Chapter 8

Hierarchical Models

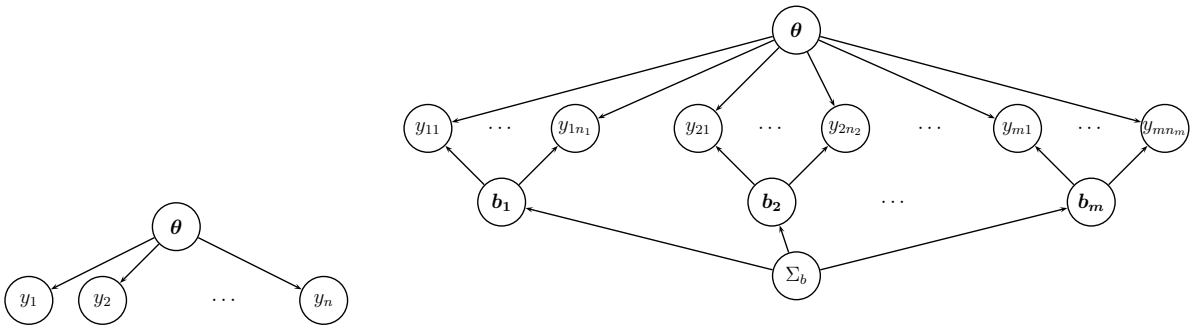
In the (generalized) linear models we've looked at so far, we've assumed that the observations are independent of each other given the predictor variables. However, there are many situations in which that type of independence does not hold. One major type of situation violating these independence assumptions is CLUSTER-LEVEL ATTRIBUTES: when observations belong to different clusters and each cluster has its own properties (different response mean, different sensitivity to each predictor). We'll now cover HIERARCHICAL (also called MULTI-LEVEL and, in some cases MIXED-EFFECTS) models, which are designed to handle this type of mutual dependence among datapoints. Common instances in which hierarchical models can be used include:

- Observations related to linguistic behavior are clustered at the level of the speaker, and speaker-specific attributes might include different baseline reading rates, differential sensitive to construction difficulty, or preference for one construction over another;

- Different sentences or even words may have idiosyncratic differences in their ease of understanding or production, and while we may not be able to model these differences, we may be able model the fact that there is incidental variation at the sentence or word level;

- Education-related observations (e.g., vocabulary size) of students have multiple levels of clustering: multiple measurements may be taken from a given student, multiple students may be observed from a class taught by a given teacher, multiple teachers may teach at the same school, multiple schools may be in the same city, and so forth.

This chapter introduces hierarchical models, building on the mathematical tools you have acquired throughout the book. This chapter makes considerably heavier use of Bayesian-style thinking and techniques than the previous chapter; this would be a good time to review marginalization (Section 3.2), Bayesian prediction and parameter estimation (Section 4.4), approximate posterior inference (Section 4.5), and confidence intervals (Chapter 5).



(a) A non-hierarchical model

(b) A simple hierarchical model, in which observations are grouped into m clusters

Figure 8.1: Non-hierarchical and hierarchical models

8.1 Introduction

The core idea behind the hierarchical model is illustrated in Figure 8.1. Figure 8.1a depicts the type of probabilistic model that we have spent most of our time with thus far: a model family has parameters $\boldsymbol{\theta}$, which determine a probability distribution over outcomes, and a set of observations $\boldsymbol{y}$ arises as a collection of independent draws from this distribution. Figure 8.1b illustrates the simplest type of hierarchical model: observations fall into a number of CLUSTERS, and the distribution over outcomes is determined jointly by (i) parameters $\boldsymbol{\theta}$ shared across clusters, and (ii) parameters $\boldsymbol{b}$ which are shared among observations within a cluster, but may be different across clusters. Crucially, there is a second probability distribution, parameterized by $\Sigma_{\boldsymbol{b}}$, over the cluster-specific parameters $\boldsymbol{b}$ themselves. In Figure 8.1b, there are m clusters, and for each cluster i there have been n_i observations $y_{i1}, \ldots, y_{in_i}$ made. All else being equal, we can expect that observations within a single cluster will tend to look more like each other than like observations in other clusters.

Let us consider a simple case in which the distribution from which the observations y_{ij} are drawn is characterized by just a few parameters—for example, they may be normally distributed, in which case the parameters are the mean and the variance. One natural type of clustering would be for each cluster to have its own mean μ_i but for all the clusters to have the same variance σ_y^2. In the terminology of Figure 8.1b, we would classify the μ_i as cluster-specific parameters (in the $\boldsymbol{b}$ nodes) and the variance σ_y^2 as a shared parameter (in the $\boldsymbol{\theta}$ node). In order to complete this probabilistic model, we would need to specify the distribution over the cluster-specific μ_i. We might make this distribution normal as well, which requires two parameters of its own: a global mean μ and variance $\sigma_{\boldsymbol{b}}^2$ (these would live in the Σ_b node). We can write this specification compactly as follows:

$$\begin{aligned} \mu_i &\sim \mathcal{N}(\mu, \sigma_b^2) \\ y_{ij} &\sim \mathcal{N}(\mu_i, \sigma_y^2) \end{aligned} \tag{8.1}$$

Equivalently, we could consider the cluster-specific parameters to be *deviations* from the

overall mean μ. In this approach, we would consider μ as a shared $\boldsymbol{\theta}$ parameter, and the mean of the deviations would be 0. We can compactly specify this version of the model as:

$$\begin{aligned} b_i &\sim \mathcal{N}(0, \sigma_b^2) \\ \mu_i &= \mu + b_i \\ y_{ij} &\sim \mathcal{N}(\mu_i, \sigma_y^2) \end{aligned} \tag{8.2}$$

The specifications in Equations 8.1 and 8.2 describe exactly the same family of probabilistic models. The advantage of the former specification is that it is more compact. The advantage of the latter specification is that the cluster-specific parameters are more directly interpretable as deviations "above" or "below" the overall average μ. In addition, the latter specification leads to a nice connection with our discussion of linear models in Section 6.2. We can describe the same model as follows:

$$y_{ij} = \mu + \underbrace{b_i}_{\sim\mathcal{N}(0,\sigma_b^2)} + \underbrace{\epsilon_{ij}}_{\sim\mathcal{N}(0,\sigma_y^2)} \tag{8.3}$$

That is, an individual observation y_{ij} is the sum of the overall mean μ, a normally-distributed cluster-level deviation b_i, and a normally-distributed observation-level deviation ϵ_{ij}.

Let us now consider a concrete example with slightly greater complexity. Suppose that a phonetician is interested in studying the distribution of the pronunciation of the vowel [ɑ], recruits six native speakers of American English, and records each speaker once a week for fifteen weeks. In each case, the phonetician computes and records the F1 and F2 formants of the pronounced syllable. Now, no two recordings will be exactly alike, but different individuals will tend to pronounce the syllable in different ways—that is, there is both within-individual and between-individual variation in F1 formant from recording to recording. Let us assume that inter-speaker (cluster-level) variation and inter-trial (observation-level) variation are both multivariate-normal. If we denote the ⟨F1,F2⟩ value for the jth recording of speaker i as y_{ij}, then we could write our model as follows:

$$\begin{aligned} b_i &\sim \mathcal{N}(\mathbf{0}, \Sigma_b) \\ \mu_i &= \mu + b_i \\ y_{ij} &\sim \mathcal{N}(\mu_i, \Sigma_y) \end{aligned} \tag{8.4}$$

where $\mathbf{0}$ is the vector $\langle 0, 0\rangle$.

The only difference between the models in Equations 8.2 and 8.4 is that whereas the former is univariate, the latter is multivariate: b_i is distributed around zero according to some covariance matrix Σ_b, and the y_{ij} are distributed around μ_i according to another covariance matrix Σ_y. Both the univariate and multivariate models (Equations 8.2 and 8.4) have precisely the structure of Figure 8.1b, with μ and Σ_y being the shared parameters θ. For the

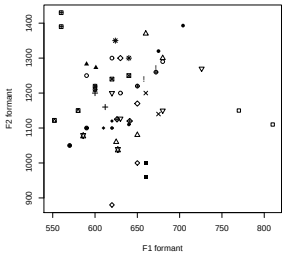


Figure 8.2: Empirically observed male adult speaker means for first and second formants of [ɑ]

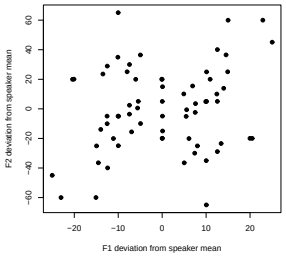


Figure 8.3: Empirically observed deviations from speaker means

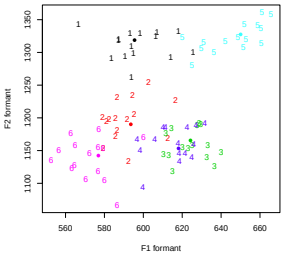


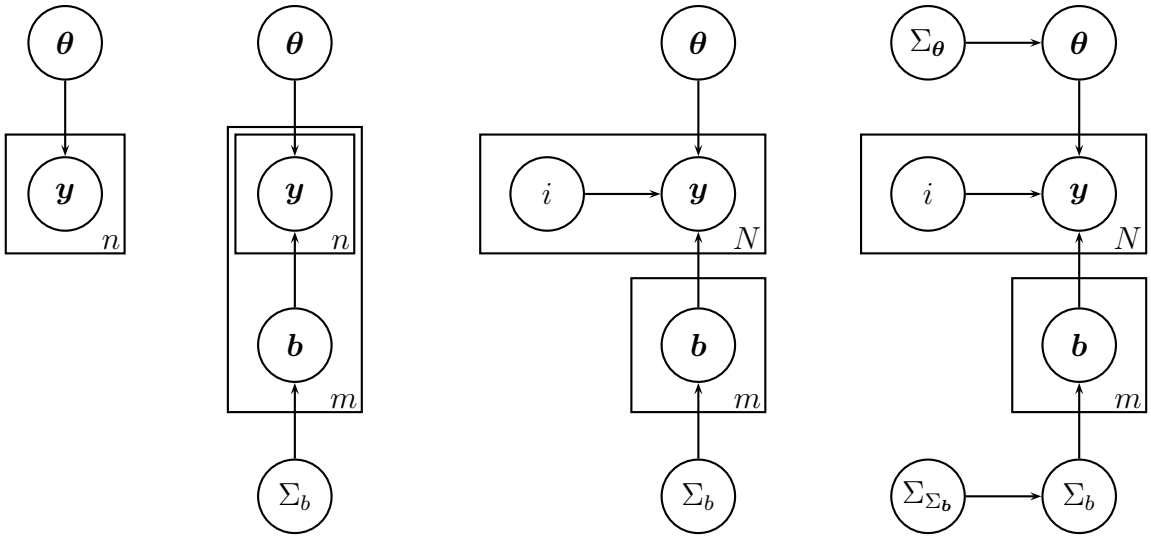
Figure 8.4: Simulated formant data for six speakers. Speaker-specific means $\mu + b_i$ are given as filled circles.

moment, to estimate the model parameters we use the simple expedient of using the sample mean and covariance of speaker-averages of adult-male data due to Peterson and Barney (1952) (shown in Figure 8.2) to estimate μ, and Σ_b, and the covariance of deviations from speaker means (shown in Figure 8.3) to estimate Σ_y. Figure 8.4 gives sample data generated from this model. The individual speakers correspond to clusters of trial-level observations. Note how there is considerable intra-cluster variation but the variation between clusters is at least as large.

8.2 Parameter estimation in hierarchical models

The previous section outlines what is essentially the complete probabilistic theory of hierarchical models. However, the problems of statistical inference within hierarchical models require more discussion. Before we dive into these issues, however, it is worthwhile to introduce a more succinct graphical representation of hierarchical models than that used in Figure 8.1b. Figure 8.5a is a representation of non-hierarchical models, as in Figure 8.1a, where the individual observations y_i have been collapsed into a single node $\boldsymbol{y}$. The box labeled "n" surrounding the $\boldsymbol{y}$ node indicates that n independent events are generated at this node; the labeled box is called a PLATE. Likewise, Figure 8.5b is a representation of the class of simple hierarchical models shown in Figure 8.1b, with both individual observations y_{ij} and class-specific parameters $\boldsymbol{b_i}$ compressed into single nodes. The outer plate indicates that m independent events are generated at the $\boldsymbol{b}$ node; the inner plate (embedded in the outer plate) indicates that for the i-th of these m events, n_i sub-events are generated. Each sub-event has a "cluster identity" label—the node labeled i, which allows us to track which cluster each observation falls into—and the nodes $\boldsymbol{b}$, $\boldsymbol{\theta}$, and i jointly determine the distribution over the outcome at the $\boldsymbol{y}$ node for this sub-event.

In light of this picture, let us consider the problem of parameter estimation for a case such as the formant measurements of the previous section. We know our observations $\boldsymbol{y}$, and we also know the cluster identity variable i—that is, which individual produced each



(a) The non-hierarchical model of Figure 8.1a

(b) The simple hierarchical model of Figure 8.1b

(c) Alternative representation for Figure 8.5b, with cluster identities explicit

(d) The model of Figure 8.1b with priors over $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma_b}$

Figure 8.5: A more succinct representation of the models in Figure 8.1

observation. We do not know the shared model parameters $\boldsymbol{\theta}$ the parameters $\Sigma_{\boldsymbol{b}}$ governing cross-speaker variability, or the speaker-specific variations $\boldsymbol{b_i}$ themselves. Upon reflection, however, it should become clear that the primary goal of the study is to learn $\boldsymbol{\theta}$ and $\Sigma_{\boldsymbol{b}}$, not $\boldsymbol{b_i}$. Yet the $\boldsymbol{b_i}$ stand in the way of estimating $\Sigma_{\boldsymbol{b}}$. It might seem like a good idea to first construct point estimates of $\boldsymbol{b_i}$ and then use these estimates directly to estimate $\Sigma_{\boldsymbol{b}}$, but this approach throws away valuable information (our uncertainty about the true values of the $\boldsymbol{b_i}$, which we should take into account). How can we make inferences about our parameters of interest in a principled way?

The answer is actually quite simple: whatever technique of parameter estimation we choose, we should *marginalize* over the cluster-specific $\boldsymbol{b_i}$. This leads us to two basic approaches to parameter estimation for hierarchical models:

1. Construct point estimates of model parameters ($\widehat{\Sigma_{\boldsymbol{b}}}$ and/or $\widehat{\boldsymbol{\theta}}$) using the principle of maximum likelihood. There are actually two different maximum-likelihood approaches that have widespread currency. The first is to simultaneously choose $\widehat{\Sigma_{\boldsymbol{b}}}$ and $\widehat{\boldsymbol{\theta}}$ to maximize the likelihood, marginal over $\boldsymbol{b}$:

$$\text{Lik}(\Sigma_{\boldsymbol{b}}, \boldsymbol{\theta}; \boldsymbol{y}) = \int_{\boldsymbol{b}} P(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{b}, i) P(\boldsymbol{b}|\Sigma_{\boldsymbol{b}}) \, \mathrm{d}b \tag{8.5}$$

   We will follow common practice in simply calling this approach "Maximum Likelihood" (ML) estimation. The second approach, called Restricted Maximum Likelihood (REML), is perhaps most easily understood as placing an (improper) uniform distribution over the shared model parameters $\boldsymbol{\theta}$ and marginalizing over them (Harville, 1974),

so that we select only $\widehat{\Sigma_{\boldsymbol{b}}}$ according to the likelihood

$$\text{Lik}(\Sigma_{\boldsymbol{b}}; \boldsymbol{y}) = \int_{\boldsymbol{b},\boldsymbol{\theta}} P(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{b}, i) P(\boldsymbol{b}|\Sigma_{\boldsymbol{b}}) \, \mathrm{d}\boldsymbol{b} \, \mathrm{d}\boldsymbol{\theta} \tag{8.6}$$

On the REML approach, the parameters $\boldsymbol{\theta}$ are of secondary interest, but one would estimate them as

$$\arg\max_{\boldsymbol{\theta}} P(\boldsymbol{y}|\boldsymbol{\theta}, \widehat{\Sigma_{\boldsymbol{b}}}_{REML}, i) = \int_{\boldsymbol{b}} P(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{b}, i) P(\boldsymbol{b}|\widehat{\Sigma_{\boldsymbol{b}}}_{REML}) \, \mathrm{d}\boldsymbol{b}$$

For practical purposes, maximum-likelihood and restricted maximum-likelihood estimation often give results that are quite similar to one another when there are relatively few free parameters in $\boldsymbol{\theta}$ compared with the number in $\boldsymbol{b}$ (Dempster et al., 1981). We will return to the ML/REML distinction in Section 8.3.2.

2. Use Bayesian inference: introduce prior distributions over $\boldsymbol{\theta}$ and $\Sigma_{\boldsymbol{b}}$, and compute the (approximate) posterior distribution over $\boldsymbol{\theta}$ and $\Sigma_{\boldsymbol{b}}$. Since the introduction of priors means that $\boldsymbol{\theta}$ and $\Sigma_{\boldsymbol{b}}$ are themselves drawn from some distribution, this is actually a shift to a slightly more complex hierarchical model, shown in Figure 8.5d. $\Sigma_{\boldsymbol{\theta}}$ and $\Sigma_{\Sigma_{\boldsymbol{b}}}$, chosen by the researcher, parameterize the priors over $\boldsymbol{\theta}$ and $\Sigma_{\boldsymbol{b}}$ respectively. Via Bayes' rule, the posterior distributions of interest can be written as follows:

$$P(\Sigma_{\boldsymbol{b}}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \int_{\boldsymbol{b}} P(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{b}) P(\boldsymbol{b}|\Sigma_{\boldsymbol{b}}, i) P(\boldsymbol{\theta}|\Sigma_{\boldsymbol{\theta}}) P(\Sigma_{\boldsymbol{b}}|\Sigma_{\Sigma_{\boldsymbol{b}}}) \, \mathrm{d}b \tag{8.7}$$

Note that the posterior distribution looks very similar to the unnormalized likelihood of Equation (8.5) above. The only difference is, as always, the presence of the prior probability $P(\boldsymbol{\theta}|\Sigma_{\boldsymbol{\theta}})P(\Sigma_{\boldsymbol{b}}|\Sigma_{\Sigma_{\boldsymbol{b}}})$ in the posterior distribution. It's worth recalling at this point that if the prior distribution is chosen to be uniform, then the maximum-likelihood estimate is also the Bayesian maximum a-posteriori (MAP) estimate.

We'll now illustrate each of these approaches with reference to actual formant data from recordings of adult male American speakers' pronunciations of [ɑ] by Peterson and Barney (1952). The F1 data are plotted in Figure 8.6.

8.2.1 Point estimation based on maximum likelihood

We illustrate the point estimation approach by treating the F1 and F2 formats separately.[1] For each of the formants, we assume the model of Equation (8.2): that the speaker-specific deviations from the grand mean are normally distributed, and that trial-specific deviations are also normally distributed. We reiterate that there are three parameters to be estimated

[1]The R package lme4 is the state of the art in likelihood-based point estimation for a wide variety of hierarchical models, and we use it here.

in each model: the overall mean μ, the inter-speaker variance $\Sigma_{\boldsymbol{b}}$, and the intra-speaker, inter-trial variance $\Sigma_{\boldsymbol{y}}$. The maximum-likelihood estimates for these parameters can be seen in the output of each model fit:

	F1	F2
μ	630.6	1191.9
$\sigma_{\boldsymbol{b}}$	43.1	100.8
$\sigma_{\boldsymbol{y}}$	16.9	40.1

In this case, there is considerably more inter-speaker variation than intra-speaker variation. Eyeballing Figure 8.6 and comparing it with the parameter estimates above, we can see that the overall mean μ is right at the grand mean of all the observations (this has to occur because the dataset is balanced in terms of cluster size), and that nearly all of the observations lie within two cluster-level standard deviations (i.e. $\sigma_{\boldsymbol{b}}$) of the grand mean.

Conditional estimates of cluster-specific parameters

In the point-estimation approach, we have focused on the parameters of interest—$\boldsymbol{\theta}$ and $\Sigma_{\boldsymbol{b}}$—while maintaining our ignorance about $\boldsymbol{b}$ by marginalizing over it. Nevertheless, in many cases we may be interested in recovering information about $\boldsymbol{b}$ from our model. For example, although the ultimate point of the phonetic study above was to estimate inter-speaker and intra-speaker variation in the pronunciation of [ɑ], we might also be peripherally interested in making inferences about the average pronunciation behavior of the specific individuals who participated in our study. Formally, the point estimates of $\boldsymbol{\theta}$ and $\Sigma_{\boldsymbol{b}}$ determine a conditional probability distribution over $\boldsymbol{b}$. The mode of this distribution is called the BEST LINEAR UNBIASED PREDICTOR (BLUP) $\widehat{\boldsymbol{b}}$:

$$\widehat{\boldsymbol{b}} \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{b}} P(\boldsymbol{b}|\widehat{\boldsymbol{\theta}}, \widehat{\Sigma_{\boldsymbol{b}}}, \boldsymbol{y})$$

The F1 BLUPs for speakers in the current example are plotted as magenta circles in Figure 8.7.

Shrinkage

Another way of estimating speaker-specific averages would simply be to take mean recorded F1 frequency for each speaker. But these two approaches lead to different inferences. Figure 8.7 shows the deviation of each speaker's mean recorded F1 frequency from the grand mean as black squares; recall that the conditional estimate $\widehat{\boldsymbol{b}}$ are magenta circles. Notice that the conditional modes are systematically closer to zero than the means of the raw trials; this effect is more dramatic for speakers with larger deviations. This happens because the finite variance of $\boldsymbol{b}$ in the hierarchical model penalizes large deviations from the grand mean μ. This effect is called SHRINKAGE and is ubiquitous in hierarchical models.

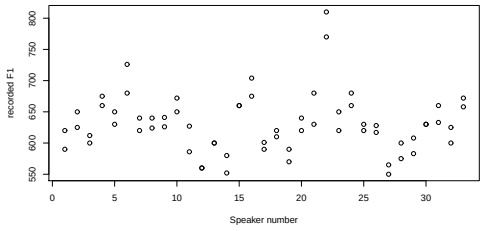


Figure 8.6: Observed F1 measurements by speaker

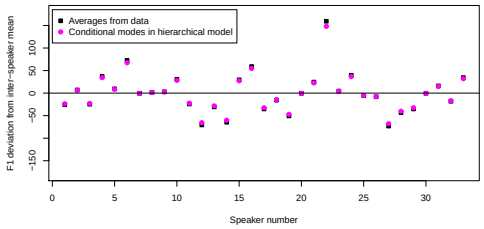


Figure 8.7: Conditional estimates of speaker-specific mean F1 frequencies, and shrinkage

8.2.2 Bayesian posterior inference in hierarchical models

We can compare the point estimates (with standard errors) that we obtained in Section 8.2.1 with posterior estimates obtained using Bayesian inference. We specify the hierarchical model as follows:

$$\begin{aligned}
\mu &\sim \mathcal{N}(0, 10^5) \\
\log \sigma_{\boldsymbol{b}} &\sim \mathcal{U}(-100, 100) \\
\log \sigma_{\boldsymbol{y}} &\sim \mathcal{U}(-100, 100) \\
b_i &\sim \mathcal{N}(0, \sigma^2_{\boldsymbol{b}}) \\
\mu_i &= \mu + b_i \\
y_{ij} &\sim \mathcal{N}(\mu_i, \sigma^2_{\boldsymbol{y}})
\end{aligned} \tag{8.8}$$

This is exactly the same model as in Equation 8.2, with the addition of three extra lines at the top specifying prior distributions over the overall mean μ, inter-speaker variability $\sigma_{\boldsymbol{b}}$, and intra-speaker variability $\sigma_{\boldsymbol{y}}$. There are two important points regarding this model specification. The first is that we are using a normal distribution with very large variance as a prior on the grand mean μ. The normal distribution is conjugate (Section 4.4.3) to the mean of a normal distribution, which has computational advantages; the large variance means that the prior is relatively uninformative, placing little constraint on our inferences about likely values of μ. The second point is how we are defining the priors on the variance parameters $\sigma_{\boldsymbol{b}}$ and $\sigma_{\boldsymbol{y}}$. Although the inverse chi-squared distribution (Section B.4) is conjugate to the variance parameter of a normal distribution, this distribution does not lend itself well to an uninformative specification. As described earlier in Section 4.5, placing a uniform distribution over the log of the standard deviation allows our prior to be uninformative in a "scale-free" sense.[2]

Using the sampling techniques described in Section 4.5, we can obtain approximate highest-posterior density confidence intervals and conditional modes (the latter being ap-

[2]Good discussion of practical choices for priors on variance parameters can be found in Gelman et al. (2004, Appendix C).

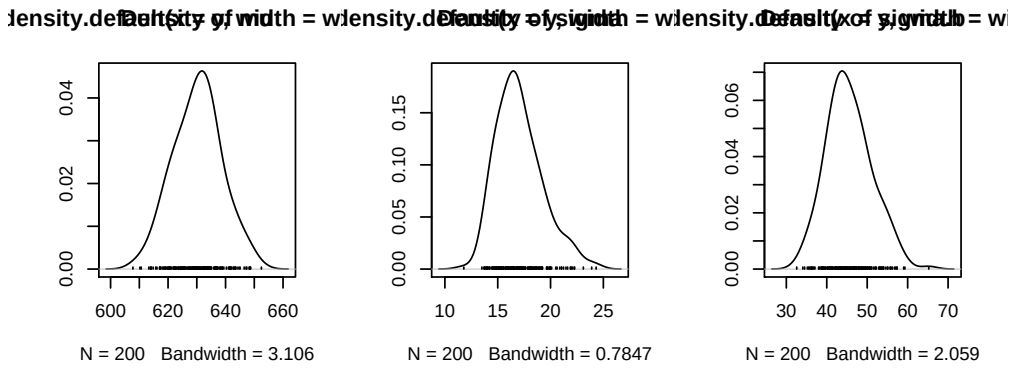


Figure 8.8: Output of MCMC sampling on hierarchical model parameters

proximated by a very narrow HPD interval), and plot estimates of the posterior (Figure 8.8). For F1 formants, these 95% HPD intervals and posterior modes are:

	lower bound	upper bound	posterior mode
μ	615.7	648.7	629.4
$\sigma_{\boldsymbol{b}}$	36	57.4	43.3
$\sigma_{\boldsymbol{y}}$	13.8	21.9	16.4

The posterior simulations are in broad agreement with the point estimates and standard errors obtained in Section 8.2.1. We leave obtaining similar results for F2 as Exercise 8.4.

8.2.3 Multivariate responses

One shortcoming of the analysis of the previous two sections is that F1 and F2 formants were analyzed separately. Correlations between F1 and F2 frequency are captured at neither the inter-speaker nor the intra-speaker level. However, there is a hint of such a correlation in the Figure 8.3. This raises the question of whether such a correlation is reliable at either level. We can address this question directly within a hierarchical model by using bivariate representations of $\boldsymbol{y}$ and $\boldsymbol{b}$. We'll illustrate this type of analysis in a Bayesian framework. The model specification looks similar to the univariate case given in Equation 8.8, but we are using different prior distributions because our normal distributions of interest are multivariate (even though they are still written as $\mathcal{N}$):

$$
\begin{aligned}
b_i &\sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{b}}) \\
\mu_i &= \mu + b_i \\
y_{ij} &\sim \mathcal{N}(\mu_i, \Sigma_{\boldsymbol{y}}) \\
\mu &\sim \mathcal{N}\left(\mathbf{0}, \left[\begin{smallmatrix} 10^5 & 0 \\ 0 & 10^5 \end{smallmatrix}\right]\right) \\
\Sigma_{\boldsymbol{b}} &\sim \mathcal{IW}\left(\left[\begin{smallmatrix} 10^5 & 0 \\ 0 & 10^5 \end{smallmatrix}\right], 2\right) \\
\Sigma_{\boldsymbol{y}} &\sim \mathcal{IW}\left(\left[\begin{smallmatrix} 10^5 & 0 \\ 0 & 10^5 \end{smallmatrix}\right], 2\right)
\end{aligned}
$$

The symbol $\mathcal{IW}$ stands for the inverse Wishart distribution, which is a widely-used prior distribution for covariance matrices; the large diagonal entries in the matrix parameterizing it signal uninformativity, and the zero off-diagonal entries signal that there is no particular prior expectation towards correlation between F1 and F2 deviations. The inverse Wishart distribution is described more completely in Section B.7.

There are eight distinct parameters in our model over which we would like to make inferences: two μ parameters, three $\Sigma_{\boldsymbol{b}}$ parameters, and three $\Sigma_{\boldsymbol{y}}$ parameters (recall from Section 3.5 that $\Sigma_{\boldsymbol{b}}$ and $\Sigma_{\boldsymbol{y}}$ have the form $\left[\begin{smallmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{smallmatrix}\right]$ but that $\sigma_{12} = \sigma_{21}$). Using BUGS once more we obtain the following 95% HPD confidence intervals from the posterior distribution:

	lower bound	higher bound	posterior mode
μ	$\langle 612.4, 1148.8 \rangle$	$\langle 641.8, 1226.3 \rangle$	$\langle 628.4, 1190.7 \rangle$
$\Sigma_{\boldsymbol{b}}$	$\begin{bmatrix} 34.4 & -0.43 \\ -0.43 & 75.8 \end{bmatrix}$	$\begin{bmatrix} 54.8 & 0.3 \\ 0.3 & 133.4 \end{bmatrix}$	$\begin{bmatrix} 42.5 & -0.03 \\ -0.03 & 98.8 \end{bmatrix}$
$\Sigma_{\boldsymbol{y}}$	$\begin{bmatrix} 13.6 & -0.06 \\ -0.06 & 31.5 \end{bmatrix}$	$\begin{bmatrix} 22.4 & 0.53 \\ 0.53 & 52.5 \end{bmatrix}$	$\begin{bmatrix} 17.1 & 0.29 \\ 0.29 & 39.7 \end{bmatrix}$

Of particular interest are the inferences about correlations between F1 and F2 formants, which are the most compelling reason to do multivariate analysis in the first place. In the above analysis, the posterior mode suggests a negative F1-F2 correlation at the inter-speaker level, but positive correlation at the intra-speaker level. However, the confidence intervals on these correlations show that this suggestion is far from conclusive.

It is instructive to compare this analysis to a more straightforward analysis of F1-F2 correlation in which inter-speaker correlation is estimated by calculating speaker means and computing a correlation coefficient on these means, and intra-speaker correlation is estimated obtained by simply subtracting out the speaker-specific mean from each observation and then calculating a correlation coefficient on the resulting residuals. For inter-speaker variation, it gives an empirical correlation coefficient of $r = -0.01$, with a 95% confidence interval of $[-0.35, 0.34]$; for intra-speaker variation, it gives an empirical correlation coefficient of $r = 0.24$, with a 95% confidence interval of $[-0.002, 0.46]$.[3] Although this approach also leads to the conclusion that there is no reliable F1-F2 correlation at either the inter-speaker

[3]This confidence interval is derived by using the transform $z = \frac{1}{2} \log \frac{1+r}{1-r}$; the resulting z is approximately normally distributed (Cohen et al., 2003, p. 45), and so a confidence interval based on the normal distribution can be calculated as described in Section 5.3.

or intra-speaker level, these confidence intervals indicate considerably more certainty in the true correlation than the Bayesian HPD intervals suggest, and the p-value for correlation at the intra-speaker level is very close to significant at 0.052. A key point here is that this latter approach based on empirical speaker means doesn't take into account the uncertainty about true speaker means, and thus leads to conclusions about inter-speaker variation of greater certainty than may be warranted. The Bayesian HPD interval takes this uncertainty into account, and is thus more agnostic about the true correlation.

8.3 Hierarchical linear models

Now we'll move on from hierarchical models of the form in Equation 8.3 to conditional models. Thus we'll be estimating distributions of the form

$$P(Y|X,i) \tag{8.9}$$

where X are the covariates (there can be many of them) and i are the cluster identities. Figure 8.9 illustrates this type of model. The only change from Figure 8.5b is the addition of the covariates X as a separate node in the graph. Once again, the primary targets of inference are typically $\boldsymbol{\theta}$ and Σ, and we'd want to marginalize over $\boldsymbol{b}$ in making our inferences about them.

We'll start with a study of hierarchical linear models. Assume that we have covariates $X_1, \ldots, X_M$ on which we want to condition Y. We can express the j-th outcome in the i-th cluster as

$$y_{ij} = \alpha + b_{i0} + (\beta_1 + b_{i1})X_1 + \cdots + (\beta_M + b_{iM})X_M + \epsilon \tag{8.10}$$

where ϵ is, as before, normally distributed. This equation means that every cluster i has a cluster-specific intercept $\alpha + b_{i0}$ and a slope parameter $\beta_k + b_{ik}$ that determines the contribution of covariate X_k to the mean outcome. In the notation of Figure 8.9, the parameters α and $\{\beta_k\}$, along with the variability $\sigma_{\boldsymbol{y}}$ governing ϵ, are $\boldsymbol{\theta}$ parameters, shared across clusters, whereas the b_{ik} parameters are specific to cluster i. Figure 8.10 shows a slightly more nuanced picture illustrating how the predicted mean mediates the influence of covariates and cluster identity on the outcome; here, only α and $\{\beta_k\}$ are $\boldsymbol{\beta}$ parameters. Equation 8.10 describes the most general case, where all predictors have both shared parameters and cluster-specific parameters. However, the models can be constrained such that some predictors have only shared parameters and some others have only cluster-specific parameters.

8.3.1 Fitting and drawing inferences from a hierarchical linear model: practice

We'll illustrate the utility of hierarchical linear models with a simple instance in which the covariates are categorical. Stanford (2008) investigated variability in the low lexical tone

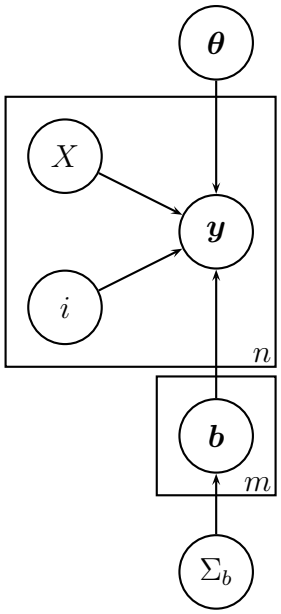


Figure 8.9: A conditional hierarchical model for probability distributions of the form $P(Y|X,i)$

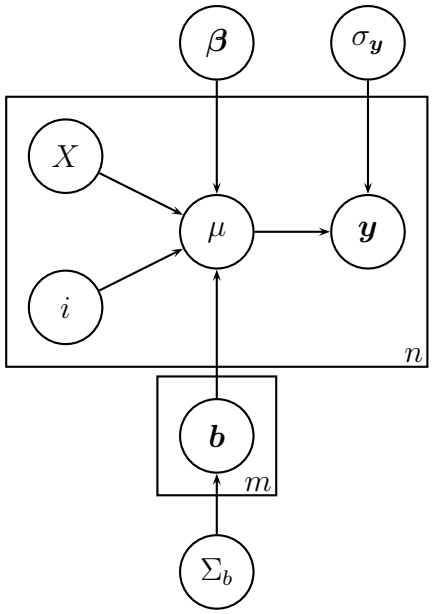
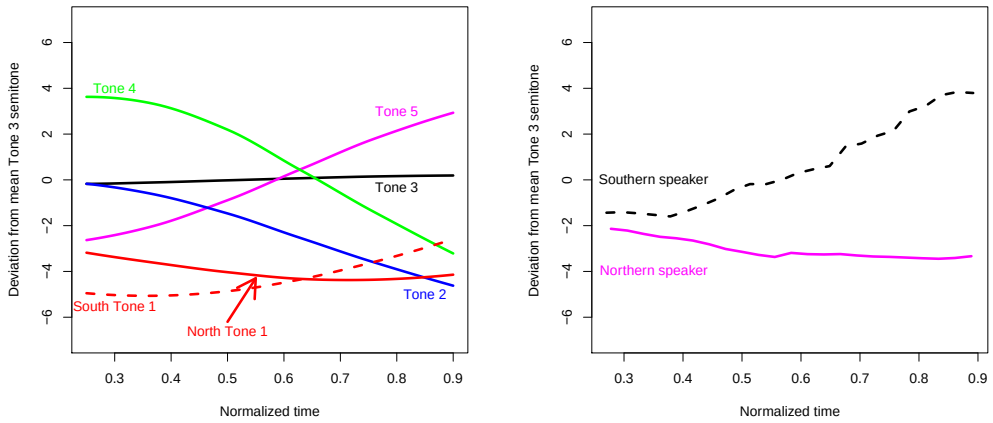


Figure 8.10: A structure more specific to hierarchical linear models, where the influence of cluster identity i and covariates X on the outcome y is only on the predicted mean μ. The shared parameters $\boldsymbol{\theta}$ from Figure 8.9 are explicitly represented here as $\boldsymbol{\beta}$ and $\sigma_{\boldsymbol{y}}$.

contour of Sui, a minority language in Southwest China. The Sui practice clan exogamy: a wife and husband must originate from different clans, and the wife immigrates to the husband's village. Both Northern and Southern Sui clan dialects have six tones, but they have different low-tone (Tone 1) pitch contours. Figure 8.11 illustrates the mean contours for five of the six Sui tones, along with sample tone contours taken from individual recordings of one southern Sui and one northern Sui speaker (who lived in their home village). According to Stanford (2008), the difference in this tone contour is audible but subtle, and the Sui do not mention it as a hallmark of the tone differences between northern and southern clan dialects. Stanford investigated two questions:

1. whether this tone contour can be reliably measured; and
2. whether immigrant Sui speakers adopt the lexical tone contour of their husband's clan, or keep their original tone contour.

We begin with question 1. Stanford observed that in tone 1, from the temporal midway point of each tone to the end, the mean tone contour is fairly straight, but it tends to rise for Southern speakers whereas it stays flat for Northern speakers (Figure 8.11a). Therefore one difference between northern and southern tone 1 contour may be characterizable by the *slope*



(a) Mean Sui tone contours

(b) Two specific trial Sui tone 1 contours for the word *jun*

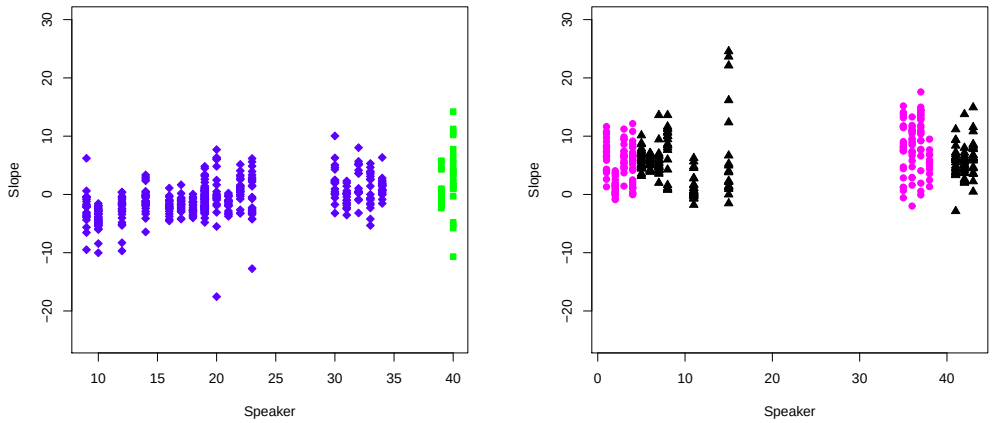
Figure 8.11: Sui tone contours

(in the sense of linear regression) of the tone contour in this part of the syllable. Figure 8.12 plots the distribution of tone contour slopes for each individual trial for Northern-origin and Southern-origin speakers. There is an apparent trend for Northern speakers to have lower slopes. However, there is also an apparent trend for different speakers of each origin to have idiosyncratically different slopes. We could deal with this nesting structure through analysis of variance with speaker as a random factor (Section 6.6), but the data are unbalanced, which is not ideal for analysis of variance. Lack of balance presents no fundamental difficulty for a hierarchical linear model, however.

In our first model of this dataset we include (i) effects on all observations of speaker origin, northward migration, and southward migration; (ii) speaker-specific idiosyncracies in average tone contour; and, of course, (iii) trial-level variability in tone contour. This linear model is thus specified as follows:

$$y_{ij} = \alpha + \beta_1 SO + \beta_2 MN + \beta_3 MS + \underbrace{b_i}_{\sim\mathcal{N}(0,\sigma_b^2)} + \underbrace{\epsilon_{ij}}_{\sim\mathcal{N}(0,\sigma_y^2)} \tag{8.11}$$

where SO is speaker origin (valued 1 for southern origin and 0 for northern origin), MN is migration north and MS migration south (each 1 if the speaker has migrated in that direction, 0 otherwise), and $\boldsymbol{b}$ is a normally-distributed speaker-level slope deviation distributed as $\mathcal{N}(0, \sigma_{\boldsymbol{b}})$. The data to which the model is fitted are shown in Figure 8.12. A maximum-likelihood fit of the parameters for this model is given in Table 8.1. For the shared model parameters (what are often called the "fixed effects" in the "mixed-effects" parlance), considering for the moment only the parameter estimates (and ignoring the standard errors and t statistics) we see that in the maximum-likelihood fit southern origin and northward migration are associated with more sharply upward-sloping tone contour, as visualized in



(a) Northern-origin slope distribution (b) Southern-origin slope distribution

Figure 8.12: Distribution of slopes for speakers originating northern and southern clans. Southward and northward migrants are plotted in green squares and magenta circles respectively.

Figure 8.11, and southward migration is associated with slightly more downward-sloping tone contour. As far as the parameter $\sigma_{\boldsymbol{b}}$ governing inter-speaker variability is concerned, note that its MLE is slightly larger than that of trial-level variability $\sigma_{\boldsymbol{y}}$ and that both these standard deviations are less than half the size of the effect of speaker origin, suggesting that while inter-speaker variability is considerable, the difference between northern-origin and southern-origin speakers is large even compared to this.

8.3.2 Hypothesis testing in hierarchical linear models

Of course, it would also be desirable to measure our certainty in the presence of the effects we just described from reading off the maximum-likelihood estimates in Table 8.1. We begin with the question of how can we assess the contribution of inter-speaker variability in this model. In a frequentist paradigm this can be done via model comparison between models fitted with and without inter-speaker variability, using the likelihood ratio test (Section 5.4.4). REML-fitted likelihoods are generally considered preferable to ML-fitted likelihoods for this purpose (e.g., Morrell, 1998), but in general the p-values obtained by the likelihood-ratio test for models differing in the number of parameters governing inter-cluster variability ("random-effects" structure, or $\Sigma_{\boldsymbol{b}}$) are CONSERVATIVE (Stram and Lee, 1994; Pinheiro and Bates, 2000; Baayen et al., 2008), meaning that the true p-value will generally be smaller (i.e. more significant) than the p-value obtained by consulting the χ^2 distribution. Conservativity is good in the sense that an obtained significant p-value can be trusted, but dangerous in the sense that a large (i.e. insignificant) p-value is not necessarily grounds to exclude inter-cluster variability from the model. Whether it is a better idea to err

	$\hat{\beta}_{ML}$	$SE(\hat{\beta}_{ML})$	t_{ML}	$\hat{\beta}_{REML}$	$SE(\hat{\beta}_{REML})$	t_{REML}
$\sigma_{\boldsymbol{b}}$	3.35			3.81		
$\sigma_{\boldsymbol{y}}$	3.07			3.07		
Intercept	-0.72	0.47	-1.53	-0.72	0.5	-1.44
SO	7.21	0.83	8.67	7.21	0.88	8.17
MN	2.38	1.44	1.65	2.38	1.53	1.56
MS	-0.82	0.94	-0.87	-0.82	1	-0.82

Table 8.1: Shared parameter estimates ($\hat{\beta}$), standard errors ($SE(\hat{\beta})$), and t statistics (defined as $\frac{\hat{\beta}}{SE(\hat{\beta})}$) for the Sui tone model defined in Equation (8.11). Note that standard errors are not appropriate for estimates of $\sigma_{\boldsymbol{b}}$ or $\sigma_{\boldsymbol{y}}$, as these are not normally distributed.

on the side of including or excluding parameters for inter-cluster variability when in doubt will depend on the precise goals of one's modeling, and we will have more to say about it in Section XXX.

To illustrate this type of hypothesis test on cluster-level parameters, we construct a "null hypothesis" version of our original Sui tone model from Equation (8.11) differing only in the absence of idiosyncratic speaker-level variability $\boldsymbol{b}$:

$$y_{ij} = \alpha + \beta_1 SO + \beta_2 MN + \beta_3 MS + \underbrace{\epsilon_{ij}}_{\sim\mathcal{N}(0,\sigma_y^2)} \tag{8.12}$$

and we can call this new model M_0 and the original model M_1. The log-likelihood of the REML fit of M_0 turns out to be -2403.8 whereas the log-likelihood for M_1 is -2306.2. Consulting twice the difference of these log-likelihoods against the χ^2_1 distribution, we find that the improvement of M_1 over M_0 is extremely unlikely under M_0 ($p \ll 0.001$). Broadly consistent with the visual picture in Figure 8.12 and with the results in Table 8.1, there is extremely strong evidence that speakers vary idiosyncratically in their average tone 1 contours, above and beyond the other sources of cross-speaker variability in the model (origin and migration).

We now move to the question of hypothesis testing involving shared model parameters ("fixed effects" structure, or θ). Perhaps surprisingly, how to test for significance of an effect of a shared model parameter is a matter of some current controversy. In principle, one could use the likelihood ratio test to compare a more complex model with a simpler model. However, it has been argued that the likelihood ratio test will lead to ANTI-CONSERVATIVE p-values (i.e. the true p-value is less significant than the p-value obtained by consulting the χ^2 distribution) for comparison of models with the same cluster-level parameters but different shared parameters (Pinheiro and Bates, 2000, pp. 87–92).[4] This leaves two approaches currently in vogue. On the first approach, a single model is fitted with the method of maximum

[4]As a caveat, it is not clear to this author that the anti-conservativity involved is appreciable unless the number of total parameters in the model is quite large relative to the number of observations, which quite often is *not* the case for linguistic datasets.

likelihood, and for the shared parameter of interest, the parameter estimate and its standard error are used to obtain a p-value based on the t statistic, just as in standard linear regression (Section 6.4; for testing multiple parameters simultaneously, an F-test is used). This approach itself carries a degree of controversy involving how many degrees of freedom the relevant t distribution should be assumed to have. As a rule of thumb, however, if there are many more observations than model parameters, the t distribution is generally taken to be approximated by the standard normal distribution (see also Section B.5). This is illustrated in Table 8.1, which gives the MLEs, standard errors, and resulting t statistics for our four parameters of interest. Recall that the standard normal distribution has just over 95% of its probability mass in the interval [-2,2] (e.g., Section 5.3), so that finding $|t| > 2$ is roughly a $p < 0.05$ result. Thus we conclude from the ML fit.

The second approach currently in vogue is to use Bayesian inference and bypass the classical hypothesis testing paradigm altogether. Instead, one can estimate a Bayesian confidence region (Section 5.1) on the shared parameters θ, by sampling from the posterior distribution over $\boldsymbol{\theta}$ using Markov Chain Monte Carlo (MCMC) sampling as discussed in Section 4.5, earlier in this chapter, and in Appendix **??**. Here we illustrate this approach by sampling from the posterior in Figure 8.10 over $P(\boldsymbol{\beta}|\boldsymbol{y}, \widehat{\Sigma_{\boldsymbol{b}}}, \widehat{\sigma_{\boldsymbol{y}}})$, performed by the `lme4` package and giving us the following 95% HPD confidence intervals and posterior modes:

	lower bound	upper bound	posterior mode
α	-1.56	0.17	-1.32
SO	5.72	8.58	7.45
MN	-0.06	5.05	2.46
MS	-2.36	1.06	-0.89

On the Bayesian interpretation, we can be over 95% certain that the true parameter estimate for the effect of being from the south (with respect to the reference level of clan origin, the north) is positive. It has recently become popular in the psycholinguistics literature to call the largest value q such that $1 - \frac{q}{2}$ of the posterior probability mass on a parameter θ lies on one side of zero a "MCMC-based p-value" for θ (Baayen, 2008). Although this value q is certainly a useful heuristic for assessing the strength of the evidence supporting a meaningful role for θ in the model, it is also worth keeping in mind that this value q is NOT a p-value in the traditional Neyman-Pearson paradigm sense of the term.

Simultaneous assessment of significance of multiple parameters

We now turn to another question: whether migration into a new clan has a reliable effect on tone 1 slope. If it did, then the theoretically sensible prediction would be that migration of a southern-origin woman to the north should tend to lower the slope, and migration of a northern woman to the south should tend to raise the slope. To test this possibility, we can consult model M_1, whose parameters associated with variables MN and MS encode this effect. Looking at Table 8.1, we see that both coefficients associated with migration are consistent with the theoretical prediction. There is considerable uncertainty about each parameter (as indicated by the t-value), but what we would really like to do is to assess the overall explanatory benefit accrued by introducing them together. The conceptually simplest

way to do this is to use the F statistic from model comparison between M_1 and a simpler model M_0' in which effects of MN and MS are absent. Computing an F-test between these two models we obtain an F statistic of 1.743. In order to evaluate statistical significance, we need to choose which F distribution should serve as the reference distribution, but as noted earlier, there is some controversy as to how many degrees of freedom to use in the *denominator* for such an F statistic. However, we can play the devil's advocate momentarily and ask how much evidence would exist for an effect of migration in the most optimistic interpretation. The maximum degrees of freedom for the F statistic denominator is the number of observations minus the number of parameters in the full model, or 887. The cumulative distribution function for $F_{2,887}$ at 1.743 is 0.824, hence the best-case p-value is 0.176, and the overall effect is thus marginal at best.

8.3.3 Heteroscedasticity across clusters

One noteworthy thing about Figure 8.12 is that some speakers clearly have more inter-trial variability than others (compare, for example, speakers 15 and 11). This presence of inter-cluster differences in residual variability is called HETEROSCEDASTICITY. (The lack of heteroscedasticity—when residual intra-cluster variability is the same for all clusters—is called HOMOSCEDASTICITY.) Although the differences are not particularly severe in this case, we can still investigate whether they affect our inferences by incorporating them into our model. Conceptually speaking, this is a minor change to the structure of the model as depicted in Figure 8.9: the residual variance σ_y moves from being a shared θ parameter to being a cluster-specific $\boldsymbol{b}$ parameter. We present a Bayesian analysis (once again because methods for point-estimation are not readily available), with the following model:

$$\begin{aligned}
\alpha, \beta_{\{1,2,3\}} &\sim \mathcal{N}(0, 10^5) \\
\mu_i &= \alpha + \beta_1 SO + \beta_2 MN + \beta_3 MS + b_i \\
b_i &\sim \mathcal{N}(0, \sigma_b) \\
y_{ij} &\sim \mathcal{N}(\mu_i, \sigma_{\boldsymbol{y},i}) \\
\log \sigma_b &\sim \mathcal{U}(-100, 100) \\
\log \sigma_{\boldsymbol{y},i} &\sim \mathcal{U}(-100, 100)
\end{aligned}$$

Approximate 95% HPD confidence intervals and posterior modes obtained from sampling look as follows:

	lower bound	upper bound	posterior mode
α	-1.68	-0.12	-1.04
SO	5.82	8.65	6.78
MN	-2.81	0.95	-1.41
MS	-1.09	4.92	2.88

Comparing these results with the point-estimate results obtained in the previous section, we see that failing to account for heteroscedasticity doesn't qualitatively change the conclusions

of the model: there is still a strong association of clan origin with tone 1 slope, and there are no reliable effects of migration. Once again, similar inferences from multiple model specifications should strengthen your confidence in the conclusions obtained.

8.3.4 Multiple clusters per observation

One of the most exciting new developments in hierarchical modeling has been improvement in the computational treatment of cases where there are multiple classes of cluster to which each observation belongs. Consider the typical psycholinguistics experiment in speech perception or language comprehension, where each observation is derived from a particular participant reading or listening to a particular experimental stimulus which appears in a certain form. For example, the self-paced reading experiment of Rohde et al. (2011), described in Section 6.6.6 involved 58 subjects each reading 20 sentences (ITEMS), where each sentence could appear in any of four possible forms. The sample sentence is repeated below:

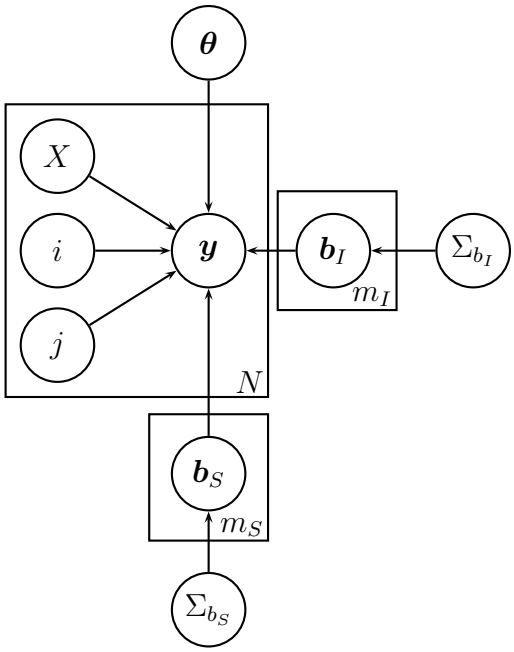
(1) John {detests/babysits} the children of the musician who {is/are} generally arrogant and rude.

where there are two experimentally manipulated predictors: the type of verb used (implicit causality (IC) or non-IC), and the level of relative-clause attachment (high or low). This corresponds to a more complex hierarchical model structure, shown in Figure 8.13. In this figure, there are two cluster identity nodes i and j; the subject-specific effects for the i-th subject are denoted by $\boldsymbol{b}_{S,i}$, and the item-specific effects for the j-th item are denoted by $\boldsymbol{b}_{I,j}$. This type of model is conceptually no different than the simpler hierarchical models we have dealt with so far. We illustrate by replicating the analysis of variance performed in Section 6.6.6 using a hierarchical linear model. Because the interaction between RC attachment level and verb type is of major interest to us, it is critical to us that we use an appropriate contrast coding (Section **??**), using predictors V to represent verb type, with values 0.5 and -0.5 for IC and non-IC verb types, and A to represent attachment level, with values 0.5 and -0.5 for high and low attachment. We will allow different subject- and item-specific effects for each of the four possible conditions C. The hierarchical model can be compactly written as follows:

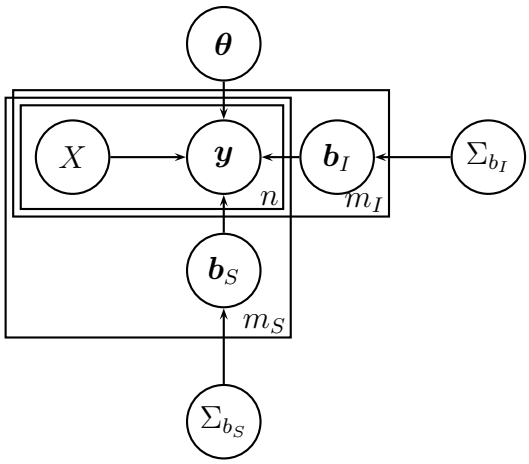
$$\begin{aligned}
\boldsymbol{b}_{S,i} &\sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{b}_S}) \\
\boldsymbol{b}_{I,j} &\sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{b}_I}) \\
\mu_{ij} &= \alpha + \beta_V V + \beta_A A + \beta_{VA} VA + \boldsymbol{b}_{S,i} C + \boldsymbol{b}_{I,j} C \\
y_{ijk} &\sim \mathcal{N}(\mu_{ij}, \sigma_{\boldsymbol{y}})
\end{aligned} \tag{8.13}$$

We'll start by using point estimation of the model parameters using (unrestricted) maximum likelihood. There is a considerable amount of detail in the resulting model so we present the complete printed representation of the fitted model from `R`:

```
Linear mixed model fit by maximum likelihood
Formula: rt ~ V * A + ((C - 1) | subj) + ((C - 1) | item)
```



(a) Crossed-clustering graphical model in which cluster identities are explicitly represented



(b) Crossed-clustering graphical model in which plates are overlapping/nested and cluster identity variables are implicit

Figure 8.13: A conditional hierarchical model for probability distributions of the form $P(Y|X,i,j)$, with observations cross-classified into two classes of clusters. The two graphical models are equivalent; in Figure 8.13a plates are non-overlapping and cluster identity variables are explicitly represented, whereas in Figure 8.13b cluster plates are overlapping, the observation plate is nested in both, and cluster identity variables are implicit

```
   Data: d
   AIC   BIC logLik deviance REMLdev
 12527 12648  -6239    12477   12446
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 subj     CIC high    11971.19 109.413
          CIC low     18983.98 137.782  0.909
          CnonIC high 23272.40 152.553  0.883 0.998
          CnonIC low  16473.41 128.349  0.988 0.963 0.946
 item     CIC high      657.90  25.650
          CIC low       563.35  23.735  1.000
          CnonIC high  6062.73  77.864  0.271 0.271
          CnonIC low   3873.88  62.241  0.991 0.991 0.399
 Residual             38502.59 196.221
Number of obs: 919, groups: subj, 55; item, 20
```

$$\Sigma_{\boldsymbol{b}_S} = \begin{bmatrix} 109.41 & 0.91 & 0.88 & 0.99 \\ 0.91 & 137.78 & 1 & 0.96 \\ 0.88 & 1 & 152.55 & 0.95 \\ 0.99 & 0.96 & 0.95 & 128.35 \end{bmatrix} \qquad \Sigma_{\boldsymbol{b}_I} = \begin{bmatrix} 25.65 & 1 & 0.27 & 0.99 \\ 1 & 23.73 & 0.27 & 0.99 \\ 0.27 & 0.27 & 77.86 & 0.4 \\ 0.99 & 0.99 & 0.4 & 62.24 \end{bmatrix}$$

Parameter	Associated Predictor	$\hat{\beta}_{ML}$	$SE[\hat{\beta}]_{ML}$	t_{ML}
α	Intercept	470.48	20.6	22.84
β_V	Verb type (Implicit Causality=0.5, not=-0.5)	-33.71	16.78	-2.01
β_A	Relative clause (RC) attachment (high=0.5, not=-0.5)	-0.42	15.69	-0.03
β_{VA}	Verb type/RC attachment interaction	-85.31	35	-2.44

Table 8.2: Cross-classified cluster (experimental participant and item) hierarchical linear model for Rohde et al. (2011) self-paced reading study. For ease of interpretation, $\Sigma_{\boldsymbol{b}_S}$ and $\Sigma_{\boldsymbol{b}_I}$ are presented with standard deviations on the diagonal entries and correlation coefficients on the non-diagonal entries.

```
Fixed effects:
            Estimate Std. Error t value
(Intercept) 470.4823    20.6029  22.836
V           -33.7094    16.7787  -2.009
A            -0.4173    15.6899  -0.027
V:A         -85.3055    34.9994  -2.437
```

The negative estimates for β_V and β_A indicate that the IC-verb and high-attachment conditions respectively are associated with faster reading time at this region, though the effect of attachment level is very small. VA is positive in the IC/high and non-IC/low conditions, so the negative estimate for β_{VA} indicates that reading times at this region are faster in these two conditions.

We now turn to assessment of statistical significance. Once again resorting to our rule of thumb that with many more observations than estimated parameters (13 versus 919 in our case), the t statistic is distributed approximately as the standard normal, we see that the hierarchical analysis leads us to the same conclusions as the ANOVA of Section 6.6.6: there is a reliable interaction between verb type and RC attachment level in reading times at the first spillover region (*generally* in Example I). Importantly, this result is now obtained from a single hypothesis test rather than from separate by-subjects and by-items tests as has been the tradition in psycholinguistics for the past three decades.[5]

[5]Single hypothesis tests can also be done with separate by-subjects and by-items ANOVAs using the *min-F* test introduced by Clark (1973), this test is quite conservative and in practice is very rarely used.

Interpreting model parameter estimates

The ability to do a single hypothesis test when observations belong to multiple cross-cutting clusters is an advantage of the hierarchical analysis approach. However, where the hierarchical approach really shines is in obtaining a single model that can be inspected, interpreted, and used. From the subject-level variance parameters, we can see that there is much more intersubjective variation in reading speed for each condition (the subject-intercept standard deviations range from 111 to 154) than there is inter-item variability (standard deviations between 26 and 81). The residual trial-level variability is larger than both the subject- and item-level variability put together. Beyond this, however, there is something else worth noticing. The correlations between subject-level parameters for each condition are extremely high. That is, when a subject reads slowly in one condition, he or she reads slowly in *all* conditions. The correlations between item-level parameters are also high, except that there is much lower correlation between the implicit-causality, high-attachment condition and the rest of the conditions. This result could be illuminating if item-level parameter estimates were extracted and compared with the experimental materials themselves.

Turning to the shared parameters ("fixed effects"), we see that there is an overall 34-millisecond speed advantage for the implicit-causality verbs at this point, but no real overall effect of attachment level. Above and beyond these main effects, there is a large (85-ms) advantage for high-attaching RCs after IC verbs over low-attaching RCs after non-IC verbs. These estimates comport fairly well with those derived from the per-condition means obtained in Section 6.6.6, but the estimates obtained here put inter-subject and inter-item variation on equal footing, and are obtained automatically as part of the model-fitting process.

Fully Bayesian analysis

We can try a similar analysis using fully Bayesian techniques rather than the point estimate. We'll present a slightly simpler model in which the speaker- and item-specific variations do not depend on condition; this simpler type of model is often called a model with *random subject- and item-specific intercepts* in the literature. A Bayesian version of the more complex model of the previous section is left to the reader (see Exercise 8.8). Here is the model specification:

$$
\begin{aligned}
\alpha, \beta_{\{1,2,3\}} &\sim \mathcal{N}(0, 10^5) \\
\log \sigma_{b_S} &\sim \mathcal{U}(-100, 100) \\
\log \sigma_{b_I} &\sim \mathcal{U}(-100, 100) \\
\log \sigma_{b_y} &\sim \mathcal{U}(-100, 100) \\
b_{S,i} &\sim \mathcal{N}(0, \sigma_{b_S}) \\
b_{I,j} &\sim \mathcal{N}(0, \sigma_{b_I}) \\
\mu_{ij} &= \alpha + \beta_1 V + \beta_2 A + \beta_3 VA + b_{S,i} + b_{I,j} \\
y_{ijk} &\sim \mathcal{N}(\mu_{ij}, \sigma_y)
\end{aligned}
$$

Note the close resemblance to the previous model specification in Equation 8.13; the two differences are the addition of the top four lines representing priors over the shared parameters α and $\beta_{\{1,2,3\}}$, and the simplified $b_{S,i}$ and $b_{I,j}$ since we only have subject- and item-specific intercepts now.

Sampling from the posterior gives us the following 95% HPD confidence intervals and posterior modes for the effects of V, A, and the interaction VA:

	lower bound	upper bound	posterior mode
V	-56.87	-1.43	-41.75
A	-26.34	23.79	-5.69
VA	-136.52	-20.65	-96.3

Once again, there is broad agreement between the point estimates obtained earlier in this section and the Bayesian HPD confidence intervals. Most notably, there is (a) strong evidence of an overall trend toward faster reading in this region for the IC verbs; and (b) even stronger evidence for an interaction between IC verb type and attachment level. One should believe an apparent trend in the data more strongly if the same trend is confirmed across multiple statistical analyses, as is the case here.

8.4 Hierarchical generalized linear models

We now shift from linear models to the broader case of generalized linear models, focusing on logit models since they (along with linear models) are the most widely used GLM in the study of language. We move from generalized linear models (GLMs) to hierarchical GLMs by adding a stochastic component to the linear predictor (c.f. Equation 6.1):

$$
\eta = \alpha + (\beta_1 + b_{i1})X_1 + \cdots + (\beta_n + b_{in})X_n \tag{8.14}
$$

and assume that the cluster-specific parameters $\boldsymbol{b}$ themselves follow some distribution parameterized by $\Sigma_{\boldsymbol{b}}$.

We then follow the rest of the strategy laid out in Section 6.1 for constructing a GLM: choosing a link function $\eta = l(\mu)$, and then choosing a function for noise around μ.

8.4.1 Hierarchical logit models

In a hierarchical logit model, we simply embed the stochastic linear predictor in the binomial error function (recall that in this case, the predicted mean μ corresponds to the binomial parameter π):

$$P(y;\mu) = \binom{n}{yn}\mu^{yn}(1-\mu)^{(1-y)n} \qquad \text{(Binomial error distribution)} \tag{8.15}$$

$$\log\frac{\mu}{1-\mu} = \eta \qquad \text{(Logit link)} \tag{8.16}$$

$$\mu = \frac{e^{\eta}}{1+e^{\eta}} \qquad \text{(Inverse logit function)} \tag{8.17}$$

8.4.2 Fitting and interpreting hierarchical logit models

As with hierarchical linear models, the likelihood function for a multi-level logit model must marginalize over the cluster-level parameters $\boldsymbol{b}$ (Equation 8.5). We can take either the maximum-likelihood approach or a Bayesian approach. Unlike the case with hierarchical linear models, however, the likelihood of the data $P(\boldsymbol{y}|\theta,\Sigma_{\boldsymbol{b}})$ (marginalizing over cluster-level parameters $\boldsymbol{b}$) cannot be evaluated exactly and thus the MLE must be approximated (Bates, 2007, Section 9). The tool of choice for approximate maximum-likelihood estimation is once again the `lme4` package in `R`.[6]

8.4.3 An example

We return to the dataset of Bresnan et al. (2007), illustrated by the alternation

(2) Susan gave *toys* **to the children**. (PP realization of recipient)

(3) Susan gave **the children** *toys*. (NP realization of recipient)

To illustrate the approach, we construct a model with the length, animacy, discourse accessibility, pronominality, and definiteness of both the recipient and theme arguments as predictors, and with verb as a random effect. We use log-transformed length predictors (see Section 6.7.4 for discussion).

Defining the model

We arbitrarily denote length, animacy, discourse status, pronominality, and definiteness of the theme with the variables L_T, A_T, S_T, P_T, D_T respectively, and those properties of the

[6] The recommended approximations to maximum likelihood are Laplacian approximation (see, e.g., Robert and Casella (2004, Section 3.4)) and adaptive Gaussian quadrature; the former is available in `lme4` and the recommended default, but the latter isn't (yet).

	$\hat{\beta}$	$SE(\hat{\beta})$	z
$\sigma_{\boldsymbol{b}}$	2.33		
Intercept	2.32	0.66 3.51	
log Recipient Length	1.31	0.15	8.64
log Theme Length	-1.17	0.11	-10.97
Recipient Animacy	2.14	0.25	8.43
Theme Animacy	-0.92	0.5	-1.85
Recipient Discourse Status	1.33	0.21	6.44
Theme Discourse Status	-1.76	0.27	-6.49
Recipient Pronominality	-1.54	0.23	-6.71
Theme Pronominality	2.2	0.26	8.37
Recipient Definiteness	0.8	0.2	3.97
Theme Definiteness	-1.09	0.2	-5.49

Figure 8.14: Point estimate for a hierarchical logit model of the dative alternation

recipient as L_R, A_R, S_R, P_R, D_R.[7] We can now write our hierarchical model as follows:

$$\begin{aligned}
b_i &\sim N(0, \sigma_b) \\
\eta_i &= \alpha + \beta_{L_T} L_T + \beta_{A_T} A_T + \beta_{S_T} S_T + \beta_{P_T} P_T + \beta_{D_T} D_T \\
&\quad + \beta_{L_R} L_R + \beta_{A_R} A_R + \beta_{S_R} S_R + \beta_{P_R} P_R + \beta_{D_R} D_R + b_i \\
\pi_i &= \frac{e_i^\eta}{1 + e_i^\eta} \\
y_{ij} &\sim Binom(1, \pi_i)
\end{aligned} \tag{8.18}$$

(Note that we could equally specify the last line as a Bernoulli distribution: $y_{ij} \sim Bern(\pi)$.) We arbitrarily consider prepositional-object (PO) realization of the recipient as the "successful" outcome (with which positive contributions to the linear predictor η will be associated). Approximate maximum-likelihood estimation gives us the following parameter estimates: The fixed-effect coefficients, standard errors, and Wald z-values can be interpreted as normal in a logistic regression (Section 6.7.1). It is important to note that there is considerable variance in verb-specific preferences for PO versus DO realizations. The scale of the random effect is that of the linear predictor, and if we consult the logistic curve we can see that a standard deviation of 2.33 means that it would be quite typical for the magnitude of this random effect to be the difference between a PO response probability of 0.1 and 0.5.

We now turn our attention to the shared model parameters. The following properties are associated with PO outcomes:

[7] To simplify model interpretation, I have reduced the tripartite distinction of Bresnan et al. of discourse status as `given`, `accessible`, and `new` into a binary distinction of `given` versus `new`, with `accessible` being lumped together with `new`.

- Longer recipients
- Inanimate recipients
- Discourse-new recipients
- Non-pronominal recipients
- Indefinite recipients

- Shorter themes
- Animate themes
- Discourse-old themes
- Pronominal themes
- Definite themes

There is a clear trend here: those properties of the theme that favor PO outcomes are the reverse of those properties that favor DO outcomes. This raises the linguistically interesting possibility that there is a unified set of principles that applies to word ordering preferences in the English postverbal domain and which is sensitive to high-level properties of constituents such as length, discourse status, and so forth, but *not* to specific combinations of these properties with the grammatical functions of the constituents. This possibility is followed up on in Exercise 8.10.

Inferences on cluster-specific parameters

Because of this considerable variance of the effect of verb, it is worth considering the inferences that we can make regarding verb-specific contributions to the linear predictor. One way of doing this would be to look at the conditional modes of the distribution on verb-specific effects $\boldsymbol{b}$, that is to say the BLUPs (Section 8.2.1). There is a disadvantage to this approach, however: there is no easy way to assess our degree of confidence in the conditional modes. Another option is to use a fully Bayesian approach and plot posterior modes (of $P(\boldsymbol{b}|y, \Sigma_{\sigma_b}, \Sigma_\theta)$, which is different from the BLUPs) along with confidence intervals. This is the approach taken in Figure 8.15, using uniform priors on all the shared parameters as well as on $\log \sigma_{\boldsymbol{b}}$. On the labels axis, each verb is followed by its SUPPORT: the number of instances in which it appears in the `dative` dataset. For most verbs, we do not have enough information to tell whether it truly has a preference (within the model specification) toward one realization or the other. However, we do have reliable inferences some verbs: for the most part, those with large support and/or with posterior modes far from 0.[8] We can see that *tell*, *teach*, *charge*, and *show* are strongly biased toward the double-object construction, whereas *loan*, *bring*, *sell*, and *take* are biased toward the prepositional-object construction.

These results are theoretically interesting because the dative alternation has been at the crux of a multifaceted debate that includes:

- whether the alternation is meaning-invariant;
- if it is not meaning-invariant, whether the alternants are best handled via constructional or lexicalist models;

[8] This is not the whole story, however: comparing *deny* with *promise*, the former has both larger support and a more extreme posterior mode, but it is the latter that has an HPD confidence interval that is closer to not including 0.

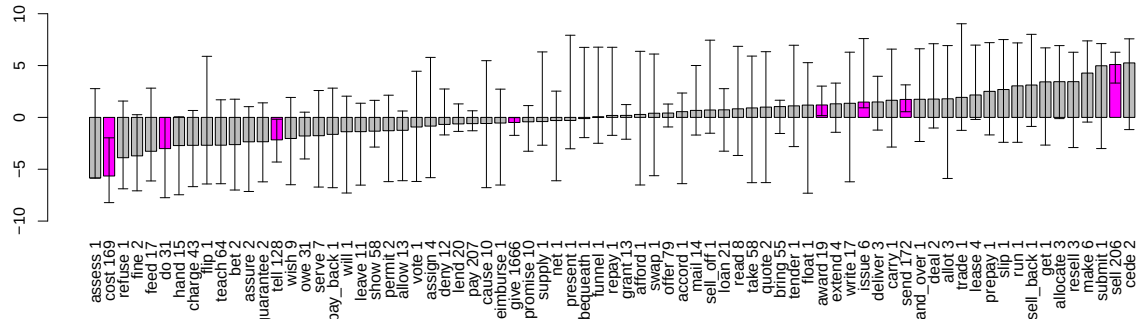


Figure 8.15: Verb-specific preferences in analysis of the `dative` dataset. 95% HPD confidence intervals are plotted on each preference. Verbs for which the 95% HPD interval is entirely on one side of the origin have their bars plotted in magenta.

- whether verb-specific preferences observable in terms of raw frequency truly have their locus at the verb, or can be explained away by other properties of the individual clauses at issue.

Because verb-specific preferences in this model play such a strong role despite the fact that many other factors are controlled for, we are on better footing to reject the alternative raised by the third bullet above that verb-specific preferences can be entirely explained away by other properties of the individual clauses. Of course, it is always possible that there are other explanatory factors correlated with verb identity that will completely explain away verb-specific preferences; but this is the nature of any type of scientific explanation. (This is also a situation where controlled, designed experiments can play an important role by eliminating the correlations between predictors.)

8.4.4 Model comparison & hypothesis testing

The framework for hypothesis testing in hierarchical generalized linear models is similar overall to that for hierarchical linear models as described in Section 8.3.2. For model comparisons involving the same shared-parameter structure but nested cluster-specific parameter structures, likelihood-ratio tests are conservative. For the assessment of the significance of a single shared parameter estimate $\hat{\beta}_i$ against the null hypothesis that $\beta_i = 0$, the Wald z-statistic (Section 6.8.1), which is approximately standard-normal distributed under the null hypothesis, can be used. In Figure 8.14, for example, the z-statistic for log of recipient length is $\frac{1.31}{0.15} = 8.64$, which is extremely unlikely under the null hypothesis.

To simultaneously assess the significance of the contribution of more than one shared parameter to a model, a likelihood-ratio test is most appropriate. Once again, however, in hierarchical models this test may be anti-conservative, as described in Section 8.3.2, so caution should be used in interpreting apparently significant results. As an example of how this test can be useful, however, let us consider an alternative model of the dative alternation in

which the tripartite discourse status of recipients and themes into given, accessible, and new (Collins, 1995). This difference introduces two new parameters into the model. Twice the difference in the log-likelihoods of the original and the updated model is 3.93; the cumulative distribution function for χ^2_2 at this point is 0.86, giving us a best-case p-value of 0.14, so we can safely state that we don't have sufficient evidence to adopt the tripartite distinction over the bipartite given/new distinction used earlier.

8.4.5 Assessing the overall performance of a hierarchical logit model

We conclude the chapter with a brief discussion of assessing overall performance of hierarchical logit models. As with any probabilistic model, the data likelihood is an essential measure of model quality, and different candidate models can be compared by assessing the likelihood they assign to a dataset. Cross-validated or held-out likelihood can be used to alleviate concerns about overfitting (Section 2.11.5). It is also useful to visualize the model's performance by plotting predictions against empirical data. When assessing the fit of a model whose response is continuous, a plot of the residuals is always useful. This is not a sensible strategy for assessing the fit of a model whose response is categorical. Something that is often done instead is to plot *predicted probability* against *observed proportion* for some binning of the data. This is shown in Figure 8.16 for 20 evenly-spaced bins on the x-axis, with the point representing each bin of size proportionate to the number of observations summarized in that point. There is a substantial amount of information in this plot: the model is quite certain about the expected outcome for most observations, and the worst-outlying bins are between predicted probability of 0.7 and 0.8, but contain relatively little data. Producing this visualization could be followed up by examining those examples to see if they contain any important patterns not captured in the model. Probability-proportion plots can also be constructed using cross-validation (Exercise 8.13).

8.5 Further Reading

Hierarchical models are an area of tremendous activity in statistics and artificial intelligence. There is good theoretical coverage (and some examples) of hierarchical generalized linear models in Agresti (2002, Chapter 12). Pinheiro and Bates (2000) is an important book on theory and practice for linear and non-linear hierarchical models from the frequentist perspective. There is also a bit of `R`-specific coverage in Venables and Ripley (2002, Section 10.4) which is useful to read as a set of applied examples, but the code they present uses penalized quasi-likelihood estimation and this is outdated by `lme4`. A more recent and comprehensive text for hierarchical regression models is Gelman and Hill (2007), which focuses the Bayesian perspective but is practically oriented, and includes coverage of both `lme4` and `BUGS`. At the time of writing, this is probably the single best place to turn to when learning the practicalities of working with hierarchical models for the analysis of complex datasets.

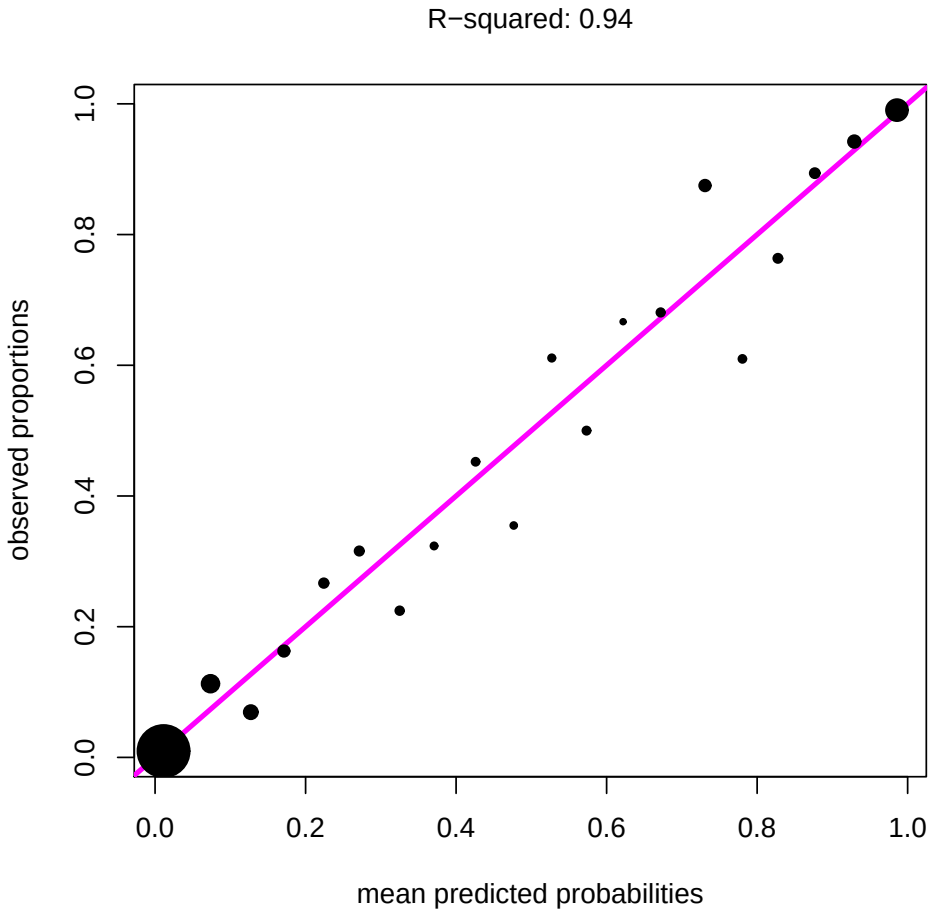


Figure 8.16: The fit between predicted probabilities and observed proportions for each 5% of predicted probability for our model

8.6 Exercises

Exercise 8.1: Defining a hierarchical model

The F1 formant levels of children's vowels from Peterson & Barney's dataset tend not to be normally distributed, but are often right-skewed (Figure 8.17). Define a hierarchical probabilistic model of F1 formant values for the vowel [a] in which the speaker-specific variation component b_i are log-normally distributed—that is, if $\log b_i = x_i$, then the x_i are normally distributed with some mean and standard deviation. Write the hierarchical model in the style of Equation 8.2. Choose parameters for your model by hand that generate data that looks qualitatively like that of Figure 8.17.

Exercise 8.2: Restricted maximum likelihood and testing "fixed effects"

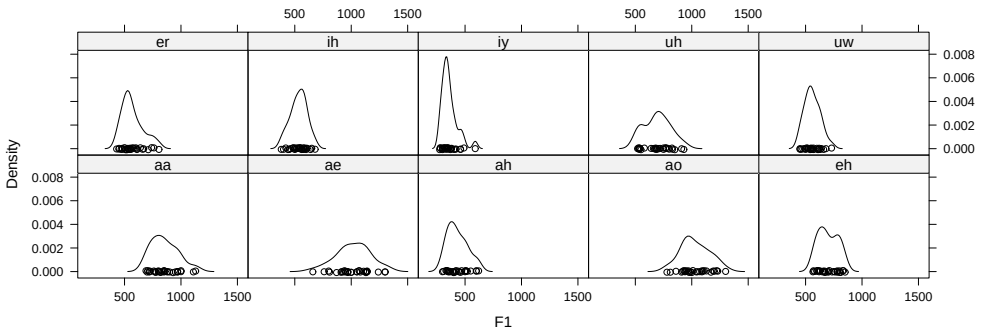


Figure 8.17: Density plots of children's F1 formant recordings from Peterson & Barney

In Section 8.3.2 I stated that one should not use ordinary (unrestricted) maximum likelihood, not REML, in frequentist test on models differing in what is often called "fixed-effects" structure (i.e., models differing only in the shared parameters $\boldsymbol{\theta}$). Can you think of any reasons why REML comparisons would not be a good idea for this purpose?

Exercise 8.3: Shrinkage

Consider the simple point estimate hierarchical model of [ɑ] first-formant (F1) data from Peterson and Barney (1952) obtained in Section 8.2.1. In this model, calculate the joint log-likelihood of $\widehat{\boldsymbol{b}}$ and $\boldsymbol{y}$—that is, of speaker-specific averages and trial-specific observations—using as $\widehat{\boldsymbol{b}}$ (i) the average recorded F1 frequency for each speaker, and (ii) the conditional modes, or BLUPs, obtained from the hierarchical model. Is the joint log-likelihood higher in (i) or (ii)?

Exercise 8.4

Replicate the posterior inference procedure in Section 8.2.2 for F2 formant frequencies from Peterson and Barney (1952).

Exercise 8.5: Priors on $\sigma_{\boldsymbol{y}}$

In Section 8.2.2, we used a prior that was locally uniform on the log of the variance parameter $\sigma_{\boldsymbol{y}}$. Another alternative would be to use a prior that was locally uniform on $\sigma_{\boldsymbol{y}}$ itself. Check to see how much of a difference, if any, this alternative would have on posterior inference over μ and $\Sigma_{\boldsymbol{b}}$. (You probably will want to do several simulations for each type of model to get a sense of how much variation there is over samples in each case.

Exercise 8.6: Hierarchical Bayesian linear models

Use `R` and `BUGS` together to replicate the analysis of Section 8.3.2, but with homoscedastic intra-cluster variation. Do the Bayesian confidence intervals on effects of northward migration, southward migration, and clan origin look similar to those obtained using point estimation?

Exercise 8.7: Hypothesis testing for random effects

Using the method of maximum likelihood via `lmer()`, together with the likelihood-ratio test, conduct a hypothesis test for the implicit-causality data of Section 8.3.4 on whether subject- and item-specific effects of experimental condition (as opposed to just subject- and item-specific intercepts) significantly improve the likelihood of a hierarchical model. That is, build a model where the only cluster-specific parameters are by-subject and by-item intercepts, and compare its log-likelihood to the full model in which all experimental conditions interact. What are your conclusions?

Exercise 8.8: Hierarchical Bayesian linear models with crossed effects

Replicate the fully Bayesian analysis presented in Section 8.3.4, but with condition-specific inter-subject and inter-item variation. Write down the complete model specification, implement it with JAGS, and compare the resulting inferences about model parameters with those from the point-estimate analysis and from the simpler Bayesian model. Be sure to consider both inferences about shared parameters and about the parameters governing cross-cluster variation!

Exercise 8.9: Transforms of quantitative predictors in mixed-effect models

Re-run the `dative.glmm` regression from Section 8.4.3, but use raw constituent lengths rather than log-transformed lengths. Compute cross-validated likelihoods (Section 2.11.5) to compare the performance of this model and of the original model? Which approach yields higher log-likelihood?

Exercise 8.10: Simplifying a model based on linguistic principles and model comparison

Define a simpler version of the model in Section 8.4.3 in which the effects of each of constituent length, animacy, discourse status, pronominality, and definiteness must be equal and opposite for recipients versus themes. Fit the model using approximate maximum-likelihood estimation, and compare it to the original model using the likelihood ratio test. Can you safely conclude that the effects of these factors truly are equal and opposite? (**Hint:** the easiest way to construct the simpler model is to define new quantitative predictors that express the summed influence of the property from both recipient and theme; see Exercise 6.11. Also keep in mind that the likelihood-ratio test can be anti-conservative for shared ("fixed-effect") parameters in hierarchical models.)

Exercise 8.11: Mixed-effects logit models and magnitudes of parameter estimates

Unlike with mixed-effects linear models, accounting for a cluster-level variable in a mixed-effects logit model can systematically change the magnitude of the parameter estimates for fixed effects. (**Warning:** this may be a pretty hard problem.)

1. Re-run the `dative.glmm` regression from Section 8.4.3 as a standard logistic regression model, completely omitting the random effect of verb, but replacing it with a fixed effect of the verb's Semantic Class). Do the magnitudes of most of the fixed-effect coefficients (intercept excluded) increase or decrease? Can you think of any reason why this would happen?

2. Test your intuitions by constructing a simple population with two clusters (call the factor `Cluster` with levels `C1,C2`) and a single two-level fixed effect (call the factor `Treatment` with levels `A,B`). Assume the underlying model involves the following linear predictor:

$$\eta = -1 + 2X + bZ$$

where X is a dummy indicator variable that is active when treatment level is `B`, Z is the cluster variable, and b is 1 for cluster 1 and -1 for cluster 2. Generate a dataset consisting of the following cell counts:

	C1	C2
A	250	250
B	250	250

using a logit model, and fit (1) a standard logistic regression with only `Treatment` as a fixed effect, and (2) a mixed-effects logistic regression. How does the estimation of the fixed effect change between models (1) and (2)?

3. Repeat your controlled experiment from (b) above, except this time use linear models (classic and mixed-effects) where the noise has standard deviation 1. Does the same change in the estimated effect of `Treatment` occur as in the logit model?

Exercise 8.12: Different priors on verb-specific effects

1. Use kernel density estimation to estimate the posterior density of the BLUPs for verbs in the dative alternation from the hierarchical logit model of Section 8.4.3. Plot this estimated density, and overlay on it a normal density with mean 0 and standard deviation $\hat{\sigma}_{\boldsymbol{b}}$. Do the BLUPs look like they follow this normal distribution? Choose another distribution for $\boldsymbol{b}$ in this model and implement it in BUGS. (Be forewarned that this is conceptually relatively simple but computationally rather intensive; loading the model and sampling from the posterior are likely to take several minutes or hours even on a new processor.) Plot the posterior inferences on $\boldsymbol{b}$ as in Figure 8.15 and compare them to the results of the original model. Are the results considerably different?

Exercise 8.13: Cross-validated likelihood and probability-proportion plot

Reconstruct Figure 8.16 using ten-fold cross-validation to obtain predicted probabilities. Does the resulting fit look better? Worse? Also compare the cross-validated likelihood to the original model's likelihood. Does the model seem substantially overfit?