Setting up the computation in R and Bugs

We can compute λ at each level using the errors `e.y, e.a, e.b` defined at the end of Section 21.5 for computing R^2. Once the Bugs model has been fit, the pooling factors can be computed in R as follows:

R code

```
lambda.y <- 1 - var (apply (e.y, 2, mean)) / mean (apply (e.y, 1, var))
lambda.a <- 1 - var (apply (e.a, 2, mean)) / mean (apply (e.a, 1, var))
lambda.b <- 1 - var (apply (e.b, 2, mean)) / mean (apply (e.b, 1, var))
```

Discussion

The proportion of variance explained (21.8) and the pooling factor (21.14) can be easily calculated at each stage of a multilevel model. In general, R^2 will be informative wherever regression predictors (including group indicators) are present, and λ will be relevant at the hierarchical stages of the model. The measures of explained variance and partial pooling conveniently summarize the fit at each level of the model and the degree to which estimates are pooled toward their population models. Together, they clarify the role of predictors at different levels of a multilevel model. They can be derived from a common framework of comparing variances at each level of the model, which also means that they do not require the fitting of additional null models.

Expressions (21.8) and (21.14) are closely related to the usual definitions of adjusted R^2 in simple linear regression and pooling in balanced one-way hierarchical models. From this perspective, they unify the data-level concept of R^2 and the group-level concept of pooling or shrinkage, and also generalize these concepts to account for uncertainty in the variance components. Further, as illustrated for the radon application, they can help us understand more complex multilevel models.

Other challenges include defining explained variance and partial pooling factors for generalized linear models, either on the scale of the data or of the latent parameters.

21.7 Adding a predictor can *increase* the residual variance!

Multilevel models can behave in ways that are unexpected from the perspective of classical statistics. We illustrate with the radon model from Chapter 12. We first fit a stripped-down multilevel model for the home radon levels, including the county-level uranium predictor but no individual-level predictors (not even the floor of measurement); thus,

$$\text{model 1:} \quad \begin{array}{rcl} y_i & \sim & \text{N}(\alpha_{j[i]}, \sigma_y^2) \\ \alpha_j & \sim & \text{N}(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2). \end{array}$$

Fitting this model to the Minnesota radon data yields estimated variance components $\sigma_y = 0.80, \sigma_\alpha = 0.12$.

We then add the house-level floor indicator x_i:

$$\text{model 2:} \quad \begin{array}{rcl} y_i & \sim & \text{N}(\alpha_{j[i]} + \beta x_i,\, \sigma_y^2) \\ \alpha_j & \sim & \text{N}(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2), \end{array}$$

yielding new estimates of $\sigma_y = 0.76, \sigma_\alpha = 0.16$. The house-level standard deviation has decreased—which makes sense since we have added a predictor at that level—but the variation at the county level has *increased*, which is a surprise. In classical

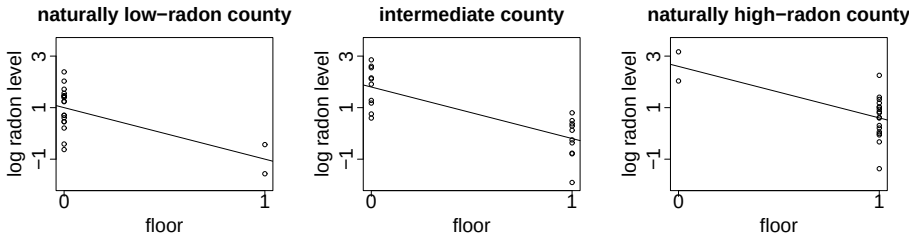

Figure 21.9 *Hypothetical data from three counties illustrating how adding an individual-level predictor can decrease the group-level variance. In each county, radon levels are higher in homes with basements. The county with low natural radon levels has more homes with basements, and the county with high natural radon levels has fewer homes with basements. As a result, the average radon level in the three counties is identical, but when floor of measurement is (appropriately) included as a predictor, the counties appear more different.*

regression models, the residual variance can only go down, not up, when a predictor is added.[6]

What is going on? After some thought, we realized that in model 2, the counties with more basements happened to have higher county coefficients α_j. In model 1, some of the variation in county radon levels was canceled by an opposite variation in the proportion of basements. The increased between-county variance in model 2 indicates true variation among counties that happened to be masked by the first model.

Figure 21.9 shows a hypothetical extreme version of this situation: the three counties have identical average radon levels (thus, $\sigma_\alpha = 0$ for model 1, which has no basement predictor), but only because the naturally low-radon county has many basements and the naturally-high radon county has few basements. (Such a pattern can happen, for example, if low-radon areas have sandy soil in which basements are easy to dig, with high-radon areas having rocky soil where basements are less commonly built.) Model 2, which controls for basements, reveals the true underlying variation among the counties, and thus σ_α increases.

This pattern, caused by correlation between individual-level variables and group-level errors, does not occur in classical regression. When it occurs in multilevel regression, the model fit can be improved by including the average of x as a group-level predictor (in this example, the proportion of houses in the county that have basements). When the county-level basement proportion is added as a group-level predictor, its coefficient is estimated at -0.41 (with a standard error of 0.2), and the estimated residual standard deviations at the data and county levels are 0.76 and 0.14.

For the radon problem, the county-level basement proportion is difficult to interpret directly but rather serves as a proxy for underlying variables (for example, the type of soil that is prevalent in the county).

In other settings, especially in social science, individual averages that are used as group-level predictors are often interpreted as "contextual effects." For example, in the police stops example in Section 15.1, one might suspect that police behavior in a precinct is influenced by the ethnic composition of the local residents. However, we must be suspicious of this sort of conclusion without further information. As the radon example illustrates, it is possible to have between-level correlations without

[6] We ignore the minor increase in the variance estimate that corresponds to reducing the degrees of freedom by 1 and thus dividing by $n - k - 1$ instead of $n - k$ in the variance calculation.

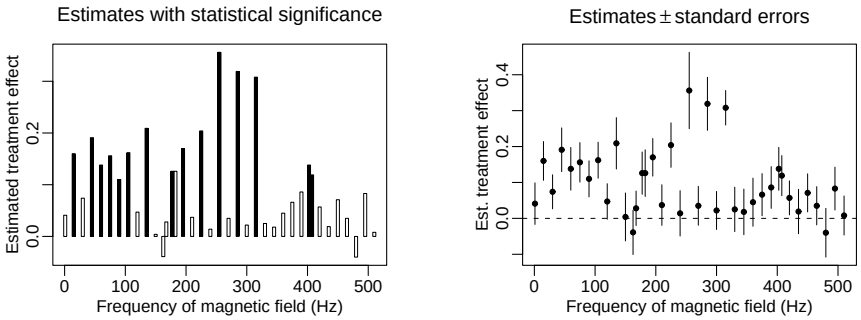

Figure 21.10 *(a) Estimated effects of electromagnetic fields on calcium efflux from chick brains, shaded to indicate different levels of statistical significance, adapted from Blackman et al. (1988). A separate experiment was performed at each frequency. (b) Same results presented as estimates ± standard errors. As discussed in the text, the first plot, with its emphasis on statistical significance, is misleading.*

the need for a "contextual" story. As usual in these models, we try to be careful to state the regression results in a predictive rather than causal manner ("the counties with more basements tend to have lower radon levels, after controlling for the basement statuses of the individual houses").

21.8 Multiple comparisons and statistical significance

A meta-analysis of a set of randomized experiments

In the wake of concerns about the health effects of low-frequency electric and magnetic fields, an experiment was performed to measure the effect of electromagnetic fields at various frequencies on the functioning of chick brains. At each of several frequencies of electromagnetic fields (1 Hz, 15 Hz, 30 Hz, ..., 510 Hz), a randomized experiment was performed to estimate the effect of exposure, compared to a control condition of no electromagnetic field. The researchers reported, for each frequency, the estimated treatment effect (the average difference between treatment and control measurements) and the standard error (that is, $\sqrt{\sigma_T^2/n_T + \sigma_C^2/n_C}$; see Section 2.3).

In the article reporting this study, the estimates at the different frequencies were summarized by their statistical significance, as we illustrate in Figure 21.10a by using different shading for results that are more than 2.3 standard errors from zero (that is, statistically significant at the 99% level), between 2.0 and 2.3 standard errors from zero (statistically significant at the 95% level), and so forth. The researchers used this sort of display to hypothesize that one process was occurring at 255, 285, and 315 Hz (where effects were highly significant), another at 135 and 225 Hz (where effects were only moderately significant), and so forth. The estimates are all of relative calcium efflux, so that an effect of 0.1, for example, corresponds to a 10% increase compared to the control condition.

In the chick-brain experiment, the researchers made the common mistake of using statistical significance as a criterion for separating the estimates of different effects. As we discuss in Section 2.5, this approach does not make sense. At the very least, it is more informative to show the estimated treatment effect and standard error at each frequency, as in Figure 21.10b.