CHAPTER 11

Multilevel structures

As we illustrate in detail in subsequent chapters, multilevel models are extensions of regression in which data are structured in groups and coefficients can vary by group. In this chapter, we illustrate basic multilevel models and present several examples of data that are collected and summarized at different levels. We start with simple grouped data—persons within cities—where some information is available on persons and some information is at the city level. We then consider examples of repeated measurements, time-series cross sections, and non-nested structures. The chapter concludes with an outline of the costs and benefits of multilevel modeling compared to classical regression.

11.1 Varying-intercept and varying-slope models

With grouped data, a regression that includes indicators for groups is called a *varying-intercept model* because it can be interpreted as a model with a different intercept within each group. Figure 11.1a illustrates with a model with one continuous predictor x and indicators for $J = 5$ groups. The model can be written as a regression with 6 predictors or, equivalently, as a regression with two predictors (x and the constant term), with the intercept varying by group:

$$\text{varying-intercept model:}\quad y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i.$$

Another option, shown in Figure 11.1b, is to let the slope vary with constant intercept:

$$\text{varying-slope model:}\quad y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i.$$

Finally, Figure 11.1c shows a model in which both the intercept and the slope vary by group:

$$\text{varying-intercept, varying-slope model:}\quad y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i.$$

The varying slopes are interactions between the continuous predictor x and the group indicators.

As we discuss shortly, it can be challenging to estimate all these α_j's and β_j's, especially when inputs are available at the group level. The first step of multilevel modeling is to set up a regression with varying coefficients; the second step is to set up a regression model for the coefficients themselves.

11.2 Clustered data: child support enforcement in cities

With multilevel modeling we need to go beyond the classical setup of a data vector y and a matrix of predictors X (as shown in Figure 3.6 on page 38). Each level of the model can have its own matrix of predictors.

We illustrate multilevel data structures with an observational study of the effect of city-level policies on enforcing child support payments from unmarried fathers. The treatment is at the group (city) level, but the outcome is measured on individual families.

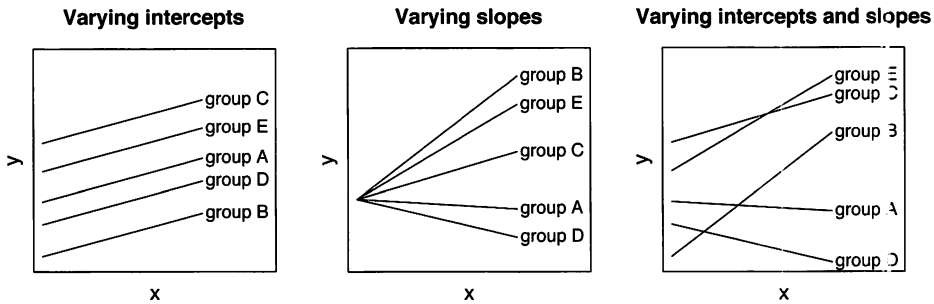

Figure 11.1 *Linear regression models with (a) varying intercepts ($y = \alpha_j + \beta x$), (b) varying slopes ($y = \alpha + \beta_j x$), and (c) both ($y = \alpha_j + \beta_j x$). The varying intercepts correspond to group indicators as regression predictors, and the varying slopes represent interactions between x and the group indicators.*

ID	dad age	mom race	informal support	city ID	city name	enforce intensity	benefit level	city indicators 1	2	···	20
1	19	hisp	1	1	Oakland	0.52	1.01	1	0	···	0
2	27	black	0	1	Oakland	0.52	1.01	1	0	···	0
3	26	black	1	1	Oakland	0.52	1.01	1	0	···	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
248	19	white	1	3	Baltimore	0.05	1.10	0	0	···	0
249	26	black	1	3	Baltimore	0.05	1.10	0	0	···	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
1366	21	black	1	20	Norfolk	−0.11	1.08	0	0	···	1
1367	28	hisp	0	20	Norfolk	−0.11	1.08	0	0	···	1

Figure 11.2 *Some of the data from the child support study, structured as a single matrix with one row for each person. These indicators would be used in classical regression to allow for variation among cities. In a multilevel model they are not necessary, as we code cities using their index variable ("city ID") instead. We prefer separating the data into individual-level and city-level datasets, as in Figure 11.3.*

Studying the effectiveness of child support enforcement

Cities and states in the United States have tried a variety of strategies to encourage or force fathers to give support payments for children with parents who live apart. In order to study the effectiveness of these policies for a particular subset of high-risk children, an analysis was done using a sample of 1367 noncohabiting parents from the Fragile Families study, a survey of unmarried mothers of newborns in 20 cities. The survey was conducted by sampling from hospitals which themselves were sampled from the chosen cities, but here we ignore the complexities of the data collection and consider the mothers to have been sampled at random (from their demographic category) in each city.

To estimate the effect of child support enforcement policies, the key "treatment" predictor is a measure of enforcement policies, which is available at the city level. The researchers estimated the probability that the mother received informal support, given the city-level enforcement measure and other city- and individual-level predictors.

ID	dad age	mom race	informal support	city ID
1	19	hisp	1	1
2	27	black	0	1
3	26	black	1	1
⋮	⋮	⋮	⋮	⋮
248	19	white	1	3
249	26	black	1	3
⋮	⋮	⋮	⋮	⋮
1366	21	black	1	20
1367	28	hisp	0	20

city ID	city name	enforce-ment	benefit level
1	Oakland	0.52	1.01
2	Austin	0.00	0.75
3	Baltimore	−0.05	1.10
⋮	⋮	⋮	⋮
20	Norfolk	−0.11	1.08

Figure 11.3 *Data from the child support study, structured as two matrices, one for persons and one for cities. The inputs at the different levels are now clear. Compare to Figure 11.2.*

A data matrix for each level of the model

Figure 11.2 shows the data for the analysis as it might be stored in a computer package, with information on each of the 1367 mothers surveyed. To make use of the multilevel structure of the data, however, we need to construct *two* data matrices, one for each level of the model, as Figure 11.3 illustrates. At the left is the person-level data matrix, with one row for each survey respondent, and their cities are indicated by an index variable; at the right is the city data matrix, giving the name and other information available for each city.

At a practical level, the two-matrix format of Figure 11.3 has the advantage that it contains each piece of information exactly once. In contrast, the single large matrix in Figure 11.2 has each city's data repeated several times. Computer memory is cheap so this would not seem to be a problem; however, if city-level information needs to be added or changed, the single-matrix format invites errors.

Conceptually, the two-matrix, or multilevel, data structure has the advantage of clearly showing which information is available on individuals and which on cities. It also gives more flexibility in fitting models, allowing us to move beyond the classical regression framework.

Individual- and group-level models

We briefly outline several possible ways of analyzing these data, as a motivation and lead-in to multilevel modeling.

Individual-level regression. In the most basic analysis, informal support (as reported by mothers in the survey) is the binary outcome, and there are several individual- and city-level predictors. Enforcement is considered as the treatment, and a logistic regression is used, also controlling for other inputs. This is the starting point of the observational study.

Using classical regression notation, the model is $\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta)$, where X includes the constant term, the treatment (enforcement intensity), and the other predictors (father's age and indicators for mother's race at the individual level; and benefit level at the city level). X is thus constructed from the data matrix of Figure 11.2. This individual-level regression has the problem that it ignores city-level variation beyond that explained by enforcement intensity and benefit level, which are the city-level predictors in the model.

city ID	city name	enforce-ment	benefit level	# in sample	avg. age	prop. black	proportion with informal support
1	Oakland	0.52	1.01	78	25.9	0.67	0.55
2	Austin	0.00	0.75	91	25.8	0.42	0.54
3	Baltimore	−0.05	1.10	101	27.0	0.86	0.67
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	Norfolk	−0.11	1.08	31	27.4	0.84	0.65

Figure 11.4 *City-level data from child support study (as in the right panel of Figure 11.3), also including sample sizes and sample averages from the individual responses.*

Group-level regression on city averages. Another approach is to perform a city-level analysis, with individual-level predictors included using their group-level averages. Figure 11.4 illustrates: here, the outcome, y_j, would be the average total support among the respondents in city j, the enforcement indicator would be the treatment, and the other variables would also be included as predictors. Such a regression—in this case, with 20 data points—has the advantage that its errors are automatically at the city level. However, by aggregating, it removes the ability of individual predictors to predict individual outcomes. For example, it is possible that older fathers give more informal support—but this would not necessarily translate into average father's age being predictive of more informal support at the city level.

Individual-level regression with city indicators, followed by group-level regression of the estimated city effects. A slightly more elaborate analysis proceeds in two steps: first fitting a logistic regression to the individual data y given individual predictors (in this example, father's age and indicators for mother's race) along with indicators for the 20 cities. This first-stage regression then has 22 predictors. (The constant term is *not* included since we wish to include indicators for all the cities; see the discussion at the end of Section 4.5.)

The next step in this two-step analysis is to perform a *linear* regression at the city level, considering the estimated coefficients of the city indicators (in the individual model that was just fit) as the "data" y_j. This city-level regression has 20 data points and uses, as predictors, the city-level data (in this case, enforcement intensity and benefit level). Each of the predictors in the model is thus included in one of the two regressions.

The two-step analysis is reasonable in this example but can run into problems when sample sizes are small in particular groups, or when there are interactions between individual- and group-level predictors. Multilevel modeling is a more general approach that can include predictors at both levels at once.

Multilevel models

The multilevel model looks something like the two-step model we have described, except that both steps are fitted at once. In this example, a simple multilevel model would have two components: a logistic regression with 1369 data points predicting the binary outcome given individual-level predictors and with an intercept that can vary by city, and a linear regression with 20 data points predicting the city intercepts from city-level predictors. In the multilevel framework, the key link between the individual and city levels is the city indicator—the "city ID" variable in Figure 11.3, which takes on values between 1 and 20.

For this example, we would have a logistic regression at the data level:

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta + \alpha_{j[i]}), \text{ for } i = 1, \ldots, n, \tag{11.1}$$

where X is the matrix of individual-level predictors and $j[i]$ indexes the city where person i resides. The second part of the model—what makes it "multilevel"—is the regression of the city coefficients:

$$\alpha_j \sim \text{N}(U_j\gamma, \sigma_\alpha^2), \text{ for } j = 1, \ldots, 20, \tag{11.2}$$

where U is the matrix of city-level predictors, γ is the vector of coefficients for the city-level regression, and σ_α is the standard deviation of the unexplained group-level errors.

The model for the α's in (11.2) allows us to include all 20 of them in model (11.1) without having to worry about collinearity. The key is the group-level variation parameter σ_α, which is estimated from the data (along with α, β, and a) in the fitting of the model. We return to this point in the next chapter.

Directions for the observational study

The "treatment" variable in this example is not randomly applied; hence it is quite possible that cities that differ in enforcement intensities could differ in other important ways in the political, economic, or cultural dimensions. Suppose the goal were to estimate the effects of potential interventions (such as increased enforcement), rather than simply performing a comparative analysis. Then it would make sense to set this up as an observational study, gather relevant pre-treatment information to capture variation among the cities, and perhaps use a matching approach to estimate effects. In addition, good pre-treatment measures on individuals should improve predictive power, thus allowing treatment effects to be estimated more accurately. The researchers studying these child support data are also looking at other outcomes, including measures of the amity between the parents as well as financial and other support.

Along with the special concerns of causal inference, the usual recommendations of regression analysis apply. For example, it might make sense to consider interactions in the model (to see if enforcement is more effective for older fathers, for example).

11.3 Repeated measurements, time-series cross sections, and other non-nested structures

Repeated measurements

Another kind of multilevel data structure involves repeated measurements on persons (or other units)—thus, measurements are clustered within persons, and predictors can be available at the measurement or person level. We illustrate with a model fitted to a longitudinal dataset of about 2000 Australian adolescents whose smoking patterns were recorded every six months (via questionnaire) for a period of three years. Interest lay in the extent to which smoking behavior can be predicted based on parental smoking and other background variables, and the extent to which boys and girls pick up the habit of smoking during their teenage years. Figure 11.5 illustrates the overall rate of smoking among survey participants.

A multilevel logistic regression was fit, in which the probability of smoking depends on sex, parental smoking, the wave of the study, and an individual parameter

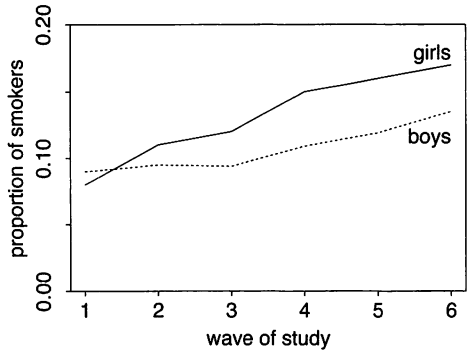

Figure 11.5 *Prevalence of regular (daily) smoking among participants responding at each wave in the study of Australian adolescents (who were on average 15 years old at wave 1).*

person		parents smoke?		wave 1		wave 2		...
ID	sex	mom	dad	age	smokes?	age	smokes?	
1	f	Y	Y	15:0	N	15:6	N	...
2	f	N	N	14:7	N	15:1	N	...
3	m	Y	N	15:1	N	15:7	Y	...
4	f	N	N	15:3	N	15:9	N	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Figure 11.6 *Data from the smoking study as they might be stored in a single computer file and read into R as a matrix,* `data`. *(Ages are in years:months.) These data have a multilevel structure, with observations nested within persons.*

for the person. For person j at wave t, the modeled probability of smoking is

$$\begin{aligned}\Pr(y_{jt}=1) &= \text{logit}^{-1}(\beta_0+\beta_1\text{psmoke}_j+\beta_2\text{female}_j+\\ &\quad +\beta_3(1-\text{female}_j)\cdot t+\beta_4\text{female}_j\cdot t+\alpha_j),\end{aligned} \tag{11.3}$$

where `psmoke` is the number of the person's parents who smoke and `female` is an indicator for females, so that β_3 and β_4 represent the time trends for boys and girls, respectively.[1]

Figures 11.6 and 11.7 show two ways of storing the smoking data, either of which would be acceptable for a multilevel analysis. Figure 11.6 shows a single data matrix, with one row for each person in the study. We could then pull out the smoking outcome $y=(y_{jt})$ in R, as follows:

R code
```
y <- data[,seq(6,16,2)]
female <- ifelse (data[,2]=="f", 1, 0)
mom.smoke <- ifelse (data[,3]=="Y", 1, 0)
dad.smoke <- ifelse (data[,4]=="Y", 1, 0)
psmoke <- mom.smoke + dad.smoke
```

and from there fit the model (11.3).

Figure 11.7 shows an alternative approach using two data matrices, one with a

[1] Alternatively, we could include a main effect for time and an interaction between time and sex, $\Pr(y_{jt}=1)=\text{logit}^{-1}(\beta_0+\beta_1\cdot\text{psmoke}_j+\beta_2\cdot\text{female}_j+\beta_3\cdot t+\beta_4\cdot\text{female}_j\cdot t+\alpha_j)$, so that the time trends for boys and girls are β_3 and $\beta_3+\beta_4$, respectively. This parameterization is appropriate to the extent that the comparison between the sexes is of interest; in this case we used (11.3) so that we could easily interpret β_3 and β_4 symmetrically.

age	smokes?	person ID	wave
15:0	N	1	1
14.7	N	2	1
15:1	N	3	1
15:3	N	4	1
⋮	⋮	⋮	⋮
15:6	N	1	2
15:1	N	2	2
15:7	Y	3	2
15:9	N	4	2
⋮	⋮	⋮	⋮

person ID	sex	parents smoke? mom	dad
1	f	Y	Y
2	f	N	N
3	m	Y	N
4	f	N	N
⋮	⋮	⋮	⋮

Figure 11.7 *Data from the smoking study, with observational data written as a single long matrix,* `obs.data`, *with person indicators, followed by a shorter matrix,* `person.data`, *of person-level information. Compare to Figure 11.6.*

row for each observation and one with a row for each person. To model these data, one could use R code such as

R code

```
y <- obs.data[,2]
person <- obs.data[,3]
wave <- obs.data[,4]
female <- ifelse (person.data[,2]=="f", 1, 0)
mom.smoke <- ifelse (person.data[,3]=="Y", 1, 0)
dad.smoke <- ifelse (person.data[,4]=="Y", 1, 0)
psmoke <- mom.smoke + dad.smoke
```

and then parameterize the model using the index i to represent individual observations, with $j[i]$ and $t[i]$ indicating the person and wave associated with observation i:

$$\begin{aligned}\Pr(y_i=1) &= \text{logit}^{-1}(\beta_0 + \beta_1 \text{psmoke}_{j[i]} + \beta_2 \text{female}_{j[i]} + \\ &\quad + \beta_3(1-\text{female}_{j[i]}) \cdot t[i] + \beta_4 \text{female}_{j[i]} \cdot t[i] + \alpha_{j[i]}). \qquad (11.4)\end{aligned}$$

Models (11.3) and (11.4) are equivalent, and both can be fit in Bugs (as we describe in Part 2B). Choosing between them is a matter of convenience. For data in a simple two-way structure (each adolescent is measured at six regular times), it can make sense to work with the double-indexed outcome variable, (y_{jt}). For a less rectangular data structure (for example, different adolescents measured at irregular intervals) it can be easier to string together a long data vector (y_i), with person and time recorded for each measurement, and with a separate matrix of person-level information (as in Figure 11.7).

Time-series cross-sectional data

In settings where overall time trends are important, repeated measurement data are sometimes called *time-series cross-sectional.* For example, Section 6.3 introduced a study of the proportion of death penalty verdicts that were overturned, in each of 34 states in the 23 years, 1973–1995. The data come at the state × year levels but we are also interested in studying variation among states and over time.

Time-series cross-sectional data are typically (although not necessarily) "rectangular" in structure, with observations at regular time intervals. In contrast, gen-

eral repeated measurements could easily have irregular patterns (for example, in the smoking study, some children could be measured only once, others could be measured monthly and others yearly). In addition, time-series cross-sectional data commonly have overall time patterns, for example, the steady expansion of the death penalty from the 1970s through the early 1990s. In this context one must consider the state-year data as clustered within states and also within years, with the potential for predictors at all three levels. We discuss such non-nested models in Section 13.5.

Other non-nested structures

Non-nested data also arise when individuals are characterized by overlapping categories of attributes. For example, consider a study of earnings given occupation and state of residence. A survey could include, say, 1500 persons in 40 job categories in 50 states, and a regression model could predict log earnings given individual demographic predictors X, 40 indicators for job categories, and 50 state indicators. We can write the model generalizing the notation of (11.1)–(11.2):

$$y_i = X_i\beta + \alpha_{j[i]} + \gamma_{k[i]} + \epsilon_i, \text{ for } i = 1, \ldots, n, \tag{11.5}$$

where $j[i]$ and $k[i]$ represent the job category and state, respectively, for person i. The model becomes multilevel with regressions for the job and state coefficients. For example,

$$\alpha_j \sim \mathrm{N}(U_j a, \sigma_\alpha^2), \text{ for } j = 1, \ldots, 40, \tag{11.6}$$

where U is a matrix of occupation-level predictors (for example, a measure of social status and an indicator for whether it is supervisory), a is a vector of coefficients for the job model, and σ_α is the standard deviation of the model errors at the level of job category. Similarly, for the state coefficients:

$$\gamma_k \sim \mathrm{N}(V_k g, \sigma_\gamma^2) \text{ for } k = 1, \ldots, 50. \tag{11.7}$$

The model defined by regressions (11.5)–(11.7) is non-nested because neither the job categories $j[i]$ nor the states $k[i]$ are subsets of the other.

As this example illustrates, regression notation can become awkward with multilevel models because of the need for new symbols (U, V, a, g, and so forth) to denote data matrices, coefficients, and errors at each level.

11.4 Indicator variables and fixed or random effects

Classical regression: including a baseline and $J - 1$ indicator variables

As discussed at the end of Section 4.5, when including an input variable with J categories into a classical regression, standard practice is to choose one of the categories as a baseline and include indicators for the other $J - 1$ categories. For example, if controlling for the $J = 20$ cities in the child support study in Figure 11.2 on page 238, one could set city 1 (Oakland) as the baseline and include indicators for the other 19. The coefficient for each city then represents its comparison to Oakland.

Multilevel regression: including all J indicators

In a multilevel model it is unnecessary to do this arbitrary step of picking one of the levels as a baseline. For example, in the child support study, one would include

indicators for all 20 cities as in model (11.1). In a classical regression these could not all be included because they would be collinear with the constant term, but in a multilevel model this is not a problem because they are themselves modeled by a group-level distribution (which itself can be a regression, as in (11.2)). We discuss on page 393 how the added information removes the collinearity that is present in the simple least squares estimate.

Fixed and random effects

The varying coefficients (α_j's or β_j's) in a multilevel model are sometimes called *random effects*, a term that refers to the randomness in the probability model for the group-level coefficients (as, for example, in (11.2) on page 241).

The term *fixed effects* is used in contrast to random effects—but not in a consistent way! Fixed effects are usually defined as varying coefficients that are not themselves modeled. For example, a classical regression including $J - 1 = 19$ city indicators as regression predictors is sometimes called a "fixed-effects model" or a model with "fixed effects for cities." Confusingly, however, "fixed-effects models" sometimes refer to regressions in which coefficients do *not* vary by group (so that they are fixed, not random).[2]

A question that commonly arises is when to use fixed effects (in the sense of varying coefficients that are unmodeled) and when to use random effects. The statistical literature is full of confusing and contradictory advice. Some say that fixed effects are appropriate if group-level coefficients are of interest, and random effects are appropriate if interest lies in the underlying population. Others recommend fixed

[2] Here we outline five definitions that we have seen of fixed and random effects:

1. Fixed effects are constant across individuals, and random effects vary. For example, in a growth study, a model with random intercepts α_i and fixed slope β corresponds to parallel lines for different individuals i, or the model $y_{it} = \alpha_i + \beta t$. Kreft and De Leeuw (1998, p. 12) thus distinguish between fixed and random coefficients.
2. Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population. Searle, Casella, and McCulloch (1992, section 1.4) explore this distinction in depth.
3. "When a sample exhausts the population, the corresponding variable is *fixed*; when the sample is a small (i.e., negligible) part of the population the corresponding variable is *random*" (Green and Tukey, 1960).
4. "If an effect is assumed to be a realized value of a random variable, it is called a random effect" (LaMotte, 1983).
5. Fixed effects are estimated using least squares (or, more generally, maximum likelihood) and random effects are estimated with shrinkage ("linear unbiased prediction" in the terminology of Robinson, 1991). This definition is standard in the multilevel modeling literature (see, for example, Snijders and Bosker, 1999, section 4.2) and in econometrics.

   In a multilevel model, this definition implies that fixed effects β_j are estimated conditional on a group-level variance $\sigma_\beta = \infty$ and random effects β_j are estimated conditional on σ_β estimated from data.

Of these definitions, the first clearly stands apart, but the other four definitions differ also. Under the second definition, an effect can change from fixed to random with a change in the goals of inference, even if the data and design are unchanged. The third definition differs from the others in defining a finite population (while leaving open the question of what to do with a large but not exhaustive sample), while the fourth definition makes no reference to an actual (rather than mathematical) population at all. The second definition allows fixed effects to come from a distribution, as long as that distribution is not of interest, whereas the fourth and fifth do not use any distribution for inference about fixed effects. The fifth definition has the virtue of mathematical precision but leaves unclear when a given set of effects should be considered fixed or random. In summary, it is easily possible for a factor to be "fixed" according to some definitions above and "random" for others. Because of these conflicting definitions, it is no surprise that "clear answers to the question 'fixed or random?' are not necessarily the norm" (Searle, Casella, and McCulloch, 1992, p. 15).

effects when the groups in the data represent all possible groups, and random effects when the population includes groups not in the data. These two recommendations (and others) can be unhelpful. For example, in the child support example, we are interested in these particular cities and also the country as a whole. The cities are only a sample of cities in the United States—but if we were suddenly given data from all the other cities, we would not want then to change our model.

Our advice (elaborated upon in the rest of this book) is to *always* use multilevel modeling ("random effects"). Because of the conflicting definitions and advice, we avoid the terms "fixed" and "random" entirely, and focus on the description of the model itself (for example, varying intercepts and constant slopes), with the understanding that batches of coefficients (for example, $\alpha_1, \ldots, \alpha_J$) will themselves be modeled.

11.5 Costs and benefits of multilevel modeling

Quick overview of classical regression

Before we go to the effort of learning multilevel modeling, it is helpful to briefly review what can be done with classical regression:

- Prediction for continuous or discrete outcomes,
- Fitting of nonlinear relations using transformations,
- Inclusion of categorical predictors using indicator variables,
- Modeling of interactions between inputs,
- Causal inference (under appropriate conditions).

Motivations for multilevel modeling

There are various reasons why it might be worth moving to a multilevel model, whether for purposes of causal inference, the study of variation, or prediction of future outcomes:

- Accounting for individual- and group-level variation in estimating *group-level* regression coefficients. For example, in the child support study in Section 11.2, interest lies in a city-level predictor (child support enforcement), and in classical regression it is not possible to include city indicators along with city-level predictors.
- Modeling variation among *individual-level* regression coefficients. In classical regression, one can do this using indicator variables, but multilevel modeling is convenient when we want to model the variation of these coefficients across groups, make predictions for new groups, or account for group-level variation in the uncertainty for individual-level coefficients.
- Estimating regression coefficients for *particular* groups. For example, in the next chapter, we discuss the problem of estimating radon levels from measurements in several counties in Minnesota. With a multilevel model, we can get reasonable estimates even for counties with small sample sizes, which would be difficult using classical regression.

One or more of these reasons might apply in any particular study.

Complexity of multilevel models

A potential drawback to multilevel modeling is the additional complexity of coefficients varying by group. We do not mind this complexity—in fact, we embrace it

in its realism—however, it does create new difficulties in understanding and summarizing the model, issues we explore in Part 3 of this book.

Additional modeling assumptions

As we discuss in the next few chapters, a multilevel model requires additional assumptions beyond those of classical regression—basically, each level of the model corresponds to its own regression with its own set of assumptions such as additivity, linearity, independence, equal variance, and normality.

We usually don't mind. First, it can be possible to check these assumptions. Perhaps more important, classical regressions can typically be identified with particular special cases of multilevel models with hierarchical variance parameters set to zero or infinity—these are the *complete pooling* and *no pooling* models discussed in Sections 12.2 and 12.3. Our ultimate justification, which can be seen through examples, is that the assumptions pay off in practice in allowing more realistic models and inferences.

When does multilevel modeling make a difference?

The usual alternative to multilevel modeling is classical regression—either ignoring group-level variation, or with varying coefficients that are estimated classically (and not themselves modeled)—or combinations of classical regressions such as the individual and group-level models described on page 239.

In various limiting cases, the classical and multilevel approaches coincide. When there is very little group-level variation, the multilevel model reduces to classical regression with no group indicators; conversely, when group-level coefficients vary greatly (compared to their standard errors of estimation), multilevel modeling reduces to classical regression with group indicators.

When the number of groups is small (less than five, say), there is typically not enough information to accurately estimate group-level variation. As a result, multilevel models in this setting typically gain little beyond classical varying-coefficient models.

These limits give us a sense of where we can gain the most from multilevel modeling—where it is worth the effort of expanding a classical regression in this way. However, there is little risk from applying a multilevel model, assuming we are willing to put in the effort to set up the model and interpret the resulting inferences.

11.6 Bibliographic note

Several introductory books on multilevel models have been written in the past decade in conjunction with specialized computer programs (see Section 1.5), including Raudenbush and Bryk (2002), Goldstein (1995), and Snijders and Bosker (1999). Kreft and De Leeuw (1998) provide an accessible introduction and a good place to start (although we do not agree with all of their recommendations). These books have a social science focus, perhaps because it is harder to justify the use of linear models in laboratory sciences where it is easier to isolate the effects of individual factors and so the functional form of responses is better understood. Giltinan and Davidian (1995) and Verbeke and Molenberghs (2000) are books on nonlinear multilevel models focusing on biostatistical applications.

Another approach to regression with multilevel data structures is to use classical estimates and then correct the standard errors to deal with the dependence in the

data. We briefly discuss the connection between multilevel models and correlated-error models in Section 12.5 but do not consider these other inferential methods, which include *generalized estimating equations* (see Carlin et al., 2001, for a comparison to multilevel models) and *panel-corrected standard errors* (see Beck and Katz, 1995, 1996).

The articles in the special issue of *Political Analysis* devoted to multilevel modeling (Kedar and Shively, 2005) illustrate several different forms of analysis of multilevel data, including two-level classical regression and multilevel modeling.

Gelman (2005) discusses difficulties with the terms "fixed" and "random" effects. See also Kreft and De Leeuw (1998, section 1.3.3), for a discussion of the multiplicity of definitions of fixed and random effects and coefficients, and Robinson (1998) for a historical overview.

The child support example comes from Nepomnyaschy and Garfinkel (2005) The teenage smoking example comes from Carlin et al. (2001), who consider several different models, including a multilevel logistic regression.

11.7 Exercises

1. The file `apt.dat` in the folder `rodents` contains data on rodent infestation in a sample of New York City apartments (see codebook `rodents.doc`). The file `dist.dat` contains data on the 55 "community districts" (neighborhoods) in the city.

   (a) Write the notation for a varying-intercept multilevel logistic regression (with community districts as the groups) for the probability of rodent infestation using the individual-level predictors but no group-level predictors.

   (b) Expand the model in (a) by including the variables in `dist.dat` as group-level predictors.

2. Time-series cross-sectional data: download data with an outcome y and predictors X in each of J countries for a series of K consecutive years. The outcome should be some measure of educational achievement of children and the predictors should be a per capita income measure, a measure of income inequality, and a variable summarizing how democratic the country is. For these countries, also create country-level predictors that are indicators for the countries' geographic regions.

   (a) Set up the data as a wide matrix of countries $\times$ measurements (as in Figure 11.6).

   (b) Set up the data as two matrices as in Figure 11.7: a long matrix with JK rows with all the measurements, and a matrix with J rows, with information on each country.

   (c) Write a multilevel regression as in (11.5)–(11.7). Explain the meaning of all the variables in the model.

3. The folder `olympics` has seven judges' ratings of seven figure skaters (on two criteria: "technical merit" and "artistic impression") from the 1932 Winter Olympics.

   (a) Construct a $7 \times 7 \times 2$ array of the data (ordered by skater, judge, and judging criterion).

   (b) Reformulate the data as a 98×4 array (similar to the top table in Figure 11.7), where the first two columns are the technical merit and artistic impression scores, the third column is a skater ID, and the fourth column is a judge ID.

(c) Add another column to this matrix representing an indicator variable that equals 1 if the skater and judge are from the same country, or 0 otherwise.

4. The folder `cd4` has CD4 percentages for a set of young children with HIV who were measured several times over a period of two years. The dataset also includes the ages of the children at each measurement.

(a) Graph the outcome (the CD4 percentage, on the square root scale) for each child as a function of time.

(b) Each child's data has a time course that can be summarized by a linear fit. Estimate these lines and plot them for all the children.

(c) Set up a model for the children's slopes and intercepts as a function of the treatment and age at baseline. Estimate this model using the two-step procedure–first estimate the intercept and slope separately for each child, then fit the between-child models using the point estimates from the first step.