustrative interpretation
Level-1 Model (See Equation 3.1)			
	π_{0i}	*Intercept* of the true change trajectory for individual i in the population.	Individual i's true value of *COG* at age 1 (i.e., his *true initial status*).
	π_{1i}	*Slope* of the true change trajectory for individual i in the population.	Individual i's yearly rate of change in true *COG* (i.e., his *true annual rate of change*).
Variance component	σ_ε^2	*Level-1 residual variance* across all occasions of measurement, for individual i in the population.	Summarizes the net (vertical) scatter of the observed data around individual i's hypothesized change trajectory.
Level-2 Model (See Equation 3.3)			
Fixed effects	γ_{00}	Population average of the level-1 intercepts, π_{0i}, for individuals with a level-2 predictor value of 0.	Population average true initial status for nonparticipants.
	γ_{01}	Population average difference in level-1 intercept, π_{0i}, for a 1-unit difference in the level-2 predictor.	Difference in population average true initial status between participants and nonparticipants.
	γ_{10}	Population average of the level-1 slopes, π_{1i}, for individuals with a level-2 predictor value of 0.	Population average annual rate of true change for nonparticipants.
	γ_{11}	Population average difference in level-1 slope, π_{1i}, for a 1-unit difference in the level-2 predictor.	Difference in population average annual rate of true change between participants and non-participants.
Variance components	σ_0^2	Level-2 residual variance in true intercept, π_{0i}, across all individuals in the population.	Population residual variance of true initial status, controlling for program participation.
	σ_1^2	*Level-2 residual variance in true slope*, π_{1i}, across all individuals in the population.	Population residual variance of true rate of change, controlling for program participation.
	σ_{01}	Level-2 residual covariance between true intercept, π_{0i}, and true slope, π_{1i}, across all individuals in the population.	Population residual covariance between true initial status and true annual rate of change, controlling for program participation.

has his or her own individual growth parameters (intercepts and slopes), different people can have their own distinct change trajectories.

Positing a level-1 submodel allows us to distinguish the trajectories of different people using just their individual growth parameters. This leap is the cornerstone of individual growth modeling because it means that we can study interindividual differences in change by studying interindividual variation in the growth parameters. Imagine a population in which each member dips into a well of possible individual growth parameter values and selects a pair—a personal intercept and a slope. These values then determine his or her true change trajectory. Statistically, we say that each person has drawn his or her individual growth parameter values from an underlying bivariate distribution of intercepts and slopes. Because each individual draws his or her coefficients from an unknown *random* distribution of parameters, statisticians often call the multilevel model for change a *random coefficients model.*

3.2.2 The Stochastic Part of the Level-1 Submodel

The *stochastic* part of the level-1 submodel appears in the second set of brackets on the right-hand side of equation 3.1. Composed of just one term, the stochastic part represents the effect of random error, ε_{ij}, associated with the measurement of individual i on occasion j. The level-1 errors appear in figure 3.2 as ε_{i1}, ε_{i2} and ε_{i3}. Each person's *true* change trajectory is determined by the structural component of the submodel. But each person's *observed* change trajectory also reflects the measurement errors. Our level-1 submodel accounts for these perturbations—the differences between the true and observed trajectories—by including random errors: ε_{i1} for individual i's first measurement occasion, ε_{i2} for individual i's second measurement occasion, and so on.

Psychometricians consider random errors a natural consequence of measurement fallibility and the vicissitudes of data collection. We think it wise to be less specific, labeling the ε_{ij} as *level-1 residuals.* For these data, each residual represents that part of child i's value of *COG* at time j not predicted by his or her age. We adopt this vaguer interpretation because we know that we can reduce the magnitude of the level-1 residuals by introducing selected time-varying predictors other than *AGE* into the level-1 submodel (as we show in section 5.3). This suggests that the stochastic part of the level-1 submodel is not just measurement error.

Regardless of how you conceptualize the level-1 errors, one thing is incontrovertible: they are *unobserved.* In ultimately fitting the level-1 submodel to data, we must invoke assumptions about the distribution of the level-1 residuals, from occasion to occasion and from person to person.

Traditional OLS regression invokes "classical" assumptions: that residuals are independently and identically distributed, with homoscedastic variance across occasions and individuals. This implies that, regardless of individual and occasion, each error is drawn independently from an underlying distribution with zero mean and an unknown residual variance. Often, we also stipulate the form of the underlying distribution, usually claiming normality. When we do, we can embody our assumptions about the level-1 residuals, ε_{ij}, by writing:

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \tag{3.2}$$

where the symbol ~ means "is distributed as," N stands for a normal distribution, and the first element in parentheses identifies the distribution's mean (here, 0) and the second element identifies its variance (here, σ_ε^2). As documented in table 3.2, the residual variance parameter σ_ε^2 captures the scatter of the level-1 residuals around each person's true change trajectory.

Of course, classical assumptions like these may be less credible in longitudinal data. When individuals change, their level-1 error structure may be more complex. Each person's level-1 residuals may be autocorrelated and heteroscedastic over time, not independent as equation 3.2 stipulates. Because the same person is measured on several occasions, any unexplained person-specific time-invariant effect in the residuals will create a correlation across occasions. So, too, the outcome may have a different precision (and reliability) for individuals at different times, perhaps being more suitable at some occasions than at others. When this happens, the error variance may differ over time and the level-1 residuals will be heteroscedastic over occasions within person. How does the multilevel model for change account for these possibilities? Although this is an important question, we cannot address it fully without further technical work. We therefore delay addressing the issues of residual autocorrelation and heteroscedasticity until chapter 4, where we show, in section 4.2, how the full multilevel model for change accommodates automatically for certain kinds of complex error structure. Later, in chapter 8, we go further and demonstrate how using covariance structure analysis to conduct analyses of change lets you hypothesize, implement, and evaluate other alternative error structures.

3.2.3 Relating the Level-1 Submodel to the OLS Exploratory Methods of Chapter 2

The exploratory OLS-fitted trajectories of section 2.2.2 may now make more sense. Although they are not fully efficient because they do not

as the *log-likelihood function*, sacrifices nothing because the values that maximize it also maximize the raw likelihood function. The transformation to logarithms simplifies the intensive numerical calculations involved because (1) the logarithm of a product is a *sum* of the separate logarithms, and (2) the logarithm of a term raised to a power is the power multiplied by the logarithm of the term. And so, since the sample likelihood contains both multiplicative and exponentiated terms, the logarithmic transformation moves the numerical maximization into a more tractable sphere, computationally speaking.

Although simpler than maximizing the likelihood function itself, maximizing the log-likelihood function also involves iteration. All software programs that provide ML estimates for the multilevel model for change use an iterative procedure. To begin, the program generates reasonable "starting" values for all model parameters, usually by applying something like the OLS methods we just rejected in chapter 2! In successive iterations, the program gradually refines these estimates as it searches for the log-likelihood function's maximum. When this search converges—and the difference between successive estimates is trivially small—the resultant estimates are output. If the algorithm does not converge (and this happens more often than you might like), you must repeat the search allowing more iterations or you must improve your model specification. (We discuss these issues in section 5.2.2.)

Once the ML estimates are found, it is relatively easy for a computer to estimate their associated sampling variation in the form of *asymptotic standard errors* (*ase*). We use the adjective "asymptotic" because, as noted earlier, ML standard errors are accurate only in large samples. Like any standard error, the *ase* measures the precision with which an estimate has been obtained—the smaller the *ase*, the more precise the estimate.

We now use maximum likelihood methods to fit the multilevel model in equations 3.1 and 3.3 to the early intervention data. Table 3.3 presents results obtained using the HLM software.[4] We first discuss the estimated fixed effects in the first four rows; in section 3.6, we discuss the estimated variance components shown in the next four rows.

3.5 Examining Estimated Fixed Effects

Empirical researchers usually conduct hypothesis tests before scrutinizing parameter estimates to determine whether an estimate warrants inspection. If an estimate is consistent with a null hypothesis of no population effect, it is unwise to interpret its direction or magnitude.

Table 3.3: Results of fitting a multilevel model for change to the early intervention data ($n = 103$)

		Parameter	Estimate	*ase*	*z*
Fixed Effects					
Initial status, π_{0i}	Intercept	γ_{00}	107.84***	2.04	52.97
	PROGRAM	γ_{01}	6.85*	2.71	2.53
Rate of change, π_{1i}	Intercept	γ_{10}	−21.13***	1.89	−11.18
	PROGRAM	γ_{11}	5.27*	2.52	2.09
Variance Components					
Level 1:	Within-person, ε_{ij}	σ^2_ε	74.24***	10.34	7.17
Level 2:	In initial status, ζ_{0i}	σ^2_0	124.64***	27.38	4.55
	In rate of change, ζ_{1i}	σ^2_1	12.29	30.50	0.40
	Covariance between ζ_{0i} and ζ_{1i}	σ_{01}	−36.41	22.74	−1.60

~ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$.
This model predicts cognitive functioning between ages 1 and 2 years as a function of (*AGE*-1) (at level-1) and *PROGRAM* (at level-2).

Note: Full ML, HLM.

Although we agree that it is wise to test hypotheses before interpreting parameters, here we reverse this sequence for pedagogic reasons, discussing interpretation in section 3.5.1 and testing in section 3.5.2. Experience convinces us that when learning a new statistical method, it is easier to understand what you are doing if you interpret parameters first and conduct tests second. This sequence emphasizes conceptual understanding over up-or-down decisions about "statistical significance" and ensures that you understand the hypotheses you test.

3.5.1 Interpreting Estimated Fixed Effects

The fixed effects parameters of the level-2 submodel—the γ's of equation 3.3—quantify the effects of predictors on the individual change trajectories. In our example, they quantify the relationship between the individual growth parameters and program participation. We interpret these estimates much as we do any regression coefficient, with one key difference: the level-2 "outcomes" that these fixed effects describe are the level-1 individual growth parameters themselves.

Until you are comfortable directly interpreting the output from software programs, we strongly recommend that you take the time to actually write down the structural portion of the fitted model before attempting to interpret the fixed effects. Although some software programs facilitate the linkage between model and estimates through

structured displays (e.g., MlwiN), others (e.g., SAS PROC MIXED) use somewhat esoteric conventions for labeling output. Substituting estimates $\hat{\gamma}$ in table 3.3 into the level-2 submodel in equation 3.3, we have:

$$\begin{aligned}\hat{\pi}_{0i} &= 107.84 + 6.85 PROGRAM_i \\ \hat{\pi}_{1i} &= -21.13 + 5.27 PROGRAM_i\end{aligned} \tag{3.5}$$

The first part of the fitted submodel describes the effects of *PROGRAM* on initial status; the second part describes its effects on the annual rates of change.

Begin with the first part of the fitted submodel, for initial status. In the population from which this sample was drawn, we estimate the true initial status (*COG* at age 1) for the average nonparticipant to be 107.84; for the average participant, we estimate that it is 6.85 points higher (114.69). The means of both groups are higher than national norms (100 for this test). The age 1 performance of participants is 6.85 points higher than that of nonparticipants. Before concluding that this differential in initial status casts doubt on the randomization mechanism, remember that the intervention started *before* the first wave of data collection, when the children were already 6 months old. This modest seven-point elevation in initial status may reflect early treatment gains attained between ages 6 months and 1 year.

Next, examine the second part of the fitted submodel, for the annual rate of change. In the population from which this sample was drawn, we estimate the true annual rate of change for the average nonparticipant to be –21.13; for the average participant, we estimate it to be 5.27 points higher (–15.86). The average nonparticipant dropped over 20 points during the second year of life; the average participant dropped over 15. The cognitive functioning of both groups of children declines over time. As we suspected when we initially examined these data, the intervention slows the rate of decline.

Another way of interpreting fixed effects is to plot fitted trajectories for prototypical individuals. Even in a simple analysis like this, which involves just one dichotomous predictor, we find it invaluable to inspect prototypical trajectories visually. For this particular multilevel model, only two prototypes are possible: a program participant (*PROGRAM* = 1) and a nonparticipant (*PROGRAM* = 0). Substituting these values into equation 3.5 yields the estimated initial status and annual growth rates for each:

$$\text{When } PROGRAM = 0: \quad \begin{aligned}\hat{\pi}_{0i} &= 107.84 + 6.85(0) = 107.84 \\ \hat{\pi}_{1i} &= -21.13 + 5.27(0) = -21.13.\end{aligned}$$

$$\text{When } PROGRAM = 1: \quad \begin{aligned}\hat{\pi}_{0i} &= 107.84 + 6.85(1) = 114.69 \\ \hat{\pi}_{1i} &= -21.13 + 5.27(1) = -15.86.\end{aligned}$$

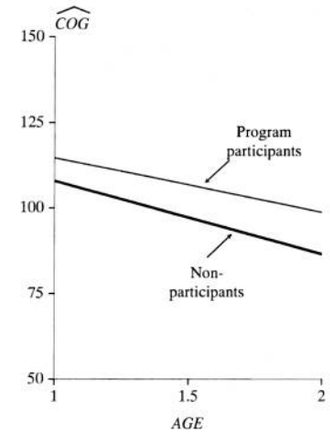


Figure 3.5. Displaying the results of a fitted multilevel model for change. Prototypical trajectories for an average program participant and nonparticipant in the early intervention data.

We use these estimates to plot the fitted individual change trajectories in figure 3.5. These plots reinforce the numeric conclusions just articulated. In comparison to nonparticipants, the average participant has a higher score at age 1 and a slower annual rate of decline.

3.5.2 Single Parameter Tests for the Fixed Effects

As in regular regression, you can conduct a hypothesis test on each fixed effect (each γ) using a single parameter test. Although you can equate the parameter value to any pre-specified value in your hypothesis test, most commonly you examine the null hypothesis that, controlling for all other predictors in the model, the population value of the parameter is 0, $H_0: \gamma = 0$, against the two-sided alternative that it is not, $H_1: \gamma \neq 0$. When you use ML methods, this test's properties are known only asymptotically (for exceptions, see note 3). You test this hypothesis for each fixed effect by computing the familiar z-statistic:

$$z = \frac{\hat{\gamma}}{ase(\hat{\gamma})}. \tag{3.7}$$

Most multilevel modeling programs provide z-statistics; if not, you can easily compute them by hand. However, care is needed because there is much looseness and inconsistency in output labels; terms like z-statistic, z-ratio, quasi-t-statistic, t-statistic, and t-ratio, which are not the same, are

change are correlated, after participation in the intervention program is accounted for.

For these data, we reject only two of these null hypotheses (each at the .001 level). The test for the level-1 residual, on σ^2_ε, suggests the existence of additional outcome variation at level-1, which may be predictable. To explain some of this remaining within-person variation, we might add suitable time-varying predictors such as the number of books in the child's home or the amount of parent-child interaction to the level-1 submodel.

The test for the level-2 residual for initial status, on σ^2_0, suggests the existence of additional variation in true initial status, π_{0i}, after accounting for the effects of program participation. This again suggests the need for additional predictors, but because this is a level-2 variance component (describing residual variation in true initial status), we would consider adding both time-invariant *and* time-varying predictors to the multilevel model.

We cannot reject the null hypotheses for the two remaining variance components. Failure to reject the null hypothesis for σ^2_1 indicates that *PROGRAM* explains all the potentially predictable variation between children in their true annual rates of change. Failure to reject the null hypothesis for σ_{01} indicates that the intercepts and slopes of the individual true change trajectories are uncorrelated—that there is no association between true initial status and true annual rates of change (once the effects of *PROGRAM* are removed). As we discuss in subsequent chapters, the results of these two tests might lead us to drop the second level-2 residual, ζ_{1i}, from our model, for neither its variance nor covariance with ζ_{0i}, is significantly different from 0.

4

Doing Data Analysis with the Multilevel Model for Change

> We are restless because of incessant change, but we would be frightened if change were stopped.
>
> —Lyman Bryson

In chapter 3, we used a pair of linked statistical models to establish the multilevel model for change. Within this representation, a level-1 submodel describes how each person changes over time and a level-2 submodel relates interindividual differences in change to predictors. To introduce these ideas in a simple context, we focused on just one method of estimation (maximum likelihood), one predictor (a dichotomy), and a single multilevel model for change.

We now delve deeper into the specification, estimation, and interpretation of the multilevel model for change. Following introduction of a new data set (section 4.1), we present a *composite* formulation of the model that combines the level-1 and level-2 submodels together into a single equation (section 4.2). The new composite model leads naturally to consideration of alternative methods of estimation (section 4.3). Not only do we describe two new methods—*generalized least squares* (GLS) and *iterative generalized least squares* (IGLS)—within each, we distinguish further between two types of approaches, the *full* and the *restricted*.

The remainder of the chapter focuses on real-world issues of data analysis. Our goal is to help you learn how to articulate and implement a coherent approach to model fitting. In section 4.4, we present two "standard" multilevel models for change that you should always fit initially in any analysis—the *unconditional means* model and the *unconditional growth* model—and we discuss how they provide invaluable baselines for subsequent comparison. In section 4.5, we discuss strategies for adding time-invariant predictors to the multilevel model for change. We then discuss methods for testing complex hypotheses (sections 4.6 and 4.7) and examining model assumptions and residuals (section 4.8). We conclude,

in section 4.9, by recovering "model-based" estimates of the individual growth trajectories that improve upon the exploratory person-by-person OLS estimates introduced in chapter 3. To highlight concepts and strategies rather than technical details, we continue to limit our presentation in several ways, by using: (1) a linear individual growth model; (2) a time-structured data set in which everyone shares the same data collection schedule; and (3) a single piece of statistical software (MLwiN).

4.1 Example: Changes in Adolescent Alcohol Use

As part of a larger study of substance abuse, Curran, Stice, and Chassin (1997) collected three waves of longitudinal data on 82 adolescents. Each year, beginning at age 14, the teenagers completed a four-item instrument assessing their alcohol consumption during the previous year. Using an 8-point scale (ranging from 0 = "not at all" to 7 = "every day"), adolescents described the frequency with which they (1) drank beer or wine, (2) drank hard liquor, (3) had five or more drinks in a row, and (4) got drunk. The data set also includes two potential predictors of alcohol use: *COA*, a dichotomy indicating whether the adolescent is a *c*hild *o*f an *a*lcoholic parent; and *PEER*, a measure of alcohol use among the adolescent's peers. This latter predictor was based on information gathered during the initial wave of data collection. Participants used a 6-point scale (ranging from 0 = "none" to 5 = "all") to estimate the proportion of their friends who drank alcohol occasionally (one item) or regularly (a second item).

In this chapter, we explore whether individual trajectories of alcohol use during adolescence differ according to the history of parental alcoholism and early peer alcohol use. Before proceeding, we note that the values of the outcome we analyze, *ALCUSE*, and of the continuous predictor, *PEER*, are both generated by computing the *square root* of the mean of participants' responses across each variable's constituent items. Transformation of the outcome allows us to assume linearity with *AGE* at level-1; transformation of the predictor allows us to assume linearity with *PEER* at level-2. Otherwise, we would need to posit nonlinear models at both levels in order to avoid violating the necessary linearity assumptions. If you find these transformations unsettling, remember that each item's original scale was arbitrary, at best. As in regular regression, analysis is often clearer if you fit a linear model to transformed variables instead of a nonlinear model to raw variables. We discuss this issue further when we introduce strategies for evaluating the tenability of the multilevel model's assumptions in section 4.8, and we explicitly introduce models that relax the linearity assumption in chapter 6.

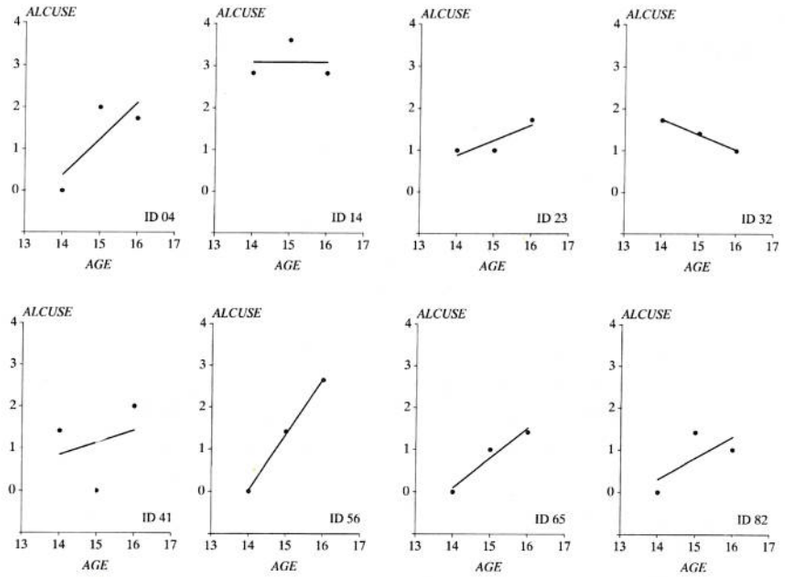


Figure 4.1. Identifying a suitable functional form for the level-1 submodel. Empirical growth plots with superimposed OLS trajectories for 8 participants in the alcohol use study.

To inform model specification, figure 4.1 presents empirical change plots with superimposed OLS-estimated linear trajectories for 8 adolescents randomly selected from the larger sample. For them, and for most of the other 74 not shown, the relationship between (the now-transformed) *ALCUSE* and *AGE* appears linear between ages 14 and 16. This suggests that we can posit a level-1 individual growth model that is linear with adolescent age $Y_{ij} = \pi_{0i} + \pi_{1i}(AGE_{ij} - 14) + \varepsilon_{ij}$, where Y_{ij} is adolescent i's value of *ALCUSE* on occasion j and AGE_{ij} is his or her age (in years) at that time. We have centered *AGE* on 14 years (the age at the first wave of data collection) to facilitate interpretation of the intercept.

As you become comfortable with model specification, you may find it easier to write the level-1 submodel using a generic variable $TIME_{ij}$ instead of a specific temporal predictor like $(AGE_{ij} - 14)$:

$$Y_{ij} = \pi_{0i} + \pi_{1i} TIME_{ij} + \varepsilon_{ij}. \qquad (4.1)$$

This representation is general enough to apply to all longitudinal data sets, regardless of outcome or time scale. Its parameters have the usual interpretations. In the population from which this sample was drawn:

- π_{0i} represents individual i's true initial status, the value of the outcome when $TIME_{ij} = 0$.
- π_{1i} represents individual i's true rate of change during the period under study.
- ε_{ij} represents that portion of individual i's outcome that is unpredicted on occasion j.

We also continue to assume that the ε_{ij} are independently drawn from a normal distribution with mean 0 and variance σ^2_ε. They are also uncorrelated with the level-1 predictor, *TIME*, and are homoscedastic across occasions.

To inform specification of the level-2 submodel, figure 4.2 presents exploratory OLS-fitted linear change trajectories for a random sample of 32 of the adolescents. To construct this display, we twice divided this subsample into two groups: once by *COA* (top panel) and again by *PEER* (bottom panel). Because *PEER* is continuous, the bottom panel represents a split at the sample mean. Thicker lines represent coincident trajectories—the thicker the line, the more trajectories. Although each plot suggests considerable interindividual heterogeneity in change, some patterns emerge. In the top panel, ignoring a few extreme trajectories, children of alcoholic parents have generally higher intercepts (but no steeper slopes). In the bottom panel, adolescents whose young friends drink more appear to drink more themselves at age 14 (that is, they tend to have higher intercepts), but their alcohol use appears to increase at a slower rate (they tend to have shallower slopes). This suggests that both *COA* and *PEER* are viable predictors of change, each deserving further consideration.

We now posit a level-2 submodel for interindividual differences in change. For simplicity, we focus only on *COA*, representing its hypothesized effect using the two parts of the level-2 submodel, one for true initial status (π_{0i}) and a second for true rate of change (π_{1i}):

$$\begin{aligned} \pi_{0i} &= \gamma_{00} + \gamma_{01} COA_i + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \gamma_{11} COA_i + \zeta_{1i}. \end{aligned} \tag{4.2}$$

In the level-2 submodel:

- γ_{00} and γ_{10}, the level-2 intercepts, represent the population average initial status and rate of change, respectively, for the child of a non-alcoholic ($COA = 0$). If both parameters are 0, the average child whose parents are non-alcoholic uses no alcohol at age 14 and does not change his or her alcohol consumption between ages 14 and 16.

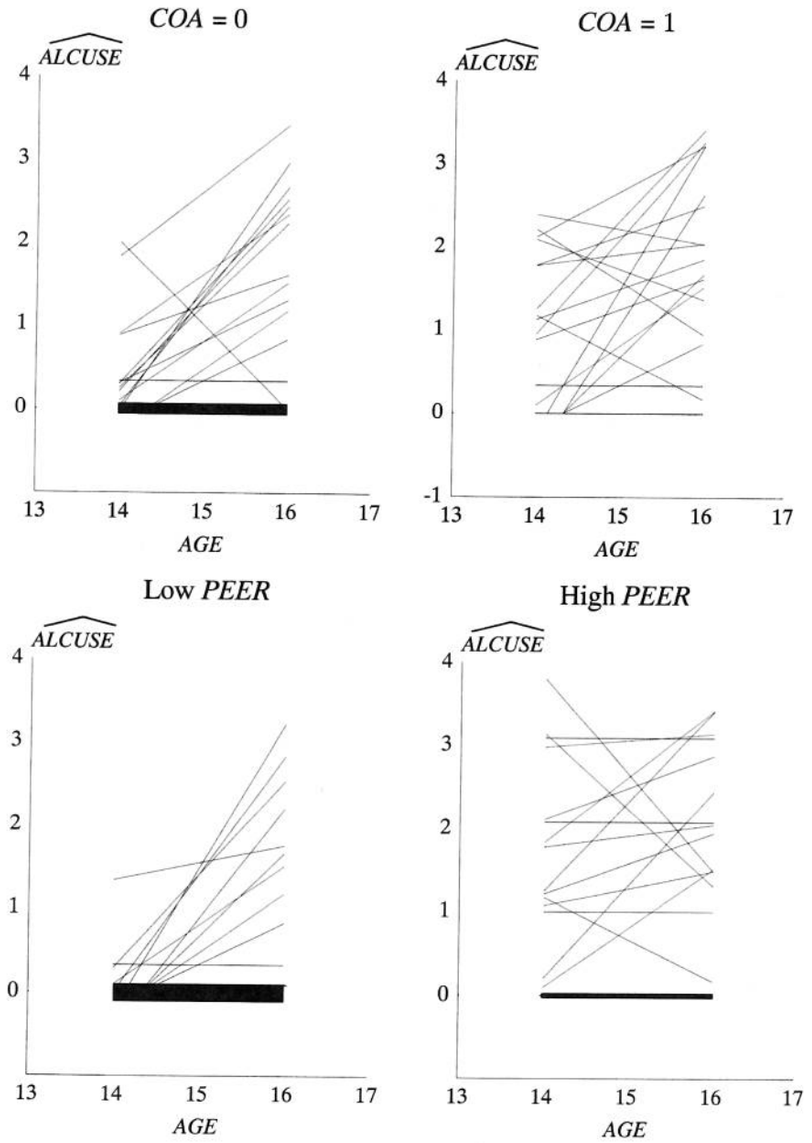


Figure 4.2. Identifying potential predictors of change by examining OLS fitted trajectories separately by levels of selected predictors. Fitted OLS trajectories for the alcohol use data displayed separately by *COA* status (upper panel) and *PEER* alcohol use (lower panel).

- γ_{01} and γ_{11}, the level-2 slopes, represent the effect of *COA* on the change trajectories, providing increments (or decrements) to initial status and rates of change, respectively, for children of alcoholics. If both parameters are 0, the average child of an alcoholic initially uses no more alcohol than the average child of a nonalcoholic and the rates of change in alcohol use do not differ as well.
- ζ_{0i} and ζ_{1i}, the level-2 residuals, represent those portions of initial status or rate of change that are unexplained at level-2. They represent deviations of the individual change trajectories around their respective group average trends.

We also continue to assume that ζ_{0i} and ζ_{1i} are independently drawn from a bivariate normal distribution with mean 0, variances σ_0^2 and σ_1^2, and covariance σ_{01}. They are also uncorrelated with the level-2 predictor, *COA*, and are homoscedastic over all values of *COA*.

As in regular regression analysis, we can modify the level-2 submodel to include other predictors—for example, replacing *COA* with *PEER* or adding *PEER* to the current model. We illustrate these modifications in section 4.5. For now, we continue with a single level-2 predictor so that we can introduce a new idea: the creation of the *composite* multilevel model for change.

4.2 The Composite Specification of the Multilevel Model for Change

The level-1/level-2 representation above is not the only specification of the multilevel model for change. A more parsimonious representation arises if you collapse the level-1 and level-2 submodels together algebraically into a single *composite* model. The composite representation, while identical to the level-1/level-2 specification mathematically, provides an alternative way of codifying hypotheses and is the specification required by many multilevel statistical software programs (including MLwiN and SAS PROC MIXED).

To derive the composite specification, first notice that any pair of linked level-1 and level-2 submodels share some common terms. Specifically, the individual growth parameters of the level-1 submodel are the outcomes of the level-2 submodel. We can therefore collapse the submodels together by substituting for π_{0i} and π_{1i} from the level-2 submodel (in equation 4.2, say) into the level-1 submodel (equation 4.1), as follows:

$$\begin{aligned} Y_{ij} &= \pi_{0i} + \pi_{1i} TIME_{ij} + \varepsilon_{ij} \\ &= (\gamma_{00} + \gamma_{01} COA_i + \zeta_{0i}) + (\gamma_{10} + \gamma_{11} COA_i + \zeta_{1i}) TIME_{ij} + \varepsilon_{ij} \end{aligned}$$

The first parenthesis contains the level-2 specification for the level-1 intercept, π_{0i}; the second parenthesis contains the level-2 specification for the level-1 slope, π_{1i}. Multiplying out and rearranging terms then yields the *composite multilevel model for change*:

$$\begin{aligned} Y_{ij} = &[\gamma_{00} + \gamma_{10} TIME_{ij} + \gamma_{01} COA_i + \gamma_{11} (COA_i \times TIME_{ij})] \\ &+ [\zeta_{0i} + \zeta_{1i} TIME_{ij} + \varepsilon_{ij}], \end{aligned} \tag{4.3}$$

where we once again use brackets to distinguish the model's structural and stochastic components.

Even though the composite specification in equation 4.3 appears more complex than the level-1/level-2 specification, the two forms are logically and mathematically equivalent. Each posits an identical set of links between an outcome (Y_{ij}) and predictors (here, *TIME* and *COA*). The specifications differ only in how they organize the hypothesized relationships, each providing valuable insight into what the multilevel model represents. The advantage of the level-1/level-2 specification is that it reflects our conceptual framework directly: we focus first on individual change and next on interindividual differences in change. It also provides an intuitive basis for interpretation because it directly identifies which parameters describe interindividual differences in initial status (γ_{00} and γ_{01}) and which describe interindividual differences in change (γ_{10} and γ_{11}). The advantage of the composite specification is that it clarifies which statistical model is actually being fit to data when the computer begins to iterate.

In introducing the composite model, we do not argue that its representation is uniformly superior to the level-1/level-2 specification. In the remainder of this book, we use both representations, adopting whichever best suits our purposes at any given time. Sometimes we invoke the substantively appealing level-1/level-2 specification; other times we invoke the algebraically parsimonious composite specification. Because both are useful, we recommend that you take the time to become equally facile with each. To aid in this process, below, we now delve into the structural and stochastic components of the composite model itself.

4.2.1 The Structural Component of the Composite Model

The structural portion of the composite multilevel model for change, in the first set of brackets in equation 4.3, may appear unusual, at least at first. Comfortingly, it contains all the original predictors—here, *COA* and *TIME*—as well as the now familiar fixed effects, γ_{00}, γ_{01}, γ_{10}, and γ_{11}. In

predictor values. Although you may be tempted to select many prototypical values for each predictor, we recommend that you limit yourself lest the displays become crowded, precluding the very interpretation they were intended to facilitate.

Prototypical values of predictors can be selected using one (or more) of the following strategies:

- *Choose substantively interesting values.* This strategy is best for categorical predictors or those with intuitively appealing values (such as 8, 12, and 16 for years of education in the United States).
- *Use a range of percentiles.* For continuous predictors without well-known values, consider using a range of percentiles (either the 25th, 50th, and 75th or the 10th, 50th, and 90th).
- *Use the sample mean ± .5 (or 1) standard deviation.* Another strategy useful for continuous predictors without well-known values.
- *Use the sample mean.* If you just want to control for the impact of a predictor rather than displaying its effect, set its value to the sample mean, yielding the "average" fitted trajectory controlling for that predictor.

Exposition is easier if you select whole number values (if the scale permits) or easily communicated fractions (e.g., ¼, ½, and ¾). When using sample data to obtain prototypical values, be sure to do the calculations on the time-invariant predictors in the original person data set, *not* the person-period data set. If you are interested in every substantive predictor in a model, display fitted trajectories for all combinations of prototypical predictor values. If you want to focus on certain predictors while statistically controlling for others, eliminate clutter by setting the values of these latter variables to their means.

The right panel of figure 4.3 presents fitted trajectories for four prototypical adolescents derived from Model E. To construct this display we needed to select prototypical values for *PEER*. Based on its standard deviation of 0.726, we chose 0.655 and 1.381, values positioned a half a standard deviation from the sample mean (1.018). For ease of exposition, we label these "low" and "high" *PEER*. Using the level-1/level-2 specification, we calculate the fitted values as follows:

PEER	*COA*	Initial status ($\hat{\pi}_{0i}$)	Rate of change ($\hat{\pi}_{1i}$)
Low	No	$-0.314 + 0.695(0.655) + 0.571(0) = 0.142$	$0.425 - 0.151(0.655) = 0.326$
Low	Yes	$-0.314 + 0.695(0.655) + 0.571(1) = 0.713$	$0.425 - 0.151(0.655) = 0.326$
High	No	$-0.314 + 0.695(1.381) + 0.571(0) = 0.646$	$0.425 - 0.151(1.381) = 0.216$
High	Yes	$-0.314 + 0.695(1.381) + 0.571(1) = 1.217$	$0.425 - 0.151(1.381) = 0.216$

The fitted trajectories of alcohol use differ by both parental history of alcoholism and peer alcohol use. At each level of *PEER*, the trajectory for children of alchoholic parents is consistently above that of children of non-alchoholic parents. But *PEER* also plays a role. Fourteen-year-olds whose friends drink more tend to drink more at that age. Regardless of parental history, the fitted change trajectory for high *PEER* is above that of low *PEER*. But *PEER* has an inverse effect on the *change* in *ALCUSE* over time. The slope of the prototypical change trajectory is about 33% lower when *PEER* is high, regardless of parental history. We note that this negative impact is not sufficient to counteract the positive early effect of *PEER*. Despite the lower rates of change, the change trajectories when *PEER* is high never approach, let alone fall below, that of adolescents whose value of *PEER* is low.

4.5.4 Recentering Predictors to Improve Interpretation

When introducing the level-1 submodel in chapter 2, we discussed the interpretive benefits of recentering the predictor used to represent time. Rather than entering time as a predictor in its raw form, we suggested that you subtract a constant from each observed value, creating variables like *AGE-11* (in chapter 2), *AGE-1* (in chapter 3), and *AGE-14* (here in chapter 4). The primary rationale for temporal recentering is that it simplifies interpretation. If we subtract a constant from the temporal predictor, the intercept in the level-1 submodel, π_{0i}, refers to the true value of Y at that particular age—11, 1, or 14. If the constant chosen represents a study's first wave of data collection, we can simplify interpretation even further by referring to π_{0i} as individual i's true "initial status."

We now extend the practice of rescaling to time-invariant predictors like *COA* and *PEER*. To understand why we might want to recenter time-invariant predictors, reconsider Model E in tables 4.1 and 4.2. When it came to the level-2 fitted intercepts, $\hat{\gamma}_{00}$ and $\hat{\gamma}_{10}$, interpretation was difficult because each represents the value of a level-1 individual growth parameter—π_{0i} or π_{1i}—when *all* predictors in the associated level-2 model are 0. If a level-2 model includes many substantive predictors or if zero is not a valid value for one or more of them, interpretation of its fitted intercepts can be difficult. Although you can always construct prototypical change trajectories in addition to direct interpretation of parameters we often find it easier to recenter the substantive predictors *before* analysis so that direct interpretation of parameters is possible.

The easiest strategy for recentering a time-invariant predictor is to subtract its sample mean from each observed value. When we center a

predictor on its sample mean, the level-2 fitted intercepts represent the *average* fitted values of initial status (or rate of change). We can also recenter a time-invariant predictor by subtracting another meaningful value—for example, 12 would be a suitable centering constant for a predictor representing years of education among U.S. residents; 100 may be a suitable centering constant for scores on an IQ test. Recentering works best when the centering constant is substantively meaningful—either because it has intuitive meaning for those familiar with the predictor *or* because it corresponds to the sample mean. Recentering can be equally beneficial for continuous and dichotomous predictors.

Models F and G in tables 4.1 and 4.2 demonstrate what happens when we center the time-invariant predictors *PEER* and *COA* on their sample means. Each of these models is equivalent to Model E, our tentative "final" model, in that all include the effect of *COA* on initial status and the effect of *PEER* on both initial status and rate of change. The difference between models is that before fitting Model F, we centered *PEER* on its sample mean of 1.018 and before fitting Model G, we also centered *COA* on its sample mean of .451. Some software packages (e.g., HLM) allow you to center predictors by toggling a switch on an interactive menu; others (e.g., MLwiN and SAS PROC MIXED) require you to create a new variable using computer code (e.g., by computing *CPEER* = *PEER* − 1.018). Our only word of caution is that you should compute the sample mean in the *person-level* data set. Otherwise, you may end up giving greater weight to individuals who happen to have more waves of data (unless the person-period data set is fully balanced, as it is here).

To evaluate empirically how recentering affects interpretation, compare the last three columns of table 4.1 and notice what remains the same and what changes. The parameter estimates for *COA* and *PEER* remain identical, regardless of recentering. This means that conclusions about the effects of predictors like *PEER* and *COA* are unaffected: $\hat{\gamma}_{01}$ remains at 0.571, $\hat{\gamma}_{11}$ remains at 0.695, and $\hat{\gamma}_{12}$ remains at −0.151 (as do their standard errors). Also notice that each of the variance components remains unchanged. This demonstrates that our conclusions about the variance components for the level-1 and level-2 residuals are also unaffected by recentering level-2 predictors.

What *does* differ across Models E, F and G are the parameter estimates (and standard errors) for the *intercepts* in each level-2 submodel. These estimates change because they represent different parameters:

- If neither *PEER* nor *COA* are centered (Model E), the intercepts represent a child of non-alchoholic parents whose peers at age 14 were totally abstinent (*PEER* = 0 and *COA* = 0).
- If *PEER* is centered and *COA* is not (Model F), the intercepts represent a child of non-alchoholic parents with an *average* value of *PEER* (*PEER* = 1.018 and *COA* = 0).
- If *both PEER* and *COA* are centered (Model G), the intercepts represent an *average* study participant—someone with *average* values of *PEER* and *COA* (*PEER* = 1.018 and *COA* = 0.451).

Of course, this last individual does not really exist because only two values of *COA* are possible: 0 and 1. Conceptually, though, the notion of an *average* study participant has great intuitive appeal.

When we center *PEER* and not *COA* in Model F, the level-2 intercepts describe an "average" child of non-alchoholic parents: $\hat{\gamma}_{00}$ estimates his or her true initial status (0.394, $p < .001$) and $\hat{\gamma}_{10}$ estimates his or her true rate of change (0.271, $p < .001$). Notice that the latter estimate is unchanged from Model B, the unconditional growth model. When we go further and center both *PEER* and *COA* in Model G, each level-2 intercept is numerically identical to the corresponding level-2 intercept in the unconditional growth model (B).[3]

Given that Models E, F, and G are substantively equivalent, which do we prefer? The advantage of Model G, in which both *PEER* and *COA* are centered, is that its level-2 intercepts are comparable to those in the unconditional growth model (B). Because of this comparability, many researchers routinely center *all* time-invariant predictors—even dichotomies—around their grand means so that the parameter estimates that result from the inclusion of additional predictors hardly change. Model E has a different advantage: because each predictor retains its original scale, we need not remember which predictors are centered and which are not. The predictor identified is the predictor included.

But both of these preferences are context free; they do not reflect our specific research questions. When we consider not just algebra but research interests—which here focus on parental alcoholism—we find ourselves preferring Model F. We base this decision on the easy interpretability of parameters for the dichotomous predictor *COA*. Not only is zero a valid value, it is an especially meaningful one (it represents children of non-alchoholic parents). We therefore see little need to center its values to yield consistency in parameter estimates with the unconditional growth model. When it comes to *PEER*, however, we have a different preference. Because it is of less substantive interest—we view it as a control predictor—we see no need *not* to center its values. Our goal is to evaluate the effects of *COA* controlling for *PEER*. By centering *PEER* at its mean, we achieve the goal of statistical control and interpretations of the level-2 intercepts are reasonable and credible. For the remainder of

this chapter, we therefore adopt Model F as our "final model." (We continue to use quotes to emphasize that even this model might be set aside in favor of an alternative in subsequent analyses.)

4.6 Comparing Models Using Deviance Statistics

In developing the taxonomy in tables 4.1 and 4.2, we tested hypotheses on fixed effects and variance components using the single parameter approach of chapter 3. This testing facilitated our decision making and helped us determine whether we should render a simpler model more complex (as when moving from Model B to C) or a more complex model simpler (as when moving from Model D to E). As noted in section 3.6, however, statisticians disagree as to the nature, form, and effectiveness of these tests. The disagreement is so strong that some multilevel software packages do not routinely output these tests, especially for variance components. We now introduce an alternative method of inference—based on the *deviance statistic*—which statisticians seem to prefer. The major advantages of this approach are that it: (1) has superior statistical properties; (2) permits composite tests on several parameters simultaneously; and (3) conserves the reservoir of Type I error (the probability of incorrectly rejecting H_0 when it is true).

4.6.1 The Deviance Statistic

The easiest way of understanding the deviance statistic is to return to the principles of maximum likelihood estimation. As described in section 3.4, we obtain ML estimates by maximizing numerically the log-likelihood function, the logarithm of the joint likelihood of observing all the sample data actually observed. The log-likelihood function, which depends on the hypothesized model and its assumptions, contains all the unknown parameters (the γ's and σ's) and the sample data. ML estimates are those values of the unknown parameters (the $\hat{\gamma}$'s and $\hat{\sigma}$'s) that maximize the log-likelihood.

As a by-product of ML estimation, the computer determines the magnitude of the log-likelihood function for this particular combination of observed data and parameter estimates. Statisticians call this number the *sample log-likelihood* statistic, often abbreviated as LL. Every program that uses ML methods outputs the LL statistic (or a transformation of it). In general, if you fit several competing models to the same data, the larger the LL statistic, the better the fit. This means that if the models you compare yield negative LL statistics, those that are *smaller* in absolute value—i.e., closer to 0—fit better. (We state this obvious point explicitly as there has been some confusion in the literature about this issue.)

The *deviance statistic* compares log-likelihood statistics for two models: (1) the *current* model, the model just fit; and (2) a *saturated* model, a more general model that fits the sample data perfectly. For reasons explained below, deviance is defined as this difference multiplied by −2:

$$\text{Deviance} = -2[LL_{\text{current model}} - LL_{\text{saturated model}}]. \tag{4.15}$$

For a given set of data, deviance quantifies *how much worse* the current model is in comparison to the best possible model. A model with a small deviance statistic is nearly as good as any you can fit; a model with a large deviance statistic is much worse. Although the deviance statistic may appear unfamiliar, you have used it many times in regression analysis, where it is identical to the residual sum of squares, $\left(\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right)$.

To calculate a deviance statistic, you need the log-likelihood statistic for the saturated model. Fortunately, in the case of the multilevel model for change, this is easy because a saturated model contains as many parameters as necessary to achieve a perfect fit, reproducing every observed outcome value in the person-period data set. This means that the maximum of its likelihood function—the probability that it will perfectly reproduce the sample data—is 1. As the logarithm of 1 is 0, the log-likelihood statistic for the saturated model is 0. We can therefore drop the second term on the right-hand side of equation 4.15, defining the deviance statistic for the multilevel model for change as:

$$\text{Deviance} = -2LL_{\text{current model}}. \tag{4.16}$$

Because the deviance statistic is just −2 times the sample log-likelihood, many statisticians (and software packages) label it −2logL or −2LL. As befits its name, we prefer models with smaller values of deviance.

The multiplication by −2 invoked during the transition from log-likelihood to deviance is more than cosmetic. Under standard normal theory assumptions, the difference in deviance statistics between a pair of nested models fit to the identical set of data has a known distribution. This allows us to test hypotheses about differences in fit between competing models by comparing deviance statistics. The resultant *likelihood ratio tests* are so named because a difference of logarithms is equal to the logarithm of a ratio.

4.6.2 When and How Can You Compare Deviance Statistics?

Deviance statistics for the seven models fit to the alcohol use data appear in table 4.1. They range from a high of 670.16 for Model A to a low of

588.69 for Model D. We caution that you cannot directly interpret their magnitude (or sign). (Also notice that the deviance statistics for Models E, F, and G are identical. Centering one or more level-2 predictors has absolutely no effect on this statistic.)

To compare deviance statistics for two models, the models must meet certain criteria. At a minimum: (1) each must be estimated using the identical data; and (2) one must be *nested* within the other. The constancy of data criterion requires that you eliminate any record in the person-period data set that is missing for any variable in *either* model. A difference of even one record invalidates the comparison. The nesting criterion requires that you can specify one model by placing *constraints* on the parameters in the other. The most common constraint is to set one or more parameters to 0. A "reduced" model is nested within a "full" model if every parameter in the former also appears in the latter.

When comparing multilevel models for change, you must attend to a third issue before comparing deviance statistics. Because these models involve two types of parameters—fixed effects (the γ's) and variance components (the σ's)—there are three distinct ways in which full and reduced models can differ: in their fixed effects, in their variance components, or in some combination of each. Depending upon the method of estimation—full or restricted ML—only certain types of differences can be tested. This limitation stems from principles underlying the estimation methods. Under FML (and IGLS), we maximize the likelihood of the sample data; under RML (and RIGLS), we maximize the likelihood of the sample *residuals*. As a result, an FML deviance statistic describes the fit of the entire model (both fixed and random effects), but a RML deviance statistic describes the fit of only its stochastic portion of the model (because, during estimation, its fixed effects are assumed "known"). This means that if you have applied FML estimation, as we have here, you can use deviance statistics to test hypotheses about any combination of parameters, fixed effects, or variance components. But if you have used RML to fit the model, you can use deviance statistics to test hypotheses only about variance components. Because RML is the default method in some multilevel programs (e.g., SAS PROC MIXED), caution is advised. Before using deviance statistics to test hypotheses, be sure you are clear about which method of estimation you have used.

Having fit a pair of models that meets these conditions, conducting tests is easy. Under the null hypothesis that the specified constraints hold, the difference in deviance statistics between a full and reduced model (often called "delta deviance" or ΔD) is distributed asymptotically as a χ^2 distribution with degrees of freedom (*d.f.*) equal to the number of independent constraints imposed. If the models differ by one parameter, you have one degree of freedom for the test; if they differ by three parameters, you have three. As with any hypothesis test, you compare ΔD to a *critical value*, appropriate for that number of degrees of freedom, rejecting H_0 when the test statistic is large.[4]

4.6.3 Implementing Deviance-Based Hypothesis Tests

Because the models in table 4.1 were fit using Full IGLS, we can use deviance statistics to compare their goodness-of-fit, whether they differ by only fixed effects (as do Models B, C, D, and E, F, G) or both fixed effects and variance components (as does Model A in comparison to all others). Before comparing two models, you must: (1) ensure that the data set has remained the same across models (it does); (2) establish that the former is nested within the latter; and (3) compute the number of additional constraints imposed.

Begin with the two unconditional models. We obtain multilevel Model A from Model B by invoking three independent constraints: $\gamma_{10} = 0$, $\sigma_1^2 = 0$, and $\sigma_{01} = 0$. The difference in deviance statistics, $(670.16 - 636.61) = 33.55$, far exceeds 16.27, the .001 critical value of a χ^2 distribution on 3 *d.f.*, allowing us to reject the null hypothesis at the $p < .001$ level that all three parameters are simultaneously 0. We conclude that the unconditional growth model provides a better fit than the unconditional means model (a conclusion already suggested by the single parameter tests for *each* parameter).

Deviance-based tests are especially useful for comparing what happens when we simultaneously add one (or more) predictor(s) to each level-2 submodel. As we move from Model B to Model C, we add *COA* as a predictor of both initial status and rate of change. Noting that we can obtain the former by invoking two independent constraints on the latter (setting both γ_{01} and γ_{11} to 0) we compare the difference in deviance statistics of $(636.61 - 621.20) = 15.41$ to a χ^2 distribution on 2 *d.f.*. As this exceeds the .001 critical value (13.82), we reject the null hypothesis that both γ_{01} and γ_{11} are simultaneously 0. (We ultimately set γ_{11} to 0 because we are unable to reject its single parameter hypothesis test in Model D. Comparing Models D and E, which differ by only this term, we find a trivial difference in deviance of 0.01 on 1 *d.f.*).

You can also use deviance-based tests to compare nested models with identical fixed effects and different random effects. Although the strategy is the same, we raise this topic explicitly for two reasons: (1) if you use restricted methods of estimation (RML or RIGLS), these are the only types of deviance comparisons you can make; and (2) they address an

important question we have yet to consider: Must the complete set of random effects appear in every multilevel model?

In every model considered so far, the level-2 submodel for each individual growth parameter (π_{0i} and π_{1i}) has included a residual (ζ_{0i} or ζ_{1i}). This practice leads to the addition of *three* variance components: σ_0^2, σ_1^2, and σ_{01}. Must all three always appear? Might we sometimes prefer a more parsimonious model? We can address these questions by considering the consequences of removing a random effect. To concretize the discussion, consider the following extension of Model F, which eliminates the second level-2 residual, ζ_{1i}:

$$
\begin{aligned}
Y_{ij} &= \pi_{0i} + \pi_{1i} TIME_{ij} + \varepsilon_{ij} \\
\pi_{0i} &= \gamma_{00} + \gamma_{01} COA_i + \gamma_{02} CPEER_i + \zeta_{0i} \\
\pi_{1i} &= \gamma_{10} + \gamma_{12} CPEER_i,
\end{aligned}
$$

and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ and $\zeta_{0i} \sim N(0, \sigma_0^2)$. In the parlance of multilevel modeling, we have "fixed" the individual growth rates, preventing them from varying randomly across individuals (although we allow them to be related to *CPEER*). Removing this one level-2 residual (remember, residuals are *not* parameters) eliminates *two* variance components (which *are* parameters): σ_1^2 and σ_{01}.

Because the fixed effects in this reduced model are identical to those in Model F, we can test the joint null hypothesis that both σ_1^2 and σ_{01} are 0 by comparing deviance statistics. When we fit the reduced model to data, we obtain a deviance statistic of 606.47 (not shown in table 4.1). Comparing this to 588.70 (the deviance for Model F) yields a difference of 18.77. As this exceeds the .001 critical value of a χ^2 distribution with 2 *d.f.* (13.82), we reject the null hypothesis. We conclude that there is residual variability in the annual rate of change in *ALCUSE* that could potentially be explained by other level-2 predictors and that we should retain the associated random effects in our model.

4.6.4 AIC and BIC Statistics: Comparing Nonnested Models Using Information Criteria

You can test many important hypotheses by comparing deviance statistics for pairs of nested models. But as you become a more proficient data analyst, you may occasionally want to compare pairs of models that are not nested. You are particularly likely to find yourself in this situation when you would like to select between alternative models that involve *different* sets of predictors.

Suppose you wanted to identify which subset of interrelated predictors best captures the effect of a single underlying construct. You might, for

example, want to control statistically for the effects of parental socioeconomic status (*SES*) on a child outcome, yet you might be unsure which combination of many possible SES measures—education, occupation, or income (either maternal or paternal)—to use. Although you could use principal components analysis to construct summary measures, you might also want to compare the fit of alternative models with different subsets of predictors. One model might use only paternal measures; another might use only maternal measures; still another might be restricted only to income indicators, but for both parents. As these models would not be nested (you cannot recreate one by placing constraints on parameters in another), you cannot compare their fit using deviance statistics.

We now introduce two ad hoc criteria that you can use to compare the relative goodness-of-fit of such models: the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978). Like the deviance statistic, each is based on the log-likelihood statistic. But instead of using the LL itself, each "penalizes" (i.e., decreases) the LL according to pre-specified criteria. The AIC penalty is based upon the number of model parameters. This is because adding parameters—even if they have no effect—will increase the LL statistic, thereby decreasing the deviance statistic. The BIC goes further. Its penalty is based not just upon the number of parameters, but also on the sample size. In larger samples, you will need a larger improvement before you prefer a more complex model to a simpler one. In each case, the result is multiplied by −2 so that the information criterion's scale is roughly equivalent to that of the deviance statistic. (Note that the number of parameters you consider in the calculations differs under full and restricted ML methods.) Under full ML, both fixed effects and variance components are relevant. Under restricted ML, as you would expect, only the variance component parameters are relevant.

Formally, we write:

$$
\begin{aligned}
\text{Information criterion} &= -2[LL - (\text{scale factor})(\text{number of model parameters})] \\
&= \text{Deviance} + 2(\text{scale factor})(\text{number of model parameters}).
\end{aligned}
$$

For the AIC, the scale factor is 1; for the BIC, it is half the log of the sample size. This latter definition leaves room for some ambiguity, as it is not clear whether the sample size should be the number of individuals under study or the number of records in the person-period data set. In the face of this ambiguity, Raftery (1995) recommends the former formulation, which we adopt here.

AICs and BICs can be compared for any pair of models, regardless of whether one is nested within another, *as long as both are fit to the identical set of data.* The model with the smaller information criterion (either AIC or BIC) fits "better." As each successive model in table 4.1 is nested within a previous one, informal comparisons like these are unnecessary. But to illustrate how to use these criteria, let us compare Models B and C. Model B involves six parameters (two fixed effects and four variance components); Model C involves eight parameters (two additional fixed effects). In this sample of 82, we find that Model B has an AIC statistic of $636.6 + 2(1)(6) = 648.6$ and an BIC of $636.6 + 2(\ln(82)/2)(6) = 663.0$, while Model C has an AIC statistic of $621.2 + 2(1)(8) = 637.2$ and an BIC of $621.2 + 2(\ln(82)/2)(8) = 656.5$. Both criteria suggest that C is preferable to B, a conclusion we already reached via comparison of deviance statistics.

Comparison of AIC and BIC statistics is an "art based on science." Unlike the objective standard of the χ^2 distribution that we use to compare deviance statistics, there are few standards for comparing information criteria. While large differences suggest that the model with the smaller value is preferable, smaller differences are difficult to evaluate. Moreover, statisticians have yet to agree on what differences are "small" or "large." In his excellent review extolling the virtues of BIC, Raftery (1995) declares the evidence associated with a difference of 0–2 to be "weak," 2–6 to be "positive," 6–10 to be "strong," and over 10 to be "very strong." But before concluding that information criteria provide a panacea for model selection, consider that Gelman and Rubin (1995) declared these statistics to be "off-target and only by serendipity manage to hit the target in special circumstances" (p. 165). We therefore offer a cautious recommendation to examine information criteria and to use them for model comparison only when more traditional methods cannot be applied.

4.7 Using Wald Statistics to Test Composite Hypotheses About Fixed Effects

Deviance-based comparisons are not the only method of testing composite hypotheses. We now introduce the Wald statistic, a generalization of the "parameter estimate divided by its standard error" strategy for testing hypotheses. The major advantage of the Wald statistic is its generality: you can test composite hypotheses about multiple effects regardless of the method of estimation used. This means that if you use restricted methods of estimation, which prevent you from using deviance-based tests to compare models with different fixed effects, you still have a means of testing composite hypotheses about sets of fixed effects.

Suppose, for example, you wanted to test whether the entire true change trajectory for a particular type of adolescent—say, a child of non-alchoholic parents with an average value of *PEER*—differs from a "null" trajectory (one with zero intercept and zero slope). This is tantamount to asking whether the average child of non-alchoholic parents drinks no alcohol at age 14 and remains abstinent over time.

To test this composite hypothesis, you must first figure out the entire set of parameters involved. This is easier if you start with a model's composite representation, such as Model F: $Y_{ij} = \gamma_{00} + \gamma_{01}COA_i + \gamma_{02}CPEER_i + \gamma_{10}TIME_{ij} + \gamma_{12}CPEER_i \times TIME_{ij} + [\zeta_{0i} + \zeta_{1i}TIME_{ij} + \varepsilon_{ij}]$. To identify parameters, simply derive the true change trajectory for the focal group, here children of non-alchoholic parents with an average value of *CPEER*. Substituting $COA = 0$ and $CPEER = 0$ we have: $E[Y_j \mid COA = 0, CPEER = 0] = \gamma_{00} + \gamma_{01}(0) + \gamma_{02}(0) + \gamma_{10}TIME_{ij} + \gamma_{12}(0) \times TIME_{ij} = \gamma_{00} + \gamma_{10}TIME_{ij}$, where the expectation notation, E[. . .], indicates that this is the *average population trajectory* for the entire $COA = 0$, $CPEER = 0$ subgroup. Taking expectations eliminates the level-1 and level-2 residuals, because—like all residuals—they average to zero. To test whether this trajectory differs from the null trajectory in the population, we formulate the composite null hypothesis:

$$H_0: \gamma_{00} = 0 \text{ and } \gamma_{10} = 0. \tag{4.17}$$

This joint hypothesis is a composite statement about an entire population trajectory, not a series of separate independent statements about each parameter.

We now restate the null hypothesis in a generic form known as a *general linear hypothesis.* In this representation, each of the model's fixed effects is multiplied by a judiciously chosen constant (an integer, a decimal, a fraction, or zero) and then the sum of these products is equated to another constant, usually zero. This "weighted linear combination" of parameters and constants is called a *linear contrast.* Because Model F includes five fixed effects—even though only two are under scrutiny here—we restate equation 4.17 as the following general linear hypothesis:

$$\begin{aligned} H_0: 1\gamma_{00} + 0\gamma_{01} + 0\gamma_{02} + 0\gamma_{10} + 0\gamma_{12} &= 0 \\ 0\gamma_{00} + 0\gamma_{01} + 0\gamma_{02} + 1\gamma_{10} + 0\gamma_{12} &= 0. \end{aligned} \tag{4.18}$$

Although each equation includes all five fixed effects, the carefully chosen multiplying constants (the *weights*) guarantee that only the two focal parameters, γ_{00} and γ_{10}, remain viable in the statement. While this

5

Treating *TIME* More Flexibly

Change is a measure of time

—Edwin Way Teale

All the illustrative longitudinal data sets in previous chapters share two structural features that simplify analysis. Each is: (1) balanced—everyone is assessed on the identical number of occasions; and (2) time-structured—each set of occasions is identical across individuals. Our analyses have also been limited in that we have used only: (1) time-invariant predictors that describe immutable characteristics of individuals or their environment (except for *TIME* itself); and (2) a representation of *TIME* that forces the level-1 individual growth parameters to represent "initial status" and "rate of change."

The multilevel model for change is far more flexible than these examples suggest. With little or no adjustment, you can use the same strategies to analyze more complex data sets. Not only can the waves of data be irregularly spaced, their number and spacing can vary across participants. Each individual can have his or her own data collection schedule and the number of waves can vary without limit from person to person. So, too, predictors of change can be time-invariant or time-varying, and the level-1 submodel can be parameterized in a variety of interesting ways.

In this chapter, we demonstrate how you can fit the multilevel model for change under these new conditions. We begin, in section 5.1, by illustrating what to do when the number of waves is constant but their spacing is irregular. In section 5.2, we illustrate what to do when the number of waves per person differs as well; we also discuss the problem of missing data, the most common source of imbalance in longitudinal work. In section 5.3, we demonstrate how to include time-varying predictors in your data analysis. We conclude, in section 5.4, by discussing why and how you can adopt alternative representations for the main effect of *TIME*.

5.1 Variably Spaced Measurement Occasions

Many researchers design their studies with the goal of assessing each individual on an identical set of occasions. In the tolerance data introduced in chapter 2, each participant was assessed five times, at ages 11, 12, 13, 14, and 15. In the early intervention data introduced in chapter 3 and the alcohol use data introduced in chapter 4, each participant was assessed three times: at ages 12, 24, and 36 months or ages 14, 15, and 16 years. The person-period data sets from these time-structured designs are elegantly balanced, with a temporal variable that has an identical cadence for everyone under study (like *AGE* in tables 2.1, and 3.1).

Yet sometimes, despite a valiant attempt to collect time-structured data, actual measurement occasions will differ. Variation often results from the realities of fieldwork and data collection. When investigating the psychological consequences of unemployment, for example, Ginexi, Howe, and Caplan (2000) designed a time-structured study with interviews scheduled at 1, 5, and 11 months after job loss. Once in the field, however, the interview times varied considerably around these targets, with increasing variability as the study went on. Although interview 1 was conducted between 2 and 61 days after job loss, interview 2 was conducted between 111 and 220 days, and interview 3 was conducted between 319 and 458 days. Ginexi and colleagues could have associated the respondents' outcomes with the *target* interview times, but they argue convincingly that the number of days since job loss is a better metric for the measurement of time. Each individual in their study, therefore, has a *unique* data collection schedule: 31, 150, and 356 days for person 1; 23, 162, and 401 days for person 2; and so on.

So, too, many researchers design their studies knowing full well that the measurement occasions may differ across participants. This is certainly true, for example, of those who use an *accelerated cohort* design in which an age-heterogeneous cohort of individuals is followed for a constant period of time. Because respondents initially vary in age, and *age*, not *wave*, is usually the appropriate metric for analysis (see the discussion of time metrics in section 1.3.2), observed measurement occasions will differ across individuals. This is actually what happened in the larger alcohol-use study from which the small data set in chapter 4 was excerpted. Not only were those 14-year-olds re-interviewed at ages 15 and 16, concurrent samples of 15- and 16-year-olds were re-interviewed at ages 16 and 17 and ages 17 and 18, respectively. The advantage of an accelerated cohort design is that you can model change over a longer temporal period (here, the five years between ages 14 and 18) using fewer waves of data. Unfortunately, under the usual conditions, the data sets

First, use theory as a guide, play your own harshest critic, and determine whether your inferences are clouded by reciprocal causation. Second, if your data allow, consider coding time-varying predictors so that their values in each record in the person-period data set refer to a *previous* point in chronological time. After all, there is nothing about the multilevel model for change that requires contemporaneous data coding. Most researchers use contemporaneous values by default. Yet it is often more logical to link *prior* status on a predictor with current status on an outcome.

For example, in their study of conduct disorder (CD) in boys, Lahey and colleagues (1995) carefully describe three ways they coded the effect of time-varying predictors representing treatment:

> In each case, the treatment was considered to be present in a given year if that form of treatment had been provided during all or part of the *previous 12 months* (emphasis added).... In addition, the analyses of treatment were repeated using the cumulative number of years that the treatment had been received as the time-varying covariate to determine whether the accumulated number of years of treatment influenced the number of CD symptoms in each year. Finally, a 1-year time-lagged analysis was conducted to look at the effect of treatment on the number of CD symptoms in the following year. (p. 90)

By linking each year's outcomes to prior treatment data, the researchers diminish the possibility that their findings are clouded by reciprocal causation. So, too, by carefully describing several alternative coding strategies, each of which describes a predictor constructed from the prior year's data, the researchers appear more credible and thoughtful in their work.

How might we respond to questions about reciprocal causation in Ginexi and colleagues' (2000) study of the link between unemployment and depression? A critic might argue that individuals whose CES-D scores decline over time are more likely to find jobs than peers whose levels remain stable or perhaps increase. If so, the observed link between re-employment and CES-D scores might result from the effects of CES-D on employment, not employment on CES-D. To rebut this criticism, we emphasize that the re-employment predictor indicates whether the person is *currently* employed at each subsequent interview. As a result, the moment of re-employment is temporally prior to the collection of CES-D scores. This design feature helps ameliorate the possibility that the observed relationship between unemployment and depression is a result of reciprocal causation. Had the CES-D and re-employment data been collected simultaneously, it would have been more difficult to marshal this argument.

Our message is simple: just because you can establish a link between a time-varying predictor and a time-varying outcome does not guarantee that the link is causal. While longitudinal data can help resolve issues of temporal ordering, the inclusion of a time-varying predictor can muddy the very issues the longitudinal models were intended to address. Moreover, as we will show in the second half of this book, issues of reciprocal causation can be even thornier when studying event occurrence because the links between outcomes and predictors are often more subtle than the examples just presented suggest. This is not to say you should not include time-varying predictors in your models. Rather, it is to say that you must recognize the issues that such predictors raise and not naively assume that longitudinal data alone will resolve the problem of reciprocal causation.

5.4 Recentering the Effect of *TIME*

TIME is the fundamental time-varying predictor. It therefore makes sense that if recentering a substantive time-varying predictor can produce interpretive advantages, so, too, should recentering *TIME*. In this section, we discuss an array of alternative recentering strategies, each yielding a different set of level-1 individual growth parameters designed to address related, but slightly different, research questions.

So far, we have tended to recenter *TIME* so that the level-1 intercept, π_{0i}, represents individual i's true *initial status*. Of course, the moment corresponding to someone's "initial status" is context specific—it might be a particular chronological age in one study (e.g., age 3, 6.5, or 13) or the occurrence of a precipitating event in another (e.g., entry into or exit from the labor force). In selecting a sensible starting point, we seek an early moment, ideally during the period of data collection, inherently meaningful for the process under study. This strategy yields level-2 submodels in which all parameters are directly and intrinsically interpretable, and it ensures that the value of *TIME* associated with the intercept, π_{0i}, falls within *TIME*'s observed range. Not coincidentally, this approach also yields a level-1 submodel that reflects everyday intuition about intercepts as a trajectory's conceptual "starting point."

Although compelling, this approach is hardly sacrosanct. Once you are comfortable with model specification and parameter interpretation, a world of alternatives opens up. We illustrate some options using data from Tomarken, Shelton, Elkins, and Anderson's (1997) randomized trial evaluating the effectiveness of supplemental antidepressant medication for individuals with major depression. The study began with an overnight

Table 5.9: Alternative coding strategies for *TIME* in the antidepressant trial

WAVE	*DAY*	*READING*	*TIME OF DAY*	*TIME*	(*TIME* – 3.33)	(*TIME* – 6.67)
1	0	8 A.M.	0.00	0.00	–3.33	–6.67
2	0	3 P.M.	0.33	0.33	–3.00	–6.33
3	0	10 P.M.	0.67	0.67	–2.67	–6.00
4	1	8 A.M.	0.00	1.00	–2.33	–5.67
5	1	3 P.M.	0.33	1.33	–2.00	–5.33
6	1	10 P.M.	0.67	1.67	–1.67	–5.00
...						
11	3	3 P.M.	0.33	3.33	0.00	–3.33
...						
16	5	8 A.M.	0.00	5.00	1.67	–1.67
17	5	3 P.M.	0.33	5.33	2.00	–1.33
18	5	10 P.M.	0.67	5.67	2.33	–1.00
19	6	8 A.M.	0.00	6.00	2.67	–0.67
20	6	3 P.M.	0.33	6.33	3.00	–0.33
21	6	10 P.M.	0.67	6.67	3.33	0.00

hospital stay for 73 men and women who were already being treated with a nonpharmacological therapy that included bouts of sleep deprivation. During the pre-intervention night, the researchers prevented each participant from obtaining any sleep. The next day, each person was sent home with a week's worth of pills (placebo or treatment), a package of mood diaries (which use a five-point scale to assess positive and negative moods), and an electronic pager. Three times a day—at 8 A.M., 3 P.M., and 10 P.M.—during the next month, respondents were electronically paged and reminded to fill out a mood diary. Here we analyze the first week's data, focusing on the participants' positive moods. With full compliance, each person would have 21 assessments. Although two people were recalcitrant (producing only 2 and 12 readings), everyone else was compliant, filling out at least 16 forms.

Table 5.9 presents seven variables that represent related, but distinct, ways of clocking time. The simplest, *WAVE*, counts from 1 to 21; although great for data processing, its cadence has little intuitive meaning because few of us divide our weeks into 21 conceptual components. *DAY*, although coarse, has great intuitive appeal, but it does not distinguish among morning, afternoon, and evening readings. One way to capture this finer information is to add a second temporal variable, such as *READING* or *TIME OF DAY*. Although the metric of the former makes it difficult to analyze, the metric of the latter is easily understood: 0 for morning readings; 0.33 for afternoon readings; 0.67 for evening readings. (We could also use a 24-hour clock and assign values that were not equidistant.) Another way to distinguish within-day readings is to create a single variable that combines both aspects of time. The next three variables, *TIME*, *TIME* – 3.33, and *TIME* – 6.67, achieve this goal. The first, *TIME*, operates like our previous temporal variables—it is centered on initial status. The others are linear transformations of *TIME*: one centered on 3.33, the study's *midpoint*, and the other centered on 6.67, the study's *final* wave.

Having created these alternative variables, we could now specify a separate set of models for each. Instead of proceeding in this tedious fashion, let us write a general model that uses a generic temporal variable (T) whose values are centered around a generic constant (c):

$$Y_{ij} = \pi_{0i} + \pi_{1i}(T_{ij} - c) + \varepsilon_{ij}. \tag{5.12a}$$

We can then write companion level-2 models for the effect of treatment:

$$\begin{aligned} \pi_{0i} &= \gamma_{00} + \gamma_{01} TREAT_i + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \gamma_{11} TREAT_i + \zeta_{1i} \end{aligned} \tag{5.12b}$$

and invoke standard normal theory assumptions for the residuals. This same model can be used for most of the temporal variables in table 5.9 (except those that distinguish only between within-day readings).

Table 5.10 presents the results of fitting this general model using the three different temporal variables, *TIME*, *TIME* – 3.33, and *TIME* – 6.67. Begin with the initial status representation of *TIME*. Because we cannot reject null hypotheses for either linear change or treatment, we conclude that: (1) on average, there is no linear trend in positive moods over time ($\hat{\gamma}_{10} = -2.42$, *n.s.*); and (2) when the study began, the groups were indistinguishable ($\hat{\gamma}_{01} = -3.11$, *n.s.*) as randomization would have us expect. The statistically significant coefficient for the effect of *TREAT* on linear change ($\hat{\gamma}_{11} = 5.54$, $p < 0.05$) indicates that the trajectories' slopes differ. The prototypical trajectories in figure 5.5 illustrate these findings. On average, the two groups are indistinguishable initially, but over time, the positive mood scores of the treatment group increase while those of the control group decline. The statistically significant variance components for the intercept ($\hat{\sigma}_0^2 = 2111.33$, $p < .001$) and linear change ($\sigma_1^2 = 63.74$, $p < .001$) indicate that that substantial variation in these parameters has yet to be explained.

What happens as we move the centering constant from 0 (initial status), to 3.33 (the study's midpoint), to 6.67 (the study's endpoint)? As expected, some estimates remain identical, while others change. The general principle is simple: parameters related to the *slope* remain stable while those related to the *intercept* differ. On the stable side, we obtain

Table 5.10: Results of using alternative representations for the main effect of *TIME* when evaluating the effect of treatment on the positive mood scores in the antidepressant trial ($n = 73$)

			Temporal predictor in level-1 model		
		Parameter	*TIME*	(*TIME* − 3.33)	(*TIME* − 6.67)
Fixed Effects					
Level-1 intercept, π_{0i}	Intercept	γ_{00}	167.46***	159.40***	151.34***
			(9.33)	(8.76)	(11.54)
	TREAT	γ_{01}	−3.11	15.35	33.80*
			(12.33)	(11.54)	(15.16)
Rate of change, π_{1i}	Intercept	γ_{10}	−2.42	−2.42	−2.42
			(1.73)	(1.73)	(1.73)
	TREAT	γ_{11}	5.54*	5.54*	5.54*
			(2.28)	(2.28)	(2.28)
Variance Components					
Level-1:	within-person	σ_ε^2	1229.93***	1229.93***	1229.93***
Level-2:	In level-1 intercept	σ_0^2	2111.33***	2008.72***	3322.45***
	In rate of change	σ_1^2	63.74***	63.74***	63.74***
	Covariance	σ_{01}	−121.62*	90.83	303.28***
Goodness-of-fit					
	Deviance		12680.5	12680.5	12680.5
	AIC		12696.5	12696.5	12696.5
	BIC		12714.8	12714.8	12714.8

~$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$.
TIME is centered around initial status, middle status, and final status.

Note: Full ML, SAS PROC MIXED.

identical estimates for the linear rate of change ($\hat{\gamma}_{10} = -2.42$, *n.s.*) and the effect of treatment on that rate ($\hat{\gamma}_{11} = 5.54$, $p < 0.05$). So, too, we obtain identical estimates for the residual variance in the rate of change ($\hat{\sigma}_1^2 = 63.74$, $p < .001$) and the within-person residual variance ($\hat{\sigma}_\varepsilon^2 = 1229.93$). And, most important, the deviance, AIC and BIC statistics remain unchanged because these models *are* structurally identical.

Where these models differ is in the location of their trajectories' *anchors*, around their starting point, midpoint, or endpoint. Because the intercepts refer to these anchors, each model tests a different set of hypotheses about them. If we change c, we change the anchors, which changes the estimates and their interpretations. In terms of the general model in equations 5.12a and 5.12b, γ_{00} assesses the elevation of the population average change trajectory at time c; γ_{01} assesses the differential elevation of this trajectory at time c between groups; σ_0^2 assesses the population variance in true status at time c; and σ_{01} assesses the population

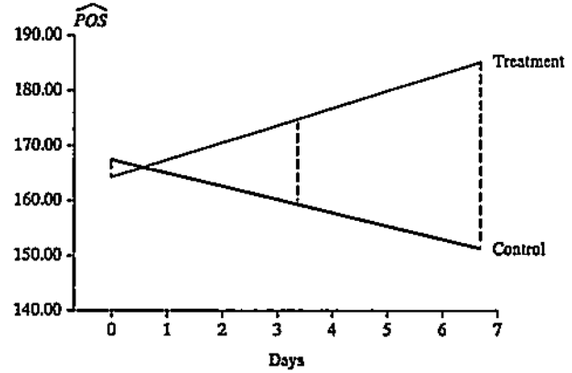


Figure 5.5. Understanding the consequences of rescaling the effect of *TIME*. Prototypical trajectories for individuals by *TREATMENT* status in the antidepressant experiment. The dashed vertical lines reflect the magnitude of the effect of *TREATMENT* if time is centered at the study's beginning (0), midpoint (3.33), and endpoint (6.67).

covariance between true status at time c and the per-unit rate of change in Y.

Although general statements like these are awkward, choice of a suitable centering constant can create simple, even elegant, interpretations. If we choose c to be 3.33, this study's midpoint, the intercept parameters assess effects at midweek. Because the treatment is still nonsignificant ($\hat{\gamma}_{01} = 15.35$, *n.s.*), we conclude that the average elevation of the two trajectories remains indistinguishable at this time. If we choose c to be 6.67, this study's endpoint, the intercept parameters assess effects at week's end. Doing so yields an important finding: Instead of reinforcing the expected nonsignificant early differences between groups, we now find a statistically significant treatment effect ($\hat{\gamma}_{01} = 33.80$, $p < .05$). After a week of antidepressant therapy, the positive mood score for the average member of the treatment group differs from that of the average member of the control group.

How can changing the centering constant for *TIME* have such a profound impact, especially since the fundamental model is unchanged? The dashed vertical lines in the prototypical plots in figure 5.5 provide an explanation. In adopting a particular centering constant, we cause the resultant estimates to describe the trajectories' behavior at that specific point in time. Changing the trajectory's anchor changes the location of the focal comparison. Of course, you could conduct *post hoc* tests of these contrasts (using methods of section 4.7) and obtain identical results. But

when doing data analysis, it is sometimes easier to establish level-1 parameters that automatically yield readymade tests for hypotheses of greatest interest. We urge you to identify a scale for *TIME* that creates a level-1 submodel with directly interpretable parameters. Initial status often works well, but there are alternatives. The midpoint option is especially useful when *total study duration* has intrinsic meaning; the endpoint option is especially useful when *final status* is of special concern.

Statistical considerations can also suggest the need to recenter *TIME*. As shown in table 5.10, a change in center can change the interpretation, and hence values, of selected random effects. Of particular note is the effect that a recentering can have on σ_{01}, the covariance between a level-1 model's intercept and slope. Not only can a recentering affect this parameter's magnitude, it can also affect its sign. In these data, the covariance between intercept and slope parameters moves from –121.62 to 90.83 to 303.28 as the centering constant changes. These covariances (and their associated variances) imply correlation coefficients of –0.33, 0.25, and 0.66, respectively. As you might imagine, were we to choose an even larger centering constant, outside the range of the data, it would be possible to find oneself specifying a model in which the correlation between parameters is close to 1.00. As Rogosa and Willett (1985) demonstrate, you can always alter the correlation between the level-1 growth parameters simply by changing the centering constant.

Understanding that the correlation between level-1 individual growth parameters can change through a change of centering constants has important analytic consequences. Recall that in section 5.2.2, we alluded to the possibility that you might encounter boundary constraints if you attempted to fit a model in which the correlation between intercept and slope is so high that iterative algorithms may not converge and you cannot find stable estimates. We now introduce the possibility that the correlation between true intercept and true slope can be so high as to preclude model fitting. When this happens, recentering *TIME* can sometimes ameliorate your problem.

There is yet another reason you might recenter time: it can sometimes lead to a simpler level-1 model. For this to work, you must ask yourself: Is there a centering constant that might totally eliminate the need for an explicit intercept parameter? If so, you could decrease the number of parameters needed to effectively characterize the process under study. This is precisely what happened in the work of Huttenlocher, Haight, Bryk, Seltzer, and Lyons (1991). Using a sample of 22 infants and toddlers, the researchers had data on the size of children's vocabularies at up to six measurement occasions between 12 and 26 months. Reasoning that there must be an age at which we expect children to have *no* words,

the researchers centered *TIME* on several early values, such as 9, 10, 11, and 12 months. In their analyses, they found that centering around age 12 months allowed them to eliminate the intercept parameter in their level-1 submodel, thereby dramatically simplifying their analyses.

We conclude by noting that there are other scales for *TIME* that alter not only a level-1 submodel's intercept but also its slope. It is possible, for example, to specify a model that uses neither a traditional intercept nor slope, but rather parameters representing initial and final status. To do so, you need to create two new temporal predictors, one to register each feature, and eliminate the stand-alone intercept term.

To fit a multilevel model for change in which the level-1 individual growth parameters refer to initial and final status, we write:

$$Y_{ij} = \pi_{0i}\left(\frac{max\ time - TIME_{ij}}{max\ time - min\ time}\right) + \pi_{1i}\left(\frac{TIME_{ij} - min\ time}{max\ time - mix\ time}\right) + \varepsilon_{ij}. \qquad (5.13a)$$

In the context of the antidepressant medication trial, in which the earliest measurement is at time 0 and the latest at time 6.67, we have:

$$Y_{ij} = \pi_{0i}\left(\frac{6.67 - TIME_{ij}}{6.67}\right) + \pi_{1i}\left(\frac{TIME_{ij}}{6.67}\right) + \varepsilon_{ij}.$$

Although it may not appear so, this model is identical to the other linear growth models; it is just that its parameters have new interpretations. This is true despite the fact that equation 5.13a contains no classical "intercept" term and *TIME* appears twice in two different predictors.

To see how the individual growth parameters in this model represent individual i's initial and final status, substitute the minimum and maximum values for *TIME* (0 and 6.67) and simplify. When *TIME* = 0, we are describing someone's initial status. At this moment, the second term of equation 5.13a falls out and the first term becomes π_{0i} so that individual i's initial status is $\pi_{0i} + \varepsilon_{ij}$. Similarly, when *TIME* = 6.67, we are describing someone's final status. At this moment, the first term of equation 5.13a falls out and the second term becomes π_{1i} so that individual i's final status is $\pi_{1i} + \varepsilon_{ij}$.

We can then specify standard level-2 submodels—for example:

$$\begin{aligned} \pi_{0i} &= \gamma_{00} + \gamma_{01} TREAT_i + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \gamma_{11} TREAT_i + \zeta_{1i} \end{aligned} \qquad (5.13b)$$

and invoke standard normal theory assumptions about the residuals. When we fit this model to data, we find the same deviance statistic we found before—12,680.5—reinforcing the observation that this model is identical to the three linear models in table 5.10. And when it comes to

the parameter estimates, notice the similarity between these and selected results in table 5.10:

$$\hat{\pi}_{0i} = 167.46 - 3.11TREAT_i$$
$$\hat{\pi}_{1i} = 151.34 + 33.80TREAT_i.$$

The first model provides estimates of initial status in the control group (167.46) and the differential in initial status in the treatment group (–3.11). The second model provides estimates of final status in the control group (151.34) and the differential in final status in the treatment group (33.80).

This unusual parameterization allows you to address questions about initial and final status simultaneously. Simultaneous investigation of these questions is superior to a piecemeal approach based on separate analyses of the first and last wave. Not only do you save considerable time and effort, you increase statistical power by using all the longitudinal data, even those collected at intermediate points in time.

6

Modeling Discontinuous and Nonlinear Change

Things have changed.

—Bob Dylan

All the multilevel models for change presented so far assume that individual growth is smooth and linear. Yet individual change can also be discontinuous or nonlinear. Patients' perceptions of their psychological well-being may abruptly shift when psychiatrists intervene and change their medications. Initial decreases in employee self-efficacy may gradually abate as new hires develop confidence with experience on the job.

This is not the first time we have confronted such possibilities. In the early intervention study of chapter 3, the trajectory of the child's cognitive development was nonlinear between infancy and age 12. To move forward and fit a model to these data, we focused on a narrower temporal period—the year of life between 12 and 24 months—in which the linearity assumption was tenable. In chapter 4, when changes in adolescent alcohol use seemed nonlinear, we transformed the outcome (and one of the predictors). Although the researchers used a nine-point scale to assess alcohol consumption, we analyzed the *square root* of scores on this scale, which yielded approximately linear change trajectories.

In this chapter, we introduce strategies for fitting models in which individual change is explicitly discontinuous or nonlinear. Rather than view these patterns as inconveniences, we treat them as substantively compelling opportunities. In doing so, we broaden our questions about the nature of change beyond the basic concepts of initial status and rate of change to a consideration of acceleration, deceleration, turning points, shifts, and asymptotes. The strategies that we use fall into two broad classes. *Empirical* strategies that let the "data speak for themselves." Under this approach, you inspect observed growth records systematically and identify a transformation of the outcome, or of *TIME*, that linearizes the

Suppose, for example, we assessed students three times a year (fall, winter, and spring) for each of three grades (third, fourth, and fifth). Rather than postulating a linear trajectory, we might hypothesize a discontinuous alternative. If we thought, for example, that students made general progress through grades, but that within a grade, there might be even steeper progress, we might postulate that: $Y_{ij} = \pi_{0i} + \pi_{1i}\,(GRADE_{ij} - 4) + \pi_{2i}SEASON_{ij} + \varepsilon_{ij}$, where both $(GRADE_{ij}-4)$ and $SEASON_{ij}$ take on the values −1, 0, and 1. In this model, π_{0i} represents individual i's true test score in the middle of fourth grade, π_{1i} represents his true rate of linear growth across grades, and π_{2i} represents any additional linear growth that occurs during the academic year. This model would yield a zigzag trajectory. Alternatively, if we thought that growth would generally be linear but that fall readings might be low because test scores may drop during the summer when children are out of school, we could postulate that: $Y_{ij} = \pi_{0i} + \pi_{1i}\,(\text{ASSESSMENT\#}_{ij} - 1) + \pi_{2i}FALL_{ij} + \varepsilon_{ij}$. In this model, π_{0i} represents individual i's true initial test score at the beginning of third grade (controlling for the fact that it is a fall assessment), π_{1i} represents his true rate of linear growth across assessments, and π_{2i} represents the potential decrement in test scores associated with the fall assessment. This model yields an underlying linear trajectory punctuated by fall-specific drops.

Please treat these examples not as explicit directives but as inspirational starting points. If you have reason to hypothesize a particular type of discontinuity, you should develop a customized model that reflects your hypothesis and not adopt an "off-the-shelf" parameterization that may not. Once you move away from the standard linear change trajectory, your options grow as does your burden of proof. Good theory and a compelling rationale should always be your guides.

6.2 Using Transformations to Model Nonlinear Individual Change

We now consider smooth, but nonlinear, individual change trajectories. Certainly the easiest strategy for fitting such models is to transform either the outcome, or *TIME*, in the level-1 submodel so that a growth model that specifies linear change in the transformed outcome or predictor will suffice. When confronted by obviously nonlinear trajectories, we usually begin with the transformation approach for two reasons. First, a straight line—even on a transformed scale—is a simple mathematical form whose two parameters have clear interpretations. Second, because the metrics of many variables are ad hoc to begin with, transformation to another ad hoc scale may sacrifice little. If the original scale lacks well-accepted intuitive anchors, you lose nothing by using a transformed alternative. It matters not whether you conduct analyses in one arbitrary world (the original metric) or another (e.g., the "square root" metric). Either metric allows you to track individuals over time and to identify predictors associated with their differential patterns of change.

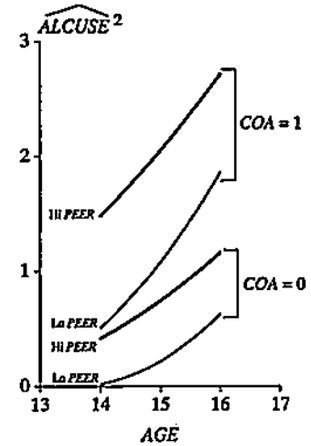


Figure 6.4. Re-expressing the prototypical trajectories in figure 4.3 for *ALCUSE* on the outcome's original scale. These prototypical trajectories are identical to those in figure 4.3 except that here we have squared the model's predicted values to reverse the effect of taking square roots before statistical analysis.

To support these assertions, reconsider the alcohol use data of chapter 4 and ask: What would those findings look like if we "de-transformed" the outcome back to its original nine-point scale? Figure 6.4 displays de-transformed trajectories based on the fitted trajectories in the final panel of figure 4.3. We obtained these trajectories by squaring the predicted values from the linear model fit to the square root data (thereby reversing the transformation, as "squaring" is the inverse of "square rooting"). As in figure 4.3, we display prototypical trajectories for children of alcoholics and nonalcoholics at low and high values of peer alcohol use. Reversing the transformation returns us to the original nine-point metric. Because this changes the scale of the vertical axis, the once linear change trajectories are now curved.

Despite this transition between metrics, the findings remain: children of alcoholics initially drink more but are no more likely to increase their drinking over time. But the "slopes" of the detransformed trajectories defy a "single number" summary. In the square root metric we originally analyzed, "annual rate of change" was meaningful because the trajectory was linear in the transformed world. But once we detransform back into the original measurement metric, the trajectory is curved and the rate of change is no longer constant over time: alcohol use increases more

rapidly as time passes. Although you might think there is a conflict between these representations, each interpretation is correct *in its own world.* In the transformed metric, change in alcohol use is linear—its rate of change is *constant* over time. In the original metric of measurement, which we enter by detransformation, change in alcohol use is nonlinear—it *accelerates* over time. Our current formulation of the multilevel model for change assumes a level-1 linear change model. If change is not linear over time, we can seek an alternative metric for the outcome, or for time, in which this assumption holds. In the transformed world, the methods we have developed work well and we violate no assumptions. We then display the findings back in the detransformed metric of the original outcome to simplify communication of the findings.

This suggests a simple general strategy for modeling nonlinear change that capitalizes on the best of both worlds. Transform the outcome (or the level-1 *TIME* predictor) so that individual change becomes linear. Fit the multilevel model for change and test hypotheses in the transformed world, then detransform and present findings back in the original metric. The key to the success of this strategy is selection of a suitable transformation, a topic to which we now turn.

6.2.1 The Ladder of Transformations and the Rule of the Bulge

You can identify a suitable transformation for "correcting" nonlinearity in longitudinal data using the same methods you use for "correcting" nonlinearity in cross-sectional data. Rather than examining a single "outcome *vs.* predictor" plot, however, you examine multiple empirical growth plots, one for each sample member, seeking a transformation that works decently for most everyone under study.

A useful aid in this process is Mosteller and Tukey's (1977) ordered list of transformations known as the ladder of powers. On the left side of figure 6.5, we present our version of their ladder for transforming a generic variable "*V*," which appears on the center rung. Transformations in the upper half of the ladder, above *V*, are positive powers greater than 1, including the *square*, the *cube*, and the *fourth power*. Transformations in the lower half, below *V*, include the *logarithm*, *fractional powers* (representing the *square root*, the *cube root*, etc.), and *negative powers* (*inverses*). When we use a transformation in the upper half of the ladder (e.g., V^2, V^3), we say we move "up" in *V*; when we use a transformation in the lower half (e.g., $LOG(V)$, $1/V$), we say we move "down" in *V*.

To identify a suitable transformation, inspect the collection of empirical growth plots and apply what Mosteller and Tukey call the *rule of the bulge*. We reprint their guidelines on the right side of figure 6.5. The idea is to match the general shape of the plots (discounting the effect of measurement error) to one of the four exemplars shown. You find linearizing transformations by moving "up" or "down" the ladder in the same direction(s) as the direction of the "bulge" in the exemplar. The arrows in figure 6.5 indicate the directions for each exemplar. In the upper left corner, the arrows point "up" in *Y* and "down" in *TIME*, suggesting that a curve with this shape can be linearized by moving "up" in *Y* (e.g., taking Y^2, Y^3 etc.) or "down" in *TIME* (e.g., taking $LOG(TIME)$, $1/TIME$ etc). In the bottom right corner, the arrows point "down" in *Y* and "up" in *TIME* suggesting that a curve with this shape can be linearized by moving "down" in *Y* or "up" in *TIME*. The further a transformation is from the ladder's center, the more dramatic its impact.

We suggest you experiment with several transformations before selecting one for analysis. The search process is hardly an exact science. Even for a single person, one transformation may not be equally successful at all points in time. As you eventually need to use the *same* transformation for everyone in your sample, selection involves some compromise so that,

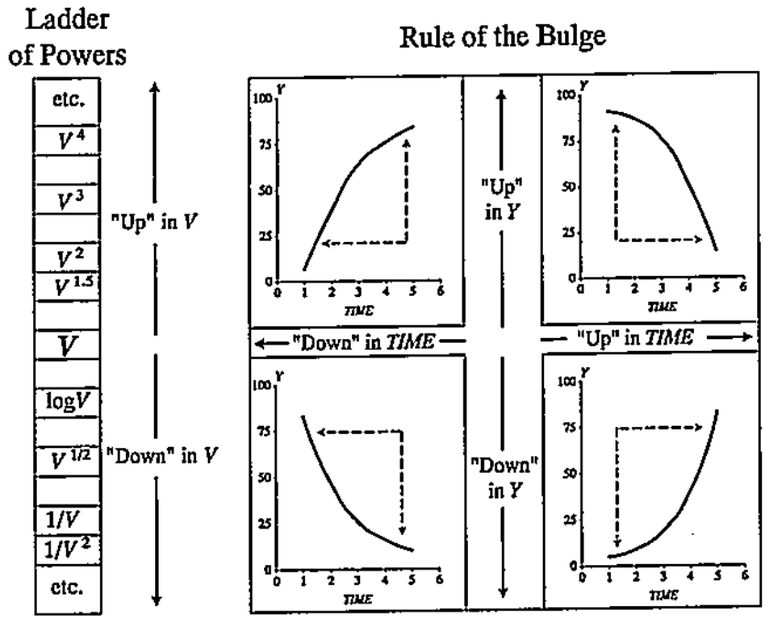


Figure 6.5. The ladder of transformations and the rule of the bulge. Guidelines for linearizing individual growth trajectories through judicious use of transformation.

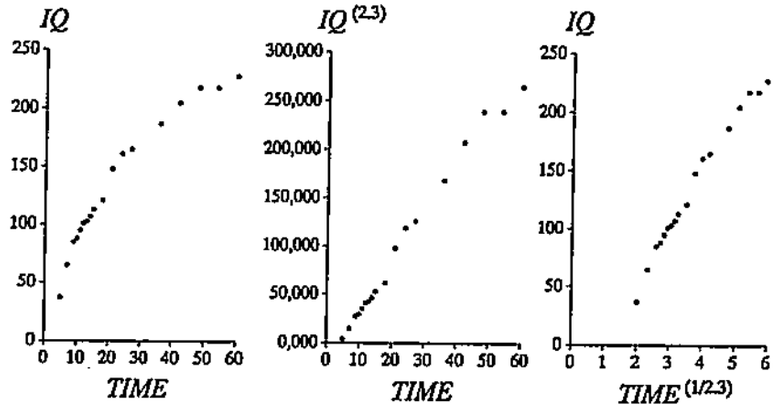


Figure 6.6. Comparing empirical growth plots for a single child in the Berkeley growth study. The left panel presents raw data; the middle panel presents the same data with the outcome, *IQ*, raised to the 2.3 power; the right panel presents the same data with the predictor, *TIME*, expressed as the 2.3th root of *AGE*.

overall, you can argue that the resultant transformed shape is linear for most everyone.

We illustrate this process in figure 6.6, which presents 20 waves of data for a single girl from the Berkeley Growth Study (Bayley, 1935). The left panel displays the child's cognitive trajectory on its original scale. Its curvilinear shape suggests that as she develops, her mental ability increases less rapidly—in other words, the curve decelerates. This matches the exemplar in the upper left corner of figure 6.5. To linearize this trajectory, we can either move "up" in Y (e.g., take Y^2, Y^3, etc.) or "down" in *TIME* (e.g., take *LOG*(*TIME*), 1/*TIME*, etc.) After trying several alternatives, we found that raising IQ to the 2.3rd power was a good compromise. The transformed trajectory appears in the middle panel. Notice its dogleg at about 20 months—a shift apparent in the original trajectory as well—which may be due to changes in the measurement method at this age. Transformation does not eliminate this discontinuity, but provides a reasonably linear change trajectory for both halves..

You can transform *either* the outcome *or TIME*, often using the inverse of the transformation that is best for the other. But applying the inverse of the "outcome" transformation to the predictor, or vice versa, will not produce the *identical* reduction in nonlinearity owing to differences in the range and scale of the variables and the presence of an intercept in the model. Such differences make it worth examining the effect of both types of transformation. For these data, taking the 2.3th *root* of age

(shown in the right panel of figure 6.6) is not as successful in linearizing the trajectory as raising the outcome to the 2.3rd power. If both transformations are equally successful, the choice is yours. If one variable is measured on an easily understood or widely accepted scale—as *TIME* usually is—we recommend that you preserve its metric by transforming its partner. Here, a transformation of the cognitive outcome is more successful in removing nonlinearity and also preserves the metric of *TIME*.

We conclude by reiterating a caution mentioned in section 2.3.1. Notice that we examine *empirical growth plots* for each sample member (or a random subset), not the aggregate trajectory formed by joining within-occasion sample averages. However tempting it is to draw inferences about the shape of individual trajectories from the shape of the aggregate, their forms may not be identical. The forms are identical when change is linear with time but they may not be when change is nonlinear. Because you do not know the shape of the true individual trajectory—if you did, you wouldn't need to do this detective work—avoid this pitfall by always using *individual* plots to identify the shape of *individual* change. (We expand upon this point in section 6.4, when we introduce truly nonlinear trajectories.)

6.3 Representing Individual Change Using a Polynomial Function of *TIME*

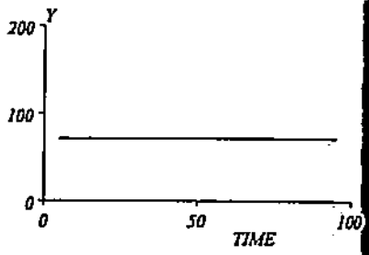
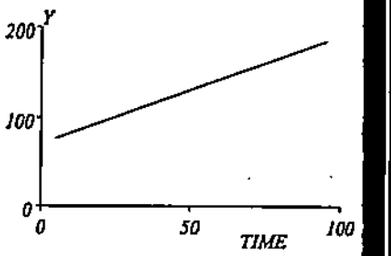
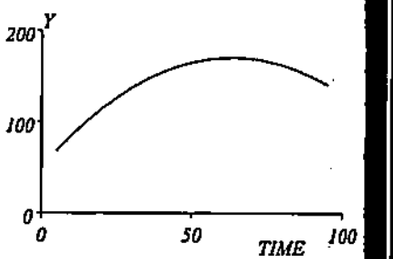
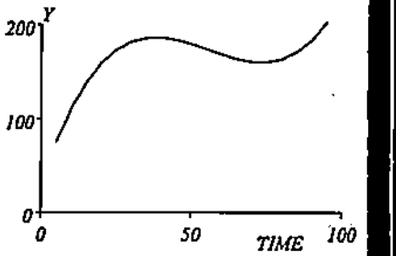
We can also model curvilinear change by including several level-1 predictors that *collectively* represent a polynomial function of time. Although the resulting *polynomial growth model* can be cumbersome, it can capture an even wider array of complex patterns of change over time.

Table 6.4 presents an ordered series of polynomial growth models. Each relates the observed value of an outcome, Y, to *TIME* for individual i on occasion j. The first column labels the trajectory; the second presents the associated level-1 model; the last illustrates the trajectory's shape for the arbitrarily selected values of Y, *TIME*, and the individual growth parameters shown in the third column. As we add higher order functions of *TIME*, the true change trajectory becomes more complex. Below, we describe how to interpret results (section 6.3.1) and select among the alternatives (sections 6.3.2 and 6.3.3).

6.3.1 The Shapes of Polynomial Individual Change Trajectories

The "no change" and "linear change" models are familiar; the remaining models, which contain quadratic and cubic functions of *TIME*, are new. For completeness, we comment on them all.

Table 6.4: A taxonomy of polynomial individual change trajectories

Shape	Level-1 model	Illustrative example: Parameter values	Illustrative example: Plot of the true change trajectory
No change	$Y_{ij} = \pi_{0i} + \varepsilon_{ij}$	$\pi_{0i} = 71$	
Linear change	$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \varepsilon_{ij}$	$\pi_{0i} = 71$ $\pi_{1i} = 1.2$	
Quadratic change	$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \pi_{2i}TIME_{ij}^2 + \varepsilon_{ij}$	$\pi_{0i} = 50$ $\pi_{1i} = 3.8$ $\pi_{2i} = -0.03$	
Cubic change	$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \pi_{2i}TIME_{ij}^2 + \pi_{3i}TIME_{ij}^3 + \varepsilon_{ij}$	$\pi_{0i} = 30$ $\pi_{1i} = 10$ $\pi_{2i} = -.2$ $\pi_{3i} = .0012$	
⋮	⋮	⋮	⋮

"No Change" Trajectory

The "no change" trajectory is known as a polynomial function of "zero order" because *TIME* raised to the 0th power is 1 (i.e., $TIME^0 = 1$). This model is tantamount to including a constant predictor, 1, in the level-1 model, as a multiplier of the sole individual growth parameter, the intercept, π_{0i}. The intercept represents the vertical elevation of the "no-change" trajectory at every point in time (71 in the example). Even though each trajectory is flat, different individuals can have different intercepts and so a collection of true "no change" trajectories is a set of vertically scattered horizontal lines. The "no change" trajectory is the level-1 submodel of the "unconditional means model" that we introduced in section 4.4.1. Here, we use the "no change" label to highlight its relationship with other polynomial trajectories.

"Linear Change" Trajectory

The "linear change" trajectory is known as a "first order" polynomial in time because *TIME* raised to the 1st power equals *TIME* itself (i.e., $TIME^1 = TIME$). Linear *TIME* is the sole predictor and the two individual growth parameters have the usual interpretations. This model allows each individual to possess a unique intercept and slope parameter that yield a collection of crisscrossing trajectories for a group of people. Associated level-2 models can link person-specific characteristics to interindividual heterogeneity in both intercept and slope.

"Quadratic Change" Trajectory

Adding $TIME^2$ to a level-1 individual growth model that already includes linear *TIME* yields a *second order* polynomial for *quadratic* change. Unlike a level-1 model that includes *only* $TIME^2$, a second order polynomial change trajectory includes two *TIME* predictors and three growth parameters (π_{0i}, π_{1i} and π_{2i}). The first two parameters have interpretations that are *similar*, but not identical, to those in the linear change trajectory; the third is new.

In the quadratic change model, π_{0i} still represents the trajectory's *intercept*, the value of *Y* when both predictors, here *TIME* and $TIME^2$, are 0. But π_{1i}, the parameter associated with *TIME*, does not represent a constant rate of change. Instead, it represents the *instantaneous rate of change* at one specific moment, when *TIME* = 0.[2] Although most people still use the "slope parameter" nomenclature, a quadratic change trajectory has no constant common slope. The rate of change changes smoothly over time. π_{2i}, the *curvature* parameter associated with level-1 predictor $TIME^2$,

describes this changing rate of change. Hypothesizing a quadratic individual change trajectory allows you to formulate level-2 questions about interindividual differences in *intercept, instantaneous rate of change,* and *curvature.*

To develop your intuition about quadratic change, examine the sample trajectory in table 6.4. It has an instantaneous rate of change of 3.8 at *TIME* 0 and a curvature of −0.03. Because π_{1i} is positive, the trajectory initially rises, with true status having the intention of increasing by 3.8 in the first unit of time. But because π_{2i} is negative, this increase does not persist. With each passing unit of time, the magnitude of the outcome's rising value diminishes. In essence, π_{1i} and π_{2i} compete to determine the value of *Y*. The quadratic term will eventually win because, for numeric reasons alone, $TIME^2$ increases more rapidly than *TIME*. So, in this example, even though the linear term suggests that *Y* increases over time, the eventual domination of the quadratic term removes more than the linear term adds and causes the trajectory to peak and then decline.

Quadratic trajectories with a single "peak" are said to be *concave* to the time axis. The peak is called the "stationary point" because the slope momentarily goes to zero before reversing direction. Quadratic curves have one stationary point. If the curvature parameter is positive, the trajectory is *convex* to the time axis, with a single "trough." Whether positive or negative, the larger the magnitude of π_{2i}, the more dramatic its effect, rendering the curvature more extreme. The moment when the quadratic trajectory curve flips over, at either a peak or a trough, is $(-\pi_{1i}/2\pi_{2i})$, which in our example is a time of $-3.8/(2(-.03)) = 63.33$.

Higher Order Change Trajectories

Adding higher powers of *TIME* increases the complexity of the polynomial trajectory. The fourth row of table 6.4 presents a third-order polynomial that includes level-1 predictors *TIME*, $TIME^2$ and $TIME^3$. A third-order polynomial has two stationary points; here, one peak and one trough. A quartic polynomial, which adds $TIME^4$ to the cubic model, has three stationary points—either two peaks and one trough or two troughs and one peak depending on the parameters' signs. A fifth-order polynomial has four stationary points; a sixth order has five. By using higher order polynomials to represent individual change, you can represent trajectories of almost any level of complexity.

Interpretation of the individual growth parameters is more complex for higher order polynomials. Even the cubic model's parameters do not represent "initial status," "instantaneous growth rate," and "curvature" as they do in a quadratic. In general, we prefer the simpler representations; we use higher order polynomials only when other approaches fail.

In the next section, we describe strategies for selecting among competing polynomial forms. Before doing so, we inject a note of reality concerning the data collection demands that these models pose. The more complex the polynomial, the more waves of data you need to collect to be able to fit the trajectory to data. In a time-structured data set, you need at least one more wave of data per person than there are individual growth parameters in the level-1 individual growth model. A level-1 linear change trajectory requires at least three waves of data. A quadratic level-1 individual growth model requires at least four; a cubic at least five. And these are only the *minimum* requirements. Greater precision and power requires more waves. In analysis as in life, nothing comes without a cost.[5]

6.3.2 Selecting a Suitable Level-1 Polynomial Trajectory for Change

We illustrate strategies for selecting a level-1 polynomial change trajectory using data on 45 children tracked from first through sixth grade as part of a larger study reported by Keiley, Bates, Dodge, and Pettit (2000). Near the end of every school year, teachers rated each child's level of externalizing behavior using Achenbach's (1991) Child Behavior Checklist. The checklist uses a three-point scale (0 = rarely/never, 1 = sometimes, 2 = often) to quantify the frequency with which the child displays 34 aggressive, disruptive, or delinquent behaviors. The outcome, *EXTERNAL*, which ranges from 0 to 68, is the sum of these 34 scores.

Figure 6.7 presents empirical growth plots for 8 children. (For now, ignore the fitted trajectories and focus on the data points.) As a group, these cases span the wide array of individual change patterns in the data. Child D displays little change over time. Child C appears to decline linearly with age (at least through fourth grade). Children A, B, and G display some type of quadratic change, but their curvatures differ. For A, the curvature parameter appears negative; for B and G, positive. Child E may have two stationary points—a trough in second grade and a peak in fifth—suggesting a cubic trajectory. Children F and H may have three stationary points—a peak, a trough, and another peak—although with only six waves, it is difficult to distinguish true quartic change from occasion-specific measurement error.

When faced with this many different patterns, which polynomial trajectory makes sense? If there is no obvious winner, we suggest that you first adopt an exploratory approach and fit separate person-specific OLS models to each person's data. This process is relatively easy, albeit tedious:

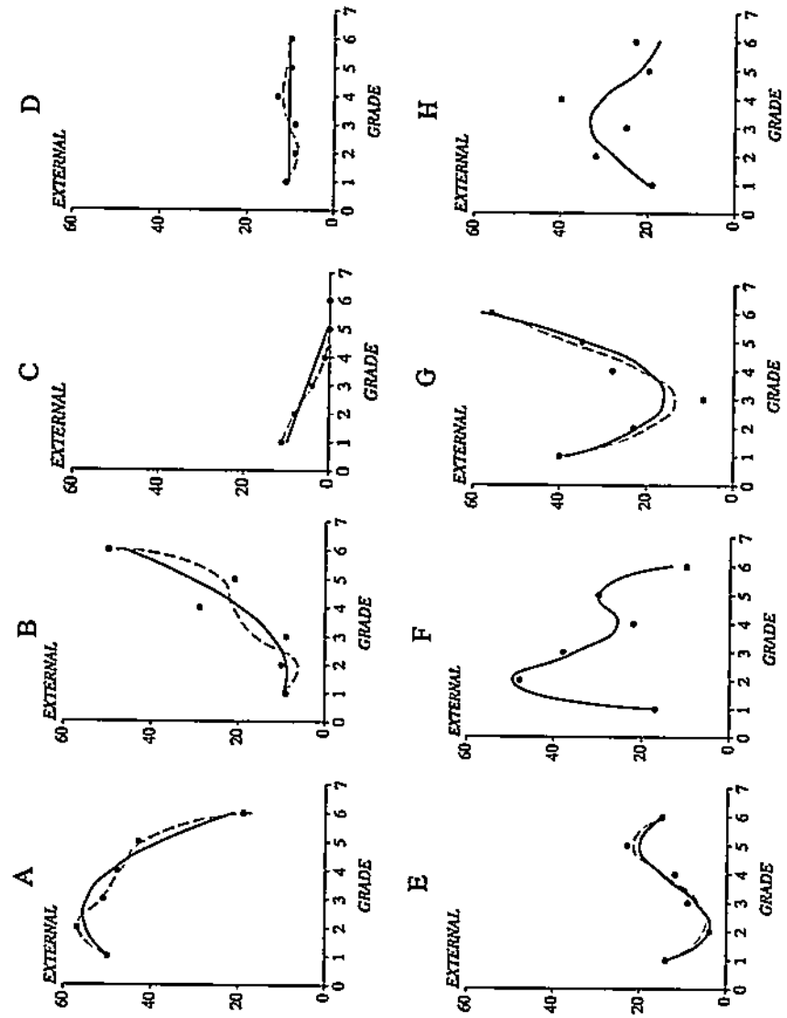


Figure 6.7. Identifying a suitable level-1 polynomial trajectory. Empirical growth plots for 8 participants in the externalizing behavior study. The solid curves represent a reasonable polynomial tailored for each child: a flat line for D; a linear trajectory for C; a quadratic trajectory for A, B, and G; a cubic for E; and a quartic for F and H. The dashed lines represent the highest order polynomial necessary—a quartic.

all you need do is create a set of temporal predictors that capture the requisite polynomial shapes—e.g., *TIME*, *TIME*^{2}, *TIME*^{3}, and so on—and then, for each child, use the set of predictors needed to represent the trajectory desired. The solid curves in figure 6.7 represent the choices just articulated: (1) "no change" for D; (2) linear for C; (3) quadratic for A, B and G; (4) cubic for E; and (5) quartic for F and H.

Comparison of observed and fitted values of *EXTERNAL* demonstrates the utility of this approach. Most of the fitted trajectories reasonably summarize each child's data record. But are ad hoc decisions like these optimal? Closer comparison of observed and fitted values suggests cause for concern. For child C, the small differences appear systematic. Perhaps he really has a quadratic trajectory with a wide flat trough that begins in sixth grade. Child H raises a different issue. Here, we fitted a quartic model (because of the two peaks and one trough), but the fitted trajectory seems quadratic (it has just one peak). Inspection of his regression results reveals that the cubic and quartic parameters are small and indistinguishable from zero. The imagined peaks and troughs may be measurement error. This suggests a need for parsimony when specifying polynomial trajectories, a recommendation we adopt more vigorously in coming sections.

But for now, we move in the opposite direction: fitting exploratory trajectories using a common more general shape. We do so for two reasons: (1) the decision-making process needed to fit custom trajectories to entire data sets can be tedious and counterproductive; and (2) we cannot easily specify a level-1 individual growth model unless we use a *common* shape for the trajectory across people. Instead of selecting a unique polynomial form for each child, we select the *highest order* polynomial needed to summarize individual change for *any child*. For the eight children in figure 6.7, we select a quartic because no child appears to need a higher order polynomial. While hardly parsimonious, a quartic can be fit easily to each child's data; the *data* will then demand the contribution of higher order terms as needed. If we use a quartic for Child A, for example, the estimated growth parameters for the cubic and quartic terms may be close, or equal, to 0.

The dashed curves in figure 6.7 display the results of fitting a quartic to each child's data. (Note that the solid lines already represent a quartic for children F and H.) This common trajectory simplifies implementation, but clearly overfits. While children E and G have fitted trajectories virtually identical to those specified using the case-specific approach, the others reveal more complex forms. Is this complexity necessary? The fitted trajectories for children A and B, which previously seemed quadratic, now have an extra "bump." And the fitted trajectory for child D, which seemed flat, now looks like a scaled down quartic!

Should you be parsimonious and potentially underestimate the trajectory's complexity or should you be cautious and potentially overfit? An answer to this question is clouded by the use of *sample* data to draw conclusions about *population* trajectories. When you inspect empirical plots, you try to account for measurement error, but this is easier said than done. For example, you may have been prepared to conclude that the true trajectories for A and B were quadratic, but now you might see them as more complex. So, too, the quartic hypothesized for child H may be unnecessarily complex. Fortunately, when exploratory analyses lead to conflicting conclusions, you can resolve this dilemma by comparing goodness-of-fit statistics across a series of models, as we now do.

6.3.3 Testing Higher Order Terms in a Polynomial Level-1 Model

Table 6.5 presents the results of fitting four models of increasing polynomial complexity at level-1 to the externalizing behavior data. Each was fit using Full IGLS in MLwiN. For simplicity, we include no other substantive predictors at either level-1 or level-2. Note, however, that as we increase the complexity of the fixed portion of the level-1 model we do add the associated random effects. We describe the rationale for this decision below as we describe the empirical results.

Let us begin with Model A, the "no-change" trajectory. The estimated grand mean is 12.96 ($p < .001$), which suggests that between first and sixth grades, the average child has a non-zero level of externalizing behavior. Examining the variance components, we find statistically significant variability both within-child (70.20, $p < .001$) and between-children (87.42, $p < .001$). We conclude that externalizing behavior varies from occasion to occasion and that children differ from each other.

Is this "no-change" trajectory adequate or should we add a linear *TIME* predictor to the level-1 individual growth model? We address this question by comparing Model A to the standard linear-change model (B). To facilitate interpretation, we express *TIME* as *GRADE*-1 so that the intercept (π_{0i}) refers to the level of externalizing behavior in first grade. We find that while the average child has a non-zero level of externalizing behavior in first grade (13.29, $p < .001$) this level does not change linearly over time on average (−0.13, *n.s.*). The statistically significant variance components ($\hat{\sigma}_0^2 = 123.52$, $p < .001$; $\hat{\sigma}_1^2 = 4.69$, $p < .01$) suggest, however, that children differ substantially from these averages. In other words, the *average* trajectory may be flat but many of the *individual* trajectories are not.

To determine whether Model B is preferable to Model A, we test the

Table 6.5: Comparison of fitting alternative polynomial change trajectories to the externalizing behavior data ($n = 48$)

	Parameter	Model A No change	Model B Linear change	Model C Quadratic change	Model D Cubic change
Fixed Effects					
Composite model	Intercept (1st grade status)	γ_{00} 12.96***	13.29***	13.97***	13.79***
	TIME (linear term)	γ_{10}	−0.13	−1.15	−0.35
	$TIME^2$ (quadratic term)	γ_{20}		0.20	−0.23
	$TIME^3$ (cubic term)	γ_{30}			0.06
Variance Components					
Level-1:	Within-person	σ_e^2 70.20***	53.72***	41.98***	40.10***
Level-2:	In 1st grade status	σ_0^2 87.42***	123.52***	107.08***	126.09***
	Linear term				
	variance	σ_1^2	4.69**	24.60*	88.71
	covar with 1st grade status	σ_{01}	−12.54*	−3.69	−51.73
	Quadratic term				
	variance	σ_2^2		1.22*	11.35
	covar with 1st grade status	σ_{02}		−1.36	22.83~
	covar with linear term	σ_{12}		−4.96*	−31.62
	Cubic term				
	variance	σ_3^2			0.08
	covar with 1st grade status	σ_{03}			−3.06~
	covar with linear term	σ_{13}			2.85
	covar with quadratic term	σ_{23}			−0.97
Goodness-of-fit					
	Deviance statistic	2010.3	1991.8	1975.8	1967.0
	AIC	2016.3	2003.8	1995.8	1997.0
	BIC	2021.9	2015.0	2014.5	2025.1

~$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$.
Model A is the "no change" trajectory; Model B is the linear change trajectory; Model C is the quadratic change trajectory; Model D is the cubic change trajectory.

Note: Full IGLS, MlwiN.

compound null hypothesis about the *set* of differences between models (in the linear growth rate, its associated variance component, and the extra covariance parameter, σ_{01}): H_0: $\gamma_{10} = 0$, $\sigma_1^2 = 0$, *and* $\sigma_{01} = 0$. As the difference in deviance statistics (18.5) far exceeds the 0.05 critical value of a χ^2 distribution on three *d.f.*, we reject H_0 and abandon the "no change" model.

You can use this same testing strategy to evaluate the impact of adding polynomial terms to the level-1 growth model, by comparing Model C to B and D to C, and so on. Before doing so, however, we draw your attention to a common dilemma that arises in model fitting. In Model B, although the variance component for linear growth (σ_1^2) is statistically significant, its associated fixed effect (γ_{10}) is not. This is not an inconsistency, but it requires interpretive care. The test for the variance component tells us that there is statistically significant variation in linear rates of change across children. The test for the fixed effect tells us that the average value of these rates is indistinguishable from 0. Yet we retain the fixed effect because the non-zero variance component suggests that we may be able to predict some of this variation with a level-2 predictor. We might find, for example, that the average slope for boys is positive and that the average slope for girls is negative. This would be of tremendous interest, even if the rates average to zero when boys and girls are pooled. We remind you that when selecting a functional form for level-1 model, you are as interested in the level-2 variance components as you are in the level-1 fixed effects.

We now compare the quadratic Model C to the linear Model B. Because we seek a level-1 individual growth model that describes the fundamental structure of these data, we include not just the additional fixed effect (for $TIME^2$) but also the required additional variance components: the population variance for curvature , σ_2^2, as well as its covariances with first grade status, σ_{02}, and linear growth, σ_{12}. To do otherwise would constrain the curvature parameter to be identical across individuals, a constraint that seems antithetical to the model-building exercise in which we are engaged. We find that the deviance statistic declines by 16.0, which exceeds the .01 critical value of a χ^2 distribution on four *d.f.* (13.27). We therefore reject the null hypothesis that all four parameters are simultaneously zero and conclude that there is potentially predictable variation in curvature across children.

Do we need to go further and adopt a cubic model? Comparison of Models D and C suggest that the answer is no. Addition of a cubic term adds one fixed effect and four random effects (σ_3^2, σ_{03}, σ_{13}, and σ_{23}), but the deviance statistic declines by only 8.8 (1975.8–1967.0), which is less than the associated .05 critical value of 11.07 (*d.f.* = 5).

We conclude that we should treat individual change in externalizing behavior as though it follows a quadratic trajectory. This conclusion, reinforced by the AIC and BIC statistics, is a realistic compromise that respects the many kinds of variation present in the data. This does not mean that *no child* follows a cubic trajectory, but rather that, overall, when individual change is hypothesized to be cubic, sufficient children end up with cubic parameters that are close enough to zero that there is too little systematic variation in this parameter to worry about. The reverse is true for the quadratic parameter: even though its average value is indistinguishable from 0, it displays sufficient variation to warrant inclusion.

Having selected a suitable polynomial individual change trajectory, model building proceeds as before, although there are extra individual growth parameters to explore. Each level-1 parameter has its own level-2 submodel. Level-1 quadratic change provides for three level-2 submodels; a level-1 cubic provides for four. To examine the effect of *FEMALE*, we would begin by postulating a level-2 association with each level-1 parameter:

$$
\begin{aligned}
Y_{ij} &= \pi_{0i} + \pi_{1i}(GRADE_{ij} - 1) + \pi_{2i}(GRADE_{ij} - 1)^2 + \varepsilon_{ij} \\
\pi_{0i} &= \gamma_{00} + \gamma_{01} FEMALE_i + \zeta_{0i} \\
\pi_{1i} &= \gamma_{10} + \gamma_{11} FEMALE_i + \zeta_{1i} \\
\pi_{2i} &= \gamma_{20} + \gamma_{21} FEMALE_i + \zeta_{2i}
\end{aligned}
$$

where

$$
\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \text{ and } \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \\ \zeta_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 \end{bmatrix} \right).
$$

Having explored a variety of gender differentials in these data, however, it turns out that none is statistically significant. *FEMALE* has no effect on first-grade status, instantaneous rate of change in first grade, or curvature.

6.4 Truly Nonlinear Trajectories

All the individual growth models described so far—including the curvilinear ones presented in this chapter—share an important mathematical property: they are *linear in the individual growth parameters*. Why do we use the label "linear" to describe trajectories that are blatantly nonlinear? The explanation for this apparent paradox is that this mathematical property depends not on the *shape* of the underlying growth trajectory but rather *where*—in which portion of the model—the nonlinearity arises. In all

previous models, nonlinearity (or discontinuity) stems from the representation of the *predictors*. To allow the hypothesized trajectory to deviate from a straight line, *TIME* is either transformed or expressed using higher order polynomial terms. In the truly nonlinear models we now discuss, nonlinearity arises in a different way—through the *parameters*.

In this section, we consider models that are not linear in the parameters. We begin, in section 6.4.1, by introducing the notion of *dynamic consistency*, a key concept for understanding the distinction between our previous models and the truly nonlinear ones we discuss here. In section 6.4.2, we illustrate a general approach for fitting truly nonlinear models by analyzing a data set in which the level-1 individual growth trajectory is hypothesized to follow a logistic curve. In section 6.4.3, we expand this approach, surveying a range of other truly nonlinear growth models including the hyperbolic, inverse polynomial, and exponential trajectories. We conclude, in section 6.4.3, by describing how researchers have historically translated substantive theories about nonlinear growth into mathematical representations that can be fit to data.

6.4.1 What Do We Mean by Truly Nonlinear Models?

To highlight the distinction between models that are linear in the parameters and those that are not, consider the following simple quadratic level-1 trajectory: $Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \pi_{2i}TIME_{ij}^2 + \varepsilon_{ij}$. We compute individual i's value of Y at time j—say at $TIME = 2$—by substituting in this value: $Y_{i2} = \pi_{0i}(1) + \pi_{1i}(2) + \pi_{2i}(2)^2 + \varepsilon_{i2}$. For reasons that will soon become apparent, we add the implicit multiplier 1 next to the intercept, π_{0i}. This addition is not essential to our explanation, but it simplifies the argument.

Ignoring ε_{i2} for a moment, individual i's hypothesized true value of Y at $TIME = 2$ is the *sum* of three quantities: $\pi_{0i}(1)$, $\pi_{1i}(2)$, and and $\pi_{2i}(2)^2$. Each has a similar form: it is an *individual growth parameter* multiplied by a *number*: π_{0i} times 1, plus π_{1i} times 2, plus π_{2i} times 2^2. All true values of Y, for all values of *TIME*, share this property—they are the *sum* of several terms, each of which is the *product* of an *individual growth parameter* and a *numerical weight* whose value is either constant (such as the "1" multiplying π_{0i}) or dependent upon the measurement occasion (such as the 2 and 2^2 multiplying π_{1i} and π_{2i}). We say that this portion of the growth model is a "weighted linear composite of the individual growth parameters" or, more simply, that true individual change is "linear in the parameters."

Individual growth models that are *linear in the parameters* have important spatial properties to which we alluded in chapter 2. These properties are apparent only at the group level—that is, when you summarize

everyone's changes using an "average trajectory." As described in section 2.3, we can derive this average trajectory in one of two ways, by computing: (1) the *curve of the averages*—estimate the average outcome on each measurement occasion and then plot a curve through these averages; or (2) *the average of the curves*—estimate the growth parameters for each individual trajectory, average these values, and then plot the result. If an individual growth model is linear in the parameters, it will not matter which approach you use because the "curve of the averages" and the "average of the curves" will be identical. In addition, the average trajectory possesses the same *functional form* (i.e., the same general *shape*) as the constituent individual trajectories: the average of a heterogeneous group of straight lines will be a straight line, the average of a heterogeneous group of quadratics will be quadratic, and so on.[4]

These two properties—(1) the coincidence between the "curve of the averages" and the "average of the curves," and (2) the equivalence in functional form between individual and average trajectories—were labeled *dynamic consistency* by Keats (1983). Many common functions are dynamically consistent, including the straight line, the quadratic, and all polynomials. If a function is linear in the parameters, it will be dynamically consistent.

The concept of dynamic consistency has two important consequences for analysis. First, it reinforces why you should never draw conclusions about the shape of an individual change trajectory from the shape of an average trajectory drawn through occasion-specific means. If true change is not dynamically consistent, your conclusions about the model's functional form will be incorrect. Second, any level-1 model that is dynamically consistent—that is any polynomial, any model with a transformed outcome, or any discontinuous model—can be fit using standard software for multilevel modeling.

Trajectories that are not dynamically consistent are less tractable. Many important level-1 models that arise from substantive theories—such as the *logistic model* for individual change that we examine next—are not linear in the parameters, and as such, are *not* dynamically consistent. We have already alluded to this possibility in section 2.3.1, when we stated that the average of a set of logistic trajectories is not a logistic but a smoothed *step-function*. We now illustrate how to proceed when the logical level-1 individual growth model is a logistic curve.

6.4.2 The Logistic Individual Growth Curve

We introduce the fitting of truly nonlinear change trajectories using data on cognitive growth collected by our colleague, Terry Tivnan (1980)