CHAPTER 13

Multilevel linear models: varying slopes, non-nested models, and other complexities

This chapter considers some generalizations of the basic multilevel regression. Models in which slopes and intercepts can vary by group (for example, $y_i = \alpha_{j[i]} + \beta_{j[i]}x_i + \cdots$, where α and β both vary by group j; see Figure 11.1c on page 238) can also be interpreted as interactions of the group index with individual-level predictors.

Another direction is non-nested models, in which a given dataset can be structured into groups in more than one way. For example, persons in a national survey can be divided by demographics or by states. Responses in a psychological experiment might be classified by person (experimental subject), experimental condition, and time.

The chapter concludes with some examples of models with nonexchangeable multivariate structures. We continue with generalized linear models in Chapters 14–15 and discuss how to fit all these models in Chapters 16–19.

13.1 Varying intercepts and slopes

The next step in multilevel modeling is to allow more than one regression coefficient to vary by group. We shall illustrate with the radon model from the previous chapter, which is relatively simple because it only has a single individual-level predictor, x (the indicator for whether the measurement was taken on the first floor).

We begin with a varying-intercept, varying-slope model including x but without the county-level uranium predictor; thus,

$$\begin{aligned} y_i &\sim \text{N}(\alpha_{j[i]} + \beta_{j[i]}x_i,\, \sigma_y^2), \text{ for } i = 1, \ldots, n \\ \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} &\sim \text{N}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right), \text{ for } j = 1, \ldots, J, \end{aligned} \tag{13.1}$$

with variation in the α_j's and the β_j's and also a between-group correlation parameter ρ. In R:

R code
```
M3 <- lmer (y ~ x + (1 + x | county))
display (M3)
```

which yields

R output
```
lmer(formula = y ~ x + (1 + x | county))
            coef.est coef.se
(Intercept)  1.46     0.05
x           -0.68     0.09
Error terms:
 Groups   Name        Std.Dev. Corr
 county   (Intercept) 0.35
          x           0.34     -0.34
```

```
 Residual             0.75
# of obs: 919, groups: county, 85
deviance = 2161.1
```

In this model, the unexplained within-county variation has an estimated standard deviation of $\hat{\sigma}_y = 0.75$; the estimated standard deviation of the county intercepts is $\hat{\sigma}_\alpha = 0.35$; the estimated standard deviation of the county slopes is $\hat{\sigma}_\beta = 0.34$; and the estimated correlation between intercepts and slopes is -0.34.

We then can type

R code
```
coef (M3)
```

to yield

R output
```
$county
   (Intercept)     x
1         1.14 -0.54
2         0.93 -0.77
3         1.47 -0.67
. . .
85        1.38 -0.65
```

Or we can separately look at the estimated population mean coefficients μ_α, μ_β and then the estimated errors for each county. First, we type

R code
```
fixef (M3)
```

to see the estimated average coefficients ("fixed effects"):

R output
```
(Intercept)           x
       1.46       -0.68
```

Then, we type

R code
```
ranef (M3)
```

to see the estimated group-level errors ("random effects"):

R output
```
   (Intercept)     x
1        -0.32  0.14
2        -0.53 -0.09
3         0.01  0.01
. . .
85       -0.08  0.03
```

We can regain the estimated intercept and slope α_j, β_j for each county by simply adding the errors to μ_α and μ_β; thus, the estimated regression line for county 1 is $(1.46 - 0.32) + (-0.68 + 0.14)x = 1.14 - 0.54x$, and so forth.

The group-level model for the parameters (α_j, β_j) allows for partial pooling in the estimated intercepts and slopes. Figure 13.1 shows the results—the estimated lines $y = \alpha_j + \beta_j x$—for the radon data in eight different counties.

Including group-level predictors

We can expand the model of (α, β) in (13.1) by including a group-level predictor (in this case, soil uranium):

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \gamma_0^\alpha + \gamma_1^\alpha u_j \\ \gamma_0^\beta + \gamma_1^\beta u_j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right), \text{ for } j = 1, \ldots, J. \qquad (13.2)$$

The resulting estimates for the α_j's and β_j's are changed slightly from what is displayed in Figure 13.1, but more interesting are the second-level models themselves, whose estimates are shown in Figure 13.2. Here is the result of fitting the model in R:

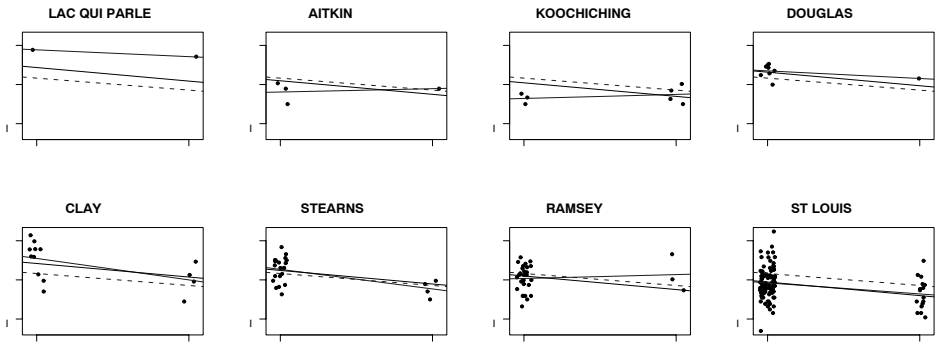

Figure 13.1 *Multilevel (partial pooling) regression lines $y = \alpha_j + \beta_j x$, displayed for eight counties j. In this model, both the intercept and the slope vary by county. The light solid and dashed lines show the no-pooling and complete pooling regression lines. Compare to Figure 12.4, in which only the intercept varies.*

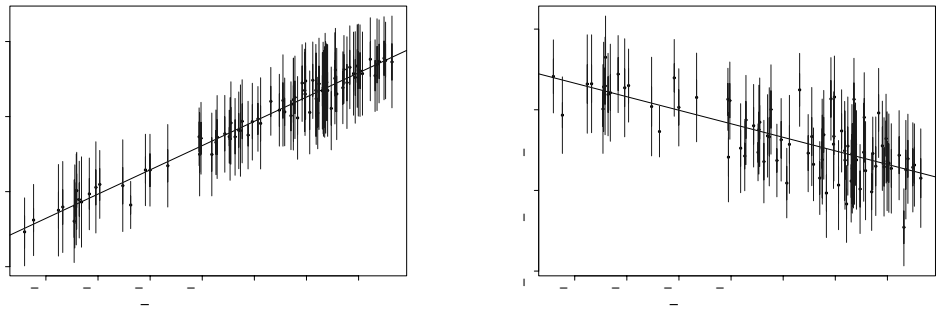

Figure 13.2 *(a) Estimates $\pm$ standard errors for the county intercepts α_j, plotted versus county-level uranium measurement u_j, along with the estimated multilevel regression line, $\alpha = \gamma_0^\alpha + \gamma_1^\alpha u$. (b) Estimates $\pm$ standard errors for the county slopes β_j, plotted versus county-level uranium measurement u_j, along with the estimated multilevel regression line, $\beta = \gamma_0^\beta + \gamma_1^\beta u$. Estimates and standard errors are the posterior medians and standard deviations, respectively. For each graph, the county coefficients roughly follow the line but not exactly; the discrepancies of the coefficients from the line are summarized by the county-level standard-deviation parameters $\sigma_\alpha, \sigma_\beta$.*

R output

```
lmer(formula = y ~ x + u.full + x:u.full + (1 + x | county))
            coef.est coef.se
(Intercept)  1.47     0.04
x           -0.67     0.08
u.full       0.81     0.09
x:u.full    -0.42     0.23
Error terms:
 Groups   Name        Std.Dev. Corr
 county   (Intercept) 0.12
          x           0.31     0.41
 Residual             0.75
# of obs: 919, groups: county, 85
deviance = 2114.3
```

The parameters $\gamma_0^\alpha, \gamma_0^\beta, \gamma_1^\alpha, \gamma_1^\beta$ in model (13.2) are the coefficients for the intercept,

`x`, `u.full`, and `x:u.full`, respectively, in the regression. In particular, the interaction corresponds to allowing uranium to be a predictor in the regression for the slopes.

The estimated coefficients in each group (from `coef(M4)`) are:

R output
```
$county
   (Intercept)     x u.full x:u.full
1         1.46 -0.65   0.81    -0.42
2         1.50 -0.89   0.81    -0.42
. . .
85        1.44 -0.70   0.81    -0.42
```

Or we can display the average coefficients (using `fixef(M4)`):

R output
```
(Intercept)           x      u.full    x:u.full
       1.47       -0.67        0.81       -0.42
```

and the group-level errors for the intercepts and slopes (using `ranef(M4)`):

R output
```
   (Intercept)     x
1        -0.01  0.02
2         0.03 -0.21
. . .
85       -0.02 -0.03
```

The coefficients for the intercept and `x` vary, as specified in the model. This can be compared to the model on page 267 in which only the intercept varies.

Going from lmer output to intercepts and slopes

As before, we can combine the average coefficients with the group-level errors to compute the intercepts α_j and slopes β_j of model (13.2). For example, the fitted regression model in county 85 is $y_i = 1.47 - 0.67x_i + 0.81u_{85} - 0.42x_iu_{85} - 0.02 - 0.03x_i$. The log uranium level in county 85, u_{85}, is 0.36, and so the fitted regression line in county 85 is $y_i = 1.73 - 0.85x_i$. More generally, we can compute a vector of county intercepts α and slopes β:

R code
```
a.hat.M4 <- coef(M4)[,1] + coef(M4)[,3]*u
b.hat.M4 <- coef(M4)[,2] + coef(M4)[,4]*u
```

Here it is actually useful to have the variable `u` defined at the county level (as compared to `u.full = u[county]` which was used in the `lmer()` call). We next consider these linear transformations algebraically.

Varying slopes as interactions

Section 12.5 gave multiple ways of writing the basic multilevel model. These same ideas apply to models with varying slopes, which can be considered as interactions between group indicators and an individual-level predictor. For example, consider the model with an individual-level predictor x_i and a group-level predictor u_j,

$$\begin{aligned} y_i &= \alpha_{j[i]} + \beta_{j[i]}x_i + \epsilon_i \\ \alpha_j &= \gamma_0^\alpha + \gamma_1^\alpha u_j + \eta_j^\alpha \\ \beta_j &= \gamma_0^\beta + \gamma_1^\beta u_j + \eta_j^\beta. \end{aligned}$$

We can re-express this as a single model by substituting the formulas for α_j and β_j into the equation for y_i:

$$y_i = \left[\gamma_0^\alpha + \gamma_1^\alpha u_{j[i]} + \eta_{j[i]}^\alpha\right] + \left[\gamma_0^\beta + \gamma_1^\beta u_{j[i]} + \eta_{j[i]}^\beta\right] x_i + \epsilon_i. \qquad (13.3)$$

This expression looks messy but it is really just a regression including various interactions. If we define a new individual-level predictor $v_i = u_{j[i]}$ (in the radon example, this is the uranium level in the county where your house is located), we can re-express (13.3) term by term as

$$y_i = a + bv_i + c_{j[i]} + dx_i + ev_ix_i + f_{j[i]}x_i + \epsilon_i.$$

This can be thought of in several ways:

- A varying-intercept, varying-slope model with four individual-level predictors (the constant term, v_i, x_i, and the interaction v_ix_i) and varying intercepts and slopes that are centered at zero.
- A regression model with $4 + 2J$ predictors: the constant term, v_i, x_i, v_ix_i, indicators for the J groups, and interactions between x and the J group indicators.
- A regression model with four predictors and three error terms.
- Or, to go back to the original formulation, a varying-intercept, varying-slope model with one group-level predictor.

Which of these expressions is most useful depends on the context. In the radon analysis, where the goal is to predict radon levels in individual counties, the varying-intercept, varying-slope formulation, as pictured in Figure 13.2, seems most appropriate. But in a problem where interest lies in the regression coefficients for x_i, u_j, and their interaction, it can be more helpful to focus on these predictors and consider the unexplained variation in intercepts and slopes merely as error terms.

13.2 Varying slopes without varying intercepts

Figure 11.1 on page 238 displays a varying-intercept model, a varying-slope model, and a varying-intercept, varying-slope model. Almost always, when a slope is allowed to vary, it makes sense for the intercept to vary also. That is, the graph in the center of Figure 11.1b usually does not make sense. For example, if the coefficient of floor varies with county, then it makes sense to allow the intercept of the regression to vary also. It would be an implausible scenario in which the counties were all identical in radon levels for houses without basements, but differed in their coefficients for x.

A situation in which a constant-intercept, varying-slope model is appropriate

Occasionally it is reasonable to allow the slope but not the intercept to vary by group. For example, consider a study in which J separate experiments are performed on samples from a common population, with each experiment randomly assigning a control condition to half its subjects and a treatment to the other half. Further suppose that the "control" conditions are the same for each experiment but the "treatments" vary. In that case, it would make sense to fix the intercept and allow the slope to vary—thus, a basic model of:

$$\begin{aligned} y_i &\sim \text{N}(\alpha + \theta_{j[i]}T_i, \sigma_y^2) \\ \theta_j &\sim \text{N}(\mu_\theta, \sigma_\theta^2), \end{aligned} \tag{13.4}$$

where $T_i = 1$ for treated units and 0 for controls. Individual-level predictors could be added to the regression for y, and any interactions with treatment could also

have varying slopes; for example,

$$\begin{aligned} y_i &\sim \text{N}\left(\alpha + \beta x_i + \theta_{1,j[i]} T_i + \beta_{2,j[i]} x_i T_i,\ \sigma_y^2\right) \\ \begin{pmatrix} \theta_{1,j} \\ \theta_{2,j} \end{pmatrix} &\sim \text{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right), \text{ for } j = 1, \ldots, J, \end{aligned} \tag{13.5}$$

The multilevel model could be further extended with group-level predictors characterizing the treatments.

Fitting in R

To fit such a model in `lmer()`, we must explicitly remove the intercept from the group of coefficients that vary by group; for example, here is model (13.4) including the treatment indicator T as a predictor:

R code
```
lmer (y ~ T + (T - 1 | group))
```

The varying slope allows a different treatment effect for each group.

And here is model (13.5) with an individual-level predictor `x`:

R code
```
lmer (y ~ x + T + (T + x:T - 1 | group))
```

Here, the treatment effect and its interaction with x vary by group.

13.3 Modeling multiple varying coefficients using the scaled inverse-Wishart distribution

When more than two coefficients vary (for example, $y_i \sim \text{N}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}, \sigma^2)$, with β_0, β_1, and β_2 varying by group), it is helpful to move to matrix notation in modeling the coefficients and their group-level regression model and covariance matrix.

Simple model with two varying coefficients and no group-level predictors

Starting with the model that begins this chapter, we can rewrite the basic varying-intercept, varying-slope model (13.1) in matrix notation as

$$\begin{aligned} y_i &\sim \text{N}(X_i B_{j[i]},\ \sigma_y^2), \text{ for } i = 1, \ldots, n \\ B_j &\sim \text{N}(M_B, \Sigma_B), \text{ for } j = 1, \ldots, J, \end{aligned} \tag{13.6}$$

where

- X is the $n \times 2$ matrix of predictors: the first column of X is a column of 1's (that is, the constant term in the regression), and the second column is the predictor x. X_i is then the vector of length 2 representing the i^{th} row of X, and $X_i B_{j[i]}$ is simply $\alpha_{j[i]} + \beta_{j[i]} x_i$ from the top line of (13.1).
- $B = (\alpha, \beta)$ is the $J \times 2$ matrix of individual-level regression coefficients. For any group j, B_j is a vector of length 2 corresponding to the j^{th} row of B (although for convenience we consider B_j as a column vector in the product $X_i B_{j[i]}$ in model (13.6)). The two elements of B_j correspond to the intercept and slope, respectively, for the regression model in group j. $B_{j[i]}$ in the first line of (13.6) is the $j[i]^{th}$ row of B, that is, the vector representing the intercept and slope for the group that includes unit i.
- $M_B = (\mu_\alpha, \mu_\beta)$ is a vector of length 2, representing the mean of the distribution of the intercepts and the mean of the distribution of the slopes.

- Σ_B is the 2×2 covariance matrix representing the variation of the intercepts and slopes in the population of groups, as in the second line of (13.1).

We are following our general notation in which uppercase letters represent matrices: thus, the vectors α and β are combined into the matrix B.

In the fitted radon model on page 279, the parameters of the group-level model are estimated at $\widehat{M}_B = (1.46, -0.68)$ and $\widehat{\Sigma}_B = \begin{pmatrix} \hat{\sigma}_a^2 & \hat{\rho}\hat{\sigma}_a\hat{\sigma}_b \\ \hat{\rho}\hat{\sigma}_a\hat{\sigma}_b & \hat{\sigma}_b^2 \end{pmatrix}$, where $\hat{\sigma_a} = 0.35$, $\hat{\sigma}_b = 0.34$, and $\hat{\rho} = -0.34$. The estimated coefficient matrix $\widehat{B}$ is given by the 85×2 array at the end of the display of `coef(M3)` on page 280.

More than two varying coefficients

The same expression as above holds, except that the 2's are replaced by K's, where K is the number of individual-level predictors (including the intercept) that vary by group. As we discuss shortly in the context of the inverse-Wishart model, estimation becomes more difficult when $K > 2$ because of constraints among the correlation parameters of the covariance matrix Σ_B.

Including group-level predictors

More generally, we can have J groups, K individual-level predictors, and L predictors in the group-level regression (including the constant term as a predictor in both cases). For example, $K = L = 2$ in the radon model that has floor as an individual predictor and uranium as a county-level predictor.

We can extend model (13.6) to include group-level predictors:

$$\begin{aligned} y_i &\sim \text{N}(X_i B_{j[i]},\, \sigma_y^2), \text{ for } i = 1, \ldots, n \\ B_j &\sim \text{N}(U_j G, \Sigma_B), \text{ for } j = 1, \ldots, J, \end{aligned} \tag{13.7}$$

where B is the $J \times K$ matrix of individual-level coefficients, U is the $J \times L$ matrix of group-level predictors (including the constant term), and G is the $L \times K$ matrix of coefficients for the group-level regression. U_j is the j^{th} row of U, the vector of predictors for group j, and so $U_j G$ is a vector of length K.

Model (13.1) is a special case with $K = L = 2$, and the coefficients in G are then $\gamma_0^\alpha, \gamma_0^\beta, \gamma_1^\alpha, \gamma_1^\beta$. For the fitted radon model on page 279, the γ's are the four unmodeled coefficients (for the intercept, `x`, `u.full`, and `x:u.full`, respectively), and the two columns of the estimated coefficient matrix $\widehat{B}$ are estimated by `a.hat` and `b.hat`, as defined by the R code on page 282.

Including individual-level predictors whose coefficients do not vary by group

The model can be further expanded by adding unmodeled individual-level coefficients, so that the top line of (13.7) becomes

$$y_i \sim \text{N}(X_i^0 \beta^0 + X_i B_{j[i]},\, \sigma_y^2), \text{ for } i = 1, \ldots, n, \tag{13.8}$$

where X^0 is a matrix of these additional predictors and β^0 is the vector of their regression coefficients (which, by assumption, are common to all the groups).

Model (13.8) is sometimes called a *mixed-effects* regression, where the β^0's and the B's are the *fixed* and *random* effects, respectively. As noted on pages 2 and 245, we avoid these terms because of their ambiguity in the statistical literature. For example, sometimes unvarying coefficients such as the β^0's in model (13.8) are called "fixed," but sometimes the term "fixed effects" refers to intercepts that vary

by groups but are not given a multilevel model (this is what we call the "no-pooling model," as pictured, for example, by the solid lines in Figure 12.2 on page 255).

Equivalently, model (13.8) can be written by folding X^0 and X into a common predictor matrix X, folding β^0 and B into a common coefficient matrix B, and using model (13.1), with the appropriate elements in Σ_B set to zero, implying no variation among groups for certain coefficients.

Modeling the group-level covariance matrix using the scaled inverse-Wishart distribution

When the number K of varying coefficients per group is more than two, modeling the correlation parameters ρ is a challenge. In addition to each of the correlations being restricted to fall between -1 and 1, the correlations are jointly constrained in a complicated way—technically, the covariance matrix Σ_β must be positive definite. (An example of the constraint is: if $\rho_{12} = 0.9$ and $\rho_{13} = 0.9$, then ρ_{23} must be at least 0.62.)

Modeling and estimation are more complicated in this jointly constrained space. We first introduce the inverse-Wishart model, then generalize to the scaled inverse-Wishart, which is what we recommend for modeling the covariance matrix of the distribution of varying coefficients.

Inverse-Wishart model. One model that has been proposed for the covariance matrix Σ_β is the *inverse-Wishart* distribution, which has the advantage of being computationally convenient (especially when using Bugs, as we illustrate in Section 17.1) but the disadvantage of being difficult to interpret.

In the model $\Sigma_B \sim \text{Inv-Wishart}_{K+1}(I)$, the two parameters of the inverse-Wishart distribution are the *degrees of freedom* (here set to $K+1$, where K is the dimension of B, that is, the number of coefficients in the model that vary by group) and the *scale* (here set to the $K \times K$ identity matrix).

To understand this model, we consider its implications for the standard deviation and correlations. Recall that if there are K varying coefficients, then Σ_B is a $K \times K$ matrix, with diagonal elements $\Sigma_{kk} = \sigma_k^2$ and off-diagonal-elements $\Sigma_{kl} = \rho_{kl}\sigma_k\sigma_l$ (generalizing models (13.1) and (13.2) to $K > 2$).

Setting the degrees-of-freedom parameter to $K+1$ has the effect of setting a uniform distribution on the individual correlation parameters (that is, they are assumed equally likely to take on any value between -1 and 1).

Scaled inverse-Wishart model. When the degrees of freedom parameter of the inverse-Wishart distribution is set to K+1, the resulting model is reasonable for the correlations but is quite constraining on the scale parameters σ_k. This is a problem because we would like to estimate σ_k from the data. Changing the degrees of freedom allows the σ_k's to be estimated more freely, but at the cost of constraining the correlation parameters.

We get around this problem by expanding the inverse-Wishart model with a new vector of scale parameters ξ_k:

$$\Sigma_B = \text{Diag}(\xi)Q\text{Diag}(\xi),$$

with the *unscaled covariance matrix* Q being given the inverse-Wishart model:

$$Q \sim \text{Inv-Wishart}_{K+1}(I).$$

The variances then correspond to the diagonal elements of the unscaled covariance

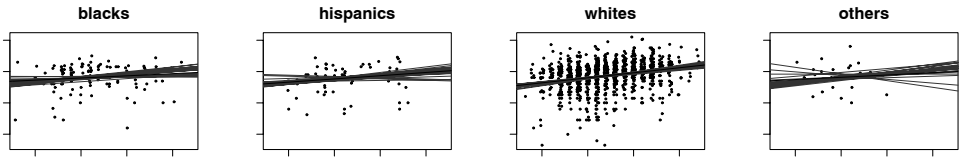

Figure 13.3 *Multilevel regression lines $y = \alpha_j + \beta_j x$ for log earnings on height (among those with positive earnings), in four ethnic categories j. The gray lines indicate uncertainty in the fitted regressions.*

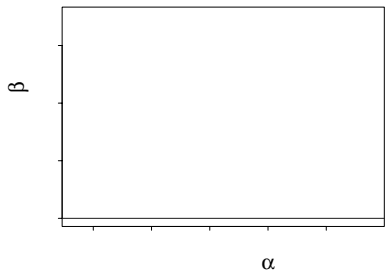

Figure 13.4 *Scatterplot of estimated intercepts and slopes (for whites, hispanics, blacks, and others), (α_j, β_j), for the earnings-height regressions shown in Figure 13.3. The extreme negative correlation arises because the center of the range of height is far from zero. Compare to the coefficients in the rescaled model, as displayed in Figure 13.7.*

matrix Q, multiplied by the appropriate scaling factors ξ:

$$\sigma_k^2 = \Sigma_{kk} = \xi_k^2 Q_{kk}, \text{ for } k = 1, \ldots, K,$$

and the covariances are

$$\Sigma_{kl} = \xi_k \xi_l Q_{kl}, \text{ for } k, l = 1, \ldots, K,$$

We prefer to express in terms of the standard deviations,

$$\sigma_k = |\xi_k| \sqrt{Q_{kk}},$$

and correlations

$$\rho_{kl} = \Sigma_{kl} / (\sigma_k \sigma_l).$$

The parameters in ξ and Q cannot be interpreted separately: they are a convenient way to set up the model, but it is the standard deviations σ_k and the correlations ρ_{kl} that are of interest (and which are relevant for producing partially pooled estimates for the coefficients in B).

As with the unscaled Wishart, the model implies a uniform distribution on the correlation parameters. As we discuss next, it can make sense to transform the data to remove any large correlations that could be expected simply from the structure of the data.

13.4 Understanding correlations between group-level intercepts and slopes

Recall that varying slopes can be interpreted as interactions between an individual-level predictor and group indicators. As with classical regression models with interactions, the intercepts can often be more clearly interpreted if the continuous

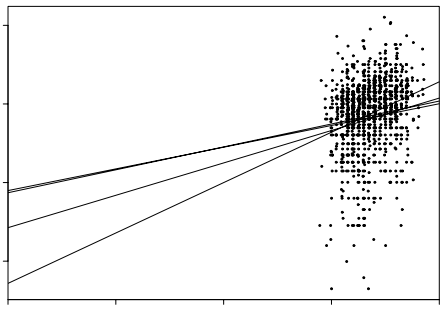

Figure 13.5 *Sketch illustrating the difficulty of simultaneously estimating α and β. The lines show the regressions for the four ethnic groups as displayed in Figure 13.3: the center of the range of x values is far from zero, and so small changes in the slope induce large changes in the intercept.*

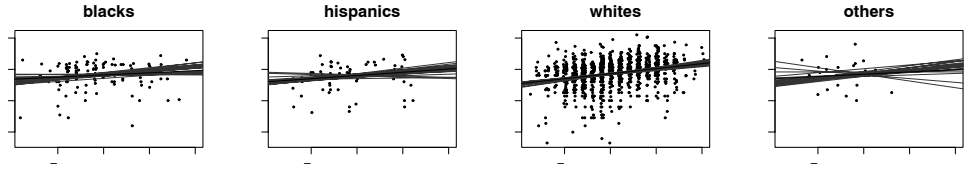

Figure 13.6 *Multilevel regression lines $y = \alpha_j + \beta_j z$, for log earnings given mean-adjusted height ($z_i = x_i - \bar{x}$), in four ethnic groups j. The gray lines indicate uncertainty in the fitted regressions.*

predictor is appropriately centered. We illustrate with the height and earnings example from Chapter 4.

We begin by fitting a multilevel model of log earnings given height, allowing the coefficients to vary by ethnicity. The data and fitted model are displayed in Figure 13.3. (Little is gained by fitting a multilevel model here—with only four groups, a classical no-pooling model would work nearly as well, as discussed in Section 12.9—but this is a convenient example to illustrate a general point.)

Figure 13.4 displays the estimates of (α_j, β_j) for the four ethnic groups, and they have a strong negative correlation: the groups with high values of α have relatively low values of β, and vice versa. This correlation occurs because the center of the x-values of the data is far from zero. The regression lines have to go roughly through the center of the data, and then changes in the slope induce opposite changes in the intercept, as illustrated in Figure 13.5.

There is nothing wrong with a high correlation between the α's and β's, but it makes the estimated intercepts more difficult to interpret. As with interaction models in classical regression, it can be helpful to subtract the average value of the continuous x before including it in the regression; thus, $y_i \sim \mathrm{N}(\alpha_{j[i]} + \beta_{j[i]} z_i,\, \sigma_y^2)$, where $z_i = x_i - \bar{x}$. Figures 13.6 and 13.7 show the results for the earnings regression: the correlation has pretty much disappeared. Centering the predictor x will not necessarily remove correlations between intercepts and slopes—but any correlation that remains can then be more easily interpreted. In addition, centering can speed convergence of the Gibbs sampling algorithm used by Bugs and other software.

We fit this model, and the subsequent models in this chapter, in Bugs (see Chap-

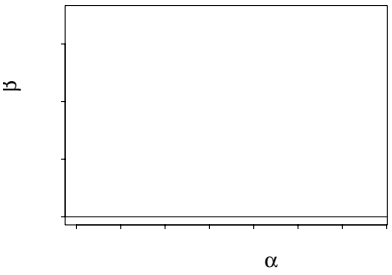

Figure 13.7 *Scatterplot of estimated intercepts and slopes, (α_j, β_j), for the regression of earnings on mean-adjusted height z, for the four groups j displayed in Figure 13.6. The coefficients are no longer strongly correlated (compare to Figure 13.4).*

ter 17 for examples of code) because, as discussed in Section 12.4, the current version of `lmer()` does not work so well when the number of groups is small—and, conversely, with these small datasets, Bugs is not too slow.

13.5 Non-nested models

So far we have considered the simplest hierarchical structure of individuals i in groups j. We now discuss models for more complicated grouping structures such as introduced in Section 11.3.

Example: a psychological experiment with two potentially interacting factors

Figure 13.8 displays data from a psychological experiment of pilots on flight simulators, with $n = 40$ data points corresponding to $J = 5$ treatment conditions and $K = 8$ different airports. The responses can be fit to a *non-nested* multilevel model of the form

$$\begin{aligned} y_i &\sim \text{N}(\mu + \gamma_{j[i]} + \delta_{k[i]}, \sigma_y^2), \text{ for } i = 1, \ldots, n \\ \gamma_j &\sim \text{N}(0, \sigma_\gamma^2), \text{ for } j = 1, \ldots, J \\ \delta_k &\sim \text{N}(0, \sigma_\delta^2), \text{ for } k = 1, \ldots, K. \end{aligned} \tag{13.9}$$

The parameters γ_j and δ_k represent treatment effects and airport effects. Their distributions are centered at zero (rather than given mean levels μ_γ, μ_δ) because the regression model for y already has an intercept, μ, and any nonzero mean for the γ and δ distributions could be folded into μ. As we shall see in Section 19.4, it can sometimes be effective for computational purposes to add extra mean-level parameters into the model, but the coefficients in this expanded model must be interpreted with care.

We can perform a quick fit as follows:

```
lmer (y ~ 1 + (1 | group.id) + (1 | scenario.id))
```
R code

where `group.id` and `scenario.id` are the index variables for the five treatment conditions and eight airports, respectively.

When fit to the data in Figure 13.8, the estimated residual standard deviations at the individual, treatment, and airport levels are $\hat{\sigma}_y = 0.23$, $\hat{\sigma}_\gamma = 0.04$, and $\hat{\sigma}_\delta = 0.32$. Thus, the variation among airports is huge—even larger than that among individual measurements—but the treatments vary almost not at all. This general pattern can be seen in Figure 13.8.

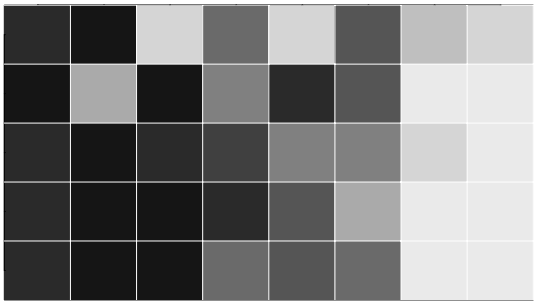

Figure 13.8 *Success rates of pilots training on a flight simulator with five different treatments and eight different airports. Shadings in the 40 cells i represent different success rates y_i, with black and white corresponding to 0 and 100%, respectively. For convenience in reading the display, the treatments and airports have each been sorted in increasing order of average success. These 40 data points have two groupings—treatments and airports—which are not nested.*

Data in matrix form

airport	treatment conditions				
1	0.38	0.25	0.50	0.14	0.43
2	0.00	0.00	0.67	0.00	0.00
3	0.38	0.50	0.33	0.71	0.29
4	0.00	0.12	0.00	0.00	0.86
5	0.33	0.50	0.14	0.29	0.86
6	1.00	1.00	1.00	1.00	0.86
7	0.12	0.12	0.00	0.14	0.14
8	1.00	0.86	1.00	1.00	0.75

Data in vector form

y	j	k
0.38	1	1
0.00	1	2
0.38	1	3
0.00	1	4
0.33	1	5
1.00	1	6
0.12	1	7
1.00	1	8
0.25	2	1
...	...	...

Figure 13.9 *Data from Figure 13.8 displayed as an array (y_{jk}) and in our preferred notation as a vector (y_i) with group indicators $j[i]$ and $k[i]$.*

Model (13.9) can also be written more cleanly as $y_{jk} \sim \text{N}(\mu+\gamma_j+\delta_k, \sigma_y^2)$, but we actually prefer the more awkward notation using $j[i]$ and $k[i]$ because it emphasizes the multilevel structure of the model and is not restricted to balanced designs. When modeling a data array of the form (y_{jk}), we usually convert it into a vector with index variables for the rows and columns, as illustrated in Figure 13.9 for the flight simulator data.

Example: regression of earnings on ethnicity categories, age categories, and height

All the ideas of the earlier part of this chapter, introduced in the context of a simple structure of individuals within groups, apply to non-nested models as well. For example, Figure 13.10 displays the estimated regression of log earnings, y_i, on height, z_i (mean-adjusted, for reasons discussed in the context of Figures 13.3–13.6), applied to the $J = 4$ ethnic groups and $K = 3$ age categories. In essence, there is a separate regression model for each age group and ethnicity combination. The multilevel model can be written, somewhat awkwardly, as a data-level model,

$$y_i \sim \text{N}(\alpha_{j[i],k[i]} + \beta_{j[i],k[i]} z_i,\, \sigma_y^2), \text{ for } i = 1, \ldots, n,$$

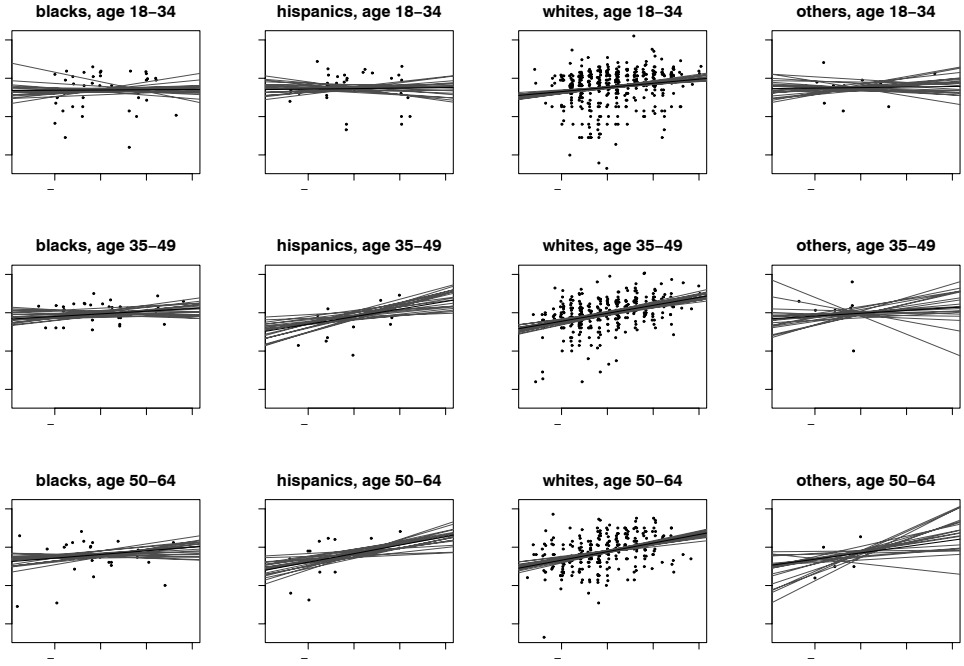

Figure 13.10 *Multilevel regression lines $y = \beta^0_{j,k} + \beta^1_{j,k} z$, for log earnings y given mean-adjusted height z, for four ethnic groups j and three age categories k. The gray lines indicate uncertainty in the fitted regressions.*

a decomposition of the intercepts and slopes into terms for ethnicity, age, and ethnicity × age,

$$\begin{pmatrix} \alpha_{j,k} \\ \beta_{j,k} \end{pmatrix} = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} + \begin{pmatrix} \gamma^{\text{eth}}_{0j} \\ \gamma^{\text{eth}}_{1j} \end{pmatrix} + \begin{pmatrix} \gamma^{\text{age}}_{0k} \\ \gamma^{\text{age}}_{1k} \end{pmatrix} + \begin{pmatrix} \gamma^{\text{eth}\times\text{age}}_{0jk} \\ \gamma^{\text{eth}\times\text{age}}_{1jk} \end{pmatrix},$$

and models for variation,

$$\begin{aligned}
\begin{pmatrix} \gamma^{\text{eth}}_{0j} \\ \gamma^{\text{eth}}_{1j} \end{pmatrix} &\sim \text{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma^{\text{eth}}\right), \text{ for } j = 1, \ldots, J \\
\begin{pmatrix} \gamma^{\text{age}}_{0k} \\ \gamma^{\text{age}}_{1k} \end{pmatrix} &\sim \text{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma^{\text{age}}\right), \text{ for } k = 1, \ldots, K \\
\begin{pmatrix} \gamma^{\text{eth}\times\text{age}}_{0jk} \\ \gamma^{\text{eth}\times\text{age}}_{1jk} \end{pmatrix} &\sim \text{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma^{\text{eth}\times\text{age}}\right), \text{ for } j = 1, \ldots, J;\ k = 1, \ldots, K.
\end{aligned}$$

Because we have included means μ_0, μ_1 in the decomposition above, we can center each batch of coefficients at 0.

Interpretation of data-level variance. The data-level errors have estimated residual standard deviation $\hat{\sigma}_y = 0.87$. That is, given ethnicity, age group, and height, log earnings can be predicted to within approximately ± 0.87, and so earnings themselves can be predicted to within a multiplicative factor of $e^{0.87} = 2.4$. So earnings cannot be predicted well at all by these factors, which is also apparent from the scatter in Figure 13.10.

Interpretation of group-level variances. The group-level errors can be separated into intercept and slope coefficients. The intercepts have estimated residual stan-

B: 257	E: 230	A: 279	C: 287	D: 202
D: 245	A: 283	E: 245	B: 280	C: 260
E: 182	B: 252	C: 280	D: 246	A: 250
A: 203	C: 204	D: 227	E: 193	B: 259
C: 231	D: 271	B: 266	A: 334	E: 338

Figure 13.11 *Data from a* 5×5 *latin square experiment studying the effects of five ordered treatments on the yields of millet crops, from Snedecor and Cochran (1989). Each cell shows the randomly assigned treatment and the observed yield for the plot.*

dard deviations of $(\widehat{\Sigma}_{00}^{\text{eth}})^{1/2} = 0.08$ at the ethnicity level, $(\widehat{\Sigma}_{00}^{\text{age}})^{1/2} = 0.25$ at the age level, and $(\widehat{\Sigma}_{00}^{\text{eth}\times\text{age}})^{1/2} = 0.11$ at the ethnicity $\times$ age level. Because we have rescaled height to have a mean of zero (see Figure 13.10), we can interpret these standard deviations as the relative importance of each factor (ethnicity, age group, and their interaction) on log earnings at the average height in the population.

This model fits earnings on the log scale and so these standard deviations can be interpreted accordingly. For example, the residual standard deviation of 0.08 for the ethnicity coefficients implies that the predictive effects of ethnic groups in the model are on the order of ± 0.08, which correspond to multiplicative factors from about $e^{-0.08} = 0.92$ to $e^{0.08} = 1.08$.

The slopes have estimated residual standard deviations of $(\widehat{\Sigma}_{11}^{\text{eth}})^{1/2} = 0.03$ at the ethnicity level, $(\widehat{\Sigma}_{11}^{\text{age}})^{1/2} = 0.02$ at the age level, and $(\widehat{\Sigma}_{11}^{\text{eth}\times\text{age}})^{1/2} = 0.02$ at the ethnicity $\times$ age level. These slopes are per inch of height, so, for example, the predictive effects of ethnic groups in the model are in the range of $\pm 3\%$ in income per inch of height. One can also look at the estimated correlation between intercepts and slopes for each factor.

Example: a latin square design with grouping factors and group-level predictors

Non-nested models can also include group-level predictors. We illustrate with data from a 5×5 latin square experiment, a design in which 25 units arranged in a square grid are assigned five different treatments, with each treatment being assigned to one unit in each row and each column. Figure 13.11 shows the treatment assignments and data from a small agricultural experiment. There are three non-nested levels of grouping—rows, columns, and treatments—and each has a natural group-level predictor corresponding to a linear trend. (The five treatments are ordered.)

The corresponding multilevel model can be written as

$$\begin{aligned}
y_i &\sim \text{N}(\mu + \beta_{j[i]}^{\text{row}} + \beta_{k[i]}^{\text{column}} + \beta_{l[i]}^{\text{treat}},\, \sigma_y^2), \text{ for } i = 1, \ldots, 25 \\
\beta_j^{\text{row}} &\sim \text{N}(\gamma^{\text{row}} \cdot (j-3), \sigma_{\beta\,\text{row}}^2), \text{ for } j = 1, \ldots, 5 \\
\beta_k^{\text{column}} &\sim \text{N}(\gamma^{\text{column}} \cdot (k-3), \sigma_{\beta\,\text{column}}^2), \text{ for } k = 1, \ldots, 5 \\
\beta_l^{\text{treat}} &\sim \text{N}(\gamma^{\text{treat}} \cdot (l-3), \sigma_{\beta\,\text{treat}}^2), \text{ for } l = 1, \ldots, 5. \qquad (13.10)
\end{aligned}$$

Thus j, k, and l serve simultaneously as values of the row, column, and treatment predictors.

By subtracting 3, we have centered the row, column, and treatment predictors at zero; the parameter μ has a clear interpretation as the grand mean of the data, with the different β's supplying deviations for rows, columns, and treatments. As with group-level models in general, the linear trends at each level potentially allow more precise estimates of the group effects, to the extent that these trends are supported by the data. An advantage of multilevel modeling here is that it doesn't force a

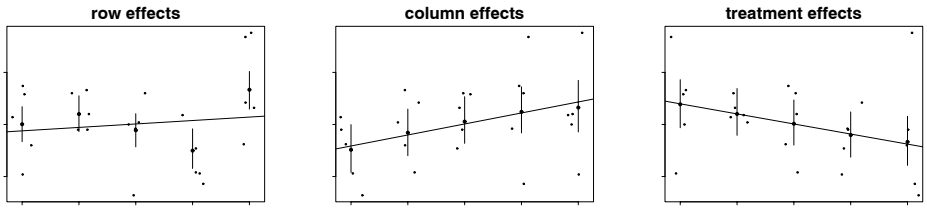

Figure 13.12 *Estimates* ± 1 *standard error for the row, column, and treatment effects for the latin square data in Figure 13.11. The five levels of each factor are ordered, and the lines display the estimated group-level regressions,* $y = \mu + \gamma^{\text{row}} \cdot (x-3)$, $y = \mu + \gamma^{\text{column}} \cdot (x-3)$, *and* $y = \mu + \gamma^{\text{treat}} \cdot (x-3)$.

choice between a linear fit and separate estimates for each level of a predictor. (This is an issue we discussed more generally in Chapter 11 in the context of including group indicators as well as group-level predictors.)

Figure 13.12 shows the estimated row, column, and treatment effects on graphs, along with the estimated linear trends. The grand mean μ has been added back to each of these observations so that the plots are on the scale of the original data. This sort of data structure is commonly studied using the analysis of variance, whose connections with multilevel models we discuss fully in Chapter 22, including a discussion of this latin square example in Section 22.5.

13.6 Selecting, transforming, and combining regression inputs

As with classical regression (see Section 4.5), choices must be made in multilevel models about which input variables to include, and how best to transform and combine them. We discuss here how some of these decisions can be expressed as particular choices of parameters in a multilevel model. The topic of formalizing modeling choices is currently an active area of research—key concerns include using information in potential input variables without being overwhelmed by the complexity of the relating model, and including model choice in uncertainty estimates. As discussed in Section 9.5, the assumption of ignorability in observational studies is more plausible when controlling for more pre-treatment inputs, which gives us a motivation to include more regression predictors.

Classical models for regression coefficients

Multilevel modeling includes classical least squares regression as a special case. In a multilevel model, each coefficient is part of a model with some mean and standard deviation. (These mean values can themselves be determined by group-level predictors in a group-level model.) In classical regression, every predictor is either in or out of the model, and each of these options corresponds to a special case of the multilevel model.

- If a predictor is "in," this corresponds to a coefficient model with standard deviation of ∞: no group-level information is used to estimate this parameter, so it is estimated directly using least squares. It turns out that in this case the group-level mean is irrelevant (see formula (12.16) on page 269 for the case $\sigma_\alpha = \infty$); for convenience we often set it to 0.
- If a predictor is "out," this corresponds to a group-level model with group-level

mean 0 and standard deviation 0: the coefficient estimate is then fixed at zero (see (12.16) for the case $\sigma_\alpha = 0$) with no uncertainty.

Multilevel modeling as an alternative to selecting regression predictors

Multilevel models can be used to combine inputs into more effective regression predictors, generalizing some of the transformation ideas discussed in Section 4.6. When many potential regression inputs are available, the fundamental approach is to include as many of these inputs as possible, but not necessarily as independent least squares predictors.

For example, Witte et al. (1994) describe a logistic regression in a case-control study of 362 persons, predicting cancer incidence given information on consumption of 87 different foods (and also controlling for five background variables which we do not discuss further here). Each of the foods can potentially increase or decrease the probability of cancer, but it would be hard to trust the result of a regression with 87 predictors fit to only 362 data points, and classical tools for selecting regression predictors do not seem so helpful here. In our general notation, the challenge is to estimate the logistic regression of cancer status y on the 362×87 matrix X of food consumption (and the 362×6 matrix X^0 containing the constant term and the 5 background variables).

More information is available, however, because each of the 87 foods can be characterized by its level of each of 35 nutrients, information that can be expressed as an 87×36 matrix of predictors Z indicating how much of each nutrient is in each food. Witte et al. fit the following multilevel model:

$$\begin{aligned} \Pr(y_i = 1) &= \text{logit}^{-1}(X_i^0\beta^0 + X_i B_{j[i]}), \text{ for } i = 1, \ldots, 362 \\ B_j &\sim \text{N}(Z_j\gamma, \sigma_\beta^2), \text{ for } j = 1, \ldots, 87. \end{aligned} \tag{13.11}$$

The food-nutrient information in Z allows the multilevel model to estimate separate predictive effects for foods, after controlling for systematic patterns associated with nutrients. In the extreme case that $\sigma_\beta = 0$, all the variation associated with the foods is explained by the nutrients. At the other extreme, $\sigma_\beta = \infty$ would imply that the nutrient information is not helping at all.

Model (13.11) is helpful in reducing the number of food predictors from 87 to 35. At this point, Witte et al. used substantive understanding of diet and cancer to understand the result. Ultimately, we would like to have a model that structures the 35 predictors even more, perhaps by categorizing them into batches or combining them in some way. The next example sketches how this might be done; it is currently an active research topic to generally structure large numbers of regression predictors.

Linear transformation and combination of inputs in a multilevel model

For another example, we consider the problem of forecasting presidential elections by state (see Section 1.2). A forecasting model based on 11 recent national elections has more than 500 "data points"—state-level elections—and can then potentially include many state-level predictors measuring factors such as economic performance, incumbency, and popularity. However, at the national level there are really only 11 observations and so one must be parsimonious with national-level predictors. In practice, this means performing some preliminary data analysis to pick a single economic predictor, a single popularity predictor, and maybe one or two other predictors based on incumbency and political ideology.

Setting up a model to allow partial pooling of a set of regression predictors

A more general approach to including national predictors is possible using multilevel modeling. For example, suppose we wish to include five measures of the national economy (for example, change in GDP per capita, change in unemployment, and so forth). The usual approach (which we have followed in the past in this problem) is to choose one of these as the economic predictor, x, thus writing the model as

$$y_i = \alpha + \beta x_i + \cdots, \tag{13.12}$$

where the dots indicate all the rest of the model, including other state-level and national predictors, as well as error terms at the state, regional, and national levels. Here we focus on the economic inputs, for simplicity setting aside the rest of the model.

Instead of choosing just one of the five economic inputs, it would perhaps be better first to standardize each of them (see Section 4.2), orient them so they are in the same direction, label these standardized variables as $X_{(j)}$, for $j = 1, \ldots, 5$, and then average them into a single predictor, defined for each data point as

$$x_i^{\text{avg}} = \frac{1}{5}\sum_{j=1}^{5}\sum X_{ij}, \text{ for } i = 1, \ldots .n. \tag{13.13}$$

This new x^{avg} can be included in place of x as the regression predictor in (13.12), or, equivalently,

$$\begin{aligned} y_i &= \alpha + \beta x_i^{\text{avg}} + \cdots \\ &= \alpha + \frac{1}{5}\beta X_{i1} + \cdots + \frac{1}{5}\beta X_{i5} + \cdots . \end{aligned}$$

The resulting model will represent an improvement to the extent that the average of the five standardized economy measures is a better predictor than the single measure chosen before.

However, model (13.13) is limited in that it restricts the coefficients of the five separate x^j's to be equal. More generally, we can replace (13.13) by a weighted average:

$$x_i^{\text{w.avg}} = \frac{1}{5}\sum_{j=1}^{5}\gamma_j X_{ij}, \text{ for } i = 1, \ldots, n, \tag{13.14}$$

so that the data model becomes

$$\begin{aligned} y_i &= \alpha + \beta x_i^{\text{w.avg}} + \cdots \\ &= \alpha + \frac{1}{5}\gamma_1\beta X_{i1} + \cdots + \frac{1}{5}\gamma_5\beta X_{i5} + \cdots . \end{aligned} \tag{13.15}$$

We would like to estimate the relative coefficients γ_j from the data, but we cannot simply use classical regression, since this would then be equivalent to estimating a separate coefficient for each of the five predictors, and we have already established that not enough data are available to do a good job of this.

Instead, one can set up a model for the γ_j's:

$$\gamma_j \sim \text{N}(1, \sigma_\gamma^2), \text{ for } j = 1, \ldots, 5, \tag{13.16}$$

so that, in the model (13.15), the common coefficient β can be estimated classically, but the relative coefficients γ_j are part of a multilevel model. The hyperparameter σ_γ can be interpreted as follows:

- If $\sigma_\gamma = 0$, the model reduces to the simple averaging (13.14): *complete pooling*

of the γ_j's to the common value of 1, so that the combined predictor $x^{\text{w.avg}}$ is simply x^{avg}, the average of the five individual $X_{(j)}$'s.

- If $\sigma_\gamma = \infty$, there is *no pooling*, with the individual coefficients $\frac{1}{5}\gamma_j\beta$ estimated separately using least squares.
- When σ_γ is positive but finite, the γ_j's are *partially pooled*, so that the five predictors x_j have coefficients that are near each other but not identical.

Depending on the amount of data available, σ_γ can be estimated as part of the model or set to a value such as 0.3 that constrains the γ_j's to be fairly close to 1 and thus constrains the coefficients of the individual x^j's toward each other in the data model (13.15).

Connection to factor analysis

A model can include multiplicative parameters for both modeling and computational purposes. For example, we could predict the election outcome in year t in state s within region $r[s]$ as

$$y_{st} = \beta^{(0)}X^{(0)}_{st} + \alpha_1 \sum_{j=1}^{5} \beta^{(1)}_j X^{(1)}_{jt} + \alpha_2\gamma_t + \alpha_3\delta_{r[s],t} + \epsilon_{st},$$

where $X^{(0)}$ is the matrix of state $\times$ year-level predictors, $X^{(1)}$ is the matrix of year-level predictors, and γ, δ, and ϵ are national, regional, and statewide error terms. In this model, the auxiliary parameters α_2 and α_3 exist for purely computational reasons, and they can be estimated, with the understanding that we are interested only in the products $\alpha_2\gamma_t$ and $\alpha_3\delta_{r,t}$. More interestingly, α_1 serves both a computational and modeling role—the $\beta^{(1)}_j$ parameters have a common $\text{N}(\frac{1}{5}, \sigma^2_m)$ model, and α_1 has the interpretation as the overall coefficient for the economic predictors.

More generally, we can imagine K batches of predictors, with the data-level regression model using a weighted average from each batch:

$$y = X^{(0)}\beta^{(0)} + \beta_1 x^{\text{w.avg},\,1} + \cdots + \beta_k x^{\text{w.avg},\,K} + \cdots,$$

where each predictor $x^{\text{w.avg}}_k$ is a combination of J_k individual predictors x^{jk}:

$$\text{for each } k\text{: } \; x^{\text{w.avg},\,k}_i = \frac{1}{J_k}\sum_{j=1}^{J_k} \gamma_{jk}x^{jk}_i, \text{ for } i = 1, \ldots, n.$$

This is equivalent to a regression model on the complete set of available predictors, $x^{11}, \ldots, x^{J_1 1}; x^{12}, \ldots, x^{J_2 2}; \ldots, x^{1K}, \ldots, x^{J_K K}$, where the predictor x^{jk} gets the coefficient $\frac{1}{J_k}\gamma_{jk}\beta_k$. Each batch of relative weights γ is then modeled hierarchically:

$$\text{for each } k\text{: } \; \gamma_{jk} \sim \text{N}(1, \sigma^2_{\gamma\,k}), \text{ for } j = 1, \ldots, J_k,$$

with the hyperparameters $\sigma_{\gamma\,k}$ estimated from the data or set to low values such as 0.3.

In this model, each combined predictor $x^{\text{w.avg},\,k}$ represents a "factor" formed by a linear combination of the J_k individual predictors, β_k represents the importance of that factor, and the γ_{jk}'s give the relative importance of the different components.

As noted at the beginning of this section, these models are currently the subject of active research, and we suggest that they can serve as a motivation to specially tailored models for individual problems rather than as off-the-shelf solutions to generic multilevel problems with many predictors.

13.7 More complex multilevel models

The models we have considered so far can be generalized in a variety of ways. Chapters 14 and 15 discuss multilevel logistic and generalized linear models. Other extensions within multilevel linear and generalized linear models include the following:

- Variances can vary, as parametric functions of input variables, and in a multilevel way by allowing different variances for groups. For example, the model $y_i \sim \mathrm{N}(X_i\beta, \sigma_i^2)$, with $\sigma_i = \exp(X_i\gamma)$, allows the variance to depend on the predictors in a way that can be estimated from the data, and similarly, in a multilevel context, a model such as $\sigma_i = \exp(a_{j[i]} + bx_i)$ allows variances to vary by group. (It is natural to model the parameters σ on the log scale because they are restricted to be positive.)
- Models with several factors can have many potential interactions, which themselves can be modeled in a structured way, for example with larger variances for coefficients of interactions whose main effects are large. This is a model-based, multilevel version of general advice for classical regression modeling.
- Regression models can be set up for multivariate outcomes, so that vectors of coefficients become matrices, with a data-level covariance matrix. These models become correspondingly more complex when multilevel factors are added.
- Time series can be modeled in many ways going beyond simple autoregressions, and these parameters can vary by group with time-series cross-sectional data. This can be seen as a special case of non-nested groupings (for example, country $\times$ year), with calendar time being a group-level predictor.
- One way to go beyond linearity is with nonparametric regression, with the simplest version being $y_i = g(X_i, \theta) + \epsilon_i$, and the function g being allowed to have some general form (for example, cubic splines, which are piecewise-continuous third-degree polynomials). Versions of such models can also be estimated using locally weighted regression, and again can be expanded to multilevel structures as appropriate.
- More complicated models are appropriate to data with spatial or network structure. These can be thought of as generalizations of multilevel models in which groups (for example, social networks) are not necessarily disjoint, and in which group membership can be continuous (some connections are stronger than others) rather than simply "in" or "out."

We do not discuss any of these models further here, but we wanted to bring them up to be clear that the particular models presented in this book are just the starting point to our general modeling approach.

13.8 Bibliographic note

The textbooks by Kreft and De Leeuw (1998), Raudenbush and Bryk (2002), and others discuss multilevel models with varying intercepts and slopes. For an early example, see Dempster, Rubin, and Tsutakawa (1981). Non-nested models are discussed by Rasbash and Browne (2003). The flight simulator example comes from Gawron et al. (2003), and the latin square example comes from Snedecor and Cochran (1989).

Models for covariance matrices have been presented by Barnard, McCulloch, and Meng (1996), Pinheiro and Bates (1996), Daniels and Kass (1999, 2001), Daniels and Pourahmadi (2002). Boscardin and Gelman (1996) discuss parametric models

for unequal variances in multilevel linear regression. The scaled inverse-Wishart model we recommend comes from O'Malley and Zaslavsky (2005).

The models for combining regression predictors discussed in Section 13.6 appear in Witte et al. (1994), Greenland (2000), Gelman (2004b), and Gustafson and Greenland (2005). See also Hodges et al. (2005) and West (2003) on methods of including many predictors and interactions in a regression. Other work on selecting and combining regression predictors in multilevel models includes Madigan and Raftery (1994), Hoeting et al. (1999), Chipman, George, and McCulloch (2001), and Dunson (2006). The election forecasting example is discussed in Gelman and King (1993) and Gelman et al. (2003, section 15.2); see Fair (1978), Rosenstone (1983), Campbell (1992), and Wlezien and Erikson (2004, 2005) for influential work in this area.

Some references for hierarchical spatial and space-time models include Besag, York, and Mollie (1991), Waller et al. (1997), Besag and Higdon (1999), Wikle et al. (2001), and Bannerjee, Gelfand, and Carlin (2003). Jackson, Best, and Richardson (2006) discuss hierarchical models combining aggregate and survey data in public health. Datta et al. (1999) compare hierarchical time series models; see also Fay and Herriot (1979). Girosi and King (2005) present a multilevel model for estimating trends within demographic subgroups.

For information on nonparametric methods such as lowess, splines, wavelets, hazard regression, generalized additive models, and regression trees, see Hastie, Tibshirani, and Friedman (2002), and, for examples in R, see Venables and Ripley (2002). Crainiceanu, Ruppert, and Wand (2005) fit spline models using Bugs. MacLehose et al. (2006) combine ideas of nonparametric and multilevel models.

13.9 Exercises

1. Fit a multilevel model to predict course evaluations from beauty and other predictors in the `beauty` dataset (see Exercises 3.5, 4.8, and 12.6) allowing the intercept and coefficient for beauty to vary by course category:

   (a) Write the model in statistical notation.

   (b) Fit the model using `lmer()` and discuss the results: the coefficient estimates and the estimated standard deviation and correlation parameters. Identify each of the estimated parameters with the notation in your model from (a).

   (c) Display the estimated model graphically in plots that also include the data.

2. Models for adjusting individual ratings: a committee of 10 persons is evaluating 100 job applications. Each person on the committee reads 30 applications (structured so that each application is read by three people) and gives each a numerical rating between 1 and 10.

   (a) It would be natural to rate the applications based on their combined scores; however, there is a worry that different raters use different standards, and we would like to correct for this. Set up a model for the ratings (with parameters for the applicants and the raters).

   (b) It is possible that some persons on the committee show more variation than others in their ratings. Expand your model to allow for this.

3. Non-nested model: continuing the Olympic ratings example from Exercise 11.3:

   (a) Write the notation for a non-nested multilevel model (varying across skaters and judges) for the technical merit ratings and fit using `lmer()`.

(b) Fit the model in (a) using the artistic impression ratings.

(c) Display your results for both outcomes graphically.

(d) Use posterior predictive checks to investigate model fit in (a) and (b).

4. Models with unequal variances: the folder `age.guessing` contains a dataset from Gelman and Nolan (2002) from a classroom demonstration in which 10 groups of students guess the ages of 10 different persons based on photographs. The dataset also includes the true ages of the people in the photographs.

   Set up a non-nested model to these data, including a coefficient for each of the persons in the photos (indicating their apparent age), a coefficient for each of the 10 groups (indicating potential systematic patterns of groups guessing high or low), and a separate error variance for each group (so that some groups are more consistent than others).

5. Return to the CD4 data introduced from Exercise 11.4.

   (a) Extend the model in Exercise 12.2 to allow for varying slopes for the time predictor.

   (b) Next fit a model that does not allow for varying slopes but does allow for different coefficients for each time point (rather than fitting the linear trend).

   (c) Compare the results of these models both numerically and graphically.

6. Using the time-series cross-sectional dataset you worked with in Exercise 11.2, fit the model you formulated in part (c) of that exercise.