4

Doing Data Analysis with the Multilevel Model for Change

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

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

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

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

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

4.1 Example: Changes in Adolescent Alcohol Use

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

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

Figure
growth
study.

To
plots
cents
most
trans
This
linea
lesce
years
first
As
easie
of a

This
sets,
inter

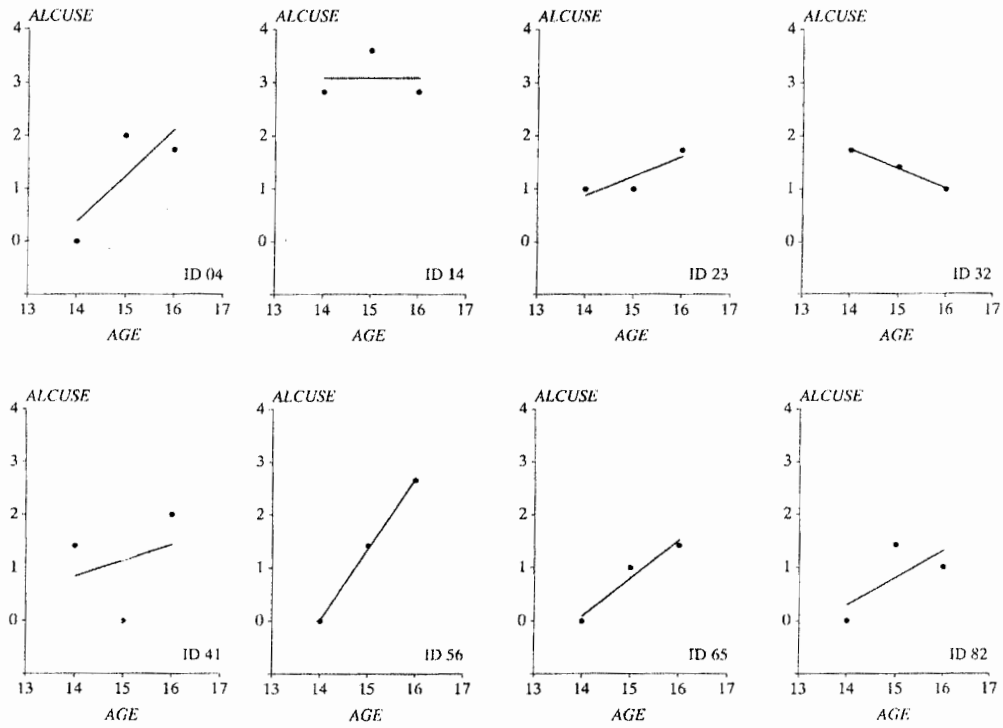


Figure 4.1. Identifying a suitable functional form for the level-1 submodel. Empirical growth plots with superimposed OLS trajectories for 8 participants in the alcohol use study.

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 $(AGE_{ij} - 14)$:

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

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.

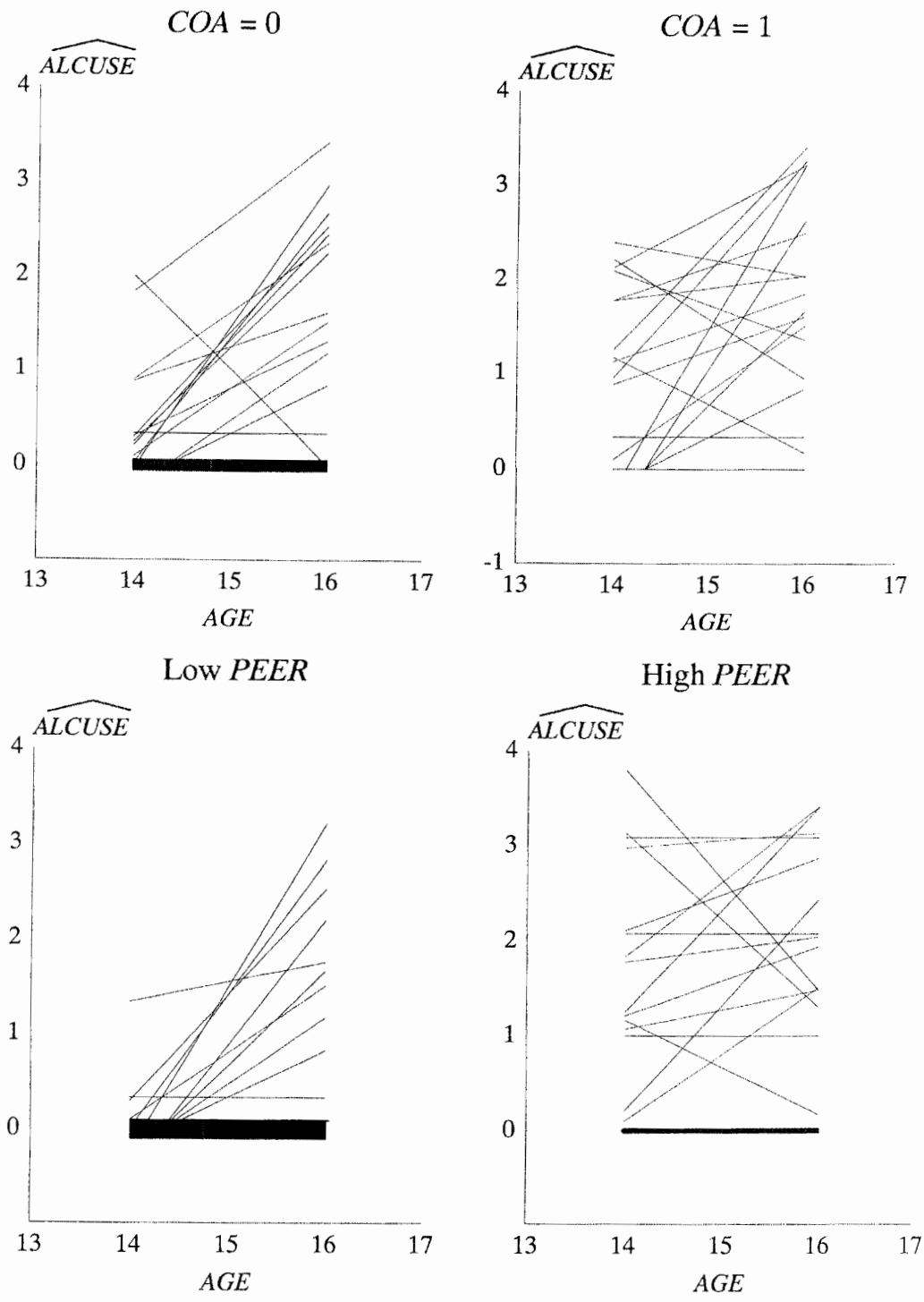


Figure 4.2. Identifying potential predictors of change by examining OLS fitted trajectories separately by levels of selected predictors. Fitted OLS trajectories for the alcohol use data displayed separately by *COA* status (upper panel) and *PEER* alcohol use (lower panel).

- 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:

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

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:

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

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 posi-

tion 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 ($COA = 0$):

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

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 distri-

bution 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

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

4.4 First Steps: Fitting Two Unconditional Multilevel Models for Change

You've articulated your research questions, created a person-period data set, conducted exploratory analyses, chosen an estimation approach, and selected a software package. Although you might be tempted to begin by fitting models that include your substantive predictors, we suggest that you first fit the two simpler models presented in this section: the *unconditional means model* (section 4.4.1) and the *unconditional growth model* (section 4.4.2). These unconditional models partition and quantify the outcome variation in two important ways: first, across people without regard to time (the unconditional means model), and second, across both people *and* time (the unconditional growth model). Their results allow you to establish: (1) whether there is systematic variation in your outcome that is worth exploring; and (2) *where* that variation resides (within or between people). They also provide two valuable baselines against which you can evaluate the success of subsequent model building, 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:

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

where we assume, as usual, that:

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

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 population 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 population 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 components, which assess the amount of outcome variation that exists at each level. Associated hypothesis tests help determine whether there is sufficient variation at that level to warrant further analysis. If a variance component 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 component 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***	0.651***	0.316***	–0.317***	–0.314***	0.394***	0.651***
			(0.096)	(0.105)	(0.131)	(0.148)	(0.146)	(0.104)	(0.080)
	COA	γ_{01}			0.743***	0.579***	0.571***	0.571***	0.571***
					(0.195)	(0.162)	(0.146)	(0.146)	(0.146)
	PEER	γ_{02}				0.694***	0.695***	0.695***	0.695***
						(0.112)	(0.111)	(0.111)	(0.111)
Rate of change, π_{2i}	Intercept	γ_{10}		0.271***	0.293***	0.429***	0.425***	0.271***	0.271***
				(0.062)	(0.084)	(0.114)	(0.106)	(0.061)	(0.061)
	COA	γ_{11}			–0.049	–0.014			
					(0.125)	(0.125)			
	PEER	γ_{12}				–0.150~	–0.151~	–0.151~	–0.151~
						(0.086)	(0.085)	(0.085)	(0.085)

Variance Components									
Level 1	Within-person	σ^2_ε	0.562***	0.337***	0.337***	0.337***	0.337***	0.337***	0.337***
			(0.062)	(0.053)	(0.053)	(0.053)	(0.053)	(0.053)	(0.053)
Level 2	In initial status	σ^2_0	0.564***	0.624***	0.488**	0.241**	0.241**	0.241**	0.241**
			(0.119)	(0.148)	(0.128)	(0.093)	(0.093)	(0.093)	(0.093)
	In rate of change	σ^2_1		0.151**	0.151*	0.139*	0.139*	0.139*	0.139*
				(0.056)	(0.056)	(0.055)	(0.055)	(0.055)	(0.055)
	Covariance	σ_{01}		−0.068	−0.059	−0.006	−0.006	−0.006	−0.006
				(0.070)	(0.066)	(0.055)	(0.055)	(0.055)	(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

	Level-1/level-2 specification		
Model	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.

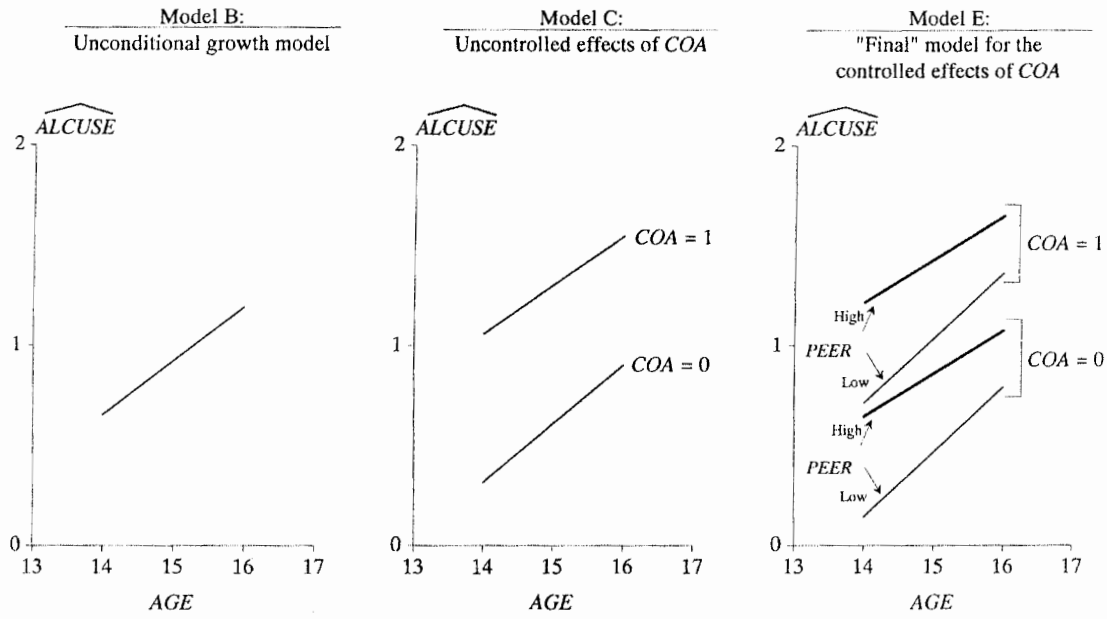


Figure 4.3. Displaying the results of fitted multilevel models for change. Prototypical trajectories from three models presented in table 4.1: Model B: the unconditional growth model, Model C: the uncontrolled effect of *COA*, Model E: the effect of *COA* controlling for *PEER*.

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_ε^2, 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 $\hat{\sigma}_\varepsilon^2$ 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_0^2 assesses the unpredicted variability in true initial status (the scatter of the π_{0i} around γ_{00}); σ_1^2 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:

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

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:

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

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 0.72 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_ε^2; the unconditional growth model provides baseline estimates of σ_0^2 and σ_1^2. Each leads to its own pseudo-R^2 statistic.

Let us begin by examining the decrease in within-person residual variance (σ_ε^2) 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:

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

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, $\hat{\sigma}_\varepsilon^2$ 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_0^2 and σ_1^2, 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:

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

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 ($\hat{\sigma}_\varepsilon^2$) 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: $\hat{\sigma}_0^2$ declines by 21.8% from Model B. Because it is still statistically significant, potentially explainable residual variation in initial status remains. While $\hat{\sigma}_1^2$ 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., $^1/_4$, $^1/_2$, and $^3/_4$). 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

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

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

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

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

- If neither *PEER* nor *COA* are centered (Model E), the intercepts represent a child of non-alchoholic parents whose peers at age 14 were totally abstinent (*PEER* = 0 and *COA* = 0).

- If *PEER* is centered and *COA* is not (Model F), the intercepts represent a child of non-alchoholic parents with an *average* value of *PEER* (*PEER* = 1.018 and *COA* = 0).
- If *both PEER* and *COA* are centered (Model G), the intercepts represent an *average* study participant—someone with *average* values of *PEER* and *COA* (*PEER* = 1.018 and *COA* = 0.451).

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

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

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

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

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

4.6 Comparing Models Using Deviance Statistics

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

4.6.1 The Deviance Statistic

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

As a by-product of ML estimation, the computer determines the magnitude of the log-likelihood function for this particular combination of observed data and parameter estimates. Statisticians call this number the *sample log-likelihood* statistic, often abbreviated as LL. Every program that uses ML methods outputs the LL statistic (or a transformation of it). In general, if you fit several competing models to the same data, the larger the LL statistic, the better the fit. This means that if the models you compare yield negative LL statistics, those that are *smaller* in absolute

value—i.e., closer to 0—fit better. (We state this obvious point explicitly as there has been some confusion in the literature about this issue.)

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

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

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

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

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

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

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

4.6.2 When and How Can You Compare Deviance Statistics?

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

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

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

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

Having fit a pair of models that meets these conditions, conducting tests is easy. Under the null hypothesis that the specified constraints hold, the difference in deviance statistics between a full and reduced model (often called "delta deviance" or ΔD) is distributed asymptotically as a χ^2 distribution with degrees of freedom (*d.f.*) equal to the number of inde-

pendent constraints imposed. If the models differ by one parameter, you have one degree of freedom for the test; if they differ by three parameters, you have three. As with any hypothesis test, you compare ΔD to a *critical value,* appropriate for that number of degrees of freedom, rejecting H_0 when the test statistic is large.[4]

4.6.3 Implementing Deviance-Based Hypothesis Tests

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Formally, we write:

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

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

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

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

4.7 Using Wald Statistics to Test Composite Hypotheses About Fixed Effects

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

based tests to compare models with different fixed effects, you still have a means of testing composite hypotheses about sets of fixed effects.

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

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

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

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

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

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

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

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: $H_0: \boldsymbol{C\gamma}' = 0$. 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 $\boldsymbol{C\gamma}' = 0$ 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 critical value 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 cannot reject the null hypothesis at any conventional level of significance ($\chi^2 = 1.44$, $d.f. = 1$). We conclude that these average true change trajectories converge by age 16. In other words, the alcohol consumption of children of non-alchoholic parents with *low CPEER* does indeed 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 deviance-based comparisons 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 you use ML methods to fit a linear regression model, for example, you assume that the errors are independent and normally distributed with constant variance. Assumptions allow you to move forward, estimate parameters, interpret results, and test hypotheses. But the validity of your conclusions rests on your assumptions' tenability. Fitting a model with untenable assumptions is as senseless as fitting a model to data that are knowingly flawed. Violations lead to biased estimates, incorrect standard errors, and erroneous inferences.

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

No analysis is complete until you examine the tenability of your assumptions. Of course, you can never be completely certain about the tenability of assumptions because you lack the very data you need to evaluate their tenability: information about the population from which your sample was drawn. Assumptions describe *true* individual change trajectories, population relationships between *true* individual growth parameters and level-2 predictors, and true errors for each person. All you can examine are the *observed* properties of *sample* quantities—*fitted* individual change trajectories, *estimated* individual growth parameters, and *sample* residuals.

Must you check the assumptions underlying every statistical model you fit? As much as we would like to say yes, reality dictates that we say no. Repetitive model checking is neither efficient nor plausible. We suggest instead that you examine the assumptions of several initial models and 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

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, $\widehat{\varepsilon}_{ij}$	level-2 residual, $\widehat{\zeta}_{0i}$	level-2 residual, $\widehat{\zeta}_{1i}$
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 $\widehat{\pi}_{0i}$. With the exception of two extreme data points, the plot of $\widehat{\pi}_{0i}$ vs. *PEER* suggests a strong linear relationship.	Because *COA* is dichotomous, there is no linearity assumption for $\widehat{\pi}_{1i}$. Plot of $\widehat{\pi}_{1i}$ vs. *PEER* suggests a weak linear relationship.
Normality. All residuals, at both level-1 and level-2, will be normally distributed.	A plot of $\widehat{\varepsilon}_{ij}$ vs. normal scores suggests normality. We find further support for normality in a plot of standardized $\widehat{\varepsilon}_{ij}$ vs *ID*, which reveals no unusual data points.	A plot of $\widehat{\zeta}_{0i}$ vs. normal scores suggests normality. So does a plot of standardized $\widehat{\zeta}_{0i}$ vs. *ID*, which reveals no unusual data points. There is slight evidence of a floor effect in the outcome.	A plot of $\widehat{\zeta}_{1i}$ 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 $\widehat{\zeta}_{1i}$ 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 $\widehat{\varepsilon}_{ij}$ vs. AGE suggests su variability at ages 14, 15, and 16.	A plot of $\widehat{\zeta}_{0i}$ 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 $\widehat{\zeta}_{1i}$ 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.

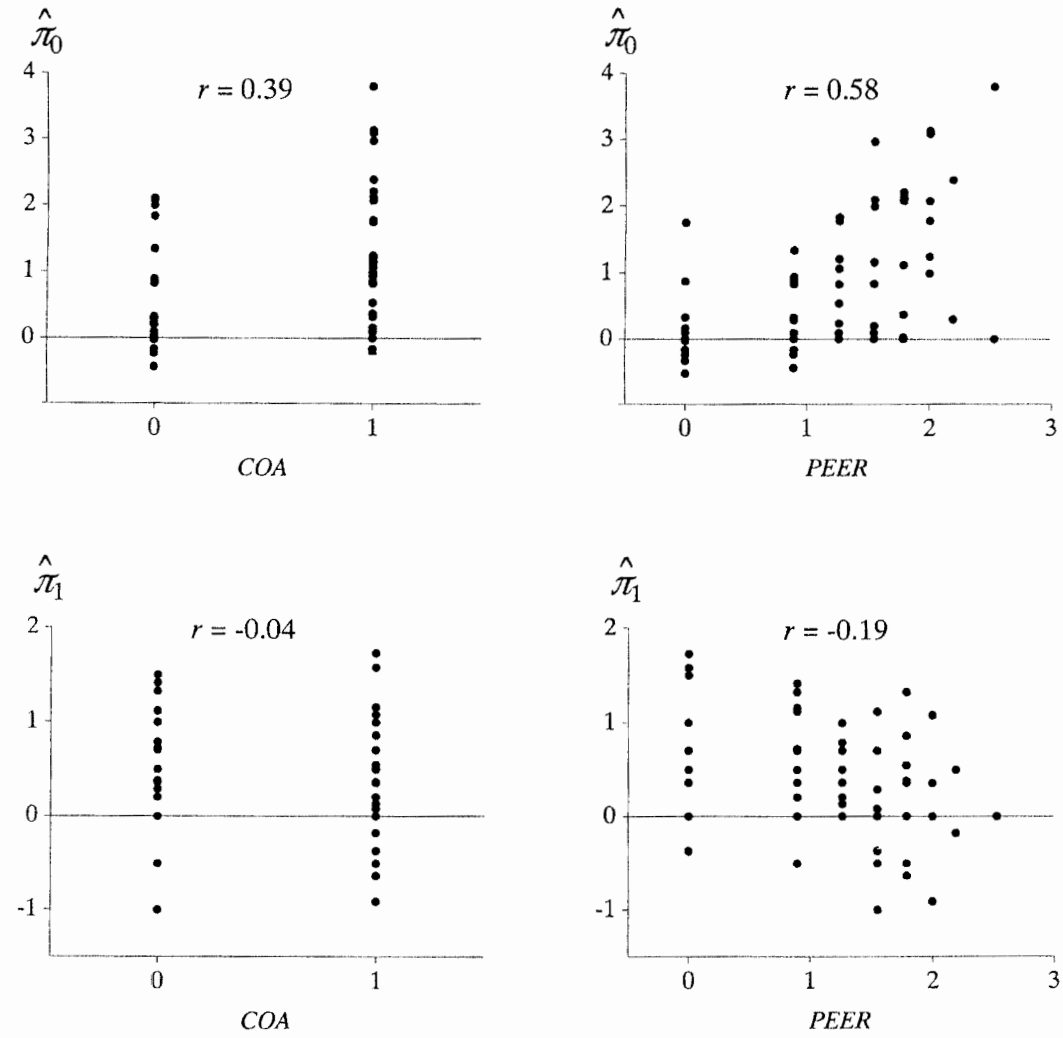


Figure 4.4. Examining the level-2 linearity assumption in the multilevel model for change. OLS estimated individual growth parameters (for the intercept and slope) plotted vs. selected predictors. Left panel is for the predictor *COA*; right panel is for the predictor *PEER*.

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, $\widehat{\varepsilon}_{ij}$, and the first level-2 residual, $\widehat{\zeta}_{0i}$. The plot for second level-2 residual, $\widehat{\zeta}_{2i}$, 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

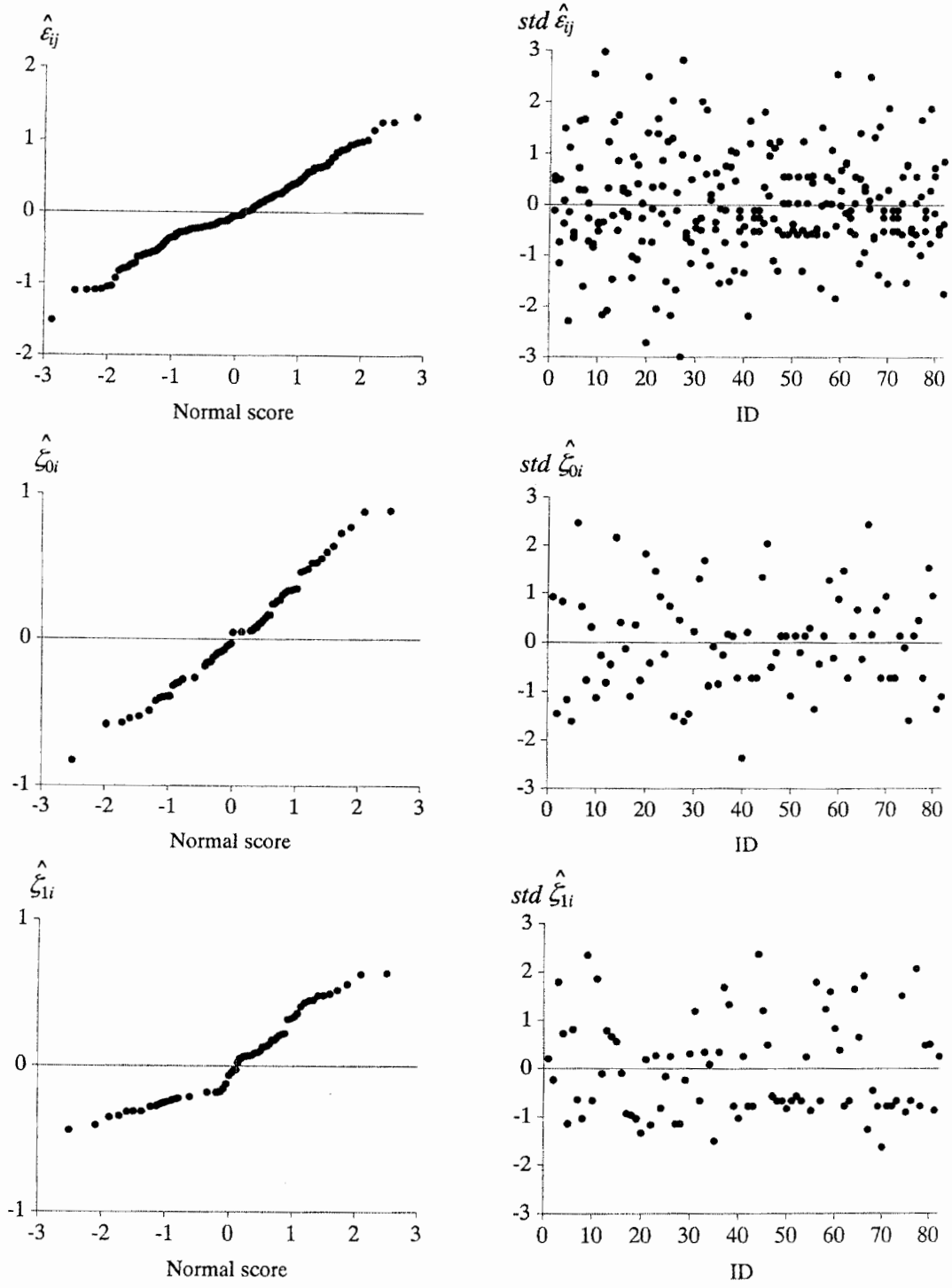


Figure 4.5. Examining normality assumptions in the multilevel model for change. Left panel presents normal probability plots for the raw residuals at level-1 and level-2. Right panel presents plots of standardized residuals at level-1 and level-2 vs. *ID* numbers.

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

Plots of *standardized* residuals—either univariate plots or bivariate plots against 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 deviations of their center (i.e., only 5% will be greater than 2). Use caution when applying this simple rule of thumb, however, because there are other distributions that are *not* normal in which about 5% of the observations also fall in these tails.

You can also plot the standardized residuals by *ID* to identify extreme individuals (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 relatively few between 2 and 3, and none beyond. Plots of standardized level-2 residuals suggest that the negative residuals tend to be smaller in magnitude, "pulled in" toward the center of both plots. This feature is most evident for the second level-2 residual, $\widehat{\zeta}_{1i}$, in the lower plot, but there is also evidence of its presence in the plot for $\widehat{\zeta}_{0i}$. Again, compression of the lower tail may result from the fact that the outcome, *ALCUSE*, has a "floor" of zero.

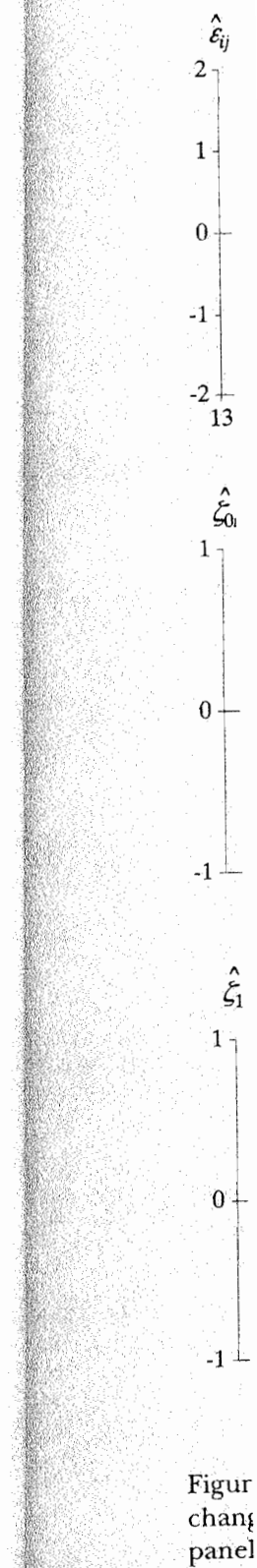
4.8.3 Checking Homoscedasticity

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

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

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

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



Figur
chan
panel

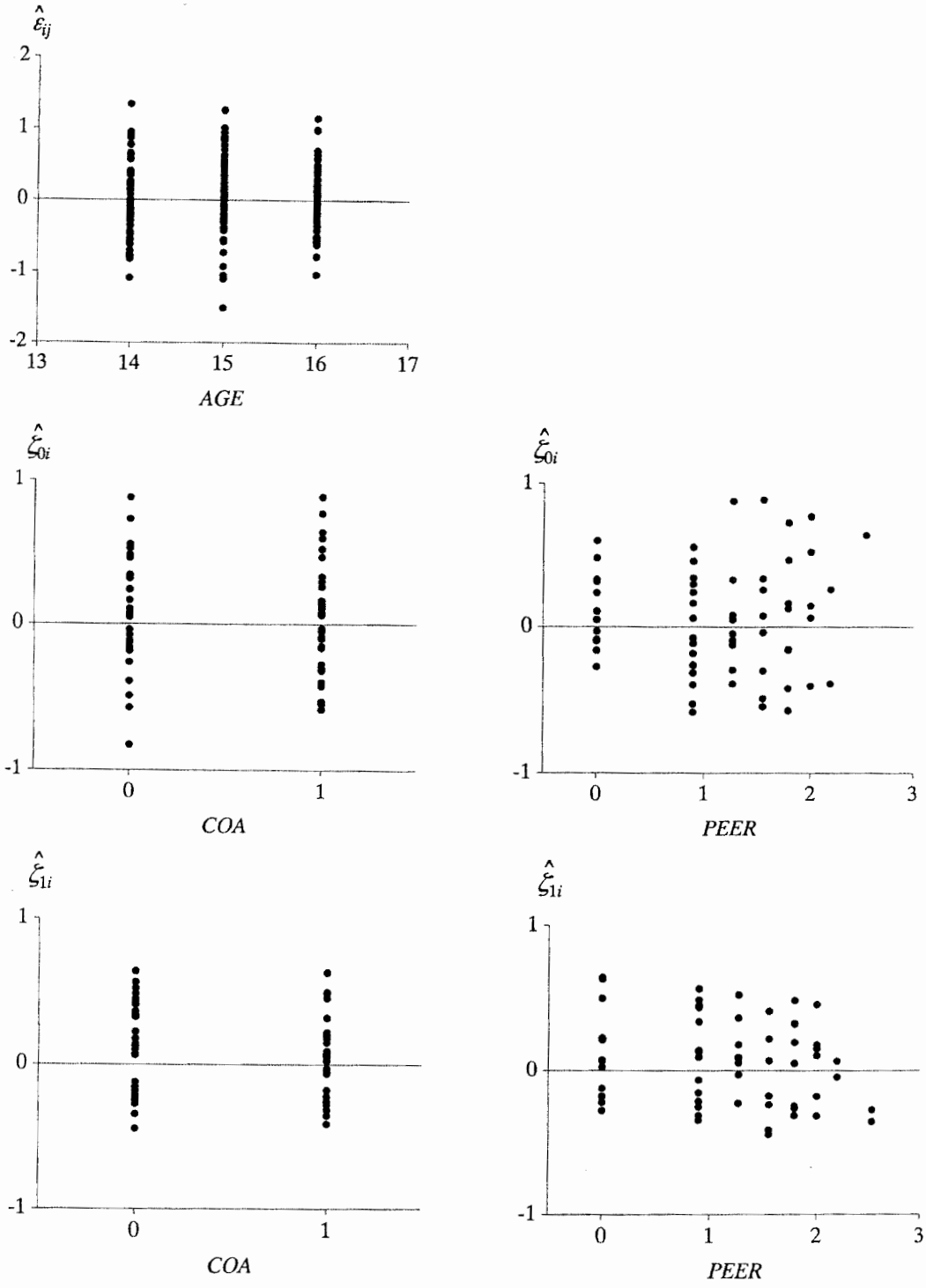


Figure 4.6. Examining the homoscedasticity assumptions in the multilevel model for change. Top panel presents raw level-1 residuals vs. the level-1 predictor *AGE*. Remaining panels present raw level-2 residuals vs. the two level-2 predictors, *COA* and *PEER*.

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, $\widehat{\zeta}_{0i}$ and $\widehat{\zeta}_{1i}$, 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:

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

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:

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

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.

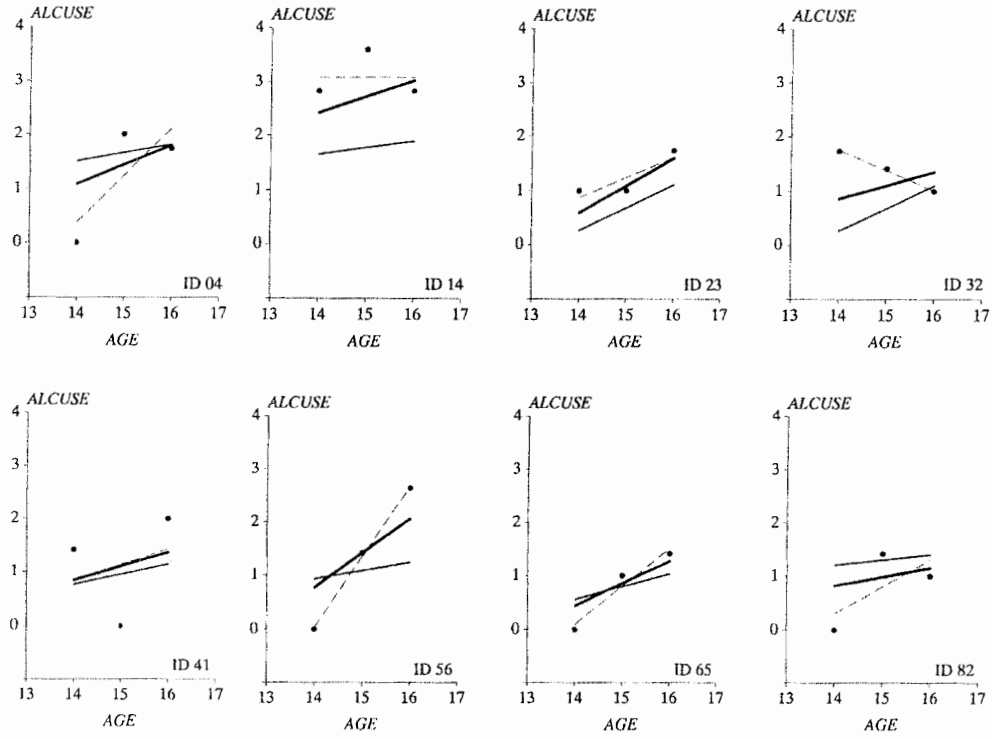


Figure 4.7. Model-based (empirical Bayes) estimates of the individual growth trajectories. Each plot presents the observed *ALCUSE* measurements (as data points), OLS fitted trajectories (dashed lines), population average trajectories (faint lines), and model-based empirical Bayes trajectories (bold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

Now notice how each model-based trajectory (in bold) falls between its 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 the 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 trajectories differ markedly from person to person, the model-based trajectories 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 trajectories are *shrunk* toward the average trajectory of that person's peer group (those with the same predictor values). This combination yields a superior, more precise, estimate.

Model-based estimates are also more precise because they require estimation of fewer parameters. In positing the multilevel model for change,

we assume that everyone shares the same level-1 residual variance, σ_ε^2. 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*. Statisticians recommend precision—indeed, increased precision is a fundamental 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, administrators 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.