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

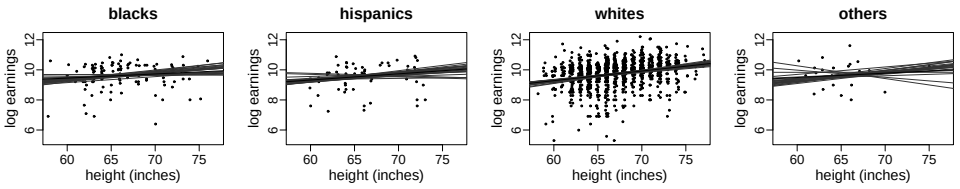

Figure 13.3 *Multilevel regression lines $y = \alpha_j + \beta_j x$ for log earnings on height (among those with positive earnings), in four ethnic categories j. The gray lines indicate uncertainty in the fitted regressions.*

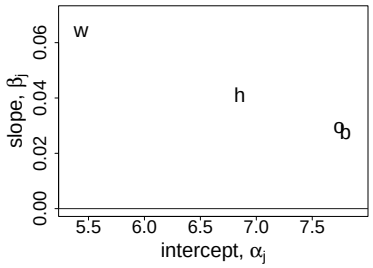

Figure 13.4 *Scatterplot of estimated intercepts and slopes (for whites, hispanics, blacks, and others), (α_j, β_j), for the earnings-height regressions shown in Figure 13.3. The extreme negative correlation arises because the center of the range of height is far from zero. Compare to the coefficients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