Applied Longitudinal Data Analysis

Modeling Change and Event Occurrence

Judith D. Singer
John B. Willett

OXFORD
UNIVERSITY PRESS
2003

OXFORD
UNIVERSITY PRESS

Oxford New York
Auckland Bangkok Buenos Aires Cape Town Chennai
Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata
Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
São Paulo Shanghai Taipei Tokyo Toronto

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Library of Congress Cataloging-in-Publication Data
Singer, Judith D.
Applied longitudinal data analysis : modeling change and event occurrence/by Judith D. Singer and John B. Willett.
p. cm.
Includes bibliographical references and index.
ISBN 0-19-515296-4
1. Longitudinal methods. 2. Social sciences—Research.
I. Willett, John B. II. Title.
H62 .S47755 2002
001.4'2—dc21 2002007055

21

Printed in the United States of America
on acid-free paper

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 their predictability based upon early childhood signs and symptoms.

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 semistructured 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 within-individual 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 a five-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.

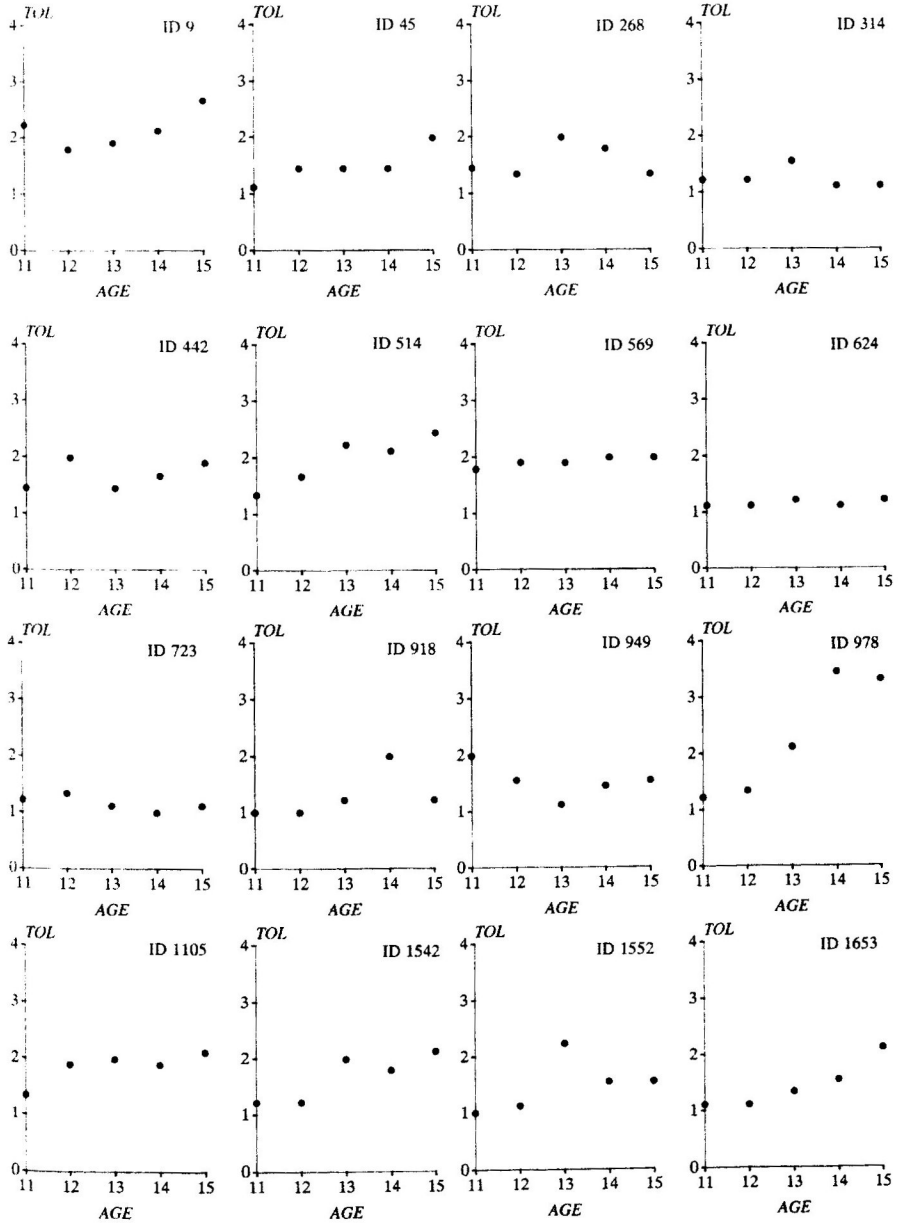


Figure 2.2. Exploring how individuals change over time. Empirical growth plots for 16 participants in the tolerance study.

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

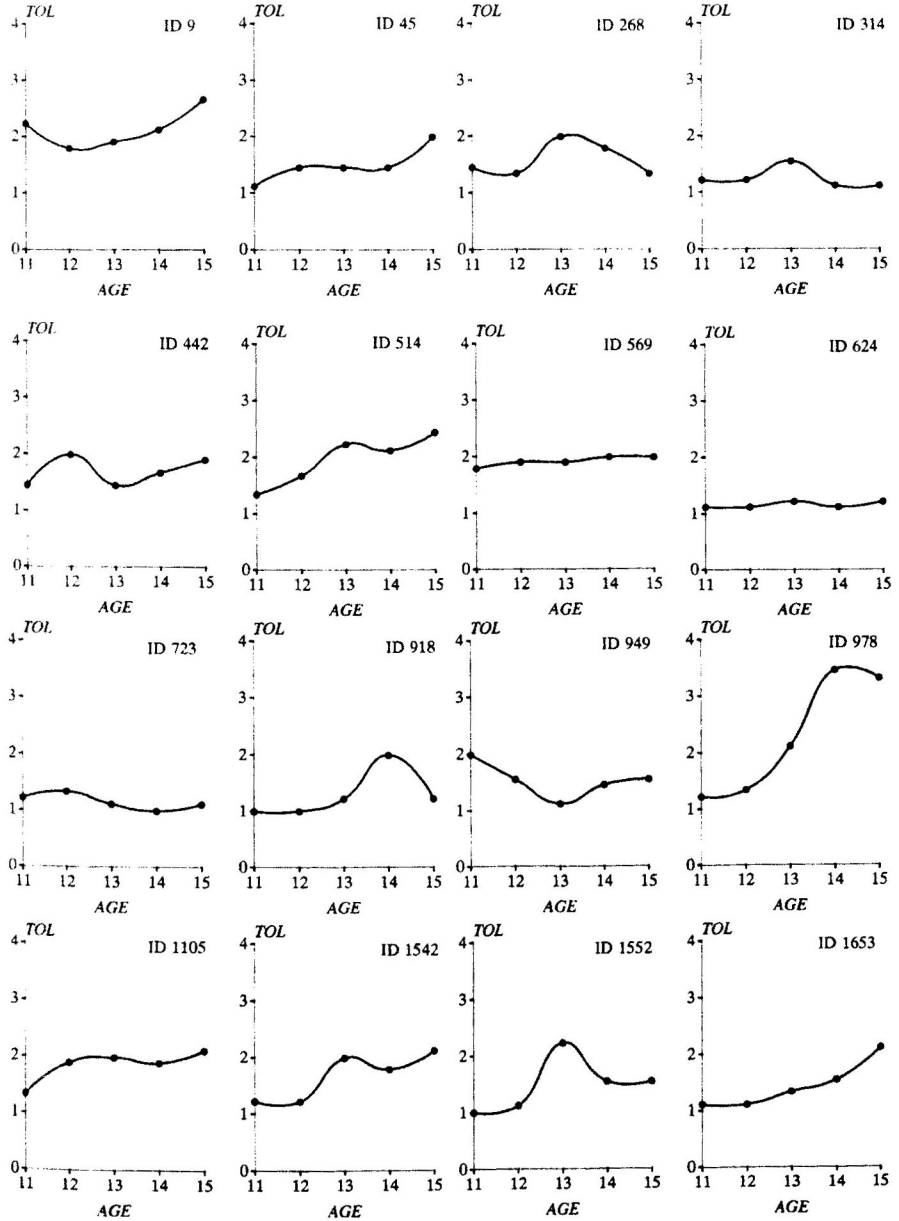


Figure 2.3. Smooth nonparametric summaries of how individuals change over time. Smooth nonparametric trajectories superimposed on empirical growth plots for participants in the tolerance study.

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 growth records and plots is nothing more than measurement error.

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 with the largest slopes become more tolerant the most rapidly. Adoles-

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

Fitted initial status

1.9	0
1.8	2
1.7	3
1.6	
1.5	4 4 8
1.4	3
1.3	1
1.2	7
1.1	2 4 8 9
1	0 3
0.9	5

Fitted rate of change

0.6	3
0.5	
0.4	
0.3	
0.2	4 5 7
0.1	2 4 5 6 7
0	2 2 5 6
-0	3 5
-0.1	0

Residual variance

.2 lo	3
.1 hi	5 7
.1 lo	0 1 1
.0 hi	5 7
.0 lo	0 0 1 3 3 3 4 4

R^2 statistic

0.8	6 8 8 9
0.7	7 8
0.6	8
0.5	
0.4	5
0.3	1 1 3
0.2	5 5
0.1	4
0	2 7

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

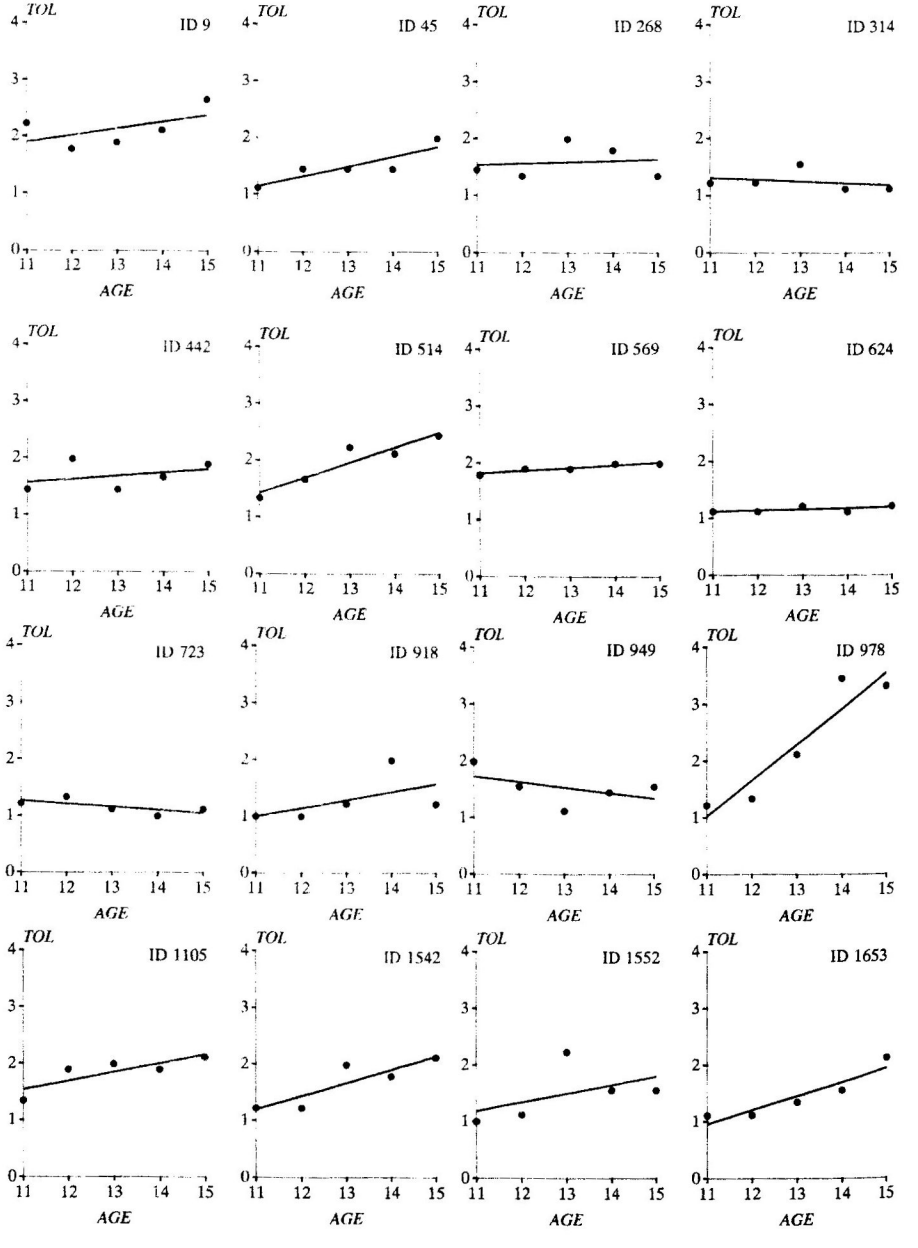


Figure 2.5. OLS summaries of how individuals change over time. Fitted OLS trajectories superimposed on empirical growth plots for participants in the tolerance study.

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.

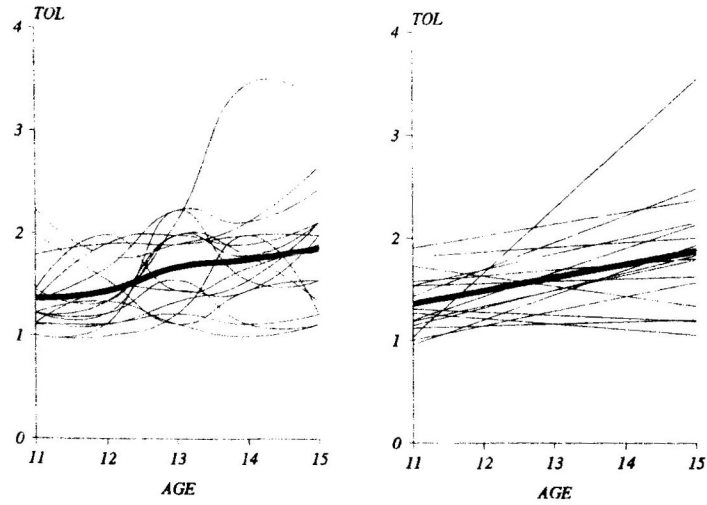


Figure 2.6. Examining the collection of smooth nonparametric and OLS trajectories across participants in the tolerance study. Panel A presents the collection of smooth nonparametric trajectories; Panel B presents the collection of fitted OLS trajectories. Both panels also present an *average change trajectory*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

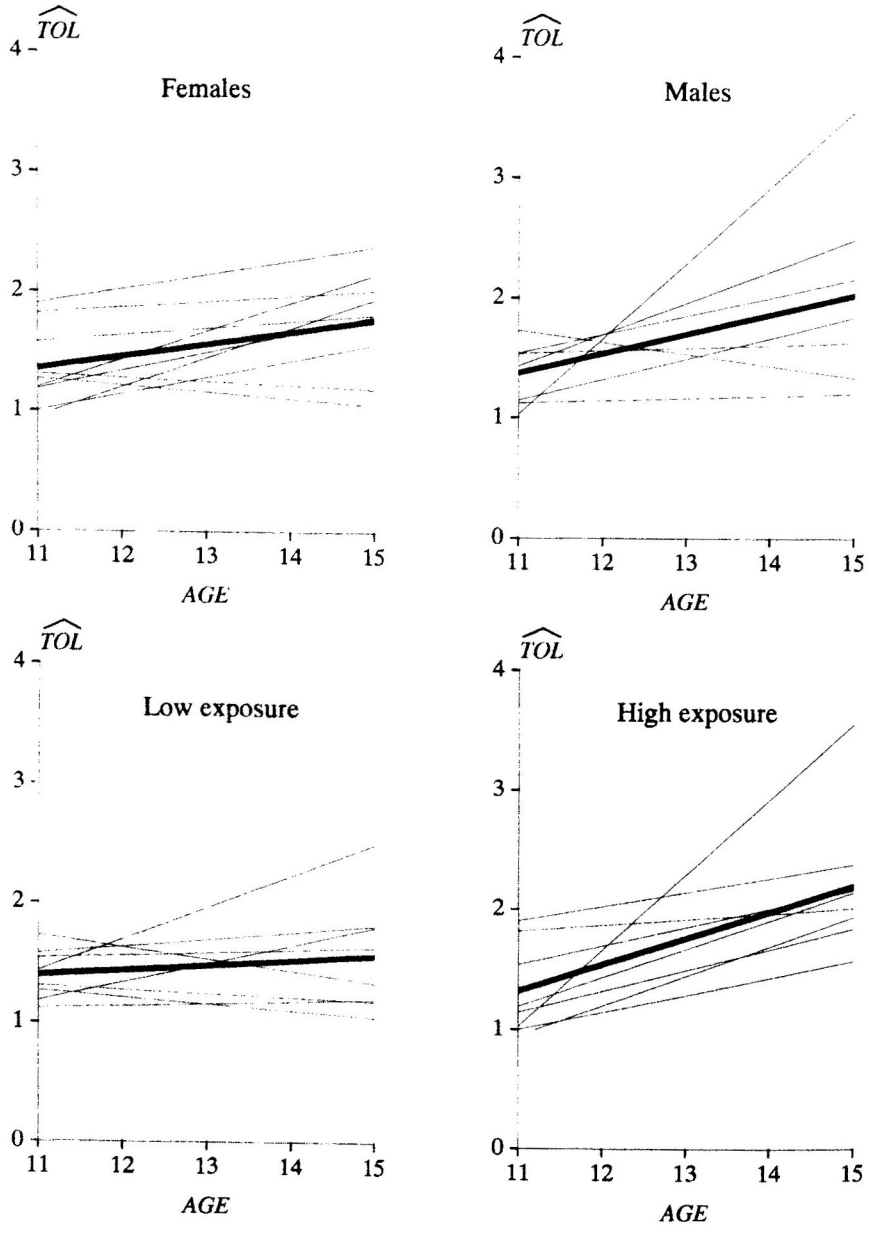


Figure 2.7. Identifying potential predictors of change by examining OLS fitted trajectories separately by levels of selected predictors. Fitted OLS trajectories for the tolerance data displayed separately by gender (upper panel) and exposure (lower panel).

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

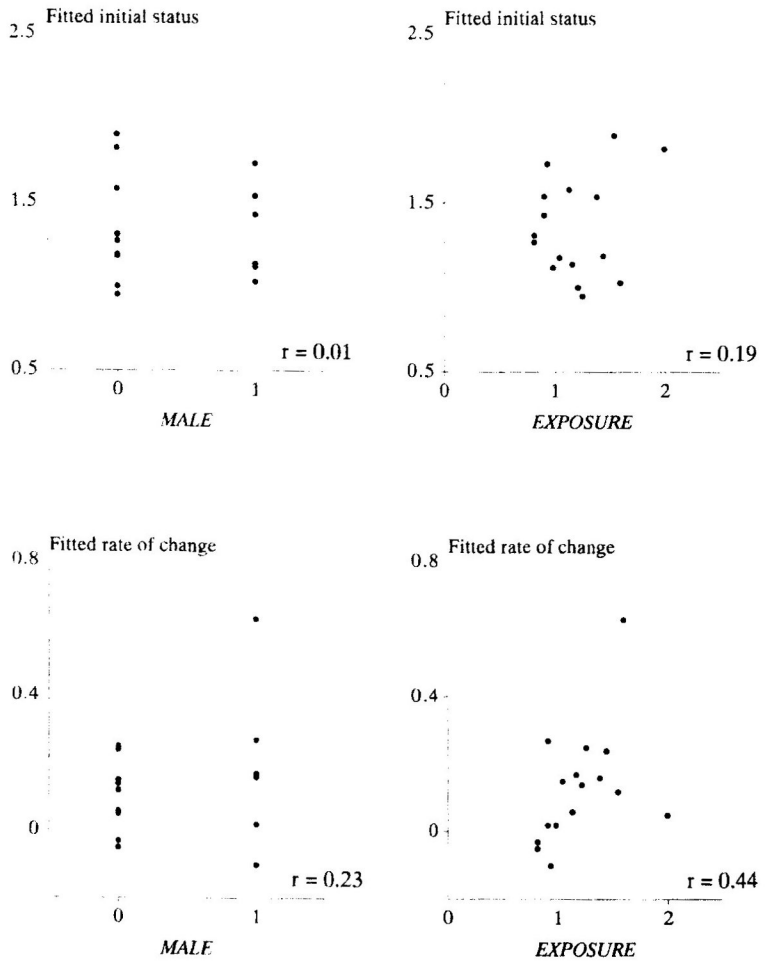


Figure 2.8. Examining the relationship between OLS parameter estimates (for initial status and rates of change) and potential predictors. Fitted OLS intercepts and slopes for the tolerance data plotted vs. two predictors: *MALE* and *EXPOSURE*.

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:

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

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 statis-

tical 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 other

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
...	...	...	...

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 per-

formance 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:

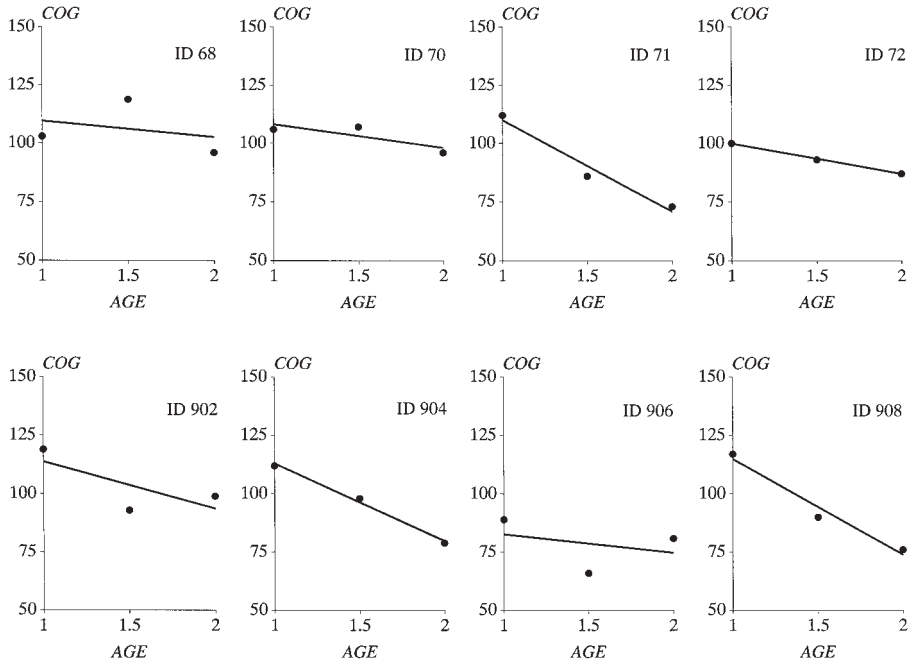


Figure 3.1. Identifying a suitable functional form for the level-1 submodel. Empirical growth plots with superimposed OLS trajectories for 8 participants in the early intervention study.

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

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

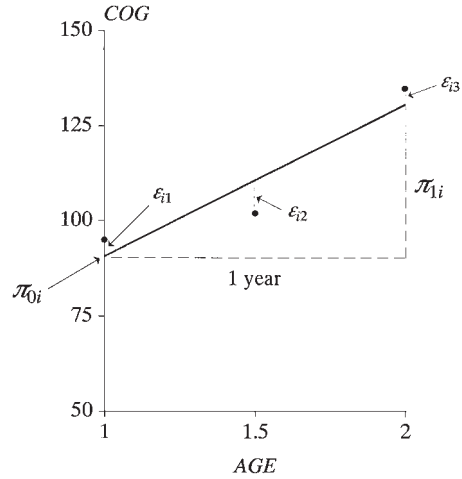


Figure 3.2. Understanding the structural and stochastic features of the level-1 individual growth model. Mapping the model in equation 3.1 onto imaginary data for child *i*, an arbitrarily selected member of the population.

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 *i*th 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)			
Individual growth parameters	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

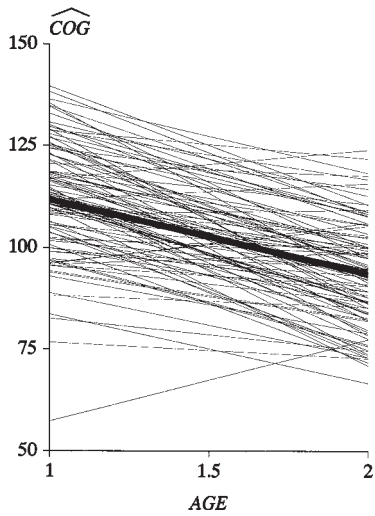
The exploratory OLS-fitted trajectories of section 2.2.2 may now make more sense. Although they are not fully efficient because they do not

properly exploit all the information present in longitudinal data, they do provide invaluable insights into the functioning of the hypothesized individual growth model. The top panel of figure 3.3 presents the results of using OLS methods to fit the level-1 submodel in equation 3.1 to the data for all 103 children (regressing *COG* on (*AGE-1*), separately by *ID*). The bottom panel presents stem and leaf displays for three summary statistics from these models: the fitted intercepts, the fitted slopes, and the estimated residual variances.

For most children, cognitive performance declines over time. For some, the decline is rapid; for others, less so. Few children show any improvement. Each fitted intercept estimates that child's true initial status; each fitted slope estimates that child's true annual rate of change during the second year of life. The fitted intercepts are centered near 110; the fitted slopes are centered near −10. This suggests that at age 1, the average child has a true cognitive level slightly above the national norm (of 100 for this test). Over time, however, most children decline (we estimate that only 7 improve).

The stem-and-leaf displays in the bottom left panel of figure 3.3 reveal great heterogeneity in fitted intercept and slope across children in the sample and suggest that not all children have identical trajectories of change. Of course, you must be cautious when interpreting the interindividual heterogeneity in change trajectories evident in figure 3.3. The between-person variation in the estimated change trajectories that you observe is necessarily inflated over the underlying interindividual variability in the unknown true change trajectories because the fitted trajectories, having been estimated from observed data, are *fallible* representations of true change. The actual variability in underlying true change will always be somewhat less than what you observe in exploratory analysis, with the magnitude of the difference depending on the quality of your outcome measurement and the efficacy of your hypothesized individual growth model.

The skewed distribution of residual variances in the bottom right panel of figure 3.3 suggests great variation in the *quality* of the OLS summaries across children (we expect the distribution of these statistics to be skewed, as they are "squared" quantities and are therefore bounded by zero below). When the residual variance is near 0, as it is for many children, the fitted trajectories are reasonable summaries of the observed data for those children. When the residual variance is larger, as it often is here, the fitted trajectories are poorer summaries: the observed values of *COG* are further away from the fitted lines, making the magnitude of the estimated level-1 residuals, and therefore the residual variance, large.



Fitted initial status	*Fitted rate of change*	*Residual variance*
14 0	2. 0	46 8
13* 5568	1*	44
13. 00134	1. 0	42
12* 5556778999	0* 79	40 00
12. 02233344	0. 134	38
11* 55667777888889	-0* 4444332	36 8
11. 000111112222233334444	-0. 99998888777765	34
10* 55666688999	-1* 4333322211000	32 3
10. 0012222244	-1. 99888877666655	30
9* 6666677799	-2* 44322211110000	28 4
9. 344	-2. 9999877776655	26 7
8* 89	-3* 443322100000	24 1444
8. 34	-3. 987	22 8
7* 7	-4* 443111	20
7.		18 3
6*		16 00011
6.		14
5* 7		12 21
		10 44433
		8 1118886666
		6 77744
		4 333844
		2 04444888833338888888
		0 000011112223333444444446666811111444 7

Figure 3.3. Observed variation in fitted OLS trajectories. Fitted OLS trajectories for participants in the early intervention study as well as stem and leaf displays for fitted initial status, fitted rate of change, residual variance.

3.3 The Level-2 Submodel for Systematic Interindividual Differences in Change

The level-2 submodel codifies the relationship between interindividual differences in the change trajectories and time-invariant characteristics of the individual. The ability to formulate this relationship using a

level-2 submodel stems from the realization that adoption of a common level-1 submodel forces people to differ only in the values of their individual growth parameters. When we use a level-1 linear change model, people can differ only in their intercepts and slopes. This allows us to recast vague questions about the relationship between "change" and predictors as specific questions about the relationship between the individual growth parameters and predictors.

Like all statistical models, the level-2 submodel describes hypothesized population processes, not sample behavior. But insights gleaned from sample data can often provide valuable insight into model formulation. In this spirit, examine the top panel of figure 3.4, which separately plots fitted OLS trajectories according to the child's program participation (program participants in the right panel, nonparticipants in the left). The average change trajectory for each group is shown in bold. Program participants tend to have higher scores at age 1 and decline less precipitously over time. This suggests that their intercepts are higher but their slopes are shallower. Also note the substantial interindividual heterogeneity *within* groups. Not all participants have higher intercepts than nonparticipants; not all nonparticipants have steeper slopes. Our level-2 model must simultaneously account for both the general patterns (here, the between-group differences in intercepts and slopes) *and* interindividual heterogeneity in patterns within groups.

What kind population model might have given rise to these patterns? The preceding discussion suggests four specific features for the level-2 submodel. First, its outcomes must be the individual growth parameters (here, π_{0i} and π_{1i} from equation 3.1). As in regular regression, where we model the population distribution of a random variable by making it an outcome, here, where we model the population distribution of the individual growth parameters, they, too, must be the outcomes. Second, the level-2 submodel must be written in separate parts, one for each level-1 growth parameter. When we use a linear change individual growth model at level-1 (as in equation 3.1), we need two level-2 submodels: one for the intercept, π_{0i}, another for the slope, π_{1i}. Third, each part must specify a relationship between an individual growth parameter and the predictor (here, *PROGRAM*). As you move across the panels in the top of figure 3.4, the value of the predictor, *PROGRAM*, shifts from 0 to 1. This suggests that each level-2 model should ascribe differences in either π_{0i} or π_{1i} to *PROGRAM* just as in a regular regression model. Fourth, each model must allow individuals who share common predictor values to vary in their individual change trajectories. This means that each level-2 submodel must allow for stochastic variation in the individual growth parameters.

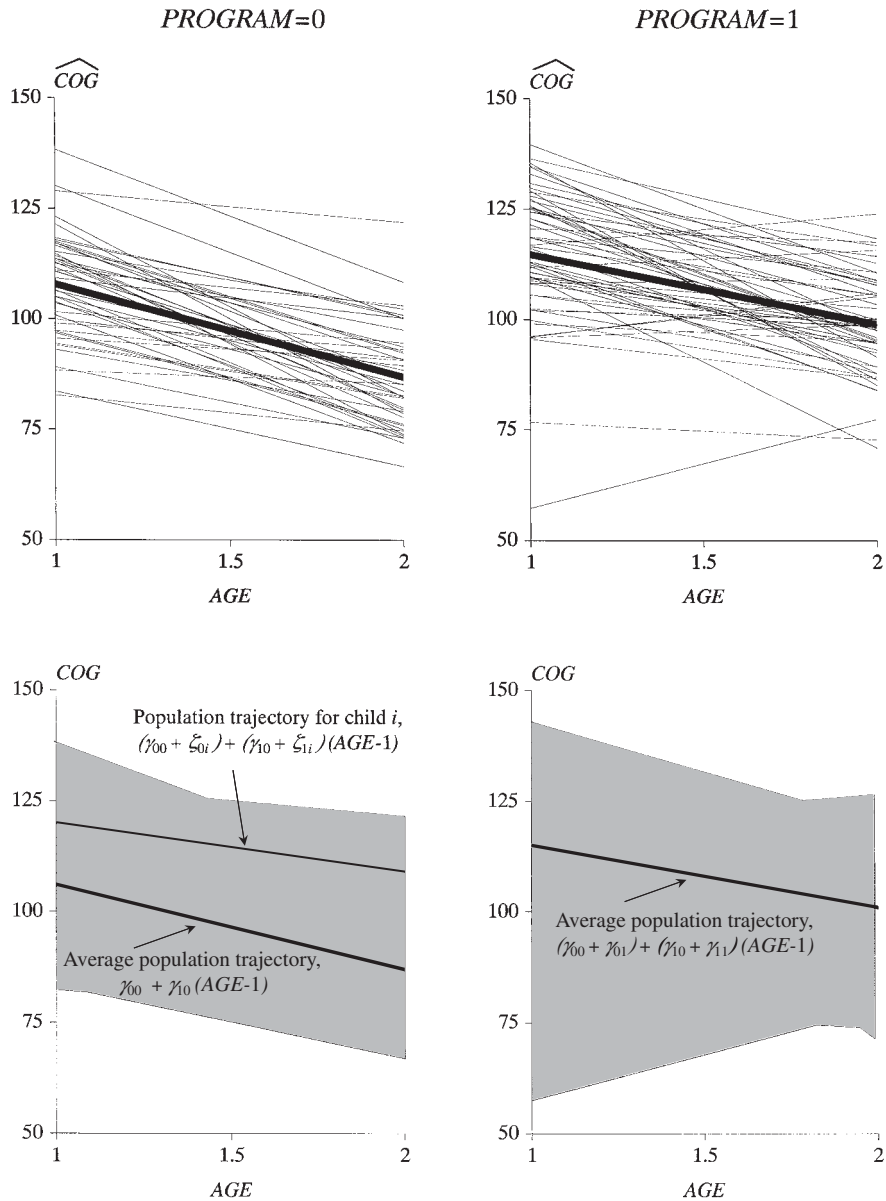


Figure 3.4. Understanding the structural and stochastic features of the level-2 submodel for inter-individual differences in change. Top panel presents fitted OLS trajectories separately by levels of the predictor *PROGRAM*. Bottom panel maps the model in equation 3.3 onto imaginary data for an arbitrary child i and the average population trajectory. The shaded portion in each of the lower panels is designed to suggest the existence of many distinct population trajectories for different children.

These considerations lead us to postulate the following level-2 submodel for these data:

$$\begin{aligned}\pi_{0i} &= \gamma_{00} + \gamma_{01}PROGRAM_i + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \gamma_{11}PROGRAM_i + \zeta_{1i}.\end{aligned} \tag{3.3}$$

Like all level-2 submodels, equation 3.3 has more than one component, each resembling a regular regression model. Taken together, the two components treat the intercept (π_{0i}) and the slope (π_{1i}) of an individual's growth trajectory as level-2 outcomes that may be associated with the predictor, *PROGRAM*. Each component also has its own residual—here, ζ_{0i} and ζ_{1i}—that permits the level-1 parameters (the π's) of one person to differ stochastically from those of others.

Although not yet apparent, the two components of this level-2 submodel have *seven* population parameters: the four regression parameters (the γ's) shown in equation 3.3 and three residual variance/covariance parameters we will soon define. All are estimated when we fit the multilevel model for change to data. We list, label, and define these parameters in the second section of table 3.2 and illustrate their action in the bottom panel of figure 3.4. We discuss their interpretation below.

3.3.1 Structural Components of the Level-2 Submodel

The structural parts of the level-2 submodel contain four level-2 parameters—γ_{00}, γ_{01}, γ_{10}, and γ_{11}—known collectively as the *fixed effects*. The fixed effects capture systematic interindividual differences in change trajectory according to values of the level-2 predictor(s). In equation 3.3, two of the fixed effects, γ_{00} and γ_{10}, are level-2 intercepts; two, γ_{01} and γ_{11}, are level-2 slopes. As in regular regression, the slopes are of greater interest because they represent the effect of predictors (here, the effect of *PROGRAM*) on the individual growth parameters. You can interpret the level-2 parameters much as you do regular regression coefficients, except that you must remember that they describe variation in "outcomes" that are themselves level-1 individual growth parameters.

The easiest way to unravel the meaning of the level-2 fixed effects is to identify a *prototypical individual* distinguished by particular predictor values, substitute those values into the level-2 submodel, and examine the consequences. To derive the postulated level-2 submodel for a prototypical nonparticipant, for example, we set *PROGRAM* to 0 in both parts of equation 3.3 to find: when $PROGRAM = 0$, $\pi_{0i} = \gamma_{00} + \zeta_{0i}$ and $\pi_{1i} = \gamma_{10} + \zeta_{1i}$. This model hypothesizes that, in the population of nonparticipants, the values of initial status and annual rate of change, π_{0i} and π_{1i}, are centered around the level-2 parameters γ_{00}, and γ_{10}. γ_{00} represents the average true initial

status (cognitive score at age 1); γ_{10} represents the average true annual rate of change. By fitting the multilevel model for change to data and estimating these parameters, we address the question: What is the average true trajectory of change in the population for children who did not receive the early intervention program? The lower left panel of figure 3.5 depicts this average population trajectory. Its intercept is γ_{00}; its slope is γ_{10}.

We repeat this process for program participants by setting *PROGRAM* to 1: in this case, $\pi_{0i} = (\gamma_{00} + \gamma_{01}) + \zeta_{0i}$ and $\pi_{1i} = (\gamma_{10} + \gamma_{11}) + \zeta_{1i}$. In the population of program participants, the values of initial status and annual rate of change, π_{0i} and π_{1i}, are centered around $(\gamma_{00} + \gamma_{01})$ and $(\gamma_{10} + \gamma_{11})$. Comparing these centers to those for nonparticipants illustrates that the level-2 parameters γ_{01} and γ_{11} capture the effects of *PROGRAM*. γ_{01} represents the hypothesized difference in average true initial status between groups; γ_{11} represents the hypothesized difference in average true annual rate of change. This allows us to think of the level-2 slopes, γ_{01} and γ_{11}, as "shifts" associated with program participation. The lower right panel of figure 3.4 depicts these shifts. If γ_{01} and γ_{11} are non-zero, the average population trajectories in the two groups differ; if they are both 0, they do not. These two level-2 slope parameters therefore address the question: What is the difference in the average trajectory of true change associated with program participation?

3.3.2 Stochastic Components of the Level-2 Submodel

Each part of the level-2 submodel contains a residual that allows the value of each person's growth parameters to be scattered around the relevant population averages. These residuals, ζ_{0i} and ζ_{1i} in equation 3.3, represent those portions of the level-2 outcomes—the individual growth parameters—that remain "unexplained" by the level-2 predictor(s). As is true for most residuals, we are interested less in their specific values than in their population variances and covariance, which we label σ_0^2, σ_1^2, and σ_{01}. You should know that labeling conventions for these population variances vary considerably across authors and statistical packages. For example, Raudenbush and Bryk (2002) label them τ_{00}, τ_{11}, and τ_{01}, while Goldstein (1995) labels them σ_{u0}^2, σ_{u1}^2, and σ_{u01}.

If child i is a member of the population of nonparticipants, *PROGRAM* takes on the value 0 and the level-2 residuals in equation 3.3 represent deviations between his or her true initial status and annual rate of change from the population average intercept and slope for nonparticipants (γ_{00} and γ_{10}). We display a trajectory for this prototypical child in the lower left panel of figure 3.4. The trajectory begins at a true initial status of $(\gamma_{00} + \zeta_{0i})$ and has a (declining) true annual rate of change of $(\gamma_{10} + \zeta_{1i})$.

Trajectories for other children can be constructed similarly by combining parameters γ_{00} and γ_{10} with other child-specific residuals. The shaded area in this panel is designed to suggest the existence of many different true trajectories, one for each nonparticipant in the population (if they could be fully enumerated). Similarly, if child i is a member of the population of participants, *PROGRAM* takes on the value 1 and the level-2 residuals in equation 3.3 represent deviations between his true initial status and annual rate of change and the population average intercept and slope for participants ($\gamma_{00} + \gamma_{01}$) and ($\gamma_{10} + \gamma_{11}$). To illustrate the heterogeneity in change for this group, the lower right panel of figure 3.4 also includes a shaded area.

Because the level-2 residuals represent deviations between the individual growth parameters and their respective population averages, their variances, σ_0^2 and σ_1^2, summarize the population variation in true individual intercept and slope around these averages. Because they describe those portions of the intercepts and slopes *left over* after accounting for the effect(s) of the model's predictor(s), they are actually *conditional* residual variances. Conditional on the presence of the model's predictors, σ_0^2 represents the population residual variance in true initial status and σ_1^2 represents the population residual variance in true annual rate of change. These variance parameters allow us to address the question: How much heterogeneity in true change remains after accounting for the effects of program participation?

When we posit a level-2 submodel, we also allow for a possible association between individual initial status and individual rates of change. Children who begin at a higher level may have higher (or lower) rates of change. To account for this possibility, we permit the level-2 residuals to be correlated. Since ζ_{0i} and ζ_{1i} represent the deviations of the individual growth parameters from their population averages, their population covariance summarizes the association between true individual intercepts and slopes. Again because of their conditional nature, the population covariance of the level-2 residuals, σ_{01}, summarizes the magnitude and direction of the association between true initial status and true annual rate of change, controlling for program participation. This parameter allows us to address the question: Controlling for program participation, are true initial status and true rate of change related?

To fit the multilevel model for change to data, we must make some assumptions about the level-2 residuals (just as we did for the level-1 residuals in equation 3.2). But because we have two level-2 residuals, we describe their underlying behavior using a *bivariate distribution.* The standard assumption is that the two level-2 residuals, ζ_{0i} and ζ_{1i}, are bivariate normal with mean 0, unknown variances, σ_0^2, and σ_1^2, and unknown

covariance, σ_{01}. We can express these assumptions compactly using matrix notation by writing:

$$\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right). \tag{3.4}$$

Matrix notation greatly simplifies the way in which we codify the model's assumptions. In broad outline, we interpret equation 3.4 in the same way we interpret the assumptions about the level-1 residuals in equation 3.2. The first matrix on the right of the equals sign in parentheses specifies the bivariate distribution's mean vector; here, we assume it to be 0 for each residual (as usual). The second matrix specifies the bivariate distribution's variance-covariance matrix, also known as the *level-2 error covariance matrix* because it captures the covariation among the level-2 residuals (or errors). Two variances, σ_0^2 and σ_1^2, appear along the diagonal, the covariance, σ_{01}, appears on the off-diagonal. Because the covariance between ζ_{0i} and ζ_{1i} is the same as the covariance between ζ_{1i} and ζ_{0i}, the off-diagonal elements are identical—that is, $\sigma_{01} = \sigma_{10}$. The complete set of residual variances and covariances—both the level-2 error variance-covariance matrix and the level-1 residual variance, σ_ε^2—is known collectively as the model's *variance components*.

3.4 Fitting the Multilevel Model for Change to Data

Until the widespread availability of software for fitting multilevel models, researchers used ad hoc strategies like those presented in chapter 2 to analyze longitudinal data: they fitted individual growth trajectories in separate within-person OLS-regression analyses and then they regressed the individual growth parameter estimates obtained on selected level-2 predictors (Willett, 1989). But as previously discussed, this approach has at least two flaws: (1) it ignores information about the individual growth parameter estimates' precision, even though we know that it varies (as seen in the varying residual variances in the bottom panel of figure 3.3); and (2) it replaces *true* individual growth parameters—the real outcomes in a level-2 submodel—with their fallible estimates. The level-2 submodels do not describe the relationship between the parameter *estimates* and predictors, but between the parameters' *true values* and predictors.

Beginning in the 1980s, several teams of statisticians began developing specialized software for fitting the multilevel model for change to data. By the early 1990s, four major packages were widely used: HLM (Bryk, Raudenbush, & Congdon, 1988), MLn (Rasbash & Woodhouse, 1995), GENMOD (Mason, Anderson, & Hayat, 1988), and VARCL (Longford, 1993).

Although the latter two are no longer supported, HLM (Raudenbush, Bryk, Cheong, & Congdon, 2001, available from http://www.ssicentral.com) and MLwiN (Goldstein, 1998, available from http://multilevel.ioe.ac.uk) continue to be modified, expanded, and upgraded regularly to handle an increasing variety of multilevel models. Several multipurpose software packages have also added multilevel routines, including SAS PROC MIXED and PROC NLMIXED (SAS Institute, 2001, http://www.sas.com), the STATA "xt" routines, such as xtreg (Stata, 2001, http://www.stata.com), and SPLUS' NLME library (Pinheiro & Bates, 2001, available from http://cm.bell-labs.com/cm/ms/departments/sia/project/nlme/). So, too, teams of statisticians continue to develop new specialty programs including BUGS (Gilks, Richardson, & Spiegelhalter, 1996, available from http://www.mrcbsu.cam.ac.uk/bugs) and MIXREG (Hedeker & Gibbons, 1996; available from http://www.uic.edu/~hedeker).

As this list suggests, you have a wide and growing array of model fitting options in the investigation of change. We ourselves have no vested interest in any particular software program and do not promote any one above the others. All have their strengths, and we use many of them in our research and in this book. At their core, each program does the same job: it fits the multilevel model for change to data and provides parameter estimates, measures of precision, diagnostics, and so on. There is also some evidence that all the different packages produce the same, or similar, answers to a given problem (Kreft & de Leeuw, 1990). So, in one sense, it does not matter which program you choose. But the packages do differ in many important ways including the "look and feel" of their interfaces, their ways of entering and preprocessing data, their model specification process, their estimation methods, their strategies for hypothesis testing, and the provision of diagnostics. These differences may lead you to decide that one piece of software is especially convenient for your work.

For now, we focus on one particular method of estimation—*maximum likelihood*—as implemented in one program, HLM (Raudenbush, Bryk, Cheong, & Congdon, 2001). In subsequent chapters, we describe other methods of estimation and we apply other statistical software, allowing us to provide advice and compare the competing approaches and packages.

3.4.1 The Advantages of Maximum Likelihood Estimation

The method of maximum likelihood (ML) is currently the most popular approach to statistical estimation. Its popularity results, in part, from its excellent performance in large random samples from well-defined target

populations. As sample size increases, ML estimates have three desirable properties: (1) they are *asymptotically unbiased (consistent)*—they converge on the unknown true values of population parameters; (2) they are *asymptotically normally distributed*—their sampling distributions are approximately normal with known variance; and (3) they are *asymptotically efficient*—their standard errors are smaller than those derived by other methods. Another advantage is that any function of ML estimates is also an ML estimate. This means that predicted growth trajectories (constructed from ML estimates of initial status and rates of change) are ML estimates of the true trajectories. All else being equal, statisticians prefer estimates that are consistent and efficient, that make use of well-established normal theory, and that can generate decent estimates of more complex quantities. Hence the appeal of ML methods.

Notice that the attractive properties of ML estimates are *asymptotic*. This means that in practice—in any actual analysis of a real sample—the properties hold only *approximately*. In large samples, they are likely to hold; in small samples, they may not.[3] To enjoy these advantages, you need a relatively large sample, and the question, how large is large, has no simple answer. Although 10 is certainly small and 100,000 is certainly large, no one can say definitively how large is large enough. In cross-sectional work, Long (1997), for example, recommends a minimum of 100 individuals and he labels sample sizes of 500 "adequate." For a general multilevel model, Snijders and Bosker (1999) consider samples of 30 or more large. Although "rules of thumb" like these provide broad guidelines, we tend to distrust them. The answer to the question "How large?" differs by context, by the particularities of different types of ML estimation, by features of the data, and by the requirements of the tests conducted. Instead we simply offer practical advice: if you use ML methods in "small" samples, treat *p*-values and confidence intervals circumspectly.

Derivation of computational formulas for ML estimation is beyond our scope or intent here. Below, we offer a heuristic explanation of what happens when you use ML methods to fit a multilevel model for change. Our goal is to lay the conceptual foundation for future chapters by explaining why ML estimates make sense and why they have such useful properties. Readers interested in mathematical details should consult Raudenbush and Bryk (2002), Goldstein (1995), or Longford (1993).

3.4.2 Using Maximum Likelihood Methods to Fit a Multilevel Model

Conceptually, maximum likelihood estimates are those guesses for the values of the unknown population parameters that maximize the

probability of observing a particular sample of data. In the early intervention study, they are those estimates of the fixed effects and variance components that make it most likely we would have observed the specific patterns of change found for these 103 children.

To derive an ML estimate for a population parameter, a statistician must first construct a *likelihood function*—an expression that describes the probability of observing the sample data as a function of the model's unknown parameters. Then, he, she, or more accurately, a computer, numerically examines the relative performance of potentially competing estimates until those that maximize the likelihood function are found. The likelihood function for the early intervention data is a function of the probability that we would observe the particular temporal pattern of *COG* values found in the person-period data set. We seek estimates of the fixed effects and variance components whose values maximize the probability of observing this specific pattern.

All likelihood functions are expressed as the product of probabilities (or probability densities). For cross-sectional data, each sample member usually contributes just one term, related to the probability that *that* person has his or her observed data. But because longitudinal data consist of several observations, one per measurement occasion, each person contributes several terms to the likelihood function, which contains as many terms as there are records in the person-period data set.

The particular term that each person contributes on each occasion depends on the specification and assumptions of the hypothesized model. The multilevel model contains structural parts (as shown in, for example, in equations 3.1 and 3.3) and stochastic parts (whose behavior is described in equations 3.2 and 3.4). The structural portion describes the true outcome value for person i on occasion j for his or her particular predictor values. It depends on the unknown values of the fixed effects. The stochastic portion—the level-1 and level-2 residuals—introduce an element of randomness into the proceedings, scattering the observations for person i on occasion j from the structurally specified value.

To derive a maximum likelihood estimate, we must also make assumptions about the *distribution* of the residuals. We have already stated assumptions in equation 3.2 for the level-1 residual, ε_{ij}, and in equation 3.4 for the two-level-2 residuals, ζ_{0i} and ζ_{1i}. Each is assumed to be normally distributed with mean 0; ε_{ij} has unknown variance, σ^2_ε; ζ_{0i} and ζ_{1i} have unknown variances, σ^2_0 and σ^2_1, and covariance, σ_{01}. We also assume that the level-2 residuals are independent of the level-1 residual and that all residuals are independent of the model's predictors.

Given a model and its underlying assumptions, a statistician can write

a mathematical expression for the distribution, or *probability density*, of the outcome. This expression has a mean determined by the model's structural parts and a variance determined by its stochastic parts. As a probability density function, it also describes the likelihood that a person with particular values of the predictors—only *PROGRAM* in equation 3.3—could have particular outcome values using a set of unknown fixed effects and variance components whose values we would like to estimate. That is, it also contains the actual data values observed for that person on that occasion.

It is a short step from here to the full sample likelihood, which we reach by exploiting the well-known multiplicative property of independent probabilities. If you toss one coin, there is a probability of .5 that it will turn up heads. If you independently toss two coins, the probability that each will turn up heads is still .5. But taken together, the probability that you will obtain two heads is only .25 ($.5 \times .5$). If you independently toss three coins, the probability of three heads declines to 0.125 ($.5 \times .5 \times .5$). Statisticians use this principle to create a full sample likelihood from the separate person-period likelihoods just developed. First they write down the value of the probability density of the outcome for each person in the data set on every occasion, thereby describing the likelihood that he or she obtained his or her particular value of the outcome on that occasion. Then they multiply these terms together, yielding an expression for the likelihood of simultaneously observing *all* the data in the person-period data set. Because each person-period likelihood is a function of the data and the unknown parameters, so is their product the full sample likelihood.

To find ML estimates of the unknown population parameters, we identify those values of the unknown parameters that maximize this product of probabilities. Conceptually, imagine a computer trying out billions of alternative estimates, multiplying them together as specified in the sample likelihood function to yield a numeric value for the likelihood, and comparing those numeric values across all of the billions of tries until those estimates that yield the maximum value of the likelihood function are found. These would be the maximum likelihood estimates for this particular problem.

Of course, an enormous numerical search like this is daunting, even with fast computers. Calculus can facilitate the search, but it cannot eliminate the difficulty of working with the products of probability densities that make up the sample likelihood function. To facilitate the search, statisticians use a simple strategy: instead of finding those values of the unknown parameters that maximize the likelihood function, they find those that maximize its logarithm. Working with this new function, known

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:

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

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

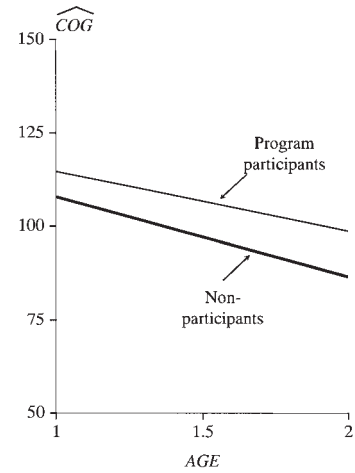


Figure 3.5. Displaying the results of a fitted multilevel model for change. Prototypical trajectories for an average program participant and nonparticipant in the early intervention data.

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

used interchangeably. In HLM, the package we used here, this statistic is labeled a "*t*-ratio." Most programs also output either an associated *p*-value or confidence interval to facilitate testing.[5]

Table 3.3 presents *z*-statistics (column 6) and approximate *p*-values (as superscripts in column 4) for testing hypotheses about the fixed effects. We reject all four null hypotheses, suggesting that each parameter plays a role in the story of the program's effect on children's cognitive development. In rejecting (at the .001 level) the null hypotheses for the two level-2 intercepts, γ_{00} and γ_{10}, we conclude that the average nonparticipant had a non-zero cognitive score at age 1 (hardly surprising!) which declined over time. In rejecting (at the .05 level) the null hypotheses for the two level-2 slopes, γ_{01} and γ_{11}, we conclude that differences between program participants and nonparticipants—in both initial status and annual rates of change—are statistically significant.

3.6 Examining Estimated Variance Components

Estimated variance and covariance components are trickier to interpret as their numeric values have little absolute meaning and there are no graphic aids to fall back on. Interpretation for a single fitted model is especially difficult as you lack benchmarks for evaluating the components' magnitudes. This increases the utility of hypothesis testing, for at least the tests provide some benchmark (against the null value of 0) for comparison.

3.6.1 Interpreting Estimated Variance Components

Variance components assess the amount of outcome variability left—at either level-1 or level-2—after fitting the multilevel model. The level-1 residual variance, σ_ε^2, summarizes the population variability in an average person's outcome values around his or her own true change trajectory. Its estimate for these data is 74.24, a number that is difficult to evaluate in absolute terms. In chapter 4, we provide strategies making relative comparisons to residual variances in other models.

The level-2 variance components summarize the between-person variability in change trajectories that remains after controlling for predictors (here, *PROGRAM*). Using the matrix notation of equation 3.4, we write:

$$\begin{bmatrix} 124.64 & -36.41 \\ -36.41 & 12.29 \end{bmatrix}.$$

Because hypothesis tests, discussed below, reveal that only one of these elements, σ_0^2, is significantly different from 0, it is the only parameter we

discuss here. But because we have no point of comparison, it is difficult to say whether its value, 124.64, is small or large. All we can say is that it quantifies the amount of residual variation in true initial status remaining after we control for program participation.

3.6.2 Single Parameter Tests for the Variance Components

Tests for variance components evaluate whether there is any remaining *residual* outcome variation that could potentially be explained by other predictors. The level of the particular variance component—either level-1 or level-2—dictates the type of predictor that might be added. In general, all the tests are similar in that they assess the evidence concerning the null hypothesis that the parameter's population value is 0, $H_0: \sigma^2 = 0$, against the alternative that it is not, $H_1: \sigma^2 \neq 0$.

There are two very different methods for conducting these hypothesis tests. In this chapter, we offer the simpler approach—the *single parameter test*. Some programs provide this test as a z-statistic—the ratio of the estimated variance component to its asymptotic standard error. Others offer the identical test by squaring the z-statistic and labeling it a χ^2 statistic on one degree of freedom. The appeal of a single parameter hypothesis test is simple. Even if you fit just one statistical model, as we have here, you can garner some insight into the variance components' relative values—at least in comparison to 0.

Unfortunately, statisticians disagree as to the nature, form, and effectiveness of these tests. Miller (1986), Raudenbush and Bryk (2002), and others have long questioned their utility because of their sensitivity to departures from normality. Longford (1999) describes their sensitivity to sample size and imbalance (unequal numbers of observations per person) and argues that they are so misleading that they should be abandoned completely. Because they can be useful for quick, albeit imprecise, assessment, we suggest you examine them only with extreme caution. In section 4.6, we present a superior method for testing hypotheses about variance components, an approach whose use we normally recommend.

Table 3.3 presents single-parameter hypothesis tests for the model's four variance/covariance components. The first three test the null hypothesis that the population variance of the level-1 residuals, σ_ε^2, is 0, that the population variance of the level-2 residuals for initial status, σ_0^2, is 0 and that the population variance of the level-2 residuals for the annual rate of change, σ_1^2, is 0. The last tests whether the covariance between the level-2 residuals for initial status and annual rates of change, σ_{01}, is 0, indicating whether true initial status and true annual rate of

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_ε^2, 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_0^2, 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_1^2 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.

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_ε^2, 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_0^2, 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_1^2 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 sum 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.

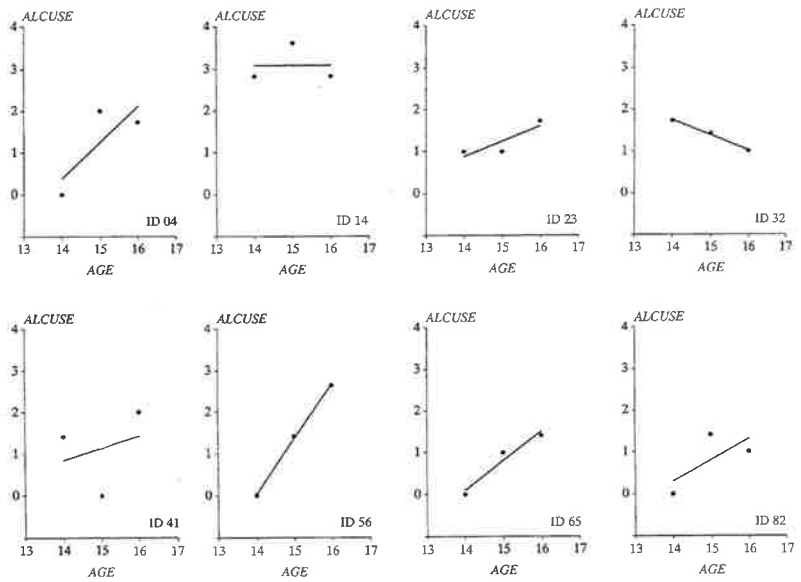


Figure 4.1. Identifying a suitable functional form for the level-1 submodel. Empirical growth plots with superimposed OLS trajectories for 8 participants in the alcohol use study.

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_ε^2. 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.

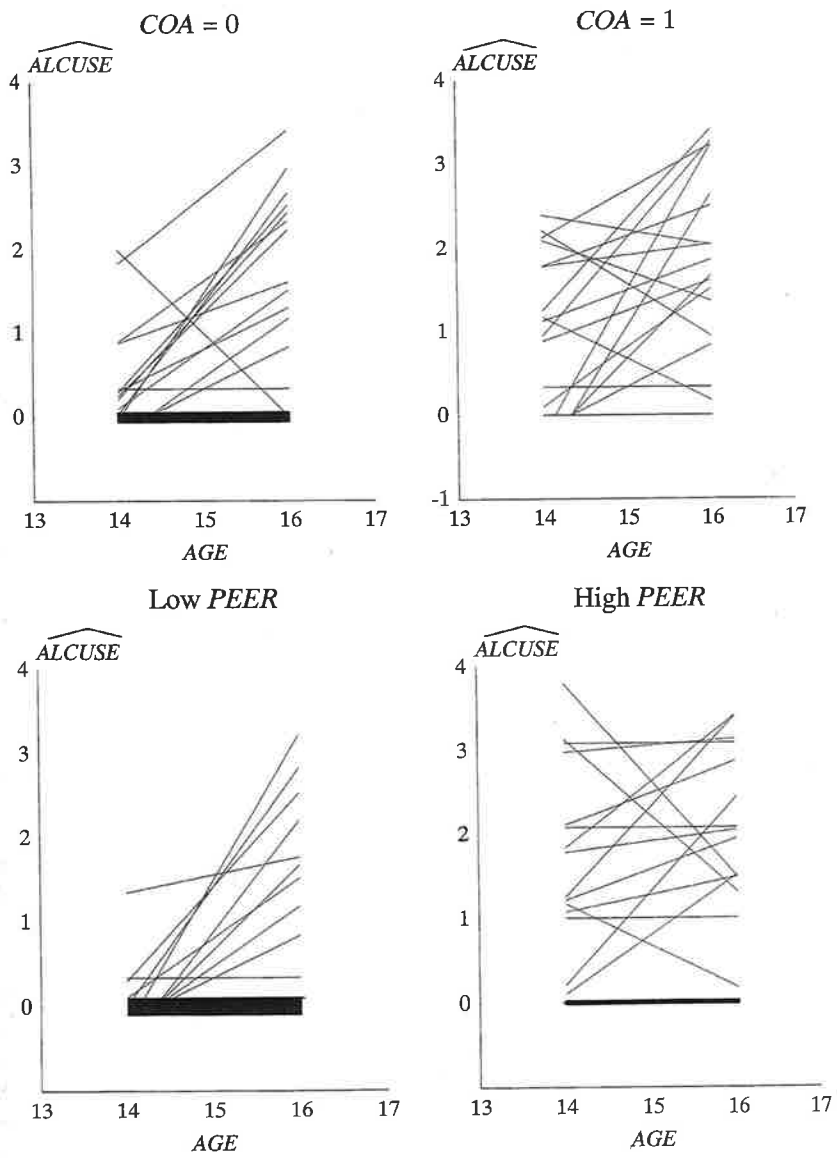


Figure 4.2. Identifying potential predictors of change by examining OLS fitted trajectories separately by levels of selected predictors. Fitted OLS trajectories for the alcohol use data displayed separately by *COA* status (upper panel) and *PEER* alcohol use (lower panel).

- 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 chapter 3, we demonstrated that the γ's describe the average change

trajectories for individuals distinguished by their level-2 predictor values: γ_{00} and γ_{10} are the intercept and slope of the average trajectory for the children of parents who are not alcoholic; $(\gamma_{00} + \gamma_{01})$ and $(\gamma_{10} + \gamma_{11})$ are the intercept and slope of the average trajectory for the children of alcoholics.

The γ's retain these interpretations in the composite model. To demonstrate this equivalence, let us substitute different values of *COA* into the model's structural portion and recover the population average change trajectories. As *COA* has only two values, 0 and 1, recovery is easy. For the children of non-alcoholic parents, we substitute 0 into equation 4.3 to find:

$$\begin{pmatrix}\text{Population average}\\ \text{trajectory for the children}\\ \text{of non-alchoholic parents}\end{pmatrix} = \gamma_{00} + \gamma_{10}TIME_{ij} + \gamma_{01}0 + \gamma_{11}(0 \times TIME_{ij})$$
$$= \gamma_{00} + \gamma_{10}TIME_{ij}, \qquad (4.4a)$$

a trajectory with intercept γ_{00} and slope γ_{10}, as indicated in the previous paragraph. For the children of alcoholic parents, we substitute in 1 to find:

$$\begin{pmatrix}\text{Population average}\\ \text{trajectory for the children}\\ \text{of alchoholic parents}\end{pmatrix} = \gamma_{00} + \gamma_{10}TIME_{ij} + \gamma_{01}1 + \gamma_{11}(1 \times TIME_{ij})$$
$$= (\gamma_{00} + \gamma_{01}) + (\gamma_{10} + \gamma_{11})TIME_{ij}, \qquad (4.4b)$$

a trajectory with intercept $(\gamma_{00} + \gamma_{01})$ and slope $(\gamma_{10} + \gamma_{11})$ also as just described.

Although their interpretation is identical, the γ's in the composite model describe patterns of change in a different way. Rather than postulating first how *ALCUSE* is related to *TIME* and the individual growth parameters, and second how the individual growth parameters are related to *COA*, the composite specification in equation 4.3 postulates that *ALCUSE* depends *simultaneously* on: (1) the level-1 predictor, *TIME*; (2) the level-2 predictor, *COA*; and (3) the *cross-level* interaction, *COA* by *TIME*. From this perspective, the composite model's structural portion strongly resembles a regular regression model with predictors, *TIME* and *COA*, appearing as main effects (associated with γ_{10} and γ_{01}, respectively) and in a *cross-level* interaction (associated with γ_{11}).

How did this cross-level interaction arise, when the level-1/level-2 specification appears to have no similar term? Its appearance arises from the "multiplying out" procedure used to generate the composite model. When we substitute the level-2 submodel for π_{1i} into its appropriate position in the level-1 submodel, the parameter γ_{11}, previously associated only with *COA*, gets multiplied by *TIME*. In the composite model, then, this parameter becomes associated with the interaction term, *COA* by *TIME*. This association makes sense if you consider the following logic. When γ_{11} is non-zero in the level-1/level-2 specification, the *slopes* of the change trajectories differ according to values of *COA*. Stated another way, the effect of *TIME* (whose effect is represented by the slopes of the change trajectories) differs by levels of *COA*. When the effects of one predictor (here, *TIME*) differ by the levels of another predictor (here, *COA*), we say that the two predictors *interact*. The cross-level interaction in the composite specification codifies this effect.

4.2.2 The Stochastic Component of the Composite Model

The *random effects* of the composite model appear in the second set of brackets in equation 4.3. Their representation is more mysterious than that of the fixed effects and differs dramatically from the simple error terms in the separate submodels. But as you would expect, ultimately, they have the same meaning under both the level-1/level-2 and composite representations. In addition, their structure in the composite model provides valuable insight into our assumptions about the behavior of residuals over time in longitudinal data.

To understand how to interpret this stochastic portion, recall that in chapter 3, we described how the random effects allow each person's true change trajectory to be scattered around the relevant population average trajectory. For example, given that the population average change trajectory for the children of non-alcoholic parents (in equation 4.4a) has intercept γ_{00} and slope γ_{10}, the level-2 residuals, ζ_{0i} and ζ_{1i}, allow individual i's trajectory to differ from this average. The true trajectory for individual i, a specific child of non-alcoholic parents, therefore has intercept $(\gamma_{00} + \zeta_{0i})$ and slope $(\gamma_{10} + \zeta_{1i})$. Once this trajectory has been determined, the level-1 residuals, ε_{ij}, then allow his or her data for occasion j to be scattered randomly about it.

We can see how the composite model represents this conceptualization by deriving the true trajectories for different individuals with specific predictor values. Using equation (4.3), we note that if adolescent i has nonalcoholic parents ($COA = 0$):

$$Y_{ij} = [\gamma_{00} + \gamma_{10}TIME_{ij} + \gamma_{01}0 + \gamma_{11}(0 \times TIME_{ij})] + [\zeta_{0i} + \zeta_{1i}TIME_{ij} + \varepsilon_{ij}]$$
$$= [\gamma_{00} + \gamma_{10}TIME_{ij}] + [\zeta_{0i} + \zeta_{1i}TIME_{ij} + \varepsilon_{ij}]$$
$$= (\gamma_{00} + \zeta_{0i}) + (\gamma_{10} + \zeta_{1i})TIME_{ij} + \varepsilon_{ij},$$

leading to a true trajectory with intercept $(\gamma_{00} + \zeta_{0i})$ and slope $(\gamma_{10} + \zeta_{1i})$ as described above. If adolescent i has an alcoholic parent $(COA = 1)$:

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

leading to a true trajectory with intercept $(\gamma_{00} + \gamma_{01} + \zeta_{0i})$ and slope $(\gamma_{10} + \gamma_{11} + \zeta_{1i})$.

A distinctive feature of the composite multilevel model is its "composite residual," the three terms in the second set of brackets on the right of equation 4.3 that combine together the level-1 residual and the two level-2 residuals:

$$\text{Composite residual: } [\zeta_{0i} + \zeta_{1i}TIME_{ij} + \varepsilon_{ij}].$$

The composite residual is not a simple sum. Instead, the second level-2 residual, ζ_{1i}, is multiplied by the level-1 predictor, *TIME*, before joining its siblings. Despite its unusual construction, the interpretation of the composite residual is straightforward: it describes the difference between the observed and the expected value of Y for individual i on occasion j.

The mathematical form of the composite residual reveals two important properties about the occasion-specific residuals not readily apparent in the level-1/level-2 specification: they can be both *autocorrelated* and *heteroscedastic* within person. As we describe briefly below, and more elaborately explain in chapter 7, these are exactly the kinds of properties that you would expect among residuals for repeated measurements of a changing outcome.

When residuals are heteroscedastic, the unexplained portions of each person's outcome have unequal variances across occasions of measurement. Although heteroscedasticity has many roots, one major cause is the effects of omitted predictors—the consequences of failing to include variables that are, in fact, related to the outcome. Because their effects have nowhere else to go, they bundle together, by default, into the residuals. If their impact differs across occasions, the residual's magnitude may differ as well, creating heteroscedasticity. The composite model allows for heteroscedasticity via the level-2 residual ζ_{1i}. Because ζ_{1i} is multiplied by *TIME* in the composite residual, its magnitude can differ (linearly, at least, in a linear level-1 submodel) across occasions. If there are systematic differences in the *magnitudes* of the composite residuals across occasions, there will be accompanying differences in residual *variance*, hence heteroscedasticity.

When residuals are autocorrelated, the unexplained portions of each person's outcome are correlated with each other across repeated occasions. Once again, omitted predictors, whose effects are bundled into the residuals, are a common cause. Because their effects may be present identically in each residual over time, an individual's residuals may become linked across occasions. The presence of the time-invariant ζ_{0i}'s and ζ_{1i}'s in the composite residual of equation 4.3 allows the residuals to be autocorrelated. Because they have only an "i" subscript (and no "j"), they feature identically in each individual's composite residual on every occasion, creating the potential for autocorrelation across time.

4.3 Methods of Estimation, Revisited

When we discussed estimation in section 3.4, we focused on the method of maximum likelihood (ML). As we suggested then, there are other ways of fitting the multilevel model for change. Below, in section 4.3.1, we describe two other methods that are extensions of the popular OLS estimation method, with which you are already familiar: *generalized least squares (GLS)* estimation and *iterative generalized least squares (IGLS)* estimation. In section 4.3.2, we delve deeper into ML methods themselves and distinguish further between two important types of ML estimation—called *full* and *restricted* maximum-likelihood estimation. Finally, in section 4.3.3, we comment on the various methods and how you might choose among them.

4.3.1 Generalized Least-Squares Estimation

Generalized least-squares (GLS) estimation is an extension of ordinary least-squares estimation that allows you to fit statistical models under more complex assumptions on the residuals. Like OLS, GLS seeks parameter estimates that minimize the sum of squared residuals.[1] But instead of requiring the residuals to be independent and homoscedastic, as OLS does, GLS allows them to be autocorrelated and heteroscedastic, as in the composite multilevel model for change.

To understand how you can use GLS to fit the composite multilevel model for change, first reconsider the inefficient exploratory OLS analyses of chapter 2. In section 2.3, our exploratory analyses actually mirrored our later level-1/level-2 specification of the multilevel model for change. To fit the model, we used OLS methods twice. First, in a set of exploratory level-1 analyses, we divided the person-period data set into person-specific chunks (by *ID*) and fit separate within-person regressions of the outcome on *TIME*. Then, in an exploratory level-2 analysis, we regressed

the resultant individual growth parameter estimates on predictors. The existence and form of the composite multilevel model for change suggests that, instead of this piecewise analysis, you could keep the person-period data set intact and regress the outcome (here, *ALCUSE*) on the predictors in the structural portion of the composite model for change (here, *TIME, COA*, and *COA* by *TIME*). This would allow you to estimate the fixed effects of greatest interest (γ_{00}, γ_{10}, γ_{01}, γ_{11}) without dividing the data set into person-specific chunks.

Were you to use OLS to conduct this regression analysis in the full person-period data set, the resultant regression coefficients (estimates of γ_{00}, γ_{10}, γ_{01}, γ_{11}) would indeed be unbiased estimates of the composite model's fixed effects. Unfortunately, their standard errors would not possess the optimal properties needed for testing hypotheses efficiently because the residuals in the stochastic portion of the composite model do not possess the "classical" assumptions of independence and homoscedasticity. In other words, the OLS approach is simply inappropriate in the full person-period data set. To estimate the fixed effects efficiently by fitting the composite model directly in the person-period data set requires the methods of GLS estimation.

This leads to a conundrum. In reality, to estimate the fixed effects in the composite model by a regression analysis in the entire person-period data set, we need GLS methods. But to conduct a GLS analysis, we need to know the shape and contents of the *true error covariance* matrix—specifically we need to know the degree of autocorrelation and heteroscedasticity that actually exists among the residuals in the population so that we can account for this error structure during GLS estimation. We cannot know these population values explicitly, as they are hidden from view; we only possess information on the sample, not the population. Hence the conundrum: to conduct an appropriate analysis of the composite multilevel model for change directly in the person-period data set we need information that we do not, indeed cannot, know.

GLS addresses this conundrum using a two-stage approach. First, fit the composite model by regressing *ALCUSE* on predictors *TIME, COA*, and *COA* by *TIME* in the full person-period data set using OLS methods and *estimate* the error covariance matrix using residuals from the OLS-fitted model. Then, refit the composite model using GLS treating the *estimated* error covariance matrix as though it were the *true* error covariance matrix. In this process, the first stage uses OLS to provide *starting values* (initial estimates) of the fixed effects. These starting values then yield predicted outcome values that allow computation of the residuals for each person on each occasion. The population error covariance matrix is then estimated using these residuals. In the second stage, compute *revised* GLS estimates of the fixed effects and associated standard errors under the assumption that the estimated error covariance matrix from the first stage is a correct representation of the population error covariance matrix of the composite model. All of this, of course, is hidden from view because the computer does it for you.

If GLS estimation with two steps is good, could GLS estimation with many steps be better? This simple question leads to an extension of GLS known as IGLS (*iterative* generalized least squares). Instead of stopping after one round of estimation and refitting, you ask the computer to implement the approach repeatedly, each time using the previous set of estimated fixed effects to re-estimate the error covariance matrix, which then leads to GLS estimates of the fixed effects that are further refined. After each round, you can ask the computer to check whether the current set of estimates is an improvement over the last. If they have not improved (as judged by criteria that you define, or the software package specifies by default), then declare that the process has *converged* and stop, outputting the estimates, their standard errors, and model goodness-of-fit statistics for your perusal.

As with all iterative procedures, the convergence of IGLS is not guaranteed. If your data set is small or severely unbalanced, or if your hypothesized model is too complex, IGLS may iterate indefinitely. To prevent this, all software packages invoke an upper limit on the number of iterations for each analysis (that you can modify, if you wish). If an IGLS analysis fails to converge after a pre-specified number of iterations, you can try again, increasing this upper limit. If it still fails to converge, the estimates may be incorrect and should be treated with caution. We illustrate the use of IGLS methods later in this chapter and discuss issues of nonconvergence in section 5.2.

4.3.2 Full and Restricted Maximum-Likelihood Estimation

Statisticians distinguish between two types of maximum likelihood estimation: *full* (FML) and *restricted* (RML). These two variants on a common theme differ in how the likelihood function is formed, which affects parameter estimation and the strategies used to test hypotheses. You must select a particular ML method *before* fitting models. Perhaps more importantly, you should understand which method your software package selects as its default (although this can usually be overridden).

Although we were not specific in chapter 3, the ML method that we described there was FML. The likelihood function described in section 3.4 assesses the joint probability of simultaneously observing all the

sample data actually obtained. The sample likelihood, a function of the data and the hypothesized model and its assumptions, contains all the unknown parameters, both the fixed effects (the γ's) and the variance components (σ_ε^2, σ_0^2, σ_1^2, and σ_{01}). Under FML, the computer computes those estimates of these population parameters that jointly maximize this likelihood.

FML estimation is not without problems. Because of the way we construct and maximize the likelihood function, FML estimates of the variance components ($\hat{\sigma}_\varepsilon^2$, $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$, and $\hat{\sigma}_{01}$) contain FML estimates of the fixed effects (the $\hat{\gamma}$'s). This means that we ignore uncertainty about the fixed effects when estimating the variance components, treating their values as known. By failing to allocate some degrees of freedom to the estimation of fixed effects, FML overstates the degrees of freedom left for estimating variance components and underestimates the variance components themselves, leading to biased estimates when samples are small (they are still asymptotically unbiased).

These concerns led statisticians to develop restricted maximum likelihood (RML; Dempster Laird & Rubin, 1977). Because both FML and RML require intensive numerical iteration when used to fit the multilevel model for change, we cannot illustrate their differences algebraically. But because similar issues arise when these methods are used to fit simpler models, including the linear regression model for cross-sectional data, we can illustrate their differences in this context where closed-form estimates *can* be written down.

We begin by describing what happens when we use FML to fit a linear regression model to cross-sectional data. Imagine using the following simple regression model to predict an outcome, Y, on the basis of p predictors, X_1 through X_p, in a sample of size n, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$, where i indexes individuals and ε_i represents the usual independent, normally distributed residual with zero mean and homoscedastic variance, σ_ε^2. If it were somehow possible to know the *true population values* of the regression parameters, the residual for individual i would be: $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})$. The FML estimator of the unknown residual variance σ_ε^2, would then be the sum of squared residuals divided by the sample size, n:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n}. \tag{4.5a}$$

Because we imagine that we *know* the population values of the regression coefficients, we need not estimate them to compute residuals, leaving n degrees of freedom for the residual variance calculation.

In practice, of course, we never know the true population values of the regression parameters; we estimate them using sample data, and so:

$$\hat{\varepsilon}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_p X_{pi}\right).$$

Substituting these estimates into equation (4.5a) yields an FML estimate of the residual variance:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n}, \tag{4.5b}$$

because functions of FML estimators, the $\hat{\beta}$'s, are themselves FML estimators.

Notice that the denominator of the FML estimated residual variance in equation 4.5b is the sample size n. Use of this denominator assumes that we still have all the original degrees of freedom in the sample to estimate this parameter. But because we estimated $(p + 1)$ regression parameters to compute the residuals, and did so with uncertainty, we used up $(p + 1)$ degrees of freedom. An *unbiased* estimate of the residual variance decreases the denominator of equation 4.5b to account for this loss:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-(p+1)}. \tag{4.5c}$$

The distinction between the estimated residual variances in equations 4.5b and 4.5c is exactly the same as that between *full* and *restricted* ML estimation in the multilevel model for change. Like RML, equation 4.5c accounts for the uncertainty associated with estimating the regression parameters (the fixed effects) before estimating the residual variance (the variance components); like FML, equation (4.5b) does not.

How are RML estimates computed? Technical work by Patterson and Thompson (1971) and Harville (1974) provides a conceptually appealing strategy. RML estimates of the variance components are those values that maximize the likelihood of observing the sample *residuals* (not the sample data). Once again, an iterative process is used. First, we estimate the fixed effects, the γ's, using some other method, often OLS or GLS. Next, as in regular regression analysis, we use the $\hat{\gamma}$'s to estimate a residual for each person on each occasion (by subtracting observed and predicted values). Under the usual assumptions about the level-1 and level-2 residuals—independence, homoscedasticity, and normality—we can write down the likelihood of observing this particular collection of "data" (that is, *residuals*), in terms of the residuals and the unknown

variance components that govern their distributions. We then take the logarithm of the restricted likelihood and maximize it to yield RML estimates of the variance components, the only unknown parameters remaining (as we have assumed that the fixed effects, the γ's, are known).

For decades, controversy has swirled around the comparative advantages of these two methods. Although Dempster et al. (1977, p. 344) declared RML to be "intuitively more correct," it has not proved to be unilaterally better than FML in practice. In their review of simulation studies that compare these methods for fitting multilevel models, Kreft and deLeeuw (1998) find no clear winner. They suggest that some of the ambiguity stems from the decreased precision that accompanies the decreased small sample bias of RML estimation.

If neither approach is uniformly superior, why belabor this distinction? An important issue is that goodness-of-fit statistics computed using the two methods (introduced in section 4.6) refer to different portions of the model. Under FML, they describe the fit of the entire model; under RML, they describe the fit of only the *stochastic* portion (the random effects). This means that the goodness-of-fit statistics from FML can be used to test hypotheses about any type of parameter, either a fixed effect or a variance component, but those from RML can be used only to test hypotheses about variance components (not the fixed effects). This distinction has profound implications for hypothesis testing as a component of model building and data analysis (as we will soon describe). When we compare models that differ only in their variance components, we can use either method. When we compare models that differ in both fixed effects and variance components, we must use full information methods. To further complicate matters, different software programs use different methods as their default option (although all can use either approach). SAS PROC MIXED, for example, uses RML by default, whereas MLwiN and HLM use FML. This means that when you use a particular statistical computer program, you must be sure to ascertain which method of ML estimation is used by default; if you prefer the alternative method—for reasons of potentially increased precision or the ability to conduct a wider array of hypothesis tests—be sure you are obtaining the desired estimates.

4.3.3 Practical Advice about Estimation

Generalized least squares and maximum likelihood estimation are not identical methods of estimation. They use different procedures to fit the model and they allow us to make different assumptions about the distribution of the random effects. We obtain GLS estimates by *minimizing* a weighted function of the residuals; we obtain ML estimates by *maximizing* a log-likelihood. Only ML estimation requires that the residuals be normally distributed. These differences imply that GLS and ML estimates of the same parameters in the same model using the same data may differ. Although you might find this disturbing, we note that two methods can yield unbiased estimates of the same population parameter but that the estimates themselves can differ. While extensive simulation studies comparing methods are still underway (Draper, 1995; Browne & Draper, 2000), limited data-based comparisons suggest that, in practice, both methods lead to similar conclusions (Kreft, de Leeuw & Kim, 1990).

There is one condition under which the correspondence between GLS and ML methods is well known: if the usual normal distribution assumptions required for ML estimation hold, GLS estimates *are* ML estimates.[2] This equivalence means that, if you are prepared to assume normality for ε and the ζ's, as we did in chapter 3, GLS estimates usually enjoy the same asymptotic unbiasedness, efficiency, and normality that ML estimates do. And since you must invoke normal theory assumptions to conduct hypothesis tests anyway, most data analysts find them compelling and easy to accept. In the remainder of the book, we therefore continue to invoke the standard normal theory assumptions when specifying the multilevel model for change.

GLS and ML are currently the dominant methods of fitting multilevel models to data. They appear in a variety of guises in different packages. Both FML and RML appear in HLM and SAS PROC MIXED. STATA xtreg uses a GLS approach. MLwiN uses IGLS and an extension of it, restricted IGLS (RIGLS), which is the GLS equivalent of RML. And new estimation approaches appear each year. This suggests that whatever we write about a particular method of estimation, or its implementation in a particular package, will soon be out of date. But if your goal is data analysis (not the development of estimation strategies), these modifications of the software are unproblematic. The educated user needs to understand the statistical model, its assumptions, and how it represents reality; the mathematical details of the method of estimation are less crucial. That said, we have three reasons for recommending that you take the time to become comfortable with both ML and GLS methods, at least at the heuristic level presented here. First, you cannot conduct credible analyses nor interpret parameter estimates without at least a conceptual understanding how the model is fit. Second, under the assumptions for which they were designed, these methods have decent statistical properties. Third, most new methods will ultimately descend from, or seek to

ctify weaknesses in, these methods. In other words, the ML and GLS ethods are here to stay.

4 First Steps: Fitting Two Unconditional Multilevel Models for Change

ou've articulated your research questions, created a person-period data et, conducted exploratory analyses, chosen an estimation approach, and elected a software package. Although you might be tempted to begin by tting models that include your substantive predictors, we suggest that ou first fit the two simpler models presented in this section: the *uncon-itional means model* (section 4.4.1) and the *unconditional growth model* section 4.4.2). These unconditional models partition and quantify the utcome variation in two important ways: first, across people without egard to time (the unconditional means model), and second, across oth people *and* time (the unconditional growth model). Their results llow you to establish: (1) whether there is systematic variation in your utcome that is worth exploring; and (2) *where* that variation resides within or between people). They also provide two valuable baselines gainst which you can evaluate the success of subsequent model build-ng, as we discuss in section 4.4.3.

4.4.1 The Unconditional Means Model

The *unconditional means model* is the first model you should always fit. Instead of describing *change* in the outcome over time, it simply describes and partitions the outcome *variation*. Its hallmark is the absence of predictors at every level:

$$\begin{aligned} Y_{ij} &= \pi_{0i} + \varepsilon_{ij} \\ \pi_{0i} &= \gamma_{00} + \zeta_{0i}, \end{aligned} \qquad (4.6a)$$

where we assume, as usual, that:

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \text{ and } \zeta_{0i} \sim N(0, \sigma_0^2). \qquad (4.6b)$$

Notice that because there is only one level-2 residual, ζ_{0i}, we assume *univariate* normality at level-2 (not *bivariate* normality, as we do when we have two level-2 residuals).

The unconditional means model stipulates that, at level-1, the true individual change trajectory for person i is completely flat, sitting at elevation π_{0i}. Because the trajectory lacks a slope parameter associated with a temporal predictor, it cannot tilt. The single part of the level-2 sub-

model stipulates that while these flat trajectories may differ in elevation, their average elevation, across everyone in the population, is γ_{00}. Any interindividual variation in elevation is not linked to predictors. Even though you hope that this model did *not* give rise to your sample data —for it is not really about *change* at all—we recommend that you always fit it first because it partitions the total *variation* in the outcome meaningfully.

To understand how this variance partition operates, notice that flat individual change trajectories are really just *means*. The true mean of Y for individual i is π_{0i}; the true mean of Y across everyone in the popula-tion is γ_{00}. Borrowing terminology from analysis of variance, π_{0i} is the *person-specific mean* and γ_{00} is the *grand mean*. The unconditional means model postulates that the *observed* value of Y for individual i on occasion j is composed of deviations about these means. On occasion j, Y_{ij} deviates from individual i's true mean (π_{0i}) by ε_{ij}. The level-1 residual is thus a "within-person" deviation that assesses the "distance" between Y_{ij} and π_{0i}. Then, for person i, his or her true mean (π_{0i}) deviates from the popula-tion average true mean (γ_{00}) by ζ_{0i}. This level-2 residual is thus a "between-person" deviation that assesses the "distance" between π_{0i} and γ_{00}.

The variance components of equation 4.6b summarize the variability in these deviations across everyone in the population: σ_ε^2 is the "within-person" variance, the pooled scatter of each person's data around his or her own mean; σ_0^2 is the "between-person" variance, the pooled scatter of the person-specific means around the grand mean. The primary reason we fit the unconditional means model is to estimate these variance com-ponents, which assess the amount of outcome variation that exists at each level. Associated hypothesis tests help determine whether there is suffi-cient variation at that level to warrant further analysis. If a variance com-ponent is zero, there is little point in trying to predict outcome variation *at that level*—there is too little variation to explain. If a variance compo-nent is non-zero, then there is some variation at that level that could potentially be explained.

Model A of table 4.1 presents the results of fitting the unconditional means model to the alcohol use data. Its one fixed effect, $\hat{\gamma}_{00}$, estimates the outcome's grand mean across all occasions and individuals. Rejection of its associated null hypothesis ($p < .001$) confirms that the average alcohol consumption of the average adolescent between ages 14 and 16 is non-zero. Squaring 0.922 (which yields 0.85) to obtain its value on the instrument's original scale, we conclude that the average adolescent does drink during these years, but not very much.

Next, examine the random effects, the major purpose for fitting this model. The estimated within-person variance, $\hat{\sigma}_\varepsilon^2$, is 0.562; the estimated

Table 4.1: Results of fitting a taxonomy of multilevel models for change to the alcohol use data ($n = 82$)

		Parameter	Model A	Model B	Model C	Model D	Model E	Model F (*CPEER*)	Model G (*CCOA* & *CPEER*)
Fixed Effects									
Initial status, π_{0i}	Intercept	γ_{00}	0.922*** (0.096)	0.651*** (0.105)	0.316*** (0.131)	−0.317*** (0.148)	−0.314*** (0.146)	0.394*** (0.104)	0.651*** (0.080)
	COA	γ_{01}			0.743*** (0.195)	0.579*** (0.162)	0.571*** (0.146)	0.571*** (0.146)	0.571*** (0.146)
	PEER	γ_{02}				0.694*** (0.112)	0.695*** (0.111)	0.695*** (0.111)	0.695*** (0.111)
Rate of change, π_{1i}	Intercept	γ_{10}		0.271*** (0.062)	0.293*** (0.084)	0.429*** (0.114)	0.425*** (0.106)	0.271*** (0.061)	0.271*** (0.061)
	COA	γ_{11}			−0.049 (0.125)	−0.014 (0.125)			
	PEER	γ_{12}				−0.150~ (0.086)	−0.151~ (0.085)	−0.151~ (0.085)	−0.151~ (0.085)
Variance Components									
Level 1	Within-person	σ^2_ε	0.562*** (0.062)	0.337*** (0.053)	0.337*** (0.053)	0.337*** (0.053)	0.337*** (0.053)	0.337*** (0.053)	0.337*** (0.053)
Level 2	In initial status	σ^2_0	0.564*** (0.119)	0.624*** (0.148)	0.488** (0.128)	0.241** (0.093)	0.241** (0.093)	0.241** (0.093)	0.241** (0.093)
	In rate of change	σ^2_1		0.151** (0.056)	0.151** (0.056)	0.139* (0.055)	0.139* (0.055)	0.139* (0.055)	0.139* (0.055)
	Covariance	σ_{01}		−0.068 (0.070)	−0.059 (0.066)	−0.006 (0.055)	−0.006 (0.055)	−0.006 (0.055)	−0.006 (0.055)
Pseudo R^2 Statistics and Goodness-of-fit									
	$R^2_{y\hat{y}}$			.043	.150	.291	.291	.291	.291
	R^2_ε			.40	.40	.40	.40	.40	.40
	R^2_0				.218	.614	.614	.614	.614
	R^2_1				.000	.079	.079	.079	.079
	Deviance		670.2	636.6	621.2	588.7	588.7	588.7	588.7
	AIC		676.2	648.6	637.2	608.7	606.7	606.7	606.7
	BIC		683.4	663.0	656.5	632.8	628.4	628.4	628.4

~ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

These models predict *ALCUSE* between ages 14 and 16 as a function of *AGE*-14 (at level-1) and various combinations of *COA* and *PEER* (at level-2). Models C, D, and E enter the level-2 predictors in their raw form; Models F and G enter the level-2 predictors in *centered* forms as indicated.

Note: MLwiN, full IGLS.

between-person variance, $\hat{\sigma}_0^2$, is 0.564. Using the single parameter hypothesis tests of section 3.6, we can reject both associated null hypotheses at the .001 level. (Although these tests can mislead—(see section 3.6.2), we use them in table 4.1 because it turns out—for these data, at least—that the conclusions are supported by the superior methods of testing presented in section 4.6.) We conclude that the average adolescent's alcohol consumption varies over time and that adolescents differ from each other in alcohol use. Because each variance component is significantly different from 0, there is hope for linking both within-person and between-person variation in alcohol use to predictors.

The unconditional means model serves another purpose: it allows us to evaluate numerically the relative magnitude of the within-person and between-person variance components. In this data set, they happen to be almost equal. A useful statistic for quantifying their relative magnitude is the *intraclass correlation coefficient*, ρ, which describes the proportion of the total outcome variation that lies "between" people. Because the total variation in Y is just the sum of the within and between-person variance components, the population intraclass correlation coefficient is:

$$\rho = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_\varepsilon^2} \tag{4.7}$$

We can estimate ρ by substituting the two estimated variance components from table 4.1 into equation (4.7). For these data, we find:

$$\hat{\rho} = \frac{0.564}{0.564 + 0.562} = 0.50,$$

indicating that half the total variation in alcohol use is attributable to differences among adolescents.

The intraclass correlation coefficient has another role as well: it summarizes the size of the residual autocorrelation in the composite unconditional means model. To understand how it does this, substitute the level-2 submodel in equation 4.6a into its level-1 submodel to yield the following composite unconditional means model:

$$Y_{ij} = \gamma_{00} + (\zeta_{0i} + \varepsilon_{ij}). \tag{4.8}$$

In this representation, Y_{ij} is composed of one fixed effect, γ_{00}, and one composite residual $(\zeta_{0i} + \varepsilon_{ij})$. Each person has a different composite residual on each occasion of measurement. But notice the difference in the subscripts of the pieces of the composite residual: while the level-1 residual, ε_{ij}, has two subscripts (i and j), the level-2 residual, ζ_{0i}, has only one (i). Each person can have a different ε_{ij} on each occasion, but has only one ζ_{0i} across every occasion. The repeated presence of ζ_{0i} in individual i's composite residual links his or her composite residuals across occasions. The error autocorrelation coefficient quantifies the magnitude of this linkage; in the unconditional means model, the error autocorrelation coefficient *is* the intraclass correlation coefficient. Thus, we estimate that, for each person, the average correlation between any pair of composite residuals—between occasions 1 and 2, or 2 and 3, or 1 and 3—is 0.50. This is quite large, and far from the zero residual autocorrelation that an OLS analysis of these data would require. We discuss the intraclass correlation coefficient further in chapter 7.

4.4.2 The Unconditional Growth Model

The next logical step is the introduction of predictor *TIME* into the level-1 submodel. Based on the exploratory analyses of section 4.1, we posit a linear change trajectory:

$$\begin{aligned} Y_{ij} &= \pi_{0i} + \pi_{1i} TIME_{ij} + \varepsilon_{ij} \\ \pi_{0i} &= \gamma_{00} + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \zeta_{1i}, \end{aligned} \tag{4.9a}$$

where we assume that

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \text{ and } \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right). \tag{4.9b}$$

Because the only predictor in this model is *TIME*, we call equation 4.9 the *unconditional growth model.*

Begin by comparing the unconditional growth model in equation 4.9a to the unconditional means model in equation 4.6a. We facilitate this comparison in table 4.2, which presents these models as well as several others we will soon fit. Instead of postulating that individual i's observed score on occasion j, Y_{ij}, deviates by ε_{ij} from his or her person-specific mean, it specifies that Y_{ij} deviates by ε_{ij} from his or her *true change trajectory.* In other words, altering the level-1 specification alters what the level-1 residuals represent. In addition, we now have a second part to the level-2 submodel that depicts interindividual variation in the rates of change (π_{1i}). But because the model includes no *substantive* predictors, each part of the level-2 submodel simply stipulates that an individual growth parameter (either π_{0i} or π_{1i}) is the sum of an intercept (either γ_{00} or γ_{10}) and a level-2 residual (ζ_{0i} or ζ_{1i}).

An important consequence of altering the level-1 specification is that the meaning of the variance components changes as well. The level-1

Table 4.2: Taxonomy of multilevel models for change fitted to the alcohol use data

Model	Level-1/level-2 specification: level-1 model	level-2 model	Composite model
A	$Y_{ij} = \pi_{0i} + \varepsilon_{ij}$	$\pi_{0i} = \gamma_{00} + \zeta_{0i}$	$Y_{ij} = \gamma_{00} + (\varepsilon_{ij} + \zeta_{0i})$
B	$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \varepsilon_{ij}$	$\pi_{0i} = \gamma_{00} + \zeta_{0i}$ $\pi_{1i} = \gamma_{10} + \zeta_{1i}$	$Y_{ij} = \gamma_{00} + \gamma_{10}TIME_{ij} + (\varepsilon_{ij} + \zeta_{0i} + \zeta_{1i}TIME_{ij})$
C	$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \varepsilon_{ij}$	$\pi_{0i} = \gamma_{00} + \gamma_{01}COA_i + \zeta_{0i}$ $\pi_{1i} = \gamma_{10} + \gamma_{11}COA_i + \zeta_{1i}$	$Y_{ij} = \gamma_{00} + \gamma_{01}COA_i + \gamma_{10}TIME_{ij} + \gamma_{11}COA_i \times TIME_{ij} + (\varepsilon_{ij} + \zeta_{0i} + \zeta_{1i}TIME_{ij})$
D	$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \varepsilon_{ij}$	$\pi_{0i} = \gamma_{00} + \gamma_{01}COA_i + \gamma_{02}PEER_i + \zeta_{0i}$ $\pi_{1i} = \gamma_{10} + \gamma_{11}COA_i + \gamma_{12}PEER_i + \zeta_{1i}$	$Y_{ij} = \gamma_{00} + \gamma_{01}COA_i + \gamma_{02}PEER_i + \gamma_{10}TIME_{ij} + \gamma_{11}COA_i \times TIME_{ij} + \gamma_{12}PEER_i \times TIME_{ij} + (\varepsilon_{ij} + \zeta_{0i} + \zeta_{1i}TIME_{ij})$
E	$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \varepsilon_{ij}$	$\pi_{0i} = \gamma_{00} + \gamma_{01}COA_i + \gamma_{02}PEER_i + \zeta_{0i}$ $\pi_{1i} = \gamma_{10} + \gamma_{12}PEER_i + \zeta_{1i}$	$Y_{ij} = \gamma_{00} + \gamma_{01}COA_i + \gamma_{02}PEER_i + \gamma_{10}TIME_{ij} + \gamma_{12}PEER_i \times TIME_{ij} + (\varepsilon_{ij} + \zeta_{0i} + \zeta_{1i}TIME_{ij})$
F	$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \varepsilon_{ij}$	$\pi_{0i} = \gamma_{00} + \gamma_{01}COA_i + \gamma_{02}CPEER_i + \zeta_{0i}$ $\pi_{1i} = \gamma_{10} + \gamma_{12}CPEER_i + \zeta_{1i}$	$Y_{ij} = \gamma_{00} + \gamma_{01}COA_i + \gamma_{02}CPEER_i + \gamma_{10}TIME_{ij} + \gamma_{12}CPEER_i \times TIME_{ij} + (\varepsilon_{ij} + \zeta_{0i} + \zeta_{1i}TIME_{ij})$
G	$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \varepsilon_{ij}$	$\pi_{0i} = \gamma_{00} + \gamma_{01}(COA_i - \overline{COA}) + \gamma_{02}CPEER_i + \zeta_{0i}$ $\pi_{1i} = \gamma_{10} + \gamma_{12}CPEER_i + \zeta_{1i}$	$Y_{ij} = \gamma_{00} + \gamma_{01}(COA_i - \overline{COA}) + \gamma_{02}CPEER_i + \gamma_{10}TIME_{ij} + \gamma_{12}CPEER_i \times TIME_{ij} + (\varepsilon_{ij} + \zeta_{0i} + \zeta_{1i}TIME_{ij})$

These models predict *ALCUSE* between ages 14 and 16 as a function of *AGE-14* (at level-1) and various combinations of *COA* and *PEER* (at level-2). Models C, D, and E enter the level-2 predictors in their raw form; Models F and G enter the level-2 predictors in *centered* forms as indicated. Results of model fitting appear in Table 4.1.

Figure 4.3. Displaying the results of fitted multilevel models for change. Prototypical trajectories from three models presented in table 4.1: Model B: the unconditional growth model, Model C: the uncontrolled effect of *COA*, Model E: the effect of *COA* controlling for *PEER*.

residual variance, σ^2_ε, now summarizes the scatter of each person's data *around his or her own linear change trajectory* (not his or her person-specific mean). The level-2 residual variances, σ^2_0 and σ^2_1, now summarize between-person variability in initial status and rates of change. Estimating these variance components allows us to distinguish level-1 variation from the two different kinds of level-2 variation and to determine whether interindividual differences in change are due to interindividual differences in true initial status or true rate of change.

Model B in table 4.1 presents the results of fitting the unconditional growth model to the alcohol use data. The fixed effects, $\hat{\gamma}_{00}$ and $\hat{\gamma}_{10}$, estimate the starting point and slope of the population average change trajectory. We reject the null hypothesis for each ($p < .001$), estimating that the average true change trajectory for *ALCUSE* has a non-zero intercept of 0.651 and a non-zero slope of +0.271. Because there are no level-2 predictors, it is simple to plot this trajectory, as we do in the left panel of figure 4.3. Although alcohol use for the average adolescent remains low, we estimate that *ALCUSE* rises steadily between ages 14 and 16, from 0.65 to 1.19. We will soon determine whether these trajectories differ systematically by parental alcoholism history or early peer alcohol use.

To assess whether there is hope for future analyses—whether there is statistically significant variation in individual initial status or rate of

change that level-2 predictors could explain—examine the variance components. By now, we hope you are beginning to see that variance components are often more interesting than fixed effects. The level-1 residual variance, σ^2_ε, summarizes the average scatter of an individual's observed outcome values around his or her own true change trajectory. If the true change trajectory is linear with age, the unconditional growth model will do a better job of predicting the observed outcome data than the unconditional means model, resulting in smaller level-1 residuals and a smaller level-1 residual variance. Comparing $\hat{\sigma}^2_\varepsilon$ in Model B to that of Model A, we find a decline of .40 (from 0.562 to 0.337). We conclude that 40% of the within-person variation in *ALCUSE* is systematically associated with linear *TIME*. Because we can reject the null hypothesis for this variance component in Model B, we also know that some important within-person variation still remains at level-1 ($p < .001$). This suggests that it might be profitable to introduce substantive predictors into the level-1 submodel. We defer discussion of level-1 substantive predictors until section 5.3 because they must be *time-varying* (not *time-invariant* like the level-2 predictors in this data set).

The level-2 variance components quantify the amount of unpredicted variation in the individual growth parameters. σ^2_0 assesses the unpredicted variability in true initial status (the scatter of the π_{0i} around γ_{00}); σ^2_1 assesses the unpredicted variability in true rates of change (the scatter of the π_{1i} around γ_{10}). Because we reject each associated null hypothesis (at $p < .001$ and $p < .01$, respectively), we conclude that there is non-zero variability in both true initial status and true rate of change. This suggests that it worth trying to use level-2 predictors to explain heterogeneity in each parameter. When we do so, these variance components—0.624 and 0.151—will provide benchmarks for quantifying the predictors' effects. We do not compare these variance components with estimates from the unconditional means model because introduction of *TIME* into the model changes their interpretation.

The population covariance of the level-2 residuals σ_{01}, has an important interpretation in the unconditional growth model. It not only assesses the relationship between the level-2 residuals, it quantifies the population covariance between true initial status and true change. This means that we can assess whether adolescents who drink more at age 14 increase their drinking more (or less) rapidly over time. Interpretation is easier if we re-express the covariance as a correlation coefficient, dividing it by the square root of the product of its associated variance components:

$$\hat{\rho}_{\pi_0\pi_1} = \hat{\rho}_{01} = \frac{\hat{\sigma}_{01}}{\sqrt{\hat{\sigma}^2_0\hat{\sigma}^2_1}} = \frac{-0.068}{\sqrt{(0.624)(0.151)}} = -0.22.$$

We conclude that the relationship between true rate of change in *ALCUSE* and its level at age 14 is negative and weak and, because we cannot reject its associated null hypothesis, possibly zero.

We can learn more about the residuals in the unconditional growth model by examining the composite specification of the multilevel model:

$$Y_{ij} = \gamma_{00} + \gamma_{10}TIME_{ij} + (\zeta_{0i} + \zeta_{1i}TIME_{ij} + \varepsilon_{ij}). \tag{4.10}$$

Each person has j composite residuals, one per occasion of measurement. The structure of the composite residual, which combines the original level-1 and level-2 residuals (with ζ_{1i} multiplied by *TIME* before being bundled into the sum), provides the anticipated heteroscedasticity and autocorrelation that longitudinal data analysis may demand.

First, we examine the variances of the composite residual. Mathematical results not presented here allow us to write the population variance of the composite residual on the jth occasion of measurement as:

$$\sigma^2_{Residual_j} = \sigma^2_0 + \sigma^2_1 TIME^2_j + 2\sigma_{01}TIME_j + \sigma^2_\varepsilon. \tag{4.11}$$

Substituting the estimated variance components from Model B in table 4.1 we have:

$$(0.624 + 0.151TIME^2_j - 0.136TIME_j + 0.337).$$

Substituting values for *TIME* at ages 14 ($TIME_1 = 0$), 15 ($TIME_2 = 1$) and 16 ($TIME_3 = 2$), we find estimated composite residual variances of 0.961, 0.976, and 1.293, respectively. While not outrageously heteroscedastic, especially for ages 14 and 15, this is beyond the bland homoscedasticity we assume of residuals in cross-sectional data.

Further mathematical results not shown here allow us to write the autocorrelation between composite residuals on occasions j and j' as:

$$\rho_{Residual_j Residual_{j'}} = \frac{\sigma^2_0 + \sigma_{01}(TIME_j + TIME_{j'}) + \sigma^2_1 TIME_j TIME_{j'}}{\sqrt{\sigma^2_{Residual_j}\sigma^2_{Residual_{j'}}}}, \tag{4.12}$$

where the residual variances in the denominator are given by equation (4.11). Substituting the estimated variance components and *TIME* into equation 4.12 yields a residual autocorrelation of 0.57 between occasions 1 and 2, 0.64 between occasions 2 and 3, and 0.44 between occasions 1 and 3. We conclude that there is substantial autocorrelation between the residuals across successive measurement occasions. We explore this behavior further in chapter 7.

4.4.3 Quantifying the Proportion of Outcome Variation "Explained"

The two unconditional models assess whether there is potentially predictable outcome variation and, if so, where it resides. For these data, the unconditional means model suggests roughly equal amounts of within-person and between-person variation. The unconditional growth model suggests that some of the within-person variation is attributable to linear *TIME* and that there is between-person variation in both true initial status and true rate of change that level-2 predictors might explain.

In multiple regression analysis, we quantify the proportion of outcome variation that a model's predictors "explain" using an R^2 (or adjusted R^2) statistic. In the multilevel model for change, definition of a similar statistic is trickier because total outcome variation is partitioned into several variance components: here, σ^2_ε, σ^2_0 and σ^2_1. As a result, statisticians have yet to agree on appropriate summaries (Kreft & deLeeuw, 1998; Snidjers & Bosker, 1994). Below, we present several *pseudo-R^2 statistics* that quantify how much outcome variation is "explained" by a multilevel model's predictors. First, we assess the proportion of *total* variation explained using a statistic similar to the traditional R^2 statistic; second, we dissect the level-1 and level-2 outcome variation using statistics similar to traditional *adjusted-R^2* statistics. These pseudo-R^2 statistics can be useful data analytic tools, as long as you construct and interpret them carefully.

An Overall Summary of Total Outcome Variability Explained

In multiple regression, one simple way of computing a summary R^2 statistic is to square the sample correlation between observed and predicted values of the outcome. The same approach can be used in the multilevel model for change. All you need do is: (1) compute a predicted outcome value for each person on each occasion of measurement; and (2) square the sample correlation between observed and predicted values. The resultant pseudo-R^2 statistic assesses the proportion of total outcome variation "explained" by the multilevel model's specific combination of predictors.

The bottom panel of table 4.1 presents this pseudo-R^2 statistic (labeled $R^2_{y\hat{y}}$) for each model fit. We calculate these statistics by correlating predicted and observed values of *ALCUSE* for each person on each occasion of measurement. For Model B, for example, the predicted values for individual i on occasion j are: $\hat{Y}_{ij} = 0.651 + 0.271\,TIME_{ij}$. As everyone in this data set has the identical set of measurement occasions (0, 1, and 2), Model B yields only three distinct predicted values:

$$\hat{Y}_{i1} = 0.651 + 0.271(0) = 0.651$$
$$\hat{Y}_{i2} = 0.651 + 0.271(1) = 0.922$$
$$\hat{Y}_{i3} = 0.651 + 0.271(2) = 1.193.$$

Across the entire person-period data set, the sample correlation between these predicted values and the observed values is 0.21, which yields a pseudo-R^2 statistic of .043. We conclude that 4.3% of the total variability in *ALCUSE* is associated with linear time. As we add substantive predictors to this model, we examine whether, and by how much, this pseudo-R^2 statistic increases.

Pseudo-R^2 Statistics Computed from the Variance Components

Residual variation—that portion of the outcome variation *unexplained* by a model's predictors—provides another criterion for comparison. When you fit a series of models, you hope that added predictors further explain unexplained outcome variation, causing residual variation to decline. The magnitude of this decline quantifies the improvement in fit. A large decline suggests that the predictors make a big difference; a small, or zero, decline suggests that they do not. To assess these declines on a common scale, we compute the *proportional reduction in residual variance* as we add predictors.

Each unconditional model yields residual variances that serve as yardsticks for comparison. The unconditional means model provides a baseline estimate of σ^2_ε; the unconditional growth model provides baseline estimates of σ^2_0 and σ^2_1. Each leads to its own pseudo-R^2 statistic.

Let us begin by examining the decrease in within-person residual variance (σ^2_ε) between the unconditional means model and unconditional growth model. As shown in table 4.1, our initial level-1 residual variance estimate, 0.562, drops to .337 in the initial model for change. As the fundamental difference between these models is the introduction of *TIME*, this pseudo-R^2 statistic assesses the proportion of within-person variation "explained by time." We compute the statistic as:

$$\text{Pseudo } R^2_\varepsilon = \frac{\hat{\sigma}^2_\varepsilon(\text{unconditional means model}) - \sigma^2_\varepsilon(\text{unconditional growth model})}{\hat{\sigma}^2_\varepsilon(\text{unconditional means model})} \tag{4.13}$$

For the alcohol use data, we have (.562 − .337)/.562 = 0.400. We conclude that 40.0% of the within-person variation in *ALCUSE* is explained by linear *TIME*. The only way of reducing this variance component further is to add time-varying predictors to the level-1 submodel. As this

data set has no such predictors, $\hat{\sigma}^2_\varepsilon$ remains unchanged in every subsequent model in table 4.1.

We can use a similar approach to compute pseudo-R^2 statistics quantifying the proportional reduction in level-2 residual variance on the addition of one or more level-2 predictors. Each level-2 residual variance component has its own pseudo-R^2 statistic. A level-1 linear change model, with two level-2 variance components, σ^2_0 and σ^2_1, has two pseudo-R^2s. Baseline estimates of these components come from the unconditional growth model. For any subsequent model, we compute a pseudo-R^2 statistic as:

$$\text{Pseudo-}R^2_\zeta = \frac{\hat{\sigma}^2_\zeta(\text{unconditional growth model}) - \sigma^2_\zeta(\text{subsequent model})}{\hat{\sigma}^2_\zeta(\text{unconditional growth model})}. \qquad (4.14)$$

Estimates of these statistics for each of the models in table 4.1 appear in the bottom of the table. We will examine these proportional declines in the next section when we evaluate the results of subsequent model fitting.

Before doing so, however, we close by identifying a potentially serious flaw with the pseudo-R^2 statistics. Unlike traditional R^2 statistics, which will always be positive (or zero), some of these statistics can be *negative*! In ordinary regression, additional predictors generally reduce the residual variance and increase R^2. Even if every added predictor is worthless, the residual variance will not change and R^2 will not change. In the multilevel model for change, additional predictors generally reduce variance components and increase pseudo-R^2 statistics. But because of explicit links among the model's several parts, you can find yourself in extreme situations in which the addition of predictors *increases* the variance components' magnitude. This is most likely to happen when all, or most, of the outcome variation is exclusively either within-individuals or between-individuals. Then, a predictor added at one level reduces the residual variance at that level but potentially *increases* the residual variance(s) at the other level. This yields negative pseudo-R^2 statistics, a disturbing result to say the least. Kreft and de Leeuw (1998, pp. 117–118) and Snijders and Bosker (1999, pp. 99–109) provide mathematical accounts of this phenomenon, explicitly calling for caution when computing and interpreting pseudo-R^2 statistics.

4.5 Practical Data Analytic Strategies for Model Building

A sound statistical model includes all necessary predictors and no unnecessary ones. But how do you separate the wheat from the chaff? We suggest you rely on a combination of substantive theory, research questions, and statistical evidence. *Never* let a computer select predictors mechanically. The computer does not know your research questions nor the literature upon which they rest. It cannot distinguish predictors of direct substantive interest from those whose effects you want to control.

In this section, we describe one data analytic path through the alcohol use data, distilling general principles from this specific case. We begin, in section 4.5.1, by introducing the notion of a *taxonomy* of statistical models, a systematic path for addressing your research questions. In section 4.5.2, we compare fitted models in the taxonomy, interpreting parameter estimates, their associated tests and pseudo-R^2 statistics. In section 4.5.3, we demonstrate how to display analytic results graphically. In section 4.5.4, we discuss alternative strategies for representing the effects of predictors. In the remaining sections of the chapter, we use these basic principles to introduce other important topics related to model building.

4.5.1 A Taxonomy of Statistical Models

A *taxonomy* of statistical models is a systematic sequence of models that, as a set, address your research questions. Each model in the taxonomy extends a prior model in some sensible way; inspection and comparison of its elements tell the story of predictors' individual and joint effects. Most data analysts iterate toward a meaningful path; good analysis does not proceed in a rigidly predetermined order.

We suggest that you base decisions to enter, retain, and remove predictors on a combination of logic, theory, and prior research, supplemented by judicious hypothesis testing and comparison of model fit. At the outset, you might examine the effect of each predictor individually. You might then focus on predictors of primary interest (while including others whose effects you want to control). As in regular regression, you can add predictors singly or in groups and you can address issues of functional form using interactions and transformations. As you develop the taxonomy, you will progress toward a "final model" whose interpretation addresses your research questions. We place quotes around this term to emphasize that we believe no statistical model is *ever* final; it is simply a placeholder until a better model is found.

When analyzing longitudinal data, be sure to capitalize on your intuition and skills cultivated in the cross-sectional world. But longitudinal analyses are more complex because they involve: (1) *multiple level-2 outcomes* (the individual growth parameters), *each* of which can be related to predictors; and (2) *multiple kinds of effects*, both fixed effects and variance

components. A level-1 linear change submodel has two level-2 outcomes; a more complex level-1 submodel may have more. The simplest strategy is to initially include each level-2 predictor simultaneously in all level-2 submodels, but as we show below, they need not remain. Each individual growth parameter can have its own predictors, and one goal of model building is to identify which predictors are important for which level-1 parameters. So, too, although each level-2 submodel can contain fixed and random effects, both are not necessarily required. Sometimes a model with fewer random effects will provide a more parsimonious representation and clearer substantive insights.

Before fitting models, take the time to distinguish between: (1) *question* predictors, whose effects are of primary substantive interest; and, (2) *control* predictors, whose effects you would like to remove. Substantive and theoretical concerns usually support the classification. For the alcohol use data, our classifications and analytic path will differ depending on our research questions. If interest centers on parental influences, *COA* is a question predictor and *PEER* a control. We would then evaluate the effect of *COA* on its own and after control for *PEER*. But if interest centers on peer influences, *PEER* is a question predictor and *COA* a control. We would then evaluate the effect of *PEER* on its own and after control for *COA*. Different classification schemes may lead to the same "final model," but they would arrive there via different paths. Sometimes, they lead to different "final models," each designed to answer its own research questions.

In what follows, we assume that research interest centers on the effects of parental alcoholism; *PEER* is a control. This allows us to adopt the analytic path illustrated in tables 4.1 and 4.2. Model C includes *COA* as a predictor of both initial status and change. Model D adds *PEER* to both level-2 models. Model E is a simplification of Model D in which the effect of *COA* on one of the individual growth parameters (the rate of change) is removed. We defer discussion of Models F and G until section 4.5.4.

4.5.2 Interpreting Fitted Models

You need not interpret every model you fit, especially those designed to guide interim decision making. When writing up findings for presentation and publication, we suggest that you identify a manageable subset of models that, taken together, tells a persuasive story parsimoniously. At a minimum, this includes the unconditional means model, the unconditional growth model, and a "final model." You may also want to present intermediate models that either provide important building blocks or tell interesting stories in their own right.

Columns 4–8 of table 4.1 present parameter estimates and associated single parameter hypothesis tests for five models in our taxonomy. (We discuss the last two models in section 4.5.4.) We recommend that you always construct a table like this because it allows you to compare fitted models systematically, describing what happens as you add and remove predictors. Sequential inspection and comparison of estimated fixed effects and variance components and their associated tests allows you to: (1) ascertain whether, and how, the variability in initial status and rate of change is gradually "explained"; and (2) identify which predictors explain what variation. Tests on the fixed effects help identify the predictors to retain; tests on the variance components help assess whether there is additional outcome variation left to predict. Integrating these conclusions helps identify the sources of outcome variation available for prediction and those predictors that are most effective in explaining that variation. As we have discussed Models A and B in section 4.3, we turn now to Model C.

Model C: The Uncontrolled Effects of COA

Model C includes *COA* as a predictor of both initial status and change. Interpretation of its four fixed effects is straightforward: (1) the estimated initial *ALCUSE* for the average child of non-alchoholic parents is 0.316 ($p < .001$); (2) the estimated differential in initial *ALCUSE* between children of alchoholic and non-alchoholic parents is 0.743 ($p < .001$); (3) the estimated rate of change in *ALCUSE* for an average child of non-alchoholic parents is 0.293 ($p < .001$); and (4) the estimated differential in the rate of change in *ALCUSE* between children of alchoholic and non-alcoholic parents is indistinguishable from 0 (–0.049, *ns*). This model provides uncontrolled answers to our research questions, suggesting that while children of alchoholic parents initially drink more than children of non-alchoholic parents, their rate of change in alcohol consumption between ages 14 and 16 does not differ.

Next examine the variance components. The statistically significant within-person variance component ($\hat{\sigma}^2_\varepsilon$) for Model C is identical to that of Model B, reinforcing the need to explore the effects of time-varying predictors (if we had some). Stability like this is expected because we added no additional level-1 predictors (although estimates can vary because of uncertainties arising from iterative estimation). The level-2 variance components, however, do change: $\hat{\sigma}^2_0$ declines by 21.8% from Model B. Because it is still statistically significant, potentially explainable residual variation in initial status remains. While $\hat{\sigma}^2_1$ is unchanged, it, too, is still statistically significant, suggesting the continued presence of

potentially explainable residual variation in rates of change. These variance components are now called *partial* or *conditional* variances because they quantify the interindividual differences in change that remain unexplained by the model's predictors. We conclude that we should explore the effects of a level-2 predictor like *PEER* because it might help explain some of the level-2 residual variation.

Failure to find a relationship between *COA* and the rate of change might lead some analysts to immediately remove this term. We resist this temptation because *COA* is our focal question predictor and we want to evaluate the full spectrum of its effects. If subsequent analyses continue to suggest that this term be removed, we can always do so (as we do, in Model E).

Model D: The Controlled Effects of COA

Model D evaluates the effects of *COA* on initial status and rates of change in *ALCUSE*, controlling for the effects of *PEER* on initial status and rate of change. Notice that the level-2 intercepts change substantially from Model C: $\hat{\gamma}_{00}$ reverses sign, from +0.316 to –0.317; $\hat{\gamma}_{10}$ increases by 50%, from 0.293 to 0.429. We expect changes like these when we add level-2 predictors to our model. This is because each level-2 intercept represents the value of the associated individual growth parameter—π_{0i} or π_{1i}—when *all* predictors in each level-2 model are 0. In Model C, which includes only one predictor, *COA*, the intercepts describe initial status and rate of change for children of non-alchoholic parents. In Model D, which includes two predictors, the intercepts describe initial status and rate of change for a subset of children of non-alchoholic parents—those for whom *PEER* also equals 0. Because we can reject the null hypothesis associated with each parameter ($p < .001$), we might conclude that children of non-alchoholic parents whose early peers do not drink have non-zero levels of alcohol consumption themselves. But this conclusion is incorrect because the fitted intercept for initial status (–0.317) is *negative* suggesting that the confidence interval for the parameter does not even reach zero from below! As *ALCUSE* cannot be negative, this interval is implausible. As in regular regression, fitted intercepts may be implausible even when they correspond to observable combinations of predictor values. We discuss strategies for improving the interpretability of the level-2 intercepts in section 4.5.4.

The remaining parameters in Model D have expected interpretations: γ_{01} and γ_{11} describe the differential in *ALCUSE* between children of alchoholic and non-alchoholic parents controlling for the effects of *PEER* and γ_{02} and γ_{12} describe the differential in *ALCUSE* for a one-unit

difference in *PEER* controlling for the effect of *COA*. Given our focus on the effects of *COA*, we are more interested in the former effects than the latter. We therefore conclude that, controlling for the effects of *PEER*: (1) the estimated differential in initial *ALCUSE* between children of alchoholic and non-alchoholic parents is 0.579 ($p < .001$); and (2) the estimated differential in the rate of change in *ALCUSE* between children of alchoholic and non-alchoholic parents is indistinguishable from 0 (–0.014, *ns*). This model provides *controlled* answers to our research questions. As before, we conclude that children of alchoholic parents initially drink more than children of non-alchoholic parents but their annual rate of change in consumption between ages 14 and 16 is no different. The magnitude of the early differential in *ALCUSE* is lower after *PEER* is controlled. At least some of the differential initially found between the two groups may be attributable to this predictor.

Next examine the associated variance components. Comparing Model D to the unconditional growth model B, we find that while $\hat{\sigma}^2_\varepsilon$ remains stable (as expected), $\hat{\sigma}^2_0$ and $\hat{\sigma}^2_1$ both decline. Taken together, *PEER* and *COA* explain 61.4% of the variation in initial status and 7.9% of the variation in rates of change. Notice that we *can* compare these random effects across models even though we *cannot* compare their fixed effects ($\hat{\gamma}_{00}$ and $\hat{\gamma}_{10}$). This is because the random effects describe the residual variance of the level-1 growth parameters—π_{0i} or π_{1i}—which retain their meaning across successive models even though the corresponding fixed effects (at level-2) do not.

Rejection of the null hypotheses associated with σ^2_0 and σ^2_1 suggests that there is further unpredicted variation in both initial status and rates of change. If our data set had included other person-level predictors, we would introduce them into the level-2 model to explain this variation. But we have no such predictors. And hypothesis tests for the parameter associated with the effect of *COA* on rate of change (γ_{11}) suggest that it need not be included in Models C or D as a predictor of change. In comparison to all other fixed effects, it is the only one whose null hypothesis cannot be rejected. We conclude that even though *COA* is our focal question predictor, we should remove this term to obtain a more parsimonious model.

Model E: A Tentative "Final Model" for the Controlled Effects of COA

Model E includes *PEER* as a predictor of both initial status and change but *COA* as a predictor of only initial status. For ease of exposition, we tentatively label this our "final model," but we hasten to add that our

decision to temporarily stop here is based on many other analyses not shown. In particular, we examined issues of functional form, including nonlinearity and interactions, and found no evidence of either (beyond that which we addressed by transforming the original outcome and predictor). We discuss issues like these in section 4.8 and in subsequent chapters as we extend the multilevel model for change.

By now, you should be able to interpret the fixed effects in Model E directly. Controlling for the effects of *PEER*, the estimated differential in initial *ALCUSE* between children of alchoholic and non-alchoholic parents is 0.571 ($p < .001$) and controlling for the effect of parental alcoholism, for each 1-point difference in *PEER*: the average initial *ALCUSE* is 0.695 higher and the average rate of change in *ALCUSE* is .151 lower. We conclude that children of alchoholic parents drink more alcohol initially than children of non-alchoholic parents but their rate of change in consumption between ages 14 and 16 is no different. We also conclude that *PEER* is positively associated with early consumption but negatively associated with the rate of change in consumption. Fourteen-year-olds whose friends drink more tend to drink more at that age, but they have a slower rate of increase in consumption over time.

Examining the random effects for Model E in comparison to Model D, we find no differences in $\hat{\sigma}^2_\varepsilon$; $\hat{\sigma}^2_0$ or $\hat{\sigma}^2_1$. This confirms that we lose little by eliminating the effect of *COA* on change. As before, rejection of all three associated null hypotheses suggests the presence of unpredicted variation that we might be able to explain with additional predictors. The population covariance of the level-2 residuals, σ_{01}, summarizes the bivariate relationship between initial status and change, controlling for the specified effects of *COA* and *PEER*; in other words, the *partial* covariance between true initial status and change. Its estimate, –0.006, is even smaller than the unconditional estimate of –0.068 in the initial model for change and its associated hypothesis test indicates that it may well be zero in the population. We conclude that, after accounting for the effects of *PEER* and *COA*, initial status and rate of change in alcohol use are unrelated.

4.5.3 Displaying Prototypical Change Trajectories

Numerical summaries are just one way of describing the results of model fitting. For longitudinal analyses, we find that graphs of fitted trajectories for prototypical individuals are more powerful tools for communicating results. These plots are especially helpful when fitted intercepts in level-2 submodels refer to unlikely or implausible combinations of predictors, as they do for Model E (as evidenced by the negative fitted intercept for the initial status model). Some multilevel software packages provide these

plots; if not, the calculations are simple and can be executed in any spreadsheet or graphics program, as shown below.

Let us begin with Model C, which includes the effect of *COA* on both initial status and change. From table 4.1, we have the following two level-2 fitted models:

$$\hat{\pi}_{0i} = 0.316 + 0.743 COA_i$$
$$\hat{\pi}_{1i} = 0.293 - 0.049 COA_i.$$

We can obtain fitted values for each group by substituting 0 and 1 for *COA*:

$$\textit{When } COA_i = 0 \quad \begin{cases} \hat{\pi}_{0i} = 0.316 + 0.743(0) = 0.316 \\ \hat{\pi}_{1i} = 0.293 - 0.049(0) = 0.293 \end{cases}$$
$$\textit{When } COA_i = 1 \quad \begin{cases} \hat{\pi}_{0i} = 0.316 + 0.743(1) = 1.059 \\ \hat{\pi}_{1i} = 0.293 - 0.049(1) = 0.244. \end{cases}$$

The average child of a non-alchoholic parent has a fitted trajectory with an intercept of 0.316 and a slope of 0.293; the average child of an alchoholic parent has a fitted trajectory with an intercept of 1.059 and a slope of 0.244.

We plot these fitted trajectories in the middle panel of figure 4.3. Notice the dramatic difference in level and trivial (nonsignificant) difference in slope. Unlike the numeric representation of these effects in table 4.1, the graph depicts both how much higher the *ALCUSE* level is at each age among children of alchoholic parents and it emphasizes the similarity in slopes.

We can also obtain fitted trajectories by working directly with the composite specification. From Model C's composite specification $\hat{Y}_{ij} = 0.316 + 0.743 COA_i + 0.293 TIME_{ij} - 0.049 COA_i \times TIME_{ij}$, we obtain the following two trajectories by substituting in the two values of *COA*:

$$\textit{When } COA_i = 0 \quad \begin{cases} \hat{Y}_{ij} = 0.316 + 0.743(0) + 0.293 TIME_{ij} - 0.049(0) TIME_{ij} \\ \hat{Y}_{ij} = 0.316 + 0.293 TIME_{ij} \end{cases}$$
$$\textit{When } COA_i = 1 \quad \begin{cases} \hat{Y}_{ij} = 0.316 + 0.743(1) + 0.293 TIME_{ij} - 0.049(1) TIME_{ij} \\ \hat{Y}_{ij} = 1.059 + 0.244 TIME_{ij}. \end{cases}$$

By working with composite model directly, we obtain fitted trajectories expressed as a function of *TIME*.

It is easy to extend these strategies to models with multiple predictors, some of which may be continuous. Instead of obtaining a fitted function for *each* predictor value, we recommend that you select *prototypical* values of the predictors and derive fitted functions for *combinations* of these

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., $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{3}{4}$). 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

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

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, $\hat{\gamma}_{02}$ 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:

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

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:

$$\text{Deviance} = -2LL_{\text{current model}}. \qquad (4.16)$$

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, 1973) 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

may seem like little more than an excessively parameterized reshuffling of symbols, its structure allows us to invoke a widely used testing strategy.

Most software programs require you to express a general linear hypotheses in matrix notation. This allows decomposition of the hypothesis into two distinct parts: (1) a matrix of multiplying constants (e.g., the 0's and 1's in equation 4.18); and (2) a vector of parameters (e.g., the γ's). To construct the matrix of multiplying constants, commonly labeled a *constraints* or *contrast matrix*, $\boldsymbol{C}$, simply lift the numbers in the general linear hypothesis equation en bloc and array them in the same order. From equation 4.18 we have:

$$\boldsymbol{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

To form the vector of fixed effects, commonly labeled the *parameter vector*, or $\boldsymbol{\gamma}$, lift the parameters in the general linear hypothesis en bloc and array them in the same order as well:

$$\boldsymbol{\gamma} = [\gamma_{00} \quad \gamma_{01} \quad \gamma_{02} \quad \gamma_{10} \quad \gamma_{12}].$$

The general linear hypothesis is formed from the product of the $\boldsymbol{C}$ matrix and the transposed $\boldsymbol{\gamma}$ vector:

$$H_0: \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{02} \\ \gamma_{10} \\ \gamma_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which can be written generically as: $H_0: \boldsymbol{C\gamma'} = \boldsymbol{0}$. For a given model, the elements of $\boldsymbol{C}$ will change from hypothesis to hypothesis but the elements of $\boldsymbol{\gamma}$ will remain the same.

Any general linear hypothesis that can be written in this $\boldsymbol{C\gamma'} = \boldsymbol{0}$ form can be tested using a Wald statistic. Instead of comparing a parameter estimate to its standard error, the Wald statistic compares the *square* of the weighted linear combination of parameters to its estimated variance. As the variance of an estimate is the square of its standard error, the Wald statistic then resembles a squared z-statistic. (Indeed, if you use a Wald statistic to test a null hypothesis about a single fixed effect, W reduces to the square of the usual z-statistic.) Under the null hypothesis and usual normal theory assumptions, W has a χ^2 distribution with degrees of freedom equal to the number of rows in the $\boldsymbol{C}$ matrix (because the number of rows determines the number of independent constraints the null hypothesis invokes). For this hypothesis, we obtain a Wald statistic of 51.01 on 2 *d.f.*, allowing us to reject the composite null hypothesis in equation 4.18 at the .001 level.

General linear hypotheses can address even more complex questions about change over time. For example, when we examined the OLS estimated change trajectories in figure 4.2, we noticed that among children of non-alchoholic parents, those with low values of *CPEER* tended to have a lower initial status and steeper slopes than those with high values of *CPEER*. We might therefore ask whether the former group "catches up" to the latter. This is a question about the "vertical" separation between these two groups' true change trajectories at some later age, say 16.

To conduct such a test, you must once again first figure out the specific parameters under scrutiny. As before, we do so by substituting appropriate predictor values into the fitted model. Setting *COA* to 0 (for the children of non-alchoholic parents) and now selecting –.363 and +.363 as the low and high values of *CPEER* (because they correspond to .5 standard deviations on either side of the centered variable's mean of 0) we have:

$$\begin{aligned} E[Y_j|COA = 0, CPEER = low] &= \gamma_{00} + \gamma_{01}(0) + \gamma_{02}(-.363) + \gamma_{10}TIME_{ij} \\ &\quad + \gamma_{12}(-.363) \times TIME_{ij} \\ &= (\gamma_{00} - .363\gamma_{02}) + (\gamma_{10} - .363\gamma_{12})TIME_{ij} \\ E[Y_j|COA = 0, CPEER = high] &= \gamma_{00} + \gamma_{01}(0) + \gamma_{02}(.363) + \gamma_{10}TIME_{ij} \\ &\quad + \gamma_{12}(.363) \times TIME_{ij} \\ &= (\gamma_{00} + .363\gamma_{02}) + (\gamma_{10} + .363\gamma_{12})TIME_{ij}. \end{aligned}$$

The predicted *ALCUSE* levels at age 16 are found by substituting $TIME = (16 - 14) = 2$ into these equations:

$$\begin{aligned} E[Y_j|COA = 0, CPEER = low] &= \gamma_{00} - .363\gamma_{02} + 2\gamma_{10} - 2(.363)\gamma_{12} \\ E[Y_j|COA = 0, CPEER = high] &= \gamma_{00} + .363\gamma_{02} + 2\gamma_{10} + 2(.363)\gamma_{12}. \end{aligned}$$

How do we express the "catching up" hypothesis? If the low *CPEER* group "catches up," the expected values of the two groups should be identical at age 16. We therefore derive the composite null hypothesis by equating their expected values:

$$\gamma_{00} - .363\gamma_{02} + 2\gamma_{10} - 2(.363)\gamma_{12} = \gamma_{00} + .363\gamma_{02} + 2\gamma_{10} + 2(.363)\gamma_{12}.$$

Simplifying yields the following constraint $\gamma_{02} + 2\gamma_{12} = 0$, which we can be re-expressed as:

$$H_0: 0\gamma_{00} + 0\gamma_{01} + 1\gamma_{02} + 0\gamma_{10} + 2\gamma_{12} = 0. \qquad (4.19)$$

Notice that unlike the composite null hypothesis in equation 4.18, which required two equations, this composite null hypothesis requires just one.

This is a result of a reduction in the number of independent constraints. Because the first hypothesis simultaneously tested *two* independent statements—one about γ_{00} and the other about γ_{10}—it required two separate equations. Because this hypothesis is just a *single* statement—albeit about two parameters, γ_{02} and γ_{12}—it requires just one. This reduction reduces the dimensions of the contrast matrix, $\boldsymbol{C}$.

We next express the composite null hypothesis in matrix form. The parameter vector, $\boldsymbol{\gamma}$, remains unchanged from equation 4.18 because the model has not changed. But because the null hypothesis has changed, the constraint matrix must change as well. Stripping off the numerical constants in equation 4.19 we have $\boldsymbol{C} = [0 \quad 0 \quad 1 \quad 0 \quad 2]$.

As expected, $\boldsymbol{C}$ is just a single row reflecting its single constraint. The composite null hypothesis is:

$$H_0: [0 \quad 0 \quad 1 \quad 0 \quad 2] \begin{bmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{02} \\ \gamma_{10} \\ \gamma_{12} \end{bmatrix} = [0],$$

which has the requisite $\boldsymbol{C\gamma'} = 0$ algebraic form. Conducting this test we find that we can reject the null hypothesis at the usual level of statistical significance ($\chi^2 = 6.23$, $p = .013$). We conclude that these average true change trajectories do not converge by age 16. In other words, the alcohol consumption of children of non-alchoholic parents with *low CPEER* does not catch up to the alcohol consumption of children of non-alchoholic parents with *high CPEER*.

Because many research questions can be stated in this form, general linear hypothesis testing is a powerful and flexible technique. It is particularly useful for conducting *omnibus tests* of several level-2 predictors so that you can assess whether sets of predictors make a difference *as a group*. If we represent a nominal or ordinal predictor using a set of indicator variables, we could use this approach to test their overall effect and evaluate pair-wise comparisons among subgroups.

Although Wald statistics can be used to test hypotheses about variance components, we suggest that you do not do so. The small-sample distribution theory necessary for these tests is poorly developed. It is only in very large samples—that is, *asymptotically*—that the distribution of a W statistic involving variance components *converges* on a χ^2 distribution as your sample size tends to infinity. We therefore do not recommend the use of Wald statistics for composite null hypotheses about variance components.

4.8 Evaluating the Tenability of a Model's Assumptions

Whenever you fit a statistical model, you invoke assumptions. When yc use ML methods to fit a linear regression model, for example, you assun that the errors are independent and normally distributed with consta variance. Assumptions allow you to move forward, estimate paramete interpret results, and test hypotheses. But the validity of your conclusio rests on your assumptions' tenability. Fitting a model with untenat assumptions is as senseless as fitting a model to data that are knowin flawed. Violations lead to biased estimates, incorrect standard errors, a erroneous inferences.

When you fit a multilevel model for change, you also invoke assun tions. And because the model is more complex, its assumptions are mc complex as well, involving both structural and stochastic features at ea level. The structural specification embodies assumptions about the tr functional form of the relationship between outcome and predicto At level-1, you specify the shape of the hypothesized individual chan trajectory, declaring it to be linear (as we have assumed so far) or nc linear (as we assume in chapter 6). At level-2, you specify the relationsl between each individual growth parameter and time-invariant predict And, as in regular regression analysis, you can specify that the leve relationship is linear (as we have so far) or more complex (nonlin discontinuous, or potentially interactive). The stochastic specificat embodies assumptions about that level's outcome (either Y_{ij} at level-1 π_{0i} and π_{1i} at level-2) that remains unexplained by the model's predict Because you know neither their nature nor value, you make assumpti about these error distributions, typically assuming univariate normalit level-1 and bivariate normality at level-2.

No analysis is complete until you examine the tenability of your assu tions. Of course, you can never be completely certain about the tenab of assumptions because you lack the very data you need to evaluate tl tenability: information about the population from which your sample drawn. Assumptions describe *true* individual change trajectories, pop tion relationships between *true* individual growth parameters and lev predictors, and true errors for each person. All you can examine are *observed* properties of *sample* quantities—*fitted* individual change traje ries, *estimated* individual growth parameters, and *sample* residuals.

Must you check the assumptions underlying every statistical model fit? As much as we would like to say yes, reality dictates that we say Repetitive model checking is neither efficient nor plausible. We sug instead that you examine the assumptions of several initial models then again in any model you cite or interpret explicitly.

We offer simple multilevel model checking strategies in the three sections below. Section 4.8.1 reviews methods for assessing functional form; although we introduced the basic ideas earlier, we reiterate them here for completeness. We then extend familiar strategies from regression analysis to comparable issues in the multilevel context: assessing normality (section 4.8.2) and homoscedasticity (section 4.8.3). Table 4.3 summarizes what you should look for at each stage of this work.

4.8.1 Checking Functional Form

The most *direct* way of examining the functional form assumptions in the multilevel model for change is to inspect "outcome versus predictors" plots at each level.

- *At level-1.* For each individual, examine empirical growth plots and superimpose an OLS-estimated individual change trajectory. Inspection should confirm the suitability of its hypothesized shape.
- *At level-2.* Plot OLS estimates of the individual growth parameters against each level-2 predictor. Inspection should confirm the suitability of the hypothesized level-2 relationships.

For the eight adolescents in figure 4.1, for example, the hypothesis of linear individual change seems reasonable for subjects 23, 32, 56 and 65, but less so for subjects 04, 14, 41, and 82. But it is hard to argue for systematic deviations from linearity for these four cases given that the departures observed might be attributable to measurement error. Inspection of empirical growth plots for the remaining adolescents leads to similar conclusions.

Examination of the level-2 assumptions is facilitated by figure 4.4, which plots OLS-estimated individual growth parameters against the two substantive predictors. In the left pair of plots, for *COA*, there is nothing to assess because a linear model is de facto acceptable for dichotomous predictors. In the right pair of plots for *PEER*, the level-2 relationships do appear to be linear (with only a few exceptions).

4.8.2 Checking Normality

Most multilevel modeling packages can output estimates of the level-1 and level-2 errors, ε_{ij}, ζ_{0i} and ζ_{1i}. We label these estimates, $\hat{\varepsilon}_{ij}$, $\hat{\zeta}_{0i}$ and $\hat{\zeta}_{1i}$, "raw residuals." As in regular regression, you can examine their behavior using exploratory analyses. Although you can also conduct formal tests for normality (using Wilks-Shapiro and Kolmogorov-Smirnov statistics, say), we prefer visual inspection of the residual distributions.

Table 4.3: Strategies for checking assumptions in the multilevel model for change, illustrated using Model F of tables 4.1 and 4.2 for the alcohol use data

Assumption and what to expect if the assumption is tenable	What we find in the alcohol use data		
	level-1 residual, $\hat{\varepsilon}_{ij}$	level-2 residual, $\hat{\zeta}_{0i}$	level-2 residual, $\hat{\zeta}_{1i}$
Shape. Linear individual change trajectories and linear relationships between individual growth parameters and level-2 predictors.	Empirical growth plots suggest that most adolescents experience linear change with age. For others, the small number of waves of data (3) makes it difficult to declare curvilinearity making the linear trajectory a reasonable approximation.	Because *COA* is dichotomous, there is no linearity assumption for π_{0i}. With the exception of two extreme data points, the plot of $\hat{\pi}_{0i}$ vs. *PEER* suggests a strong linear relationship.	Because *COA* is dichotomous, there is no linearity assumption for π_{1i}. Plot of π_{1i} vs. *PEER* suggests a weak linear relationship.
Normality. All residuals, at both level-1 and level-2, will be normally distributed.	A plot of $\hat{\varepsilon}_{ij}$ vs. normal scores suggests normality. We find further support for normality in a plot of standardized $\hat{\varepsilon}_{ij}$ vs *ID*, which reveals no unusual data points.	A plot of $\hat{\zeta}_{0i}$ vs. normal scores suggests normality. So does a plot of standardized $\hat{\zeta}_{0i}$ vs. *ID*, which reveals no unusual data points. There is slight evidence of a floor effect in the outcome.	A plot of $\hat{\zeta}_{1i}$ vs. normal scores suggests normality, at least in the upper tail. The lower tail seems compressed. We find further support for this claim when we find no unusual data points in a plot of standardized $\hat{\zeta}_{1i}$ vs. *ID*. There is also evidence of a floor effect in the outcome.
Homoscedasticity. Equal variances of the level-1 and level-2 residuals at each level of every predictor.	A plot of $\hat{\varepsilon}_{ij}$ vs. AGE suggests approximately equal variability at ages 14, 15, and 16.	A plot of $\hat{\zeta}_{0i}$ vs. *COA* suggests homoscedasticity at both values of *COA*. So does a plot vs. *PEER*, at least for values up to, and including, 2. Beyond this, there are too few cases to judge.	A plot of $\hat{\zeta}_{1i}$ vs. *COA* suggests homoscedasticity at both values of *COA*. So does a plot vs. *PEER* at least for values up to, and including, 2. Beyond this, there are too few cases to judge.

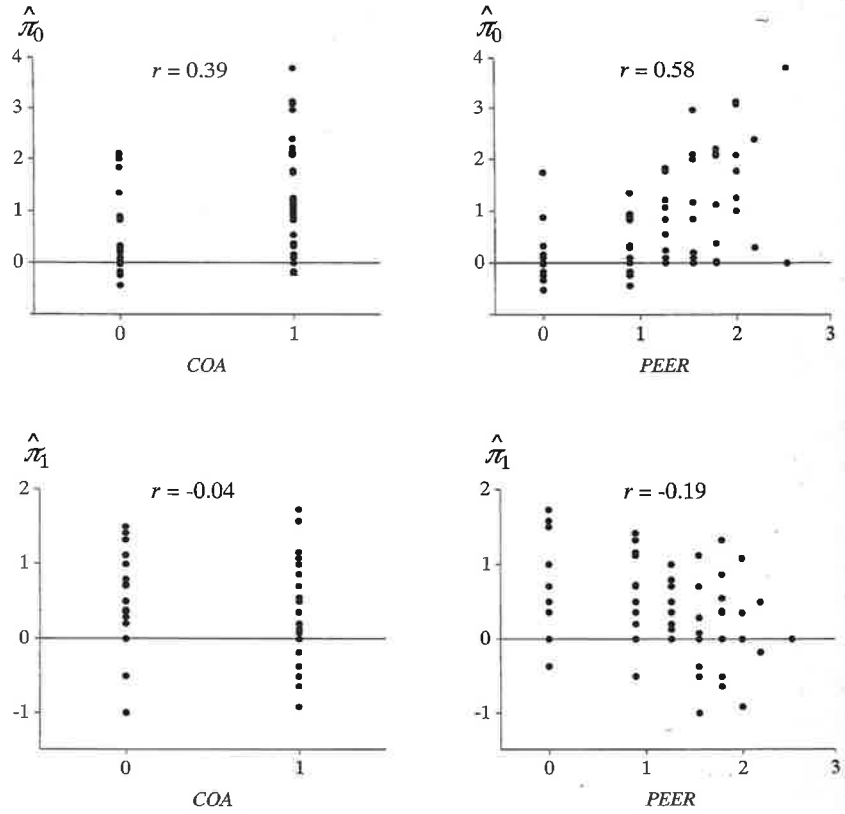


Figure 4.4. Examining the level-2 linearity assumption in the multilevel model for change. OLS estimated individual growth parameters (for the intercept and slope) plotted vs. selected predictors. Left panel is for the predictor *COA*; right panel is for the predictor *PEER*.

For each raw residual—the one at level-1 and the two at level-2—examine a *normal probability plot*, a plot of their values against their associated *normal scores*. If the distribution is normal, the points will form a line. Any departure from linearity indicates a departure from normality. As shown in the left column of figure 4.5, the normal probability plots for Model F for the alcohol use data appear linear for the level-1 residual, $\hat{\varepsilon}_{ij}$, and the first level-2 residual, $\hat{\zeta}_{0i}$. The plot for second level-2 residual, $\hat{\zeta}_{2i}$, is crooked, however, with a foreshortened lower tail falling closer to the center than anticipated. As the second level-2 residual describes unpredicted inter-individual variation in rates of change, we conclude that variability in this distribution's lower tail may be limited. This may

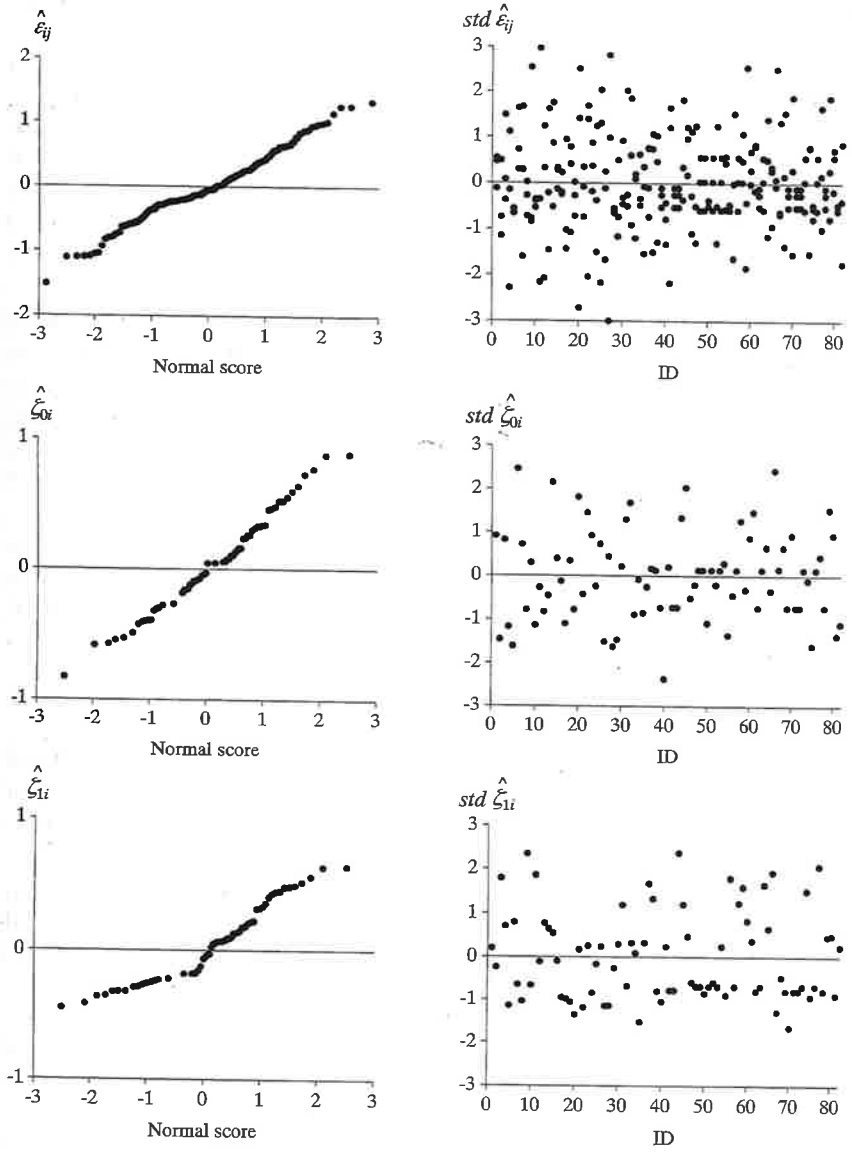


Figure 4.5. Examining normality assumptions in the multilevel model for change. Left panel presents normal probability plots for the raw residuals at level-1 and level-2. Right panel presents plots of standardized residuals at level-1 and level-2 vs. *ID* numbers.

e due to the bounded nature of *ALCUSE*, whose "floor" of zero imposes limit on the possible rates of change.

Plots of *standardized* residuals—either univariate plots or bivariate plots gainst predictors—can also provide insight into the tenability of normality assumptions. If the raw residuals are normally distributed, approximately 95% of the standardized residuals will fall within ±2 standard eviations of their center (i.e., only 5% will be greater than 2). Use aution when applying this simple rule of thumb, however, because there re other distributions that are *not* normal in which about 5% of the bservations also fall in these tails.

You can also plot the standardized residuals by *ID* to identify extreme ıdividuals (as in the right panel of figure 4.5). In the top plot, the standardized level-1 residuals appear to conform to normal theory assumptions—a large majority fall within 2 standard deviations of center, with elatively few between 2 and 3, and none beyond. Plots of standardized evel-2 residuals suggest that the negative residuals tend to be smaller in ıagnitude, "pulled in" toward the center of both plots. This feature is ıost evident for the second level-2 residual, $\hat{\zeta}_{1i}$, in the lower plot, but ıere is also evidence of its presence in the plot for $\hat{\zeta}_{0i}$. Again, compression of the lower tail may result from the fact that the outcome, *ALCUSE*, as a "floor" of zero.

4.8.3 Checking Homoscedasticity

ou can evaluate the homoscedasticity assumption by plotting raw residuals against predictors: the level-1 residuals against the level-1 predictor, ıe level-2 residuals against the level-2 predictor(s). If the assumption olds, residual variability will be approximately equal at every predictor alue. Figure 4.6 presents these plots for Model F of the alcohol use data.

The level-1 residuals, $\hat{\varepsilon}_{ij}$, have approximately equal range and variability at all ages; so, too, do the level-2 residuals plotted against *COA*. The lots of the level-2 residuals against *PEER* reveal a precipitous drop in ariability at the highest predictor values (*PEER* > 2.5), suggesting potenal heteroscedasticity in this region. But the small sample size (only 82 ıdividuals) makes it difficult to reach a definitive conclusion, so we ıtisfy ourselves that the model's basic assumptions are met.

.9 Model-Based (Empirical Bayes) Estimates of the Individual Growth Parameters

)ne advantage of the multilevel model for change is that it improves the recision with which we can estimate individual growth parameters. Yet

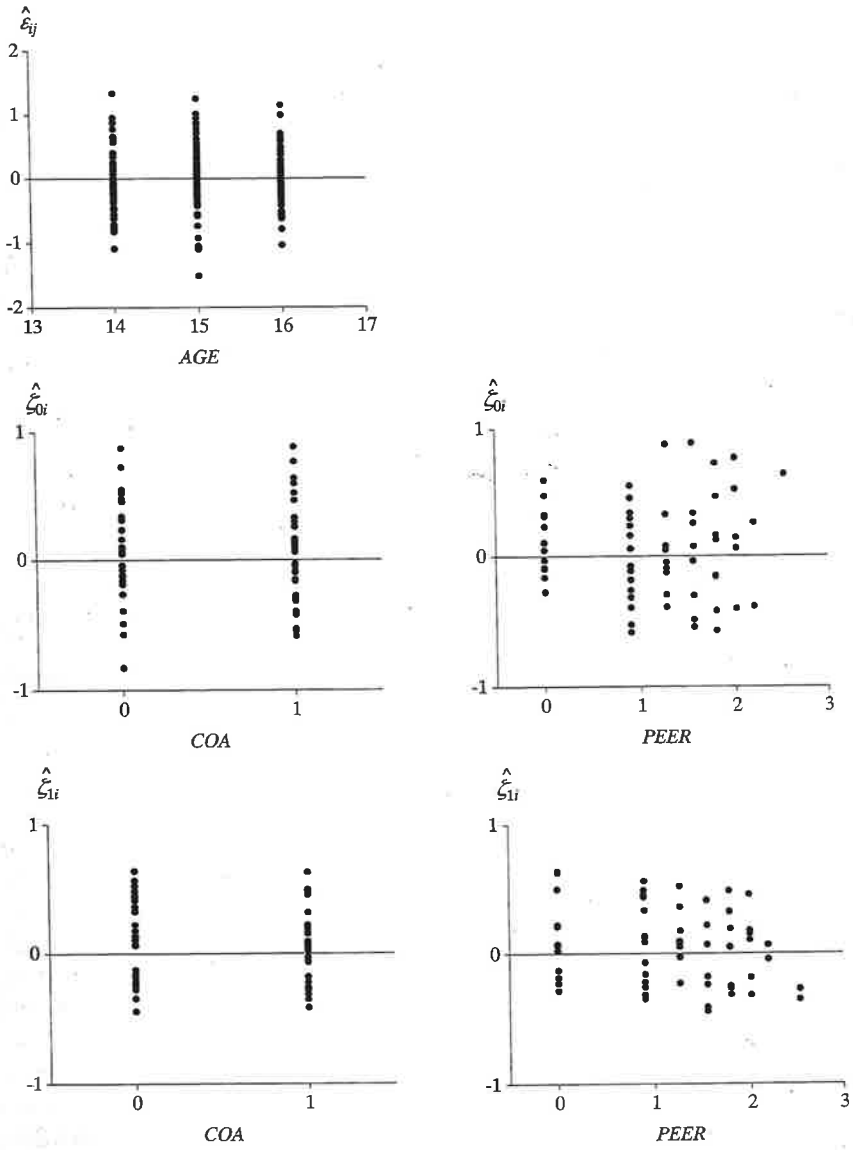


Figure 4.6. Examining the homoscedasticity assumptions in the multilevel model for change. Top panel presents raw level-1 residuals vs. the level-1 predictor *AGE*. Remaining panels present raw level-2 residuals vs. the two level-2 predictors, *COA* and *PEER*.

we have continued to display exploratory OLS estimates even though we know they are inefficient. In this section, we present superior estimates by combining OLS estimates with population average estimates derived from the fitted model. The resultant trajectories, known as *model-based* or *empirical Bayes* estimates, are usually your best bet if you would like to display individual growth trajectories for particular sample members.

There are two distinct methods for deriving model-based estimates. One is to explicitly construct a weighted average of the OLS and population average estimates. The other, which we adopt here, has closer links to the model's conceptual underpinnings: first we obtain population average trajectories based upon an individual's predictor values and second we add individual-specific information to these estimates (by using the level-2 residuals).

We begin by computing a population average growth trajectory for each person in the data set using a particular model's estimates. Adopting Model F for the alcohol use data, we have:

$$\hat{\pi}_{0i} = 0.394 + 0.571COA_i + 0.695CPEER_i$$
$$\hat{\pi}_{1i} = 0.271 - 0.151CPEER_i.$$

Substituting each person's observed predictor values into these equations yields his or her population average trajectory. For example, for subject 23, a child of an alchoholic parent whose friends at age 14 did not drink (resulting in a value of −1.018 for *CPEER*) we have:

$$\begin{aligned}\hat{\pi}_{0,23} &= 0.394 + 0.571(1) + 0.695(-1.018) = 0.257 \\ \hat{\pi}_{1,23} &= 0.271 - 0.151(-1.018) = 0.425,\end{aligned} \tag{4.20}$$

a trajectory that begins at 0.257 at age 14 and rises linearly by 0.425 each year.

This intuitively appealing approach has a drawback: it yields identical trajectories for everyone with the same specific combination of predictor values. Indeed, it is indistinguishable from the same approach used in Section 4.5.3 to obtain fitted trajectories for prototypical individuals. The trajectory in equation 4.20 represents our expectations for the *average* child of alchoholic parents whose young friends do not drink. However, what we seek here is an *individual* trajectory for this person, subject 23. His OLS trajectory does not take advantage of what we have learned from model fitting. Yet his population average trajectory does not capitalize on a key feature of the model: its explicit allowance for interindividual variation in initial status and rates of change.

The level-2 residuals, $\hat{\zeta}_{0i}$ and $\hat{\zeta}_{1i}$, which distinguish each person's growth parameters from his or her population average trajectory, provide

the missing link. Because each person has his or her own set of residuals, we can add them to the model's fitted values:

$$\begin{aligned}\tilde{\pi}_{0i} &= \hat{\pi}_{0i} + \hat{\zeta}_{0i} \\ \tilde{\pi}_{1i} &= \hat{\pi}_{1i} + \hat{\zeta}_{1i},\end{aligned} \tag{4.21}$$

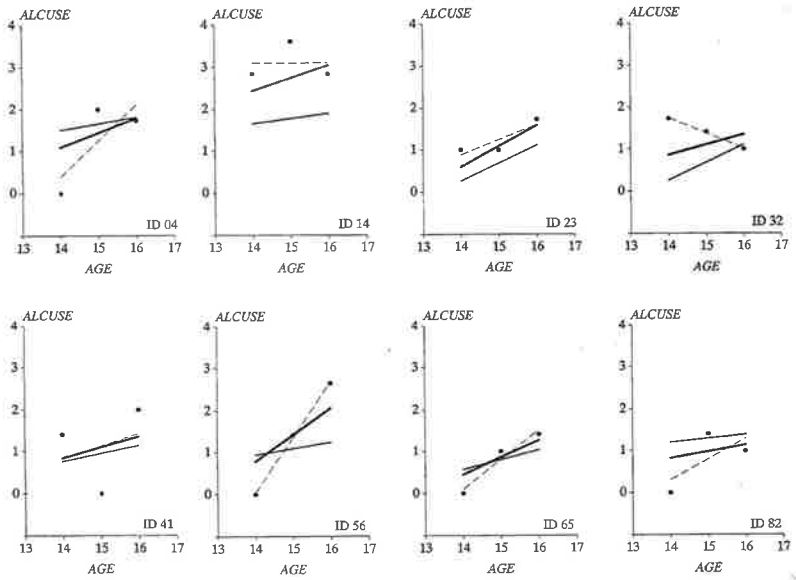
where we place a ~ over the model-based estimates to distinguish them from the population average trajectories. Adding residuals to the population averages distinguishes each person from his or her peer group (defined by his or her predictor values). Most multilevel modeling software programs routinely provide these residuals (or the model-based estimates themselves). For subject 23, for example, the child of alchoholic parents whose peers did not drink, his level-2 residuals of 0.331 and 0.075 yield the following model-based estimates of his individual growth trajectory:

$$\tilde{\pi}_{0,23} = 0.257 + 0.331 = 0.588$$
$$\tilde{\pi}_{1,23} = 0.425 + 0.075 = 0.500.$$

Notice that both of these estimates are larger than the population average values obtained above.

Figure 4.7 displays the observed data for the eight individuals depicted in figure 4.1 and adds three types of fitted trajectories: (1) OLS-estimated trajectories (dashed lines); (2) population average trajectories (faint lines); and (3) model-based individual trajectories (bold lines). First, notice that across the plots, the population average trajectories (the faint lines) are the most stable, varying the least from person to person. We expect greater stability because these are *average* trajectories for groups of individuals who share particular predictor values. People who share identical predictor values will have identical average trajectories, even though their observed outcome data may differ. Population average trajectories do not reflect the behavior of individuals and hence are likely to be the least variable.

Next examine the model-based and OLS estimates (the bold and dashed lines), each designed to provide the individual information we seek. For three adolescents, the difference between estimates is small (subjects 23, 41, and 65), but for four others (subjects 4, 14, 56, and 82) it is pronounced and for subject 32, it is profound. We expect discrepancies like these because we estimate each trajectory using a different method and they depend upon the data in different ways. This does not mean that one of them is "right" and the other "wrong." Each has a set of statistical properties for which it is valued. OLS estimates are unbiased but inefficient; model-based estimates are biased, but more precise.



igure 4.7. Model-based (empirical Bayes) estimates of the individual growth trajectories. ach plot presents the observed *ALCUSE* measurements (as data points), OLS fitted tra-:ctories (dashed lines), population average trajectories (faint lines), and model-based mpirical Bayes trajectories (bold lines).

Now notice how each model-based trajectory (in bold) falls between ts OLS and population average trajectories (the dashed and faint lines). This is a hallmark of the model-based procedure to which we alluded earlier. Numerically, the model-based estimates are weighted averages of he OLS and population average trajectories. When OLS estimates are precise, they have greater weight; when OLS estimates are imprecise, the population average trajectories have greater weight. Because OLS trajec-ories differ markedly from person to person, the model-based trajecto-ies differ as well, but their discrepancies are smaller because the population average trajectories are more stable. Statisticians use the term "borrowing strength" to describe procedures like this in which individual estimates are enhanced by incorporating information from others with whom he or she shares attributes. In this case, the model-based trajecto-ies are *shrunk* toward the average trajectory of that person's peer group (those with the same predictor values). This combination yields a supe-ior, more precise, estimate.

Model-based estimates are also more precise because they require esti-mation of fewer parameters. In positing the multilevel model for change, we assume that everyone shares the same level-1 residual variance, σ^2_ε. When we fit OLS trajectories, we estimate a separate level-1 variance for each individual in the sample. Fewer parameters in the multilevel model for change mean greater precision.

In choosing between OLS- and model-based trajectories, you must decide which criterion you value most, *unbiasedness* or *precision*. Statisti-cians recommend precision—indeed, increased precision is a funda-mental motivation for fitting the multilevel model. But as we extol the virtues of model-based estimates, we conclude with a word of caution. Their quality depends heavily on the quality of the model fit. If the model is flawed, particularly if its level-2 components are specified incorrectly, then the model-based estimates will be flawed as well.

How might you use model-based estimates like these in practice? Stage (2001) provides a simple illustration of the power of this approach in his evaluation of the relationship between first-grade reading fluency and changes in oral reading proficiency in second-graders. He began by fitting a multilevel model for change to four waves of second-grade data, demonstrating that while first-grade performance was a strong predictor of initial status it was not a statistically significant predictor of rate of change. Stage went on to compute empirical Bayes estimates of the number of words each child was able to read by the end of second grade and he compared these estimates to: (1) the number of words each child was observed to have read at the end of second grade; and (2) the number of words each child was predicted to have read on the basis of simple OLS regression analyses within child. As Stage suggests, adminis-trators might be better off assigning children to summer school programs (for remedial reading) not on the basis of observed or OLS-predicted end-of-year scores but rather on the basis of the empirical Bayes estimates, which yield more precise estimates of the child's status at the end of the year.

5

Treating *TIME* More Flexibly

Change is a measure of time

—Edwin Way Teale

l the illustrative longitudinal data sets in previous chapters share two ·uctural features that simplify analysis. Each is: (1) balanced—everyone assessed on the identical number of occasions; and (2) time-uctured—each set of occasions is identical across individuals. Our analyses ıve also been limited in that we have used only: (1) time-invariant ·edictors that describe immutable characteristics of individuals or their ıvironment (except for *TIME* itself); and (2) a representation of *TIME* at forces the level-1 individual growth parameters to represent "initial ıtus" 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 e irregularly spaced, their number and spacing can vary across participants. Each individual can have his or her own data collection schedule ıd the number of waves can vary without limit from person to person.), too, predictors of change can be time-invariant or time-varying, and ıe level-1 submodel can be parameterized in a variety of interesting ways.

In this chapter, we demonstrate how you can fit the multilevel model ır 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 irregular. In section 5.2, we illustrate what to do when the number of aves per person differs as well; we also discuss the problem of missing ata, the most common source of imbalance in longitudinal work. In ·ction 5.3, we demonstrate how to include time-varying predictors in ɔur data analysis. We conclude, in section 5.4, by discussing why and how ɔu 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

are then sparser at the earliest and latest ages, which can complicate the specification of the level-1 submodel.

In this section, we show how you can use the methods of previous chapters to analyze data sets with variably spaced measurement occasions. All you need to deal with are some minor coding issues for the temporal predictor in the person-period data set; model specification, parameter estimation, and substantive interpretation proceeds as before. To illustrate just how simple the analysis can be, we begin by discussing data sets in which the *number* of waves is constant but their *spacing* varies. We discuss data sets in which the *number* of waves varies as well in section 5.2.

5.1.1 The Structure of Variably Spaced Data Sets

We illustrate how to analyze data sets with variably spaced measurement occasions using a small sample extracted from the Children of the National Longitudinal Study of Youth (CNLSY). The data set, comprising children's scores on the reading subtest of the Peabody Individual Achievement Test (PIAT), includes three waves of data for 89 African-American children. Each child was 6 years old in 1986, the first year of data collection. During the second wave of data collection, in 1988, these children were to be 8; during the third wave, in 1990, they were to be 10. We focus here on an unconditional growth model, not the inclusion of level-2 predictors, because this second aspect of analysis remains unchanged.

Table 5.1 presents excerpts from the person-period data set. Notice that its structure is virtually identical to all person-period data sets shown so far. The only difference is that it contains *three* temporal variables denoting the passage of time: *WAVE*, *AGE*, and *AGEGRP*. Although we will include only one of these in any given model, a distinctive feature of time-unstructured data sets is the possibility of multiple metrics for clocking time (often called metameters).

WAVE is the simplest but least analytically useful of the three. Although its values—1, 2, and 3—reflect the study's design, they have little substantive meaning when it comes to addressing the research question. Because *WAVE* does not identify the child's age at each occasion, nor does it capture the chronological distance between occasions, it cannot contribute to a meaningful level-1 submodel. We mention this issue explicitly because empirical researchers sometimes postulate individual growth models using design variables like *WAVE* (or year of data collection) even though other temporal predictors are generally more compelling.

AGE is a better predictor because it specifies the child's actual age (to the nearest month) on the day each test was administered. A child like

Table 5.1: Excerpts from the person-period data set for the reading study

ID	*WAVE*	*AGEGRP*	*AGE*	*PIAT*
04	1	6.5	6.00	18
04	2	8.5	8.50	31
04	3	10.5	10.67	50
27	1	6.5	6.25	19
27	2	8.5	9.17	36
27	3	10.5	10.92	57
31	1	6.5	6.33	18
31	2	8.5	8.83	31
31	3	10.5	10.92	51
33	1	6.5	6.33	18
33	2	8.5	8.92	34
33	3	10.5	10.75	29
41	1	6.5	6.33	18
41	2	8.5	8.75	28
41	3	10.5	10.83	36
49	1	6.5	6.50	19
49	2	8.5	8.75	32
49	3	10.5	10.67	48
69	1	6.5	6.67	26
69	2	8.5	9.17	47
69	3	10.5	11.33	45
77	1	6.5	6.83	17
77	2	8.5	8.08	19
77	3	10.5	10.00	28
87	1	6.5	6.92	22
87	2	8.5	9.42	49
87	3	10.5	11.50	64
...	...	...	...	...

Note that *TIME* is clocked using three distinct variables: *WAVE*, *AGEGRP*, and *AGE*.

ID 04, who had just turned 6 at wave 1, has an *AGE* of 6.00 for that record a child like *ID* 87, who would soon turn 7, has an *AGE* of 6.92. The averag child is 6.5 years old at wave 1, as we would expect if births and testin occasions were randomly distributed. If data collection had proceede according to plan, the average child would have been 8.5 and 10.5 year old at the next two waves. Not surprisingly, actual ages varied aroun these targets. By wave 2, the youngest child had just turned 8 while th oldest was well over 9. By wave 3, the youngest child had just turned 1 while the oldest was nearly 12. Like many longitudinal studies, the CNLS suffers from "occasion creep"—over time, the temporal separation o

ccasions widens as the actual ages exceed design projections. In this data
:t, the average child is 8.9 years in wave 2 and nearly 11 years in wave 3.

The third temporal variable, *AGEGRP*, is a time-structured predictor
iat is more substantively meaningful than the design variable *WAVE*. Its
alues indicate the child's "expected age" on each measurement occasion
5.5, 8.5, and 10.5). This time-structured predictor clocks time on a scale
iat is comparable numerically to the irregularly spaced predictor *AGE*.
dding *AGEGRP* to the person-period data set allows us to demonstrate
iat the characterization of a data set as time-structured or irregular can
epend on nothing more than the *cadence* of the temporal predictor used
) postulate a model. If we postulate our model using *AGEGRP*, the data set
; time-structured; if we postulate a comparable model using *AGE*, it is not.

The multilevel model for change does not care if the individual-
pecific cadence of the level-1 predictor is identical for everyone or if it
aries from case to case. Because we fit the model using the actual
umeric values of the temporal predictor, spacing is irrelevant. We can
postulate and fit a comparable model regardless of the variable's cadence.
Of far greater importance is the choice of the functional form for the
evel-1 submodel. Should it represent linear change or a more complex
hape for the individual growth trajectory? Might this decision depend
ipon the specific temporal predictor chosen for model building?

To address these questions, figure 5.1 presents empirical change plots
vith superimposed OLS linear change trajectories for 9 children. Each
panel plots each child's *PIAT* scores twice, once for each temporal pre-
lictor. We use •'s and a dashed line when plotting by *AGE*; we use +'s and
a solid line when plotting by *AGEGRP*. With just three waves of data—
vhichever temporal predictor we use—it is difficult to argue for anything
out a linear change individual growth model.

If we can postulate a linear change individual growth model using
:ither temporal predictor, which one should we use? As argued above,
ve prefer *AGE* because it provides more precise information about the
:hild at the moment of testing. Why set this information aside just to use
:he equally spaced, but inevitably less accurate, *AGEGRP*. Yet this is what
many researchers do when analyzing longitudinal data—indeed, it is what
we did in chapters 3 and 4. There, instead of using the participant's
precise ages, we used integers: 12, 18, and 24 months for the children in
:hapter 3; 14, 15, and 16 for the teenagers in chapter 4. Although the
loss of precision may be small, as suggested by the close correspondence
between the pairs of fitted OLS trajectories in each panel of figure 5.1,
there are children for whom the differential is much larger. To inve-
stigate this question empirically, we fit two multilevel models for
change to these data: one using *AGEGRP*, another using *AGE* as the

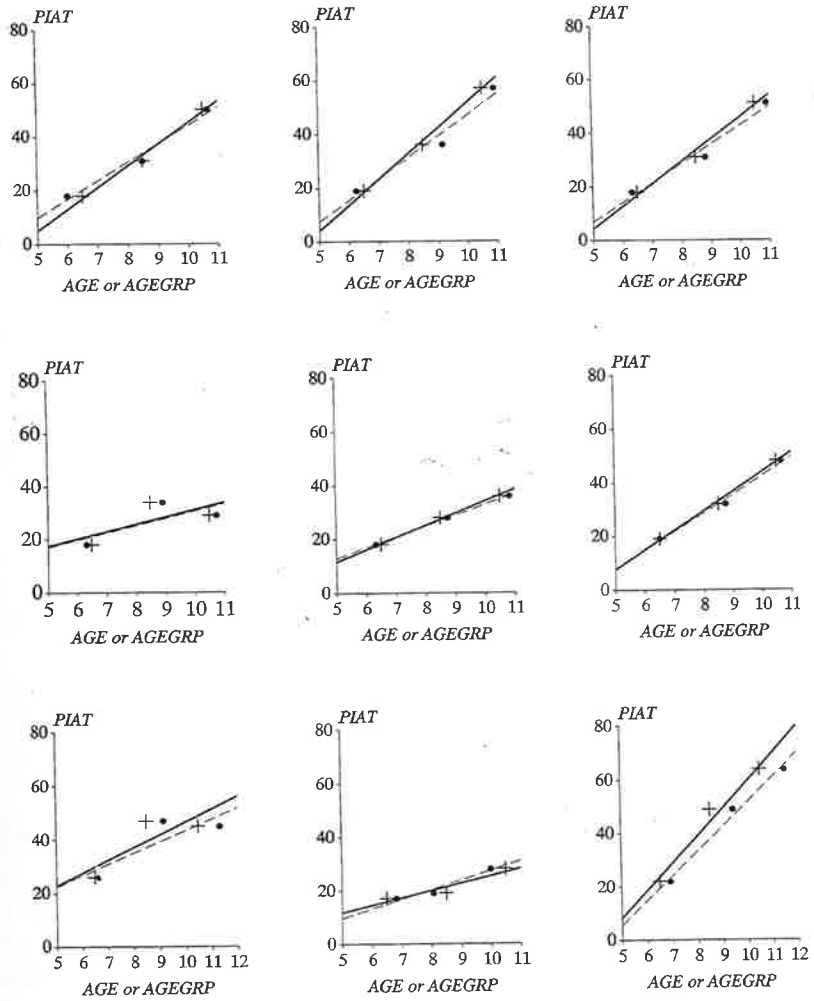


Figure 5.1. Comparing time-structured and time-unstructured representations of the effect of *TIME*. Empirical change plots with superimposed OLS trajectories for 9 participants in the reading study. The +'s and solid lines are for *TIME* clocked using the child's *target* age at data collection; the •'s and dashed lines are for *TIME* clocked using each child's observed age.

nporal predictor at level-1. Doing so allows us to demonstrate how to alyze irregularly spaced data sets *and* to illustrate the importance of essing the merits of alternative metrics for time empirically.

5.1.2 Postulating and Fitting Multilevel Models with Variably Spaced Waves of Data

gardless of which temporal representation we use, we postulate, fit, and erpret the multilevel model for change using the same strategies. apting the general specification of an unconditional growth model in uations 4.9a and 4.9b, let Y_{ij} be child i's *PIAT* score on occasion j and ME_{ij} represent either temporal variable:

$$\begin{aligned} Y_{ij} &= \pi_{0i} + \pi_{1i} TIME_{ij} + \varepsilon_{ij} \\ \pi_{0i} &= \gamma_{00} + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \zeta_{1i}, \end{aligned} \tag{5.1a}$$

ıere

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \text{ and } \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right). \tag{5.1b}$$

ve center both *AGE* and *AGEGRP* on age 6.5—the average child's age at ve 1—the parameters have the usual interpretations. In the population ɔm which this sample was drawn, γ_{00} represents the average child's true tial status (at age 6.5); γ_{10} represents the average child's annual rate of ıe change between ages 6 and 11; σ_ε^2 summarizes the within-child scatter ɔund his or her own true change trajectory; and σ_0^2 and σ_1^2 summarize the tween-child variability in initial status and annual rates of change.

Use of a generic representation $TIME_{ij}$ in the level-1 growth model ıstead of a specific representation like *AGE* − 6.5 or *AGEGRP* − 6.5) :lds these interpretations. We can postulate the same model for either edictor because $TIME_{ij}$ includes subscripts that are both person-specific and time-specific (j). If *TIME* represents *AGEGRP* − 6.5, the data set time structured; if we use *AGE* − 6.5, it is not. From a data-analytic per- ective, you just specify the relevant temporal representation to your sta- tical software. From an interpretive perspective, the distinction is moot.

Table 5.2 presents the results of fitting these two unconditional growth ɔdels to these data: the first uses *AGEGRP* − 6.5; the second uses *AGE* 6.5. Each was fit using full ML in SAS PROC MIXED. The parameter timates for initial status, $\hat{\gamma}_{00}$, are virtually identical—21.16 and 21.06—are those for the within-child variance, σ_ε^2: 27.04 and 27.45. But the nilarities stop there. For the slope parameter, γ_{10}, the estimated growth

Table 5.2: Results of using alternative representations for the main effect of *TIME* ($n = 89$) when fitting an unconditional growth model to the CNLSY reading data

			Predictor representing *TIME*	
	Parameter		*AGEGRP* − 6.5	*AGE* − 6.5
Fixed Effects				
Initial status, π_{0i}	Intercept	γ_{00}	21.1629***	21.0608***
			(0.6143)	(0.5593)
Rate of Change, π_{1i}	Intercept	γ_{10}	5.0309***	4.5400***
			(0.2956)	(0.2606)
Variance Components				
Level-1:	within-person	σ_ε^2	27.04***	27.45***
Level-2:	In initial status	σ_0^2	11.05*	5.11
	In rate of change	σ_1^2	4.40***	3.30***
Goodness-of-fit				
	Deviance		1819.8	1803.9
	AIC		1831.9	1815.9
	BIC		1846.9	1830.8

$^{\sim}p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

The first model treats the data set as time-structured by using the predictor (*AGEGRP* − 6.5); the second model treats the data set as time-unstructured by using each child's actual age at each assessment, (*AGE* − 6.5).

Note: SAS Proc Mixed, Full ML. Also note that the covariance component, σ_{01}, is estimated, but not displayed.

rate is half a point larger in a model with *AGEGRP* − 6.5 (5.03 vs. 4.54). This cumulates to a two-point differential in PIAT scores over the four years under study. So, too, the two level-2 variance components are much larger for a model with *AGEGRP* − 6.5.

Why are these estimates larger when we treat the data set as time-structured, using *AGEGRP* − 6.5 as our level-1 predictor, than when we treat it as irregular, using *AGE* − 6.5? We obtain a larger fixed effect for linear growth because *AGEGRP* associates the data for waves 2 and 3 with earlier ages (8.5 and 10.5) than observed. If we amortize the same gain over a shorter time period, the slope must be steeper. We obtain larger estimated variance components because the model with the time-structured predictor fits less well—there is more unexplained variation in initial status and growth rates—than when we associate each child's data with his or her age at testing. In other words, treating this unstructured data set as though it is time-structured introduces error into the analysis—error that we can reduce by using the child's age at testing as the temporal predictor.

We conclude that the model with *AGEGRP* as the level-1 temporal

redictor fits less well than the model with *AGE*. With the former repre-
ntation, the slope is inappropriately larger—inaccurately implying more
apid gains—and there is more unexplained variation in initial status and
ates of change. The superiority of the model with AGE as the temporal
redictor is supported by its smaller AIC and BIC statistics. The bottom
ne: never "force" an unstructured data set to be structured. If you have
everal metrics for tracking time—and you often will—investigate the
ossibility of alternative temporal specifications. Your first choice, espe-
ially if tied to design, not substance, may not always be the best.

.2 Varying Numbers of Measurement Occasions

Once you allow the spacing of waves to vary across individuals, it is a small
eap to allow their *number* to vary as well. Statisticians say that such data sets
re *unbalanced*. As you would expect, balance facilitates analysis: models
an be parameterized more easily, random effects can be estimated more
precisely, and computer algorithms will converge more rapidly.

Yet a major advantage of the multilevel model for change is that it
s easily fit to unbalanced data. Unlike approaches such as repeated
measures analysis of variance, with the multilevel modeling of change it
s straightforward to analyze data sets with varying numbers of waves of
data. To illustrate the general approach, we begin, in section 5.2.1, by
introducing a new data set in which the number of waves per person
varies widely, from 1 to 13. We extend this discussion in section 5.2.2, by
discussing implementation and estimation problems that can arise when
data are unbalanced. We conclude, in section 5.2.3, by discussing poten-
tial causes of imbalance—especially missing data—and how they can
affect statistical analysis.

5.2.1 Analyzing Data Sets in Which the Number of Waves per Person Varies

Murnane, Boudett, and Willett (1999) used data from the National Longitudinal Survey of Youth (NLSY) to track the labor-market experiences of male high school dropouts. Like many large panel studies, the NLSY poses a variety of design complications: (1) at the first wave of data collection, the men varied in age from 14 to 17; (2) some subsequent waves were separated by one year, others by two; (3) each wave's interviews were conducted at different times during the calendar year; and (4) respondents could describe more than one job at each interview. Person-specific schooling and employment patterns posed further problems. Not only could respondents drop out of school at different times and enter the

Table 5.3: Excerpts from the person-period data set for the high school dropout wage study

ID	*EXPER*	*LNW*	*BLACK*	*HGC*	*UERATE*
206	1.874	2.028	0	10	9.200
206	2.814	2.297	0	10	11.000
206	4.314	2.482	0	10	6.295
332	0.125	1.630	0	8	7.100
332	1.625	1.476	0	8	9.600
332	2.413	1.804	0	8	7.200
332	3.393	1.439	0	8	6.195
332	4.470	1.748	0	8	5.595
332	5.178	1.526	0	8	4.595
332	6.082	2.044	0	8	4.295
332	7.043	2.179	0	8	3.395
332	8.197	2.186	0	8	4.395
332	9.092	4.035	0	8	6.695
1028	0.004	0.872	1	8	9.300
1028	0.035	0.903	1	8	7.400
1028	0.515	1.389	1	8	7.300
1028	1.483	2.324	1	8	7.400
1028	2.141	1.484	1	8	6.295
1028	3.161	1.705	1	8	5.895
1028	4.103	2.343	1	8	6.900

labor force at different times, they also changed jobs at different times. To track wages on a common temporal scale, Murnane and colleagues decided to clock time from each respondent's first day of work. This allows each hourly wage to be associated with a temporally appropriate point in the respondent's labor force history. The resulting data set has an unusual temporal schedule, varying not only in spacing but length.

Table 5.3 presents excerpts from the person-period data set. To adjust for inflation, each hourly wage is expressed in constant 1990 dollars. To address the skewness commonly found in wage data and to linearize the individual wage trajectories, we analyze the natural logarithm of wages, *LNW*. Then, to express this outcome on its original scale, we take antilogs (e.g., $e^{(2.028)} = \$7.60$ per hour).

The temporal variable *EXPER* identifies the specific moment—to the nearest day—in each man's labor force history associated with each observed value of *LNW*. Notice the variability in the number and spacing of waves. Dropout 206 has three waves, for jobs held at 1.874, 2.814 and 4.314 years of experience after labor force entry. Dropout 332 has 10 waves, the first for a job held immediately after entering the labor force, the others for jobs held approximately every subsequent year. Dropout

)28 has 7 waves; the first three describe the first six months of work (at 004, 0.035, and 0.515 years). Across the full sample, 77 men have 1 or waves of data, 82 have 3 or 4, 166 have 5 or 6, 226 have 7 or 8, 240 have or 10, and 97 have more than 10. The earliest wave describes someone's rst day of work; the latest describes a job held 13 years later.

This is the first data set we have presented in which the number of aves of data varies across individuals. Some men even have fewer than ıree waves—less than the minimum articulated in previous chapters. A ıajor advantage of the multilevel model for change is that everyone can articipate in the estimation, regardless of how many waves he con- ıibutes to the data set. Even the 38 men with just 1 wave of data and the 9 with just 2 waves are included in the estimation. Although they provide ess, or no, information about within-person variation—and hence do not ontribute to variance component estimation—they can still contribute o the estimation of fixed effects where appropriate. Ultimately, each ıerson's fitted trajectory is based on a combination of his: (1) observed rajectory, and (2) a model-based trajectory determined by the values of he predictors.

You need no special procedures to fit a multilevel model for change o unbalanced data. All you need do is specify the model appropriately o your statistical software. As long as the person-period data set includes enough people with enough waves of data for the numeric algorithms to converge, you will encounter no difficulties. If the data set is severely ınbalanced, or if too many people have too few waves for the complexity ɔf your hypothesized model, problems may arise in the estimation. For ıow, we continue with this data set, which includes so many people with ɜo many waves that estimation is straightforward. We discuss strategies for identifying and resolving estimation problems in section 5.2.2.

Table 5.4 presents the results of fitting three multilevel models for change to the wage data, using full ML in SAS PROC MIXED. First examine the results for Model A, the unconditional growth model. The positive and statistically significant fixed effect for *EXPER* indicates that inflation-adjusted wages rise over time. Because the outcome, *LNW*, is expressed on a logarithmic scale, its parameter estimate, $\hat{\gamma}_{10}$, is not a linear growth rate. As in regular regression, however, transformation facilitates interpretation. If an outcome in a linear relationship, Y, is expressed as a natural logarithm and $\hat{\gamma}_{10}$ is the regression coefficient for a predictor X, then $100(e^{(\hat{\gamma}_{10})} - 1)$ is the *percentage change* in Y per unit difference in X. Because *EXPER* is calibrated in years, this transformation yields an annual percentage growth rate in wages. Computing $100(e^{(0.0457)} - 1) = 4.7$, we estimate that the average high school dropout's inflation-adjusted hourly wages rise by 4.7% with each year of labor force participation.

Table 5.4: Results of fitting a taxonomy of multilevel models for change to the high school dropout wage data ($n = 888$)

		Parameter	Model A	Model B	Model C
Fixed Effects					
Initial status, π_{0i}	Intercept	γ_{00}	1.7156***	1.7171***	1.7215***
			(0.0108)	(0.0125)	(0.0107)
	(*HGC* − 9)	γ_{01}		0.0349***	0.0384***
				(0.0079)	(0.0064)
	BLACK	γ_{02}		0.0154	
				(0.0239)	
Rate of change, π_{1i}	Intercept	γ_{10}	0.0457***	0.0493***	0.0489***
			(0.0023)	(0.0026)	(0.0025)
	(*HGC* − 9)	γ_{11}		0.0013	
				(0.0017)	
	BLACK	γ_{12}		−0.0182**	−0.0161***
				(0.0055)	(0.0045)
Variance Components					
Level-1:	within-person	σ^2_ε	0.0951***	0.0952***	0.0952***
Level-2:	In initial status	σ^2_0	0.0543***	0.0518***	0.0518***
	In rate of change	σ^2_1	0.0017***	0.0016***	0.0016***
Goodness-of-fit					
	Deviance		4921.4	4873.8	4874.7
	AIC		4933.4	4893.8	4890.7
	BIC		4962.1	4941.7	4929.0

$-p < .10$; $*p < .05$; $**p < .01$; $***p < .001$.
Model A is an unconditional growth model; Model B includes the effects of highest grade completed (*HGC* − 9) and race (*BLACK*) on both initial status and rate of change; Model C is a reduced model in which (*HGC* − 9) predicts only initial status and *BLACK* predicts only rate of change.

Note: SAS Proc Mixed, Full ML. Also note that the covariance component, σ_{01}, is estimated, but not displayed.

After specifying a suitable individual growth model, you add level-2 predictors in the usual way. The statistically significant variance components in Model A, for both initial status and rate of change, suggest the wisdom of this action. Models B and C examine the effects of two predictors: (1) the race/ethnicity of the dropout; and (b) the highest grade he completed before dropping out. Although the sample includes 438 Whites, 246 African Americans, and 204 Latinos, analyses not shown here suggest that we cannot distinguish statistically between the trajectories of Latino and White dropouts. For this reason, these models include just one race/ethnicity predictor (*BLACK*). Highest grade completed, *HGC*, is a continuous variable that ranges from 6th through 12th grade, with an average of 8.8 and a standard deviation of 1.4. To facilitate

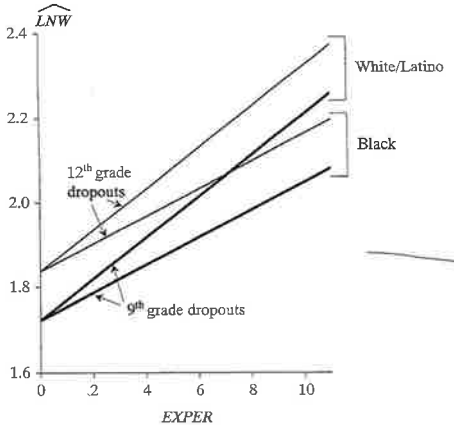


Figure 5.2. Displaying the results of a fitted multilevel model for change. Log wage trajectories from Model C of table 5.4 for four prototypical dropouts: Blacks and Whites/Latinos who dropped out in 9th and 12th grades.

interpretation, our analyses use a rescaled version, $HGC - 9$, which centers HGC around this substantively meaningful value near the sample mean (see section 4.5.4 for a discussion of centering).

Model B of Table 5.4 associates each predictor with initial status and rate of change. The estimated fixed effects suggest that $HGC - 9$ is related only to initial status while $BLACK$ is related only to the rate of change. We therefore fit Model C, whose level-2 submodels reflect this observation. The fixed effect for $HGC - 9$ on initial status tells us that dropouts who stay in school longer earn higher wages on labor force entry ($\hat{\gamma}_{01} = 0.0384$, $p < .001$), as we might expect because they are likely to have more skills than peers who left school earlier. The fixed effect for $BLACK$ on rate of change tells us that, in contrast to Whites and Latinos, the wages of Black males increase less rapidly with labor force experience ($\hat{\gamma}_{12} = -0.0161$, $p < .001$). The statistically significant level-2 variance components indicate the presence of additional unpredicted interindividual variation in both initial status and rate of change. In sections 5.3.3 and 6.1.2, we add other predictors that explain some of this remaining variation.

Figure 5.2 summarizes the effects in Model C by displaying wage trajectories for four prototypical dropouts: Blacks and Whites/Latinos who dropped out in 9th and 12th grades. We obtained these trajectories using the same two-stage process presented in section 4.5.3. We first substituted the two values of $BLACK$ (0 and 1) into Model C and then substituted in two prototypical values of $HGC - 9$ (0 and 3, to correspond to 9 and 12 years of education). The plots document the large and statistically significant effects of education and race on the wage trajectories. The longer a prospective dropout stays in school, the higher his wages on labor force entry. But race plays an important role, not on initial wages but on the rate of change. Although the average Black dropout initially earns an

hourly wage indistinguishable from the average White or Latino dropout, his annual percentage increase is lower. Controlling for highest grade completed, the average annual percentage increase is $100(e^{(0.0489)} - 1) = 5.0\%$ for Whites and Latinos in comparison to $100(e^{(0.0328)} - 1) = 3.3\%$ for Blacks. Over time, this race differential overwhelms the initial advantage of remaining in school. Beyond 7 years of labor force participation, a Black male who left school in 12th grade earns a lower hourly wage than a White or Latino male who left in 9th.

5.2.2 Practical Problems That May Arise When Analyzing Unbalanced Data Sets

We encountered no problems when fitting models to the unbalanced data in section 5.2.1. The most complex model (C) converged in just three iterations and we could estimate every parameter in the model. But if your data set is severely unbalanced, or if too few people have enough waves of data, computer iterative algorithms may not converge and you may be unable to estimate one or more variance components.

Why does imbalance affect the estimation of variance components but not fixed effects? No matter how unbalanced the person-period data set, the estimation of fixed effects is generally no more difficult than the estimation of regression coefficients in a regular linear model. To demonstrate why, let us begin with a multilevel model—for simplicity, an unconditional growth model—expressed in composite form:

$$Y_{ij} = [\gamma_{00} + \gamma_{10} TIME_{ij}] + [\zeta_{0i} + \zeta_{1i} TIME_{ij} + \varepsilon_{ij}]. \quad (5.2a)$$

If we re-express the composite error term in the second set of brackets as: $\varepsilon^*_{ij} = [\zeta_{0i} + \zeta_{1i} TIME_{ij} + \varepsilon_{ij}]$, we obtain an equivalent representation of equation 5.2a:

$$Y_{ij} = \gamma_{00} + \gamma_{10} TIME_{ij} + \varepsilon^*_{ij}. \quad (5.2b)$$

Equation 5.2b resembles a standard regression model, with γ's instead of β's and ε^*_{ij} instead of ε_{ij}. The difference is that we do not assume that the composite residuals ε^*_{ij} are independent and normally distributed with mean 0 and variance $\sigma^2_{\varepsilon^*}$. Instead we assume that their constituents—ζ_{0i}, ζ_{1i}, and ε_{ij}—follow the assumptions:

$$\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon) \text{ and } \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_0 & \sigma_{01} \\ \sigma_{10} & \sigma^2_1 \end{bmatrix} \right).$$

It is these complex assumptions—about the variance components—that complicate estimation.

Now consider the following thought experiment. Suppose we are willing

ake a simplifying assumption about the composite residuals, declaring
ı to be independent and normally distributed: $\varepsilon^*_{ij} \sim N(0,\sigma^2_{\varepsilon^*})$. This is
amount to assuming that both level-2 residuals, ζ_{0i} and ζ_{1i}, are always 0,
ould be their associated variance components (i.e., both σ^2_0 and σ^2_1 are
0). In the language of multilevel modeling, we would be *fixing* the
:ept and rate of change, making them constant across individuals.
ther each person contributed one wave or many, estimation of the two
l effects and the one variance component would then become a stan-
. regression problem. All you would need are a sufficient number of
nct values of $TIME_{ij}$ in the person-period data set—enough distinct
ts in a plot of Y_{ij} vs. $TIME_{ij}$—to identify the level-1 submodel's func-
al form. In a time-structured data set, this plot would be composed of
cal stripes, one for each measurement occasion. This is why you would
l at least three waves of data—the stripes would lie at just those three
sions. In an unstructured data set, the variable spacing of waves makes
sier to estimate fixed effects because the data points are more sepa-
l "horizontally." This allows you to relax the data minimum per
on—allowing some people to have fewer than three waves—as long as
ıave enough distinct values of $TIME_{ij}$ to estimate the fixed effects.
we are unwilling to make these simplifying assumptions—and we
:rally are—estimation of variance components can be difficult if too
y people have too few waves. Variability in the spacing of waves helps,
may not resolve the problem. Estimation of variance components
ires that enough people have sufficient data to allow quantification
ithin-person residual variation—variation in the residuals over and
e the fixed effects. If too many people have too little data, you will
nable to quantify this residual variability.
hen does the numeric task become so difficult that the variance com-
:nts cannot be estimated? We offer no rules because so many issues
nvolved, including the degree of imbalance, the complexity of the
el, the number of people with few vs. many waves, and the inclusion
ne-varying predictors (discussed in section 5.3). Suffice it to say that
ı imbalance is severe enough, numeric computer algorithms can
uce theoretically impossible values or fail to converge. Each statisti-
oftware program has its own way of informing the user of a problem;
discovered, we recommend that you be proactive and not auto-
:ally accept the default "solution" your program offers. Below, we
ıss each of the two major estimation problems.

Boundary Constraints

y population parameters have *boundary constraints*—limits beyond
h they cannot theoretically lie. Like variances and correlation

coefficients, the variance/covariance components in the multilevel model have clear boundaries: (1) a variance component cannot be negative; and (2) a covariance component, expressed in correlation form, must lie between −1 and +1. Because of the complexity of the estimation task—especially with unbalanced data—as well as the iterative nature of the computational algorithms, multilevel modeling programs occasionally generate parameter estimates that reach, or lie outside, these limits. When this happens, the program may output the implausible estimate or its boundary value (e.g., it might set a variance component to 0).

How will you know if you have encountered a boundary constraint? The warning signs differ across programs. If you use SAS PROC MIXED, the program log will note that "the G matrix [the variance-covariance matrix for the variance components] is not positive definite." By default, SAS sets the offending estimate to its boundary value. MLwiN does not provide a note; instead, it sets the offending estimate, and all associated estimates, to boundary values. If your output indicates that an estimate is exactly 0, you have likely encountered a boundary constraint. HLM will provide you with a warning message and modify its computational algorithm to avoid the problem. With all software, one clue that you may be approaching a "boundary" is if you find you need an excessive number of iterations to reach convergence.

We recommend that you never let a computer program arbitrarily make important decisions like these. Regardless of which program you use, you should be proactive about boundary constraints. Overspecification of the model's stochastic portion is the usual cause; model simplification is generally the cure. A practical solution is to compare alternative models that remove one, or more, offending random effects systematically until the model can be fit. This strategy, known as *fixing* a predictor's effect, usually resolves the problems.

We illustrate this approach using a small data set purposefully selected from the larger wage data set just analyzed. We constructed this sample for pedagogic purposes, hoping to create such extreme imbalance that boundary constraints would arise. This new data set is composed of the 124 men who had three or fewer waves of wage data: 47 men have three waves, 39 have two, and 38 have only one. The earliest value of *EXPER* is 0.002; the latest is 7.768. This data set is *not* a random sample of the original group.

Table 5.5 presents the results of fitting three models to this smaller data set; each is based upon Model C, the "final" model of table 5.4. As before, each was fit using ML in SAS PROC MIXED. In the first model, which is identical to Model C, the estimated variance component for linear growth, $\hat{\sigma}^2_1$, is exactly 0. This is a standard sign of a boundary

Table 5.5: Comparison of three alternative approaches to fitting Model C of table 5.4 to a severely unbalanced subset of the high school dropout wage data ($n = 124$)

		Parameter	A Default method	B Removing boundary constraints	C Fixing rates of change
Fixed Effects					
Initial status, π_{0i}	Intercept	γ_{00}	1.7373***	—	1.7373***
			(0.0476)		(0.0483)
	(*HGC* − 9)	γ_{01}	0.0462~	—	0.0458~
			(0.0245)		(0.0245)
Rate of change, π_{1i}	Intercept	γ_{10}	0.0516*	—	0.0518*
			(0.0211)		(0.0209)
	BLACK	γ_{12}	−0.0596~	—	−0.0601~
			(0.0348)		(0.0346)
Variance Components					
Level-1:	Within-person	σ^2_ε	0.1150***	0.1374***	0.1148***
Level-2:	In initial status	σ^2_0	0.0818**	0.0267	0.0842***
	In rate of change	σ^2_1	0.0000	−0.0072	—
Goodness-of-fit					
	Deviance		283.9	—	283.9
	AIC		297.9	—	295.9
	BIC		317.6	—	312.8

~ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$.
Model A uses the default option in SAS PROC MIXED; Model B removes boundary constraints for the variance components; Model C removes the level-2 residual for rate of change, eliminating the associated variance component (as well as the associated covariance component).

Note: SAS Proc Mixed, Full ML. Also note that the covariance component, σ_{01}, is estimated where appropriate, but not displayed.

problem, used by both SAS PROC MIXED and MLwiN. Estimates of 0 are always suspicious; here they indicate that the algorithm has encountered a boundary constraint. (Note that SAS allows the associated covariance component to be non-zero, whereas MLwiN would also set that term to 0.)

Model B in table 5.5 represents our dogged attempt to fit the specified model to data. To do so, we invoke a software option that relaxes the default boundary constraint permitting us to obtain a negative variance component. When analyzing severely unbalanced data, eliminating automatic fix-ups can help identify problems with boundary constraints. Unfortunately, in this case, the iterative algorithm does not converge (a different problem that we will soon discuss). Nevertheless, notice that the estimated variance component for rate of change at the last iteration is negative—a logical impossibility. This, too, is another sign suggesting the need for model simplification.

Model C in table 5.5 constrains the variance component for the linear growth rate, and its associated covariance component, to be 0. Notice that the deviance statistic for this model is identical to that of the first, suggesting the wisdom of fixing this parameter. This model fits no worse and involves fewer parameters (as reflected by the superior AIC and BIC statistics). This means that with this data set—which is *not* a random sample from the original—we cannot confirm the existence of any systematic residual variation in the slopes of the wage trajectories beyond the modest effect of *BLACK* shown in the final column of table 5.5.

Nonconvergence

As discussed in section 4.3, all multilevel modeling programs implement iterative numeric algorithms for model fitting. These algorithms compare fit criteria (such as the log-likelihood statistic) across successive iterations and declare convergence when the change in the fit criterion is sufficiently "small." Although the user can determine how small is "small enough," all programs have a default criterion, generally an arbitrarily small proportional change. When the criterion is met, the algorithm *converges* (i.e., stops iterating). If the criterion cannot be met in a large number of iterations, estimates should be treated with suspicion.

How many iterations are needed to achieve convergence? If your data set is highly structured and your model simple, convergence takes just a few iterations, well within the default values set by most programs. With unbalanced data sets and complex models, convergence can take hundreds or thousands of iterations although the algorithms in specialized packages (e.g., HLM and MLwiN) usually converge more rapidly than those in multipurpose programs (e.g., SAS PROC MIXED).

For every model you fit—but especially for models fit to unbalanced data—be sure to check that the algorithm has converged. In complex problems, the program's default limits on the maximum number of iterations may be too low to reach convergence. All packages allow you to increase this limit. If the algorithm still does not converge, sequentially increase the limit until it does. Some programs allow you to facilitate this search by providing "starting values" for the variance and covariance components.

No matter how many iterations you permit and no matter how much prior information you provide, there will be times when the algorithm will not converge. Nonconvergence can result from many factors, but two common causes are poorly specified models and insufficient data; their

combination can be deadly. If you need an extremely large number of iterations to fit a model to data, closely examine the variance components and determine whether you have sufficient information to warrant allowing level-2 residuals for both initial status and rates of change. (If you are fitting nonlinear models using the methods of chapter 6, scrutinize other variance components as well.) Remember that any given data set contains a finite amount of information. You can postulate a complex model, but it is not always possible to fit that model to the available data.

We conclude by noting that other problems besides boundary constraints can cause nonconvergence. One problem, easily remedied, is a variable's scale. If an outcome's values are too small, the variance components will be smaller still; this can cause nonconvergence via rounding error issues. Simple multiplication of the outcome by 100, 1000, or another factor of 10 can usually ameliorate this difficulty. Predictor scaling can also cause problems but usually you want to adjust its metric in the *opposite* direction. For a temporal predictor, for example, you might move from a briefer time unit to a longer one (from days to months or months to years) so as to increase the growth rate's magnitude. These kinds of transformations have only cosmetic effects on your essential findings. (They will change the value of the log likelihood and associated statistics, but leave the results of tests unaffected.)

5.2.3 Distinguishing among Different Types of Missingness

No discussion of imbalance is complete without a complementary discussion of its underlying source. Although some researchers build imbalance into their design, most imbalance is unplanned, owing to scheduling problems, missed appointments, attrition, and data processing errors. Further imbalance accrues if individuals who miss a wave of data collection subsequently return to the sample. For example, although the NLSY has a low annual attrition rate—less than 5% of the original sample initially leave in each of the first 13 years—many participants miss one or two waves. In their exhaustive study of NSLY attrition, MaCurdy, Mroz, and Gritz (1998) find many differences among persisters, dropouts, and returnees. Of relevance for the wage analyses just presented are the findings that attrition is higher for both the unemployed and men who once earned high wages.

Unplanned imbalance, especially when it stems from attrition or other potentially systematic sources, may invalidate your inferences. The issue is not the technical ability to fit a model but rather a substantive question about credible generalization. To probe the issues, statisticians frame the problem, not in terms of imbalance, but rather in terms of *missing data*. When you fit a multilevel model for change, you implicitly assume that each person's observed records are a random sample of data from his or her underlying true growth trajectory. If your design is sound and has no built-in bias, and everyone is assessed on every planned occasion, your observed data will meet this assumption. If one or more individuals are not assessed on one or more occasions, your observed data may not meet this assumption. In this case, your parameter estimates may be biased and your generalizations incorrect.

Notice that we use the word "may," not "will," throughout the previous paragraph. This is because missingness, in and of itself, is not necessarily problematic. It all depends upon what statisticians call the *type of missingness*. In seminal work on this topic, Little (1995), refining earlier work with Rubin (Little & Rubin, 1987), distinguishes among three types of missingness: (1) *missing completely at random* (MCAR); (2) covariate-dependent dropout (CDD); and (3) *missing at random* (MAR) (see also Schafer, 1997). As Laird (1988) demonstrates, we can validly generalize the results of fitting a multilevel model for change under all three of these missingness conditions, which she groups together under rubric *ignorable nonresponse*.

When we say that data are MCAR, we argue that the observed values are a random sample of all the values that could have been observed (according to plan), had there been no missing data. Because time-invariant predictors are usually measured when a study begins, their values are rarely missing. As a result, when a multilevel model includes no time-varying predictors, the only predictor that can be missing is *TIME* itself (when a planned measurement occasion is missed). This means that longitudinal data are MCAR if the probability of assessment on any occasion is independent of: (1) the particular time; (2) the values of the substantive predictors; and (3) the values of the outcome (which are, by definition, unobserved). For the NLSY wage data just analyzed, we can make a case for the MCAR assumption if the probability of providing wage data at any point in time is independent of the particular moment in that individual's labor force history, all other predictors, and the unobserved wage. There cannot be particular moments when a man would be unlikely to grant an interview, as would be the case if men were unwilling to do so on specific days (which seems unlikely). But missingness must also not vary systematically by an individual's wage or other potentially unobserved characteristics. MaCurdy and colleagues (1998) convincingly demonstrate that these latter two conditions are implausible for the NLSY.

The conclusion that the MCAR assumption is untenable for the NLSY

data is unsurprising as this assumption is especially restrictive—wonderful when met, but rarely so. Covariate dependent dropout (CDD) is a less restrictive assumption that permits associations between the probability of missingness and observed predictor values ("covariates"). Data can be CDD even if the probability of missingness is systematically related to either *TIME* or observed substantive predictors. For the NLSY wage data, we can argue for the validity of the CDD assumption even if there are particular moments when men are unlikely to grant interviews. Missingness can also vary by either race or highest grade completed (our two observed predictors). By including these observed predictors in the multilevel model, we deflect the possibility of bias, allowing appropriate generalization of empirical results.

The major difficulty in establishing the tenability of the MCAR and CDD assumptions is the requirement of demonstrating that the probability of missingness at any point in time is unrelated to the contemporaneous value of the associated outcome. Because this outcome is unobserved, you cannot provide empirical support as you lack the very data you need. Only a substantive argument and thought experiment will do. Any potential relationship between the unobserved outcome and the probability of missingness invalidates these assumptions. For example, if men with particularly high or low wages are less likely to participate in an NLSY interview, we cannot support either assumption. As this hypothesis is both tenable and likely, we cannot defend either assumption for the NLSY wage data (nor for many other longitudinal data sets).

Fortunately, there is an even less restrictive type of missingness—more common in longitudinal research—that still permits valid generalization of the multilevel model for change: the MAR assumption. When data are MAR, the probability of missingness can depend upon *any* observed data, for either the predictors or any outcome values. It cannot, however, depend upon any *un*observed value of either any predictor or the outcome. So if we are willing to argue that the probability of missingness in the NLSY depends only upon observed predictor values (that is, *BLACK* and *HGC*) *and* wage data, we can make a case for the MAR assumption. The allowance for dependence upon observed outcome data can account for a multitude of sins, often supporting the credibility of the MAR assumption even when MCAR and CDD assumptions seem far-fetched.

As general as it seems, you should not accept the MAR assumption without scrutiny. Greenland and Finkle (1995) examine this assumption in cross-sectional research and suggest that even it can be difficult to meet. To illustrate their point, they argue that someone's unwillingness to answer a question about sexual preference (i.e., heterosexual vs. homosexual) is likely correlated with his or her true sexual preference. We agree, but believe that there are many times when an individual's outcome values will adequately reflect such concerns. Yet even this assertion can be untrue. For example, a recovering alcoholic's willingness to continue participating in a study about abstinence is likely related to his or her ability to stay sober on each occasion. Such a systematic pattern—even if impossible to prove—invalidates the MAR assumption.

In practice, the burden of evaluating the tenability of these missingness assumptions rests with you. Any type of ignorable missingness permits valid inference; you just need to determine which seems most credible for your project. We suggest that you act as your own harshest critic—better you than the reviewers! As MAR is the least restrictive assumption, it provides the acid test. The key question is whether it is safe to assume that the probability of missingness is unrelated to unobserved concurrent outcomes (conditional on all observed outcomes). For the NLSY wage data, we can invent two plausible scenarios that undermine this assumption: If men are less likely to be interviewed at a particular wave if, at that time, they are earning especially: (1) *high* wages—because they might be less willing to take the time off from work to participate; or (2) *low* wages—because they might be less willing to reveal these low values to an interviewer. Because current wages (even unobserved) are strongly correlated with past and future wages, however, these risks are likely minimal. We therefore conclude that they are unlikely to be a major source of missingness for these data, supporting the credibility of the MAR assumption.[1]

If you cannot invoke one of these three missingness assumptions, you will need to add corrections to the multilevel model for change. Two different strategies are currently used: *selection models* and *pattern mixture models*. Under the selection approach, you build one statistical model for the "complete" data and a second model for the selection process that gave rise to the missingness. Under the pattern mixture approach, you identify a small number of missingness patterns and then fit a multilevel model stratified by these patterns. For further information, we direct your attention to the excellent papers by Hedeker and Gibbons (1997), Little (1995), and Little and Yau (1998).

5.3 Time-Varying Predictors

A time-varying predictor is a variable whose *values* may differ over time. Unlike their time-invariant cousins, which record an individual's static status, time-varying predictors record an individual's potentially differing

status on each associated measurement occasion. Some time-varying predictors have values that change naturally; others have values that change by design.

In their four-year study of how teen employment affects the amount of time adolescents spend with their families, Shanahan, Elder, Burchinal, and Conger (1996) examined the effects of three time-varying predictors: (1) the average number of hours worked per week; (2) the total amount of money earned per year; and (3) whether earnings were used for nonleisure activities (e.g., schoolbooks or savings). At age $12\frac{1}{2}$, the average adolescent spent 16.3 hours per week with his or her family; over time, this amount declined at an average annual rate of 1.2 hours per week. Teen employment had both positive and negative effects. Although teens who made more money experienced steeper declines than peers who made less, those who spent some earnings on nonleisure activities or who worked especially long hours spent *more* time, on average, with their families (although their rates of decline were no shallower). The authors conclude that: "adolescent work constitutes a potentially positive source of social development, although this depends on how its multiple dimensions—earnings, spending patterns, [and] hours . . .—fit with the adolescent's broader life course" (p. 2198).

In this section, we demonstrate how you can include time-varying predictors in the multilevel model for change. We begin, in section 5.3.1, by showing how to parameterize, interpret, and graphically display a model that includes a time-varying predictor's main effect. In section 5.3.2, we allow the *effect* of a time-varying predictor to vary over time. In section 5.3.3, we discuss how to recenter time-varying predictors so as to facilitate interpretation. We conclude, in section 5.3.4, with some words of caution. Having described the analytic opportunities that time-varying predictors afford, we raise complex conceptual issues that can compromise your ability to draw clear convincing conclusions.

5.3.1 Including the Main Effect of a Time-Varying Predictor

Conceptually, you need no special strategies to include the main effect of a time-varying predictor in a multilevel model for change. The key to understanding why this is so lies in the *structure* of the person-period data set. Because each predictor—whether time-invariant or time-varying—has its own value on *each* occasion, it matters little whether these values vary across each person's multiple records. A time-invariant predictor's values remain constant; a time-varying predictor's values vary. There is nothing more complex to it than that.

Table 5.6: Excerpts from the person-period data set for the unemployment study

ID	*MONTHS*	*CES-D*	*UNEMP*
7589	1.3142	36	1
7589	5.0924	40	1
7589	11.7947	39	1
55697	1.3471	7	1
55697	5.7823	4	1
65641	0.3285	32	1
65641	4.1068	9	0
65641	10.9405	10	0
65441	1.0842	27	1
65441	4.6982	15	1
65441	11.2690	7	0
53782	0.4271	22	1
53782	4.2382	15	0
53782	11.0719	21	1

We illustrate the general approach using data from Ginexi and colleagues' (2000) study of the effects of unemployment on depressive symptoms (mentioned briefly in section 5.1). By recruiting 254 participants from local unemployment offices, the researchers were able to interview individuals soon after job loss (within the first 2 months). Follow-up interviews were conducted between 3 and 8 months and 10 and 16 months after job loss. Each time, participants completed the Center for Epidemiologic Studies' Depression (CES-D) scale (Radloff, 1977), which asks them to rate, on a four-point scale, the frequency with which they experience each of 20 depressive symptoms. CES-D scores can vary from a low of 0 for someone with no symptoms to a high of 80 for someone in serious distress.

Just over half the sample ($n = 132$) was unemployed at every interview. Others had a variety of re-employment patterns: 62 were always working after the first interview; 41 were still unemployed at the second interview but working by the third; 19 were working by the second interview but unemployed at the third. We investigate the effect of unemployment using the time-varying predictor, *UNEMP*. As shown in the person-period data set in table 5.6, *UNEMP* represents individual i's unemployment status at each measurement occasion. Because subjects 7589 and 55697 were consistently unemployed, their values of *UNEMP* are consistently 1. Because the unemployment status of the remaining cases *changed*, their values of *UNEMP* change as well: subject 65641 was working at both the second and third interviews (pattern 1-0-0); subject 65441 was working by the third (pattern 1-1-0); and subject 53782 was working at the second

interview but unemployed again by the third (pattern 1-0-1). For any individual, *UNEMP* can be either 0 or 1 at each measurement occasion except the first (because, by design, everyone was initially unemployed).

We begin, as usual, with an unconditional growth model without substantive predictors:

$$\begin{aligned} Y_{ij} &= \pi_{0i} + \pi_{1i} TIME_{ij} + \varepsilon_{ij} \\ \pi_{0i} &= \gamma_{00} + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \zeta_{1i}, \end{aligned} \tag{5.3a}$$

where

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \text{ and } \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right). \tag{5.3b}$$

Model A of table 5.7 presents the results of fitting this model to data, where $TIME_{ij}$ indicates the number of months (to the nearest day) between the date of interview j for person i and his date of initial unemployment. On the first day of job loss ($TIME_{ij} = 0$), we estimate that the average person has a non-zero CES-D score of 17.67 ($p < .01$); over time, this level declines linearly at a rate of 0.42 per month ($p < .001$). The variance components for both initial status and rates of change are statistically significant, suggesting the wisdom of exploring the effects of person-specific predictors.

Using a Composite Specification

Because many respondents eventually find work, the unconditional growth model likely tells an incomplete story. If employment alleviates depressive symptoms, might the reemployment of half the sample explain the observed decline? If you exclusively use level-1/level-2 representations, you may have difficulty postulating a model that addresses this question. In particular, it may not be clear where—in which model—the time-varying predictor should appear. So far, person-specific variables have appeared in level-2 submodels as predictors of level-1 growth parameters. Although you might therefore conclude that substantive predictors must always appear at level-2, this conclusion would be incorrect!

The easiest way of understanding how to include a time-varying predictor is to use the composite specification of the multilevel model. It is not that we cannot include a time-varying predictor in a model written using a level-1/level-2 specification (we will soon show how to do so), but rather that it is easier to learn how these predictors' effects operate and what types of models you might fit, if you start here.

We begin with the composite specification for the unconditional growth model, formed by substituting the second and third equations in equation 5.3a into the first:

$$Y_{ij} = [\gamma_{00} + \gamma_{10} TIME_{ij}] + [\zeta_{0i} + \zeta_{1i} TIME_{ij} + \varepsilon_{ij}]. \tag{5.4}$$

As in chapter 4, we use brackets to distinguish the model's fixed and stochastic portions. Because the fixed portion in the first bracket resembles a standard regression model, we can add the main effect of the time-varying predictor, *UNEMP*, by writing:

Table 5.7: Results of fitting a taxonomy of multilevel models for change to the unemployment data ($n = 254$)

		Parameter	Model A	Model B	Model C	Model D
Fixed Effects						
Composite model	Intercept (initial status)	γ_{00}	17.6694** (0.7756)	12.6656*** (1.2421)	9.6167*** (1.8893)	11.2666*** (0.7690)
	TIME (rate of change)	γ_{10}	−0.4220*** (0.0830)	−0.2020* (0.0933)	0.1620 (0.1937)	
	UNEMP	γ_{20}		5.1113*** (0.9888)	8.5291*** (1.8779)	6.8795*** (0.9133)
	UNEMP by *TIME*	γ_{30}			−0.4652* (0.2172)	−0.3254** (0.1105)
Variance Components						
Level-1:	Within-person	σ_ε^2	68.85***	62.39***	62.03***	62.43***
Level-2:	In intercept	σ_0^2	86.85***	93.52***	93.71***	41.52***
	In rate of change	σ_1^2	0.36*	0.46**	0.45**	—
	In *UNEMP*	σ_2^2	—	—	—	40.45*
	In *UNEMP* by *TIME*	σ_3^2	—	—	—	0.71**
Goodness-of-fit						
	Deviance		5133.1	5107.6	5103.0	5093.6
	AIC		5145.1	5121.6	5119.7	5113.6
	BIC		5166.3	5146.4	5147.3	5148.9

Not sure how to compute these in R

$\sim p < .10$; $* p < .05$; $** p < .01$; $*** p < .001$.

These models predict depression scores (on the *CES-D*) in the months following unemployment as a function of the time-varying predictor *UNEMP*. Model A is an unconditional growth model (see equation 5.4). Model B adds the main effect of *UNEMP* as a fixed effect (see equation 5.5); Model C also adds the interaction between *UNEMP* and linear *TIME* (see equation 5.7). Model D allows *UNEMP* to have both fixed and random effects (see equation 5.10). Notice that we have changed the order in which the fixed effects appear to correspond to the composite specification of the model.

Note: Full ML, SAS Proc Mixed. Also note the models include all associated covariance parameters, which we do not display to conserve space.

$$Y_{ij} = [\gamma_{00} + \gamma_{10} TIME_{ij} + \gamma_{20} UNEMP_{ij}] + [\zeta_{0i} + \zeta_{1i} TIME_{ij} + \varepsilon_{ij}]. \quad (5.5)$$

The two subscripts on *UNEMP* signify its time-varying nature. In writing equation 5.5, we assume that individual i's value of Y at time j depends upon: (1) the number of months of since job loss (*TIME*); (2) his or her contemporaneous value of *UNEMP*; and (3) three person-specific residuals, ζ_{0i}, ζ_{1i}, and ε_{ij}.

What does this model imply about the time-varying predictor's main effect? Because the fixed effects, the γ's, are essentially regression parameters, we can interpret them using standard conventions:

- γ_{10} is the population average monthly rate of change in CES-D scores, controlling for unemployment status.
- γ_{20} is the population average difference, over time, in CES-D scores between the unemployed and employed.

The intercept, γ_{00}, refers to a logical impossibility: someone who is employed (*UNEMP* = 0) on the first day of job loss (*TIME* = 0). As in regular regression, an intercept can fall outside the range of the data (or theoretical possibility) without undermining the validity of the remaining parameters.

We can delve further into the model's assumptions by examining figure 5.3, which presents four average population trajectories implied by the model. As in figure 3.4, we obtained these trajectories by substituting in specific values for the substantive predictor(s). But because *UNEMP* is time-varying, we substitute in *time-varying patterns* not constant values. Since everyone was initially unemployed, *UNEMP* can take on one of four distinct patterns: (1) 1 1 1, for someone consistently unemployed; (2) 1 0 0, for someone who soon finds a job and remains employed; (3) 1 1 0, for someone who remains unemployed for a while but eventually finds a job; and (4) 1 0 1, for someone who soon finds a job only to lose it. Each pattern yields a different population trajectory, as shown in figure 5.3.

The unbroken trajectory in the upper left panel represents the predicted change in depressive symptoms for people who remain unemployed during the study. Because their values of *UNEMP* do not change, their implied average trajectory is linear. In displaying this single line, we do not mean to suggest that everyone who is consistently unemployed follows this line. The person-specific residuals, ζ_{0i} and ζ_{1i}, allow different individuals to have unique intercepts and slopes. But every true trajectory for someone who is consistently unemployed is linear, regardless of its level or slope.

The remaining trajectories in figure 5.3 reflect different patterns of temporal variation in *UNEMP*. Unlike the population trajectories in pre-

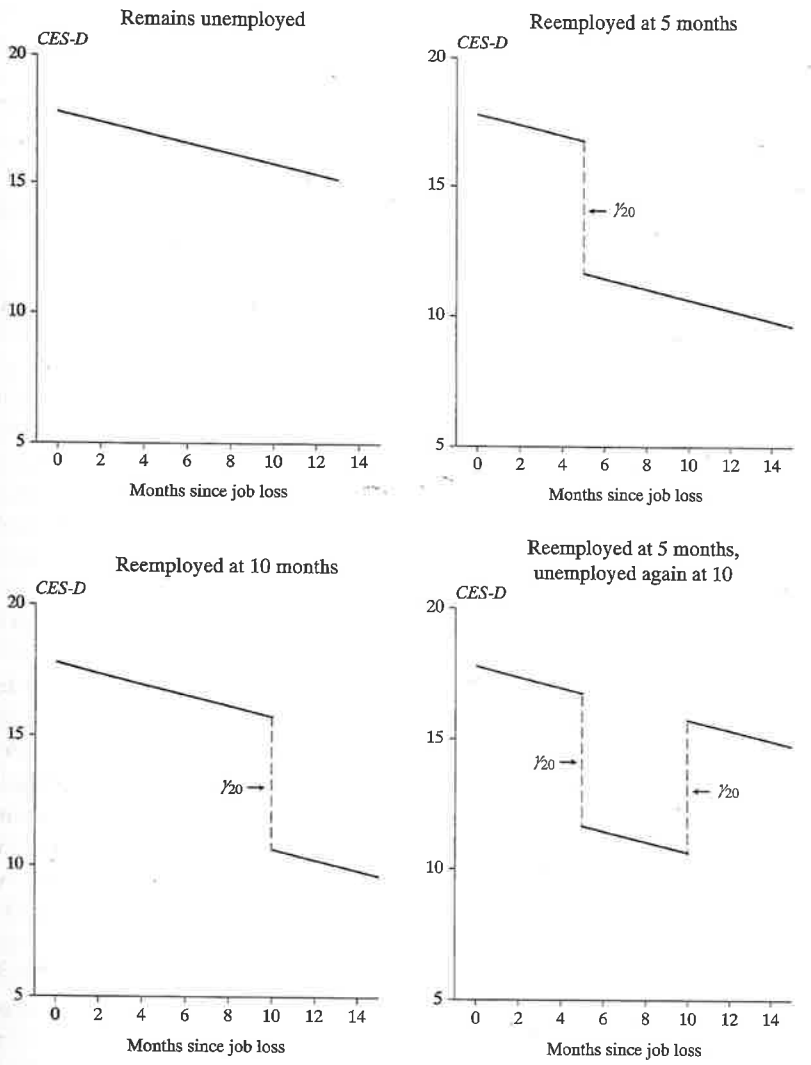


Figure 5.3. Identifying a suitable level-1 model for a time-varying predictor. Four averag population trajectories implied by equation 5.5 for the effects of time-varying unen ployment (*UNEMP*) on CES-D scores. In each panel, the magnitude of the effect of unen ployment remains constant (at γ_{20}), but because *UNEMP* is time-varying, the model implie different population average trajectories corresponding to alternative patterns of unen ployment and reemployment.

vious chapters, these are *discontinuous*. Discontinuity is a direct consequence of *UNEMP*'s dichotomous time-varying nature. The upper right panel, for the 1 0 0 pattern, presents a hypothesized population trajectory for someone who finds a job at 5 months and remain employed. The lower left panel, for the 1 1 0 pattern, presents a hypothesized trajectory for someone who finds a job at 10 months and remains employed. The lower right panel, for the 1 0 1 pattern, presents a hypothesized trajectory for someone who finds a job at 5 months only to lose it at 10.

In offering these hypothetical trajectories, we must mention two caveats. First, although we link the upper and lower segments in each panel using dashed lines, our model implies only the solid portions. We use the dashed lines to emphasize that a change in unemployment status is associated with a switch in trajectory. Second, these few trajectories are not the only ones implied by the model. As in the first panel, person-specific residuals—ζ_{0i} and ζ_{1i}—suggest the existence of many other discontinuous trajectories, each with its own intercept and slope. But because the model constrains the effect of *UNEMP* to be constant, the *gap* between trajectories—for any individual—will be identical, at γ_{20}, the parameter associated with *UNEMP*. (We relax this assumption in section 5.3.2.)

Model B of table 5.7 presents the results of fitting this model to data. The parameter estimate for TIME, $\hat{\gamma}_{10}$, suggests that the monthly rate of decline in CES-D, while still statistically significant, has been cut in half (to 0.20 from 0.42 in Model A). This suggests that reemployment explains some of the observed decline in CES-D scores. This conclusion is reinforced by: (1) the large statistically significant effect of *UNEMP*—the average CES-D score is 5.11 points higher ($p < .001$) among the unemployed; and (2) the poorer fit of Model A in comparison to Model B—the difference in deviance statistics is 25.5 on the addition of one parameter ($p < .001$) and the AIC and BIC statistics are much lower as well. (We discuss the variance components later in this section.)

The left panel of figure 5.4 displays prototypical trajectories for Model B. Rather than present many different discontinuous trajectories reflecting the wide variety of transition times for *UNEMP*, we present just two continuous trajectories: the upper one for someone consistently unemployed; the lower one for someone consistently employed after 3.5 months. Displaying only two trajectories reduces clutter and highlights the most extreme contrasts possible. Because of this study's design, we start the fitted trajectory for *UNEMP* = 0 at 3.5 months, the earliest time when a participant could be interviewed while working. To illustrate what would happen were we to extrapolate this trajectory back to *TIME* = 0, we include the dashed line. Because the model includes only the main effect of *UNEMP*, the two fitted trajectories are constrained to be parallel.

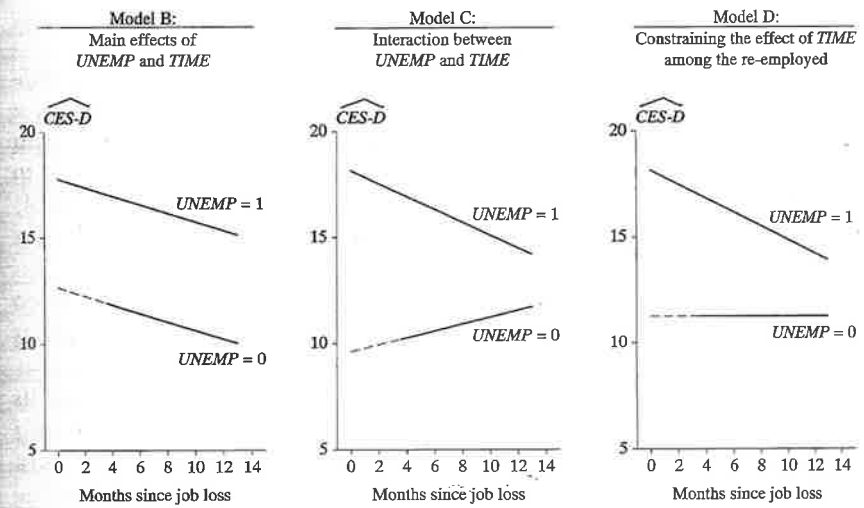


Figure 5.4. Displaying the results of fitted multilevel models for change that include a time-varying predictor. Prototypical trajectories from three models presented in table 5.7: Model B—the main effect of *UNEMP* and *TIME*, Model C—the interaction between *UNEMP* and *TIME*, and Model D—which constrains the effect of *TIME* to be 0 among the reemployed.

How do these two fitted trajectories display the main effect of unemployment status in Model B? Had the study followed just two static groups—the consistently unemployed and the consistently employed—these two trajectories would be the only ones implied by the model. But because *UNEMP* is time-varying, Model B implies the existence of many more depression trajectories, one for each possible *pattern* of unemployment/employment. Where are these additional trajectories? We find it helpful to think of the extremes shown as a conceptual *envelope* encompassing all discontinuous trajectories implied by the model. If *UNEMP* remains constant, an individual stays on one depression trajectory; if *UNEMP* changes, an individual shifts trajectories. As everyone in this study is unemployed at the first interview, everyone begins on the top trajectory. Those who find new jobs drop to the lower trajectory. Those who remain employed stay there. Those who lose their new jobs return to the upper trajectory. Conceptually, envision many dashed vertical lines running from the upper trajectory to the bottom (and back again) for individuals who change employment status. The set of these trajectories, which fall within the envelope shown, represent the complete set of prototypes implied by the model.

Using a Level-1/Level-2 Specification

Having included a time-varying predictor under the composite specification, we now show how you can specify the identical model using a level-1/level-2 specification. This representation provides further insight into how time-varying predictors' effects operate; it also allows you to include time-varying predictors using software packages (e.g., HLM) that require a level-1/level-2 specification of the multilevel model for change.

To derive the level-1/level-2 specification that corresponds to a given composite specification, you proceed backwards. In other words, just as we can substitute level-2 submodels into a level-1 submodel to form a composite specification, so, too, can we *decompose* a composite model into its constituent level-1 and level-2 parts. Because the time-specific subscript j can appear only in a level-1 model, all time-varying predictors must appear in at level-1. We therefore write the level-1 submodel for the composite main effects model in equation 5.5 as:

$$Y_{ij} = \pi_{0i} + \pi_{1i}TIME_{ij} + \pi_{2i}UNEMP_{ij} + \varepsilon_{ij}. \tag{5.6a}$$

Person-specific predictors that vary over time appear at level-1, not level-2. If you have no time-invariant predictors, as here, the accompanying level-2 models are brief:

$$\begin{aligned} \pi_{0i} &= \gamma_{00} + \zeta_{0i} \\ \pi_{1i} &= \gamma_{10} + \zeta_{1i} \\ \pi_{2i} &= \gamma_{20}. \end{aligned} \tag{5.6b}$$

You can verify that substituting these level-2 models into the level-1 model in equation 5.6a yields the composite specification in equation 5.5. To add the effects of time-invariant predictors, you include them, as usual, in the level-2 submodels.

Notice that the third equation in equation 5.6b, for π_{2i}, the parameter for *UNEMP*, includes no level-2 residual. All the multilevel models fit so far have invoked a similar constraint—that the effect of a person-specific predictor is constant across population members. Time-invariant predictors require this assumption because they have no *within*-person variation to allow for a level-2 residual. But for time-varying predictors we could easily modify the last model in equation 5.6b to be:

$$\pi_{2i} = \gamma_{20} + \zeta_{2i}. \tag{5.6c}$$

This allows the effect of *UNEMP* to vary randomly across individuals in the population. Adding this residual relaxes the assumption that the gap between postulated trajectories in figure 5.3 is constant. To fit the new model to data, we revise the distributional assumptions for the residuals as presented in equation 5.3b. Commonly, we expand the assumption of multivariate normality to include all three level-2 residuals:

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \text{ and } \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \\ \zeta_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 \end{bmatrix} \right). \tag{5.6d}$$

Notice that in adding one extra residual, ζ_{2i}, we add three extra variance components: σ_2^2, σ_{20} and σ_{21}.

Just because we *can* add these terms to our model does not mean that we should. Before doing so, we must decide whether the additional parameters are: (1) necessary; and (2) estimable using the available data. To address the first issue, consider whether the effect of employment on CES-D scores, controlling for time, *should* vary randomly across individuals. Before answering yes, remember that we are talking about *random* variation. If we expect the effect of unemployment to vary *systematically* across people, we can add substantive predictors that reflect this hypothesis. The question here is whether we should go further and add a residual that allows the effect of *UNEMP* to vary randomly. To be sure, much of our caution stems from concerns about the second point—the ability to estimate the additional parameters. With three (and sometimes fewer) measurement occasions per person, we often lack sufficient data to estimate additional variance components. Indeed, if we attempt to fit this more elaborate model, we encounter boundary constraints (as described in section 5.2.2). We therefore suggest that you resist the temptation to automatically allow the effects of time-varying predictors to vary at level-2 unless you have good reason, and sufficient data, to do so. (We will soon do so in section 5.3.2.)

As your models become more complex, we offer some practical advice (born of the consequences of the failure to follow it). When including time-varying predictors, we suggest that you write out the entire model before specifying your choice to a computer package. We suggest this extra step because it is not always obvious which random effects to include. In equation 5.6b, for example, the level-2 submodels require the first two parameters to be random and the third to be fixed. In other words, to fit this model you must use what appears to be an *inconsistent* set of level-2 submodels. As in many aspects of longitudinal analysis, the default or "standard" specifications may not yield the model *you* want to fit.

Time-Varying Predictors and Variance Components

In section 4.5.2, we discussed how the magnitude of variance components generally change on the inclusion of time-invariant predictors: (1) the level-1 variance component, σ^2_ε, remains relatively stable because time-invariant predictors cannot explain much within-person variation; but (2) the level-2 variance components, σ^2_0 and σ^2_1, will decline if the time-invariant predictors "explain" some of the between-person variation in initial status or rates of change, respectively. Time-varying predictors, in contrast, can affect all three variance components because they vary both within- *and* between-persons. And although you can interpret a decrease in the magnitude of the level-1 variance component, changes in level-2 variance components may not be meaningful, as we now show.

The general principles can be illustrated simply using Models A and B in table 5.7. Adding *UNEMP* to the unconditional growth model (Model A) reduces the magnitude of the within-person variance component, σ^2_ε, by 9.4% (from 68.85 to 62.39). Using strategies from section 4.4.3, equation 4.13, we conclude that time-varying unemployment status explains just over 9% of the variation in CES-D scores. This interpretation is straightforward because the time-varying predictor is added to the level-1 model, reducing the magnitude of the level-1 residual, ε_{ij}.

But ascribing meaning to observed changes in the level-2 variance components σ^2_0 and σ^2_1 can be nearly impossible. As we move from Model A to B both estimates *increase!* Although we alluded to this possibility in section 4.4.3, this is first example in which we observe such a pattern. The explanation for this seeming paradox—that changes in level-2 variance components do not assess the effects of time-varying predictors—lies in the associated level-1 submodel. When you add a time-varying predictor, as either a main effect or an interaction, you *change* the meaning of the individual growth parameters because:

- The intercept parameter, π_{0i}, now refers to the value of the outcome when *all* level-1 predictors, not only *TIME* but also the time-varying predictor, are zero.
- The slope parameter, π_{1i}, is now a *conditional* rate of change, controlling for the effects of the time-varying predictor.

Altering the population quantity that each parameter represents alters the meaning of the associated level-2 variance component. Hence, it makes no sense to compare the magnitude of these variance components across successive models.

This means that you must rely on changes in the time-varying predictors fixed effects, and associated goodness-of-fit statistics, when deciding whether to retain a time-varying predictor in your model. As tempting as it is to compute the percentage reduction in a variance component associated with the inclusion of a time-varying predictor, there is no consistently meaningful way of doing so.

5.3.2 Allowing the Effect of a Time-Varying Predictor to Vary over Time

Might unemployment status also affect the trajectory's slope? In previous chapters, we initially associated predictors with both initial status *and* rates of change. Yet because Model B includes only the *main* effects of *TIME* and *UNEMP*, the trajectories are constrained to be parallel.

There are many ways to specify a model in which the trajectories' slopes vary by unemployment status. The easiest approach, and the one we suggest you begin with, is to add the cross-product—here, between *UNEMP* and *TIME*—to the main effects model:

$$Y_{ij} = [\gamma_{00} + \gamma_{10}TIME_{ij} + \gamma_{20}UNEMP_{ij} + \gamma_{30}UNEMP_{ij} \times TIME_{ij}] + [\zeta_{0i} + \zeta_{1i}TIME_{ij} + \varepsilon_{ij}]. \quad (5.7)$$

Notice the close resemblance between this and the composite model that includes an interaction between a time-*invariant* predictor and *TIME* (shown in equation 4.3). The differences between the two are purely cosmetic: (1) the substantive predictor (here *UNEMP* and there *COA*) has an additional subscript j to indicate that it is time-varying; and (2) different subscripts reference the relevant fixed effects (the γ's).

Model C of table 5.7 presents the results of fitting this model to data. The interaction between *TIME* and *UNEMP* is statistically significant ($\hat{\gamma}_{30} = -0.46$, $p < .05$). As with all interactions, we can interpret this effect in two ways: (1) the effect of unemployment status on CES-D scores varies over time; and (2) the rate of change in CES-D scores over time differs by unemployment status. Rather than delve into these interpretations, we draw your attention to the prototypical trajectories for this model displayed in the middle panel of figure 5.4. Here we find an unexpected pattern: while CES-D scores decline among the unemployed, the *reverse* is found among the re-employed—their CES-D scores appear to increase! The parameter estimate for the main effect of *TIME*, $\hat{\gamma}_{10} = 0.16$, suggests why we observe this anomaly—it is not statistically significant (it is even smaller than its standard error, 0.19). Although we estimate a non-zero rate of change among the re-employed, we might have obtained this estimate even if the true rate of change in the population was zero.

This suggests that it might be wise to constrain the trajectory among the re-employed to be flat, with a slope of 0, while allowing the trajectory

mong the unemployed to decline over time. Were we fitting a standard egression model, we might achieve this goal by removing the main effect f *TIME*:

$$Y_{ij} = [\gamma_{00} + \gamma_{20}UNEMP_{ij} + \gamma_{30}UNEMP_{ij} \times TIME_{ij}] + [\zeta_{0i} + \zeta_{1i}TIME_{ij} + \varepsilon_{ij}]. \quad (5.8)$$

Had we fit this model to data and obtained fitted trajectories by unemployment status we would find: when $UNEMP = 0$, $\hat{Y}_{ij} = \hat{\gamma}_{00}$, when $UNEMP = 1$, $\hat{Y}_{ij} = (\hat{\gamma}_{00} + \hat{\gamma}_{20}) + \hat{\gamma}_{30}TIME_{ij}$.

This model's structural portion yields trajectories with the desired properties: (1) for the employed, we would have a flat line at level $\hat{\gamma}_{00}$; and (2) for the unemployed, we would have a slanted line, with intercept $\hat{\gamma}_{00} + \hat{\gamma}_{20}$ and slope $\hat{\gamma}_{30}$.

We do not fit this model, however, because of the lack of congruence between its structural and stochastic portions. Comparing the elements in the two sets of brackets in equation 5.8, notice that the model includes: (1) a random effect for *TIME*, ζ_{1i}, but no corresponding main effect (we removed γ_{10} from the model when we removed the main effect of *TIME*); and (2) a fixed effect for the *UNEMP* by *TIME* interaction (γ_{30}) and no corresponding random effect. We therefore postulate an alternative model in which the fixed and random effects are better aligned:

$$Y_{ij} = [\gamma_{00} + \gamma_{20}UNEMP_{ij} + \gamma_{30}UNEMP_{ij} \times TIME_{ij}] + [\zeta_{0i} + \zeta_{3i}UNEMP_{ij} \times TIME_{ij} + \varepsilon_{ij}]. \quad (5.9)$$

Notice that the interaction term, *UNEMP* by *TIME*, appears as both a fixed and a random effect. But when we attempt to fit the model in equation 5.9 to data, we find that its AIC and BIC statistics are larger (worse) than that of Model C (we cannot conduct a formal test because this model is not fully nested within the other, nor do we present the results in table 5.7).

It might appear, then, that Model C is preferable. But before reaching this conclusion, we revisit a question raised in the previous section: Should the effect of *UNEMP* be constant across the population? When we previously attempted to allow this effect to vary randomly (by augmenting Model B, which included the main effect of *TIME*) we could not fit the model to data. But having constrained the model's structural portion so that the trajectory among the re-employed is flat, we notice an inconsistency in equation 5.8: it allows the intercept among the employed, γ_{00}, to vary randomly (through the inclusion of the residual, ζ_{0i}) but not the increment to this intercept associated with unemployment, γ_{20} (there is no corresponding residual, ζ_{2i}). Why should we allow the flat level of the trajectory among the re-employed to vary and *not* allow the increment to this flat level (which yields the the intercept among the unemployed) to vary randomly as well? Perhaps the fit of the model in equation 5.9 is poorer than Model C because of this unrealistically stringent constraint on the random effects.

We address this supposition by fitting Model D:

$$Y_{ij} = [\gamma_{00} + \gamma_{20}UNEMP_{ij} + \gamma_{30}UNEMP_{ij} \times TIME_{ij}] + [\zeta_{0i} + \zeta_{2i}UNEMP_{ij} + \zeta_{3i}UNEMP_{ij} \times TIME_{ij} + \varepsilon_{ij}], \quad (5.10)$$

which allows each fixed effect to have an associated random effect. The results of fitting this model are shown in the final column of table 5.7 and are graphed in the right panel of figure 5.4. Immediately upon layoff, the average unemployed person in the population has a CES-D score of 18.15 (=11.27 + 6.88). Over time, as they acclimate to their new status, the average unemployed person's CES-D scores decline at a rate of −0.33 per month ($p < .01$). CES-D scores among those who find a job are lower (by as much as 6.88 if the job is found immediately after layoff or as little as 2.97 if 12 months later (14.24 − 11.27). Once a formerly unemployed individual finds a job and keeps it, we find no evidence of systematic change in CES-D scores over time. We believe that this model provides a more realistic representation of the patterns of change in CES-D scores over time than Model C. Not only is it substantively compelling, its AIC statistic is superior (and its BIC nearly equivalent) even though it includes several additional parameters (the extra variance components shown in table 5.7 as well as the extra covariance components not shown).

We hope that this example illustrates how you can test important hypotheses about time-varying predictors' effects and investigate even more ways in which outcomes might change over time (here, how CES-D scores change not just with time but also re-employment). As we will show in chapter 6, the ability to include time-varying predictors opens up a world of analytic opportunities. Not only can level-1 individual growth models be smooth and linear, they can also be discontinuous and curvilinear. This allows us to postulate and fit level-1 submodels that better reflect our hypotheses about the population processes that give rise to sample data and assess the tenability of such hypotheses with data. But to adequately build a foundation for pursuing those types of analyses, we must consider other issues that arise when working with time-varying predictors, and we do so by beginning with issues of centering.

5.3.3 Recentering Time-Varying Predictors

In chapter 4, when discussing interpretation of parameters associated with time-invariant predictors, we introduced the practice of recentering:

subtracting a constant from a predictor's values to alter its parameter's meaning. In some analyses, we subtracted a predictor's overall sample mean (known as *grand-mean centering*); in others, we subtracted a substantively interesting value (such as 9 for *highest grade completed*). We now describe similar strategies you can use with time-varying predictors.

To concretize the discussion, let us return to the wage data for high school dropouts summarized in table 5.4. We can express Model C in composite form by writing: $Y_{ij} = [\gamma_{00} + \gamma_{10}TIME_{ij} + \gamma_{01}(HGC_i - 9) + \gamma_{12}BLACK_i \times TIME_{ij}] + [\zeta_{0i} + \zeta_{1i}TIME_{ij} + \varepsilon_{ij}]$. As did the original researchers, we now introduce the possibility that wages might be affected by a time-varying predictor, *UERATE*, the unemployment rate in the local geographic area:

$$Y_{ij} = [\gamma_{00} + \gamma_{10}TIME_{ij} + \gamma_{01}(HGC_i - 9) + \gamma_{12}BLACK_i \times TIME_{ij} + \gamma_{20}UERATE_{ij}] + [\zeta_{0i} + \zeta_{1i}TIME_{ij} + \varepsilon_{ij}]. \quad (5.11)$$

We restrict attention to the main effect of *UERATE* because extensive analysis suggests that its effect on log wages does not vary over time.

Adapting recentering strategies outlined in section 4.5.4 for time-invariant predictors, we could include *UERATE* in several different ways, each using one of the following:

- Its raw values
- Deviations around its grand mean *in the person-period data set* (7.73)
- Deviations from another meaningful constant (say, 6, 7 or 8, common unemployment rates during the time period under study)

Each strategy would lead to virtually identical conclusions. Were we to fit the model in equation 5.11 using each, we would find identical parameter estimates, standard errors, and goodness-of-fit statistics *with just one exception*: for the intercept, γ_{00}. Inspecting equation 5.11 clarifies why this is so. As in regression, adding a main effect does not alter the meaning of the model's remaining parameters. If *UERATE* is expressed on its raw scale, γ_{00} estimates the average log wage on the first day of work (*EXPER* = 0) for a black male who dropped out in ninth grade ($HGC - 9 = 0$) and who lives in an area with *no* unemployment (*UERATE* = 0). If *UERATE* is grand-mean centered, γ_{00} estimates the average log-wage for a comparable male who lives in an area with an "*average*" unemployment rate. But because this "average" would be computed in the person-period data set, in which both the measurement occasions and number of waves vary across people, it may not be particularly meaningful.

Table 5.8: Results of adding three alternative representations of the time-varying predictor for local area unemployment rate (*UERATE*) to Model C of table 5.4 for the high school dropout wage data ($n = 888$)

		Parameter	Model A: centered at 7	Model B: within-person centering	Model C: time-1 centered
Fixed Effects					
Initial status, π_{0i}	Intercept	γ_{00}	1.7490*** (0.0114)	1.8743*** (0.0295)	1.8693*** (0.0260)
	$(HGC - 9)$	γ_{01}	0.0400*** (0.0064)	0.0402*** (0.0064)	0.0399*** (0.0064)
	UERATE	γ_{20}	−0.0120*** (0.0018)	−0.0177*** (0.0035)	−0.0162*** (0.0027)
	Deviation of *UERATE* from centering value	γ_{30}		−0.0099*** (0.0021)	−0.0103*** (0.0019)
Rate of change, π_{1i}	Intercept	γ_{10}	0.0441*** (0.0026)	0.0451* (0.0027)	0.0448*** (0.0026)
	BLACK	γ_{12}	−0.0182*** (0.0045)	−0.0189*** (0.0045)	−0.0183*** (0.0045)
Variance Components					
Level-1:	within-person	σ^2_ε	0.0948***	0.0948***	0.0948***
Level-2:	In initial status	σ^2_0	0.0506***	0.0510***	0.0503***
	In rate of change	σ^2_1	0.0016***	0.0016***	0.0016***
Goodness-of-fit					
	Deviance		4830.5	4827.0	4825.8
	AIC		4848.5	4847.0	4845.8
	BIC		4891.6	4894.9	4893.7

$-p < .10$; $*p < .05$; $**p < .01$; $***p < .001$.
Model A adds (*UERATE* − 7); Model B centers *UERATE* at each person's mean; Model C centers *UERATE* around each person's value of *UERATE* at his first measurement occasion.

Note: SAS Proc Mixed, Full ML. Also note that the covariance component, σ_{01}, is estimated, but not displayed.

We therefore often prefer recentering time-varying predictors *not* around the grand-mean but rather around a substantively meaningful constant—here, say 7. This allows γ_{00} to describe the average log-wage for someone whose local area has a 7% unemployment rate. The results of fitting this last model appear in the first column of table 5.8. As in section 5.2.1, we can interpret this parameter estimate by computing $100(e^{(-0.0120)} - 1) = -1.2$. We conclude that each one-percentage point difference in local area unemployment rate is associated with wages that are 1.2 percent lower.

Given that centering has so little effect on model interpretation, you may wonder why we raise this issue. We do so for three reasons: (1) the topic receives much attention in the multilevel literature (see, e.g., Kreft et al., 1995; Hofmann & Gavin, 1998); (2) some computer programs tempt analysts into recentering their predictors through the availability of simple toggle switches on an interactive menu; and (3) there are still other meaningful ways of recentering. Not only can you recenter around a *single* constant, you can recenter around *multiple constants*, one per person. It is this approach, also known as within-context or group-mean centering, to which we now turn.

The general idea behind within-context centering is simple: instead of representing a time-varying predictor using a single variable, decompose the predictor into multiple constituent variables, which, taken together, separately identify specific sources of variation in the outcome. Of the many ways of decomposing a time-varying predictor, two deserve special mention:

- *Within-person centering*: include the *average* unemployment rate for individual i, $\overline{UERATE}_{i0}$, as well as the deviation of each period's rate from this average, $(UERATE_{ij} - \overline{UERATE}_{i0})$.
- *Time-1 centering*: include *time-1*'s unemployment rate for individual i, $UERATE_{i1}$, as well as the deviation of each subsequent rate from this original value, $(UERATE_{ij} - UERATE_{i1})$.

Within-context centering provides *multiple* ways of representing a time varying predictor. Under within-person centering, you include a time-invariant *average* value and deviations from that average; under time-1 centering, you include the time-invariant *initial* value and deviations from that starting point. In both cases, as well as in the many other possible versions of within-context centering, the goal is to represent the predictor in a way that provides greater insight into its effects. (Of course, within-person centering raises interpretive problems of endogeneity, discussed in the following section.)

The last two columns of table 5.8 present the results of fitting the multilevel model for change with *UERATE* centered within-person (Model B) and around time-1 (Model C). Each contributes a particular insight into the negative effect of local unemployment on dropouts' wages. Model B reveals an association between wages and two aspects of the unemployment: (1) its average over time—the lower the average rate, the lower the wage; and (2) its relative magnitude, at each point in time, in comparison to this average. Model C demonstrates that wages are also associated with two other aspects of the time-varying unemployment rates: (1) their *initial* value, when the dropout first enters the labor force; and (2) the *increment* or *decrement*, at each subsequent point in time, from that initial value. Is either of these centered options clearly superior to the raw variable representation? Given that we cannot compare deviance statistics (because no model is nested within any other), comparison of AIC and BIC statistics suggests that all three are roughly comparable, with BIC giving the nod to Model A and AIC the nod to Model C.

These strategies for representing the effect of a time-varying predictor are hardly the only options. We offer them primarily in the hope that they will stimulate your thinking about substantively interesting ways of representing predictors' effects. We find routine recommendations to always, or never, center unconstructive. We prefer instead to recommend that you think carefully about which representations might provide the greatest insight into the phenomenon you are studying.

5.3.4 An Important Caveat: The Problem of Reciprocal Causation

Most researchers get very excited by the possibility that a statistical model could represent the relationship between changing characteristics of individuals and their environments, on the one hand, and individual outcomes on the other. We now dampen this enthusiasm by highlighting interpretive difficulties that time-varying predictors can present. The problem, known generally as *reciprocal causation* or *endogeneity*, is the familiar "chicken and egg" cliché: if X is correlated with Y, can you conclude that X *causes* Y or is it possible that Y *causes* X?

Many, but not all, time-varying predictors are subject to these problems. To help identify which are most susceptible, we classify time-varying predictors into four groups: *defined*, *ancillary*, *contextual*, and *internal*[2]. In the context of individual growth modeling, classification is based on the degree to which a predictor's values at time t_{ij} are: (1) assignable *a priori*; and (2) potentially influenced by the study participant's contemporaneous outcome. The more "control" a study participant has over his or her predictor values, the more clouded your inferences.

A time-varying predictor is *defined* if, in advance of data collection, its values are predetermined for everyone under study. Defined predictors are impervious to issues of reciprocal causation because no one—not the study participants nor the researchers—can alter their values. Most defined predictors are themselves functions of time. All representations of *TIME* are defined because their values depend solely on a record's time-period. Time-varying predictors that reflect other periodic aspects of time—such as season (fall, summer, etc) or anniversary (anniversary month, nonanniversary month)—are defined because once the metric

for time is chosen, so, too, are their values. Predictors whose values are set by an external schedule are also defined. If Ginexi and colleagues (2000) added a variable representing each person's time-varying unemployment benefits, its values would be defined because payments reflect a uniform schedule. Similarly, when comparing the efficacy of time-varying drug tapering regimens in a randomized trial, an individual's dosage is defined if the researcher determines the entire dosing schedule *a priori*. Different people may take different doses at different times, but if the schedule is predetermined, the predictor is defined.

A time-varying predictor is *ancillary* if its values cannot be influenced by study participants because they are determined by a stochastic process totally external to them. We use the term "stochastic process" to emphasize that, unlike a defined predictor, an ancillary predictor can behave erratically over time. Ancillary predictors are impervious to issues of reciprocal causation because no one involved in the study directly affects their values. Most ancillary predictors assess potentially changing characteristics of the physical or social environment in which respondents live. In his study of marital dissolution, for example, South (1995) divided the United States into 382 local marriage markets and used census data to create a time-varying predictor assessing the availability of spousal alternatives in each market. His *availability index* contrasted the number of unmarried persons "locally available" to the respondent with the number of unmarried persons "locally available" to the respondent's spouse. As no respondent could be part of the local marriage market (because all were married), this predictor is ancillary. If some *were* part of the local market (as they would be in a study of marital *initiation*), this predictor would be approximately ancillary because: (1) the contribution of any individual to the index would be negligible (given that the smallest marriage market included over a half million people); and (2) few individuals move to a particular area because of the availability of spousal alternatives. Following this logic, the local area unemployment rate just used in the high school dropout wage analysis is approximately ancillary. Other ancillary predictors include weather (Young, Meaden, Fogg, Cherin, & Eastman, 1997) and treatment, if randomly assigned.

A *contextual* time-varying predictor also describes an "external" stochastic process, but the connection between units is closer—between husbands and wives, parents and children, teachers and students, employers and employees. Because of this proximity, contextual predictors can be influenced by an individual's contemporaneous outcome values; if so, they are susceptible to issues of reciprocal causation. To assess whether reciprocal causation is a problem, you must analyze the particular situation. For example, in their 30-year study of the effects of parental divorce on mental health, Cherlin, Chase-Lansdale and McRae (1998) included time-varying predictors denoting whether children had experienced a parental divorce during four developmental phases: 7–10, 11–15, 16–22, and 23–33. These contextual time-varying predictors are unlikely to create interpretive problems because it is doubtful that someone's level of emotional problems would influence either the occurrence or the timing of a parental divorce. But in their three-year study of the link between the quality of childcare centers and children's early cognitive and language development, Burchinal et al. (2000) face a thornier problem. Because parents may *choose* particular childcare centers precisely because they emphasize particular skills, observed links between center quality and child development may be due to a link between development and quality, not quality and development. If such criticisms seem reasonable, we suggest that you treat a contextual time-varying predictor as if it were internal, and address issues of reciprocal causation in ways we now describe.

Internal time-varying predictors describe an individual's potentially changeable status over time. Some describe *psychological* states (mood or satisfaction), while others describe *physical* states (respiratory function, blood levels), *social* states (married/unmarried, working/unemployed), or other personal attributes. In their four-year study of adolescent smoking, for example, Killen, Robinson, Haydel, et al. (1997) annually assessed dozens of internal predictors ranging from counts of the number of friends who smoke and the frequency of drinking to the adolescent's height and weight. And in their four-year study of conduct disorder in boys, Lahey, McBurnett, Loeber, & Hart (1995) collected annual data on receipt of various kinds of psychological treatment, both in-patient and out-patient, medication and talk therapy.

Internal time-varying predictors raise serious interpretive dilemmas. Isn't it reasonable to argue, for example, that as teens start smoking, they increase the number of friends who smoke, increase their frequency of drinking, and lose weight? So, too, isn't it possible that as a child's behavior worsens a parent may be more likely to initiate psychotherapy? Although the causal link may be from predictor to outcome, it may also run the opposite way. Some readers may believe that longitudinal data—and the associated statistical models—should resolve such concerns. But resolution of the directional arrow is more difficult. As long as a model links *contemporaneous* information about time-varying predictors and outcomes, we effectively convert a longitudinal problem into a cross-sectional one, fully burdened by questions of reciprocal causation.

Given the conceptual appeal of internal and contextual time-varying predictors, what should you do? We have two concrete recommendations.

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 his 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 in the placebo group ($\hat{\gamma}_{10} = -2.42$, *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

5.10: Results of using alternative representations for the main effect of *TIME* when ıting the effect of treatment on the positive mood scores in the antidepressant trial '3)

			Temporal predictor in level-1 model		
	Parameter		*TIME*	(*TIME* – 3.33)	(*TIME* – 6.67)
Effects					
·l	Intercept	γ_{00}	167.46***	159.40***	151.34***
ɔrcept, π_{0i}			(9.33)	(8.76)	(11.54)
	TREAT	γ_{01}	−3.11	15.35	33.80*
			(12.33)	(11.54)	(15.16)
ɔf	Intercept	γ_{10}	−2.42	−2.42	−2.42
ıge, π_{1i}			(1.73)	(1.73)	(1.73)
	TREAT	γ_{11}	5.54*	5.54*	5.54*
			(2.28)	(2.28)	(2.28)
ıce Components					
·1:	within-person	σ^2_ε	1229.93***	1229.93***	1229.93***
2:	In level-1 intercept	σ^2_0	2111.33***	2008.72***	3322.45***
	In rate of change	σ^2_1	63.74***	63.74***	63.74***
	Covariance	σ_{01}	−121.62*	90.83	303.28***
ness-of-fit					
	Deviance		12680.5	12680.5	12680.5
	AIC		12696.5	12696.5	12696.5
	BIC		12714.8	12714.8	12714.8

10; * $p < .05$; ** $p < .01$; *** $p < .001$.
is centered around initial status, middle status, and final status.
'ull ML, SAS PROC MIXED.

dentical estimates for the linear rate of change in the placebo group ($\hat{\gamma}_{10} = -2.42$, *n.s.*) and the effect of treatment on that rate ($\hat{\gamma}_{11} = 5.54$, $p < 0.05$). ɔo, too, we obtain identical estimates for the residual variance in the rate ɔf change ($\hat{\sigma}^2_1 = 63.74$, $p < .001$) and the within-person residual variance $\hat{\sigma}^2_\varepsilon = 1229.93$). 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 ntercepts refer to these anchors, each model tests a different set of ıypotheses about them. If we change c, we change the anchors, which hanges the estimates and their interpretations. In terms of the general nodel in equations 5.12a and 5.12b, γ_{00} assesses the elevation of the popılation average change trajectory at time c; γ_{01} assesses the differential ·levation of this trajectory at time c between groups; σ^2_0 assesses the popılation variance in true status at time c; and σ_{01} assesses the population

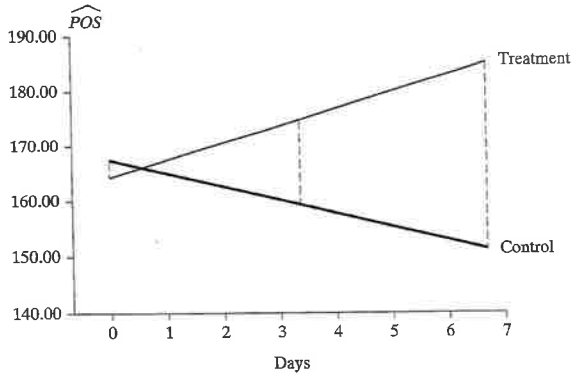


Figure 5.5. Understanding the consequences of rescaling the effect of *TIME*. Prototypical trajectories for individuals by *TREATMENT* status in the antidepressant experiment. The dashed vertical lines reflect the magnitude of the effect of *TREATMENT* if time is centered at the study's beginning (0),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

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

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:

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

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