9

Introduction to Hierarchical Models

One of the important features of a Bayesian approach is the relative ease with which hierarchical models can be constructed and estimated using Gibbs sampling. In fact, one of the key reasons for the recent growth in the use of Bayesian methods in the social sciences is that the use of hierarchical models has also increased dramatically in the last two decades.

Hierarchical models serve two purposes. One purpose is methodological; the other is substantive. Methodologically, when units of analysis are drawn from clusters within a population (communities, neighborhoods, city blocks, etc.), they can no longer be considered independent. Individuals who come from the same cluster will be more similar to each other than they will be to individuals from other clusters. Therefore, unobserved variables may induce statistical dependence between observations within clusters that may be uncaptured by covariates within the model, violating a key assumption of maximum likelihood estimation as it is typically conducted when independence of errors is assumed. Recall that a likelihood function, when observations are independent, is simply the product of the density functions for each observation taken over all the observations. However, when independence does not hold, we cannot construct the likelihood as simply. Thus, one reason for constructing hierarchical models is to compensate for the biases—largely in the standard errors—that are introduced when the independence assumption is violated. See Ezell, Land, and Cohen (2003) for a thorough review of the approaches that have been used to correct standard errors in hazard modeling applications with repeated events, one class of models in which repeated measurement yields hierarchical clustering.

In addition to the methodological need for hierarchical models, substantively we may believe that there are differences in how predictors in a regression model influence an outcome of interest across clusters, and we may wish to model these differences. In other words, the influence of predictors may be *context-dependent,* a notion that is extremely important and relevant to a social scientific—especially sociological—understanding of the world. For example, the emergence of hierarchical modeling in education research occurred

because there is a natural nesting of students within classes (and classes within schools, schools within communities, and so on), and grades, test performance, etc. may be dependent on teacher quality, making students in one class different from those in another class. In other words, student performance may be dependent on the teacher—the environmental context of classes.

In this chapter, I discuss simple hierarchical models in general as well as hierarchical linear regression models. I conclude the chapter with a brief discussion of terminological issues that make hierarchical modeling seem mysterious and complicated. I recommend Gelman et al. (1995) for an in-depth exposition of the Bayesian approach to a variety of hierarchical models, both the simple hierarchical models discussed in the next section as well as hierarchical regression models discussed later in the chapter. I recommend Raudenbush and Bryk (2002) and Snijders and Bosker (1999) for thorough coverage of the classical approach to hiearchical linear regression models.

9.1 Hierarchical models in general

Hierarchical models are models in which there is some sort of hierarchical structure to the parameters and potentially to the covariates if the model is a regression model. I begin by discussing the simpler case in which the model of interest is not a regression model with covariates, but rather is simply hierarchical in the parameters.

Recall that Bayes' Theorem is often expressed as:

$$\underbrace{p(\theta \mid \text{data})}_{\text{posterior}} \propto \underbrace{p(\text{data} \mid \theta)}_{\text{likelihood}} \times \underbrace{p(\theta)}_{\text{prior}}$$

This equation itself reveals a simple hierarchical structure in the parameters, because it says that a posterior distribution for a parameter is equal to a conditional distribution for data under the parameter (first level) multiplied by the marginal (prior) probability for the parameter (a second, higher, level). Put another way, the posterior distribution is the prior distribution weighted by the observed information.

This hierarchical structure of the parameters need not stop at one higher level; instead, the conditioning structure in theory can continue *ad infinitum*. For instance, suppose we have a model that contains an added layer of hierarchy. Suppose we have J observations within each of G groups: $y_{11}, \ldots, y_{J1}, y_{12}, \ldots, y_{J2}, \ldots, y_{1G}, \ldots, y_{JG}$, and we assume that the data are distributed within groups according to some distribution Q with parameter θ, but that each group has its own parameter (θ_g). Thus:

$$y_{ig} \sim Q(\theta_g).$$

Suppose we assume further that these parameters θ_g arise from a common distribution W with parameter γ (this parameter is called a "hyperparameter"). So:

$$\theta_g \sim W(\gamma).$$

Finally, assume γ has some vague distribution like a uniform:

$$\gamma \sim U(-100, 100).$$

A posterior distribution for all unknown parameters would then be (after substituting the densities Q and W into the conditional structure below):

$$p(\gamma, \theta|y) \propto p(y \mid \theta, \gamma)p(\theta \mid \gamma)p(\gamma).$$

To see how this hierarchical structure "works," notice that the last two terms here $[p(\theta \mid \gamma)p(\gamma)]$, when multiplied together, yield a joint distribution for γ and θ $[p(\theta\,, \gamma)]$. Thus, we are left with a marginal joint distribution for the two parameters, which is then multiplied by a sampling density for the data $[p(y \mid \theta\,, \gamma)]$. Bayes' theorem tells us that the multiple of this marginal joint density for the parameters and the sampling density for the data, given the parameters, yields a posterior density for all of the parameters.

Ultimately we might not be interested much in the posterior distributions for the group level parameters (θ_g), but rather in the posterior distribution for the hyperparameter γ that structures the distribution of the group level parameters. In other words, we may be interested only in the marginal distribution for γ:

$$p(\gamma|y) \propto \int p(y|\theta, \gamma)p(\theta|\gamma)p(\gamma)d\theta.$$

As we have discussed throughout the last several chapters, this integration is performed stochastically via MCMC methods as we sample from the conditional posterior distributions for each parameter.

This result demonstrates the simplicity with which a Bayesian approach can handle hierarchical structure in data or parameters. We could very easily, if desired, add subsequent layers to the structure, and we can also break each layer of the structure into regression components.

9.1.1 The voting example redux

In Chapter 3, I illustrated Bayes' Theorem with a voting example from 2004 pre-election polls. In that example, we considered the posterior probability that Kerry would win the election in Ohio using the most recent poll as the current data and data from three previous polls as prior information. We assumed a binomial likelihood function/sampling density for the current polling data (x) given the proportion of voters who would vote for Kerry (K),

and we used a beta distribution as the prior for K, with the number of votes for Kerry and Bush in the previous polls being represented by the parameters α and β, respectively. To summarize, our posterior density was:

$$p(K|\alpha,\beta,X) \propto \underbrace{K^{556}(1-K)^{511}}_{\substack{\text{current data}\\ \text{(likelihood)}}} \underbrace{K^{941}(1-K)^{1007}}_{\substack{\text{previous poll data}\\ \text{(prior)}}}.$$

In the original example I noted that, although the four polls we used appeared to show some trending, complete data from all available polls from various polling organizations did not suggest any trending, justifying our combination of the previous pollng data into a single prior distribution for α and β. As an alternative approach, without trending, the polls could be considered as separate samples drawn from the same population, each one providing conditionally independent information regarding the parameters α and β. In that case, we could consider that each poll's results were the result of a unique, poll-specific parameter K_i, with the K_i being random realizations from the beta distribution with *hyperparameters* α and β. This approach recasts the voting example as a hierarchical model with the following structure:

$$p(\alpha,\beta,K|X) \propto \underbrace{p(X|K)}_{\text{likelihood}} \underbrace{p(K|\alpha,\beta)}_{\text{prior}} \underbrace{p(\alpha,\beta)}_{\text{hyperprior}}.$$

Here, and throughout the remainder of the chapter, I suppress notation in the conditional distributions when a particular quantity does not directly depend on a higher level parameter. For example, the likelihood function here ultimately depends on the hyperparameters α and β; however, it only depends on these parameters through the prior for K, and so, I do not spell out the complete likelihood as $p(X|K,\alpha,\beta)$.

The likelihood portion of the model is the product of the sampling densities for the four polls:

$$p(X|K) \propto \prod_{i=1}^{4} K_i^{x_i}(1-K_i)^{n_i-x_i}.$$

The prior densities for each K $(K_1 \ldots K_4)$ are beta densities; their product is the full prior density:

$$p(K|\alpha,\beta) \propto \prod_{i=1}^{4} \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) K_i^{\alpha-1}(1-K_i)^{\beta-1}.$$

Finally, we must establish hyperpriors for the hyperparameters α and β. However, before we consider the form of the hyperprior, let's consider the full expression for the posterior density:

$$p(\alpha, \beta, K|x) \propto$$

$$\left(\prod_{i=1}^{4} K_i^{x_i}(1-K_i)^{n_i-x_i}\right)\left(\prod_{i=1}^{4}\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)K_i^{\alpha-1}(1-K_i)^{\beta-1}\right)p(\alpha,\beta).$$

We can simplify this posterior distribution by combining like products as follows:

$$p(\alpha, \beta, K|x) \propto$$

$$\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^4\left(\prod_{i=1}^{4} K_i^{x_i+\alpha-1}(1-K_i)^{n_i-x_i+\beta-1}\right)p(\alpha,\beta). \tag{9.1}$$

The key difference between the current approach and as it was presented in the original example in Chapter 3 is that the current data were assumed to be simply the most recent polling data, and the previous three polls were combined and assumed to be fixed quantities representing the values of α and β. Under the current approach, in contrast, the previous polling data—rather than being treated as fixed prior information—are also considered to arise from a random process governed by the hyperparameters α and β. When these parameters were assumed to be fixed, the posterior density only involved the single parameter K. Now, however, the full posterior involves each K_i in addition to α and β. Before, the leading expression involving the gamma function $[\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta))]$ could be dropped as a normalizing constant, because α and β were, in fact, constant. However, under the hierarchical approach they are now considered random variables, and terms involving them cannot simply be dropped. Indeed, although the individual K parameters are still of interest, interest centers primarily on α and β, which are thought to be the population parameters governing the proportion of voters who would vote for Kerry and which drive each individual poll result.

A Gibbs sampling strategy, then, should involve sampling the α, β, and each K from their conditional posterior distributions. The conditional posterior distributions for each K, after eliminating terms in the posterior in Equation 9.1 that do not involve them, are easily seen to be beta distributions with parameters $A = x_i + \alpha$ and $B = n_i - x_i + \beta$:

$$p(K_i|\alpha, \beta, x_i) \propto K_i^{x_i+\alpha-1}(1-K_i)^{n_i-x_i+\beta-1}.$$

The conditional posterior distributions for α and β are not as simple. Consider the posterior for α. If we eliminate terms not involving α, the posterior for α is:

$$\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^4\prod_{i=1}^{4} K_i^{x_i+\alpha-1}p(\alpha,\beta).$$

This posterior can be simplified considerably if we use a "trick" to allow the combination of the exponents. If we take the log and exponentiate simultaneously, we obtain:

$$\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^4 \exp\left\{\ln\left(\prod_{i=1}^4 K_i^{x_i+\alpha-1}\right)\right\} p(\alpha,\beta).$$

The exponents can be brought down in front of the logarithm, the product of the logs become sums, and we obtain:

$$\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^4 \exp\left\{\sum_{i=1}^4 (x_i+\alpha-1)\ln K_i\right\} p(\alpha,\beta).$$

At this point, we can expand the summation, distribute the three terms in front of the logarithms, and group like terms. We can also again remove terms that do not involve α. We are left with:

$$p(\alpha|\beta,K,x) \propto \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^4 \exp\left\{\alpha\left(\sum_{i=1}^4 \ln K_i\right)\right\} p(\alpha,\beta).$$

A similar strategy reveals that the posterior density for β is:

$$p(\beta|\alpha,K,x) \propto \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^4 \exp\left\{\beta\left(\sum_{i=1}^4 \ln(1-K_i)\right)\right\} p(\alpha,\beta).$$

What remains is the specification of the prior density $p(\alpha,\beta)$. Ideally, we may like a prior that is relatively noninformative. However, in this particular example, we must be careful, because these conditional posterior densities are not of known forms and, with too vague of a prior, will not be proper.

Recall that the hyperparameters α and β of the beta distribution can be viewed as prior successes and failures, respectively, and are therefore constrained to be nonnegative. In the example in Chapter 3, we fixed these parameters at constants to represent the successes/failures from the first three surveys in Ohio. Now, in contrast, we want to specify distributions for them. An appropriate distribution that would constrain these parameters to be nonnegative is the gamma distribution, which itself has two parameters, say C and D. If we assume that α and β have independent prior distributions, then $p(\alpha,\beta) = p(\alpha)p(\beta)$, and we can assign each a gamma distribution prior:

$$\begin{aligned} p(\alpha) &\propto \alpha^{C_\alpha-1}\exp\left(-D_\alpha\alpha\right) \\ p(\beta) &\propto \beta^{C_\beta-1}\exp\left(-D_\beta\beta\right). \end{aligned}$$

This hyperprior yields the following conditional posterior for α:

$$p(\alpha|\beta,K,x,C_\alpha,D_\alpha) \propto \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^4 \alpha^{C_\alpha-1}\exp\left\{\alpha\left(\sum_{i=1}^4 \ln K_i - D_\alpha\right)\right\}.$$

A comparable result can be obtained for β. All that remains is to specify values for C and D in each hyperprior.

Given parameters C and D, the mean of a gamma distribution is equal to C/D, and the variance is equal to C/D^2. We may choose to set these parameters at values that reflect our prior knowledge. Numerous previous polls throughout the country had showed the race to be virtually a dead heat, and so, we may choose comparable values of C and D for both prior distributions. The typical poll conducted throughout the fall by different polling organizations consisted of about 500 or so potential voters, roughly half of which were expected to vote for Kerry. So, we may choose values of C and D such that $C/D = 250$. We can capture prior uncertainty in this estimate by specifying the variance to be large. For example, if we choose a standard deviation to be equal to 100, then $C/D^2 = 10,000$, and so $C = 6.25$ and $D = .025$. To evaluate the influence of the hyperparameter specification, I varied these parameters and conducted several runs of the Gibbs sampler, as discussed below.

Below is a hybrid Gibbs sampler/MH algorithm for simulating the parameters of the model. Although the K parameters, conditional on the data and values for α and β, can be drawn directly from beta distributions, the α and β hyperparameters are not known forms and must therefore be simulated using MH steps:

```
#MCMC algorithm for hierarchical beta-binomial model

a=matrix(10,100000);b=matrix(10,100000); acca=0; accb=0
y=matrix(c(556,346,312,284),4); n=matrix(c(1067,685,637,628),4)
k=matrix((y)/n,m,4,byrow=T)

apost<-function(f,g,k){
 post=4*(lgamma(f+g)-lgamma(f)-lgamma(g)) + f * sum(log(k))
 post=post+(6.25-1)*log(f)-(f*.025)
 return(post)
}
bpost<-function(f,g,k){
 post=4*(lgamma(f+g)-lgamma(f)-lgamma(g)) + g * sum(log(1-k))
 post=post+(6.25-1)*log(g)-(g*.025)
 return(post)
}

for(i in 2:100000){
 #draw a
 a[i]=a[i-1]+rnorm(1,0,20)
 if(a[i]>0){
  acca=acca+1
  newpost=apost(a[i],b[i-1],k[i-1,])
  oldpost=apost(a[i-1],b[i-1],k[i-1,])
  if(log(runif(1,min=0,max=1))>(newpost-oldpost))
```

```
  {a[i]=a[i-1]; acca=acca-1}
 }
if(a[i]<0){a[i]=a[i-1]}

#draw b
b[i]=b[i-1]+rnorm(1,0,20)
if(b[i]>0){
 accb=accb+1
 newpost=bpost(a[i],b[i],k[i-1,])
 oldpost=bpost(a[i],b[i-1],k[i-1,])
 if(log(runif(1,min=0,max=1))>(newpost-oldpost))
  {b[i]=b[i-1]; accb=accb-1}
 }
if(b[i]<0){b[i]=b[i-1]}

#draw k from beta distributions
k[i,]=rbeta(4,(y+a[i]),(n-y+b[i]))

if(i%%10==0){print(c(i,a[i],b[i],acca/i,accb/i))}
}
```

This program is fairly straightforward. First, matrices are established for the α and β parameters, and acceptance rate variables are also constructed for monitoring the MH steps used to simulate them. Next, the data, including votes for Kerry (`y`), poll sizes (`n`), and proportions favoring Kerry (`k`), are established. The next two program blocks are functions that evaluate the conditional log-posterior densities for α and β, respectively, given values of these parameters, the previous value for the observed sample proportions, and a prior distribution (the second line of each function is the hyperprior).

The program then proceeds to simulate 100,000 draws from the posterior for all the parameters. The α and β parameters are drawn using MH steps. Candidates are generated from normal proposals with a standard deviation set to produce an approximate acceptance rate of 50%. Once a candidate is generated, the log-posterior is evaluated at the candidate values for these parameters and the previous values. I have structured these blocks so that the candidate parameter is assumed to be accepted and is evaluated for rejection. If the candidate is less than 0, or the log of the uniform draw exceeds the ratio of the log-posterior at the current versus previous values, the current value of the parameter is reset to the previous value, and the acceptance tally is reduced by one. Once values of these parameters have been drawn, each K_i parameter is drawn from the appropriate beta distribution.

The key parameters of interest in the model include the individual survey proportions ($K_1 \ldots K_4$) and the population proportion implied by the α and β parameters, which is equal to $\alpha/(\alpha+\beta)$. Table 9.1 shows posterior summaries of these parameters under a variety of specifications for C and D in the hyperpriors for α and β. The first four columns of the table show the gamma distribution hyperprior specifications for the α and β parameters of the prior

distribution. These values for the hyperpriors were chosen to examine how sensitive the posterior inferences are to prior specification.

The first two columns show the mean and standard deviation of the gamma hyperprior distribution for α, respectively; the third and fourth columns show the mean and standard deviation of the hyperprior for β. Recall from above that the mean of the gamma distribution for α can be considered as previous votes for Kerry, and the variance/standard deviation of this distribution can be viewed as a measure of our uncertainty in this number of previous votes. Similarly, the mean of the gamma distribution for β can be considered as previous votes for Bush, and its standard deviation reflects our uncertainty in this number. Thus, the first specification implies that previous polls have shown an equal—and small—number of votes for both candidates, and the relatively large standard deviation of each (10) suggests that we are not very certain of these numbers.

Thus, the first row shows the posterior inference when the prior information is fairly weak. That is, this hyperprior specification implies that we have prior information equivalent to 10 previous votes for Kerry and 10 for Bush, with a fairly large standard deviation reflecting considerable uncertainty about these numbers of votes. In contrast, the final hyperprior specification implies that we have prior information equivalent to 2,500 votes for Kerry and 500 votes for Bush, and that our confidence in these numbers is relatively strong (standard deviation of only 50, compared with the number of prior votes).

The bottom two rows of the table show the results under two alternative approaches to the hierarchical approach discussed here. The first row at the bottom shows the results obtained if the four polls are analyzed independently; the second shows the results obtained if the data from all polls are pooled and given a noninformative prior distribution—an equivalent approach to treating the most recent polling data as the current data and the earlier three polls as prior information (see Chapter 3).

Overall, all the hyperprior specifications lead to similar posterior inference for the prior distribution mean $\alpha/(\alpha+\beta)$ and for each of the polls, with the exception of the most informative specification which shows heavy favoritism for Kerry (2,500 prior votes versus 500). Under that specification, the posterior mean for the second level beta prior distribution is pulled strongly away from the mean implied by the polling data and toward the prior.

A couple of comments are warranted regarding these results. First, notice that pooling the data led to a posterior mean of .497 for Kerry's proportion of the vote, and that a similar proportion was obtained using $\alpha/(\alpha+\beta)$ in the hierarchical model, except for the final one with the strongest and most unbalanced hyperprior. However, although the posterior means are comparable, the posterior standard deviation for this proportion tended to be much larger under the hierarchical approach. The reason for this result is that, under the hierarchical approach, the distribution for $\alpha/(\alpha+\beta)$ captures the range of the survey specific K_i parameters, each of which contains its own variability. Under the pooled-data approach, on the other hand, three of the K_i are assumed

Table 9.1. Results of hierarchical model for voting example under different gamma hyperprior specifications.

Gamma Priors				Posterior Inferences				
α		β						
$\frac{C}{D}$	$\sqrt{\frac{C}{D^2}}$	$\frac{C}{D}$	$\sqrt{\frac{C}{D^2}}$	$\frac{\alpha}{\alpha+\beta}$	K_1	K_2	K_3	K_4
10	10	10	10	.493(.048)	.520(.015)	.505(.019)	.490(.019)	.454(.019)
100	100	100	100	.493(.021)	.516(.014)	.502(.017)	.491(.018)	.463(.018)
250	100	250	100	.494(.015)	.513(.014)	.501(.016)	.491(.016)	.470(.017)
250	100	100	100	.494(.016)	.514(.014)	.501(.016)	.491(.017)	.469(.017)
2500	50	500	50	.586(.008)	.572(.010)	.574(.010)	.572(.010)	.567(.010)
Separate Models				NA	.521(.015)	.505(.019)	.490(.020)	.452(.020)
Pooled Data				.497(.009)	NA	NA	NA	NA

Note: The hyperpriors are gamma distributions for both α and β. The hyperparameters C and D in each gamma distribution were set to produce the means and standard deviations shown (C/D and $\sqrt{C/D^2}$, respectively). The posterior quantities are the posterior mean of the beta prior distribution, $\alpha/(\alpha+\beta)$, and the posterior means for each of the sample proportions (posterior standard deviations are in parentheses).

be known, fixed quantities, reducing variability in the overall mean. Second, notice that it is generally the case that the variability for each K_i parameter is smaller than that obtained under the separate-models approach. The reason for this result is that, by combining all samples into a single, hierarchical model, each K_i distribution "borrows strength" from the common linkage of all the polls provided by the hyperparameters α and β.

9.2 Hierarchical linear regression models

The example in the previous section shows a basic hierarchical model in which the model parameters, but not the data, were structured hierarchically—all of the data were measured at the same level (individual polls). It is common in social science research, however, to have hierarchical structure to the data, that is, to have variables collected at different levels. In these cases, social scientists often turn to hierarchical models to capture variation at different levels of analysis. Because these models involve variables measured at different levels, they are sometimes called "multilevel models." Most commonly, individuals are nested within physical or geographic units, or time-specific measures are nested within individuals. As a few examples of the former type of nesting, consider students within classrooms or individuals within neighborhoods. As an example of the latter type of nesting, consider a panel study

in which individuals are measured repeatedly across time. In such a case, the "group" is the individual, and the time-specific measures are nested within the individual. The examples here will follow this latter format—time-specific measures nested within individuals—although the underlying concepts of hierarchy are identical.

I discuss several types of such hierarchical regression models, beginning with an example that evaluates the extent to which Internet usage influences income using a two-wave panel study.[1] These data are from the 2000 and 2001 Current Population Survey Computer Use and Internet Supplement. This supplement measured, among other variables, individual use of computers and the Internet in 2000 and again in 2001 and allows us to examine the relationship between Internet usage and wages across a brief, but important, period of time when availability of broadband Internet connectivity was exploding. Wages in these examples have been transformed to 1982 dollars and are recoded into log-dollars per hour for additional analyses not presented here.

At the end of the chapter, I turn to an example that examines factors that influence health trajectories for individuals across age using a four-wave study (the National Health Epidemiologic Follow-up Surveys) discussed in previous chapters.

9.2.1 Random effects: The random intercept model

Generally, the goal of hierarchical modeling is to determine the extent to which factors measured at different levels influence an outcome using a typical regression modeling framework. OLS regression, however, is inappropriate, because of the lack of independence of errors for observations within groups. Thus, an alternative model must be developed to compensate for this lack of independence.

The foundation for the hierarchical regression model is the simple random effects model. Assume, as an example, that we observe a collection of individuals twice over a two-year period and ask their income at each point in time. It is most likely the case that each individual's income changes only slightly over the time period, and so, we could model the data such that each individual receives his/her own intercept (or mean). In equation form:

$$y_{it} = \alpha_i + e_{it},$$

with $\alpha_i \sim N(\alpha, \tau^2)$ and $e_{it} \sim N(0, \sigma^2)$. This specification shows that the outcome of interest (income; y) is considered a function of "variables" measured at two different levels: α_i is an individual (group) level variable, and e_{it} is a time-specific (individual) random error term.

An alternative, but equivalent, way to specify this model is to use probability notation. This approach clarifies the hierarchical nature of the model:

[1] I thank Bart Bonikowski and Paul DiMaggio for allowing me to use their Internet/income data in the examples.

$$
\begin{aligned}
y_{it} &\sim N(\alpha_i, \sigma^2) \\
\alpha_i &\sim N(\alpha, \tau^2) \\
\alpha &\sim N(m, s^2) \\
\tau^2 &\sim IG(a, b) \\
\sigma^2 &\sim IG(c, d).
\end{aligned}
$$

This specification says that an individual's time-specific income is a random normal variable with a mean equal to an individual-specific mean and some variance. The second equation shows that the individual-specific means themselves come from a (normal) distribution with a mean equal to some population mean and some variance. Finally, the last three equations specify hyperprior distributions for the population grand mean α, the population variance (around the mean) τ^2, and the error variance σ^2. The hyperprior distribution for the population mean is specified here to be normal, with parameters m and s^2; Without prior knowledge, these parameters should be specified to make the hyperprior vague (e.g., say $m = 0$ and $s^2 = 10,000$). The hyperprior distributions for the population variance and the error variance are inverse gamma distributions, with parameters a and b and c and d, respectively. Once again, without prior information, these parameters should be fixed to make the hyperprior vague.

In addition to being a simple random effects model, this model is sometimes called a "random intercept model," because the model can be viewed as a regression model with each α_i considered a group-specific intercept term arising from a (normal) probability distribution (at this point, with no covariates included).

To implement a Gibbs sampler for this model, we first need to construct the posterior distribution. The posterior distribution for this model is straightforward to derive following the hierarchical modeling structure using conditional distributions presented at the beginning of the chapter. The parameters of interest in the posterior distribution are the individual α_i, the population mean α, its variance τ^2, and the residual variance σ^2, and so our posterior density is:

$$
p(\alpha, \tau^2, \alpha_i, \sigma^2 | Y) \propto p(Y|\alpha_i, \sigma^2) p(\alpha_i|\alpha, \tau^2) p(\alpha|m, s^2) p(\tau^2|c, d) p(\sigma^2|a, b).
$$

To complete the specification of the posterior distribution, we simply need to replace each term with its actual distribution. As discussed above, the data are assumed to be normally distributed, and so the likelihood term is:

$$
p(Y|\alpha_i, \sigma^2) \propto \prod_{i=1}^{n} \prod_{t=1}^{2} \frac{1}{\sqrt{\sigma^2}} \exp\left\{ -\frac{(y_{it} - \alpha_i)^2}{2\sigma^2} \right\}.
$$

The distribution for each α_i is also normal and is:

$$p(\alpha_i|\alpha,\tau^2) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{\tau^2}} \exp\left\{-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right\}.$$

The remaining terms are hyperprior distributions for the population mean (α), population random effects variance (τ^2), and residual variance (σ^2). As mentioned above, α is assumed to come from a normal distribution with parameters m and s^2, and the two variance parameters are assumed to come from inverse gamma distributions with parameters a and b and c and d, respectively. This implies the following joint hyperprior distribution:

$$p(\alpha|m,s^2)p(\tau^2|a,b)p(\sigma^2|c,d) \propto$$
$$\frac{1}{\sqrt{s^2}} \exp\left\{-\frac{(\alpha-m)^2}{2s^2}\right\} \times \frac{1}{(\tau^2)^{a+1}} \exp\left\{-b/(\tau^2)\right\} \times \frac{1}{(\sigma^2)^{c+1}} \exp\left\{-d/(\sigma^2)\right\}.$$

The full posterior, then, is simply the product of these three terms—the likelihood, prior, and hyperprior distributions. Although the posterior distribution can be simplified considerably by carrying out the multiplication of exponentials and combining like terms, it is simpler to derive the conditionals for the Gibbs sampler by leaving the posterior written as is. For the Gibbs sampler, we need the conditional distributions for each of the parameters; deriving them from the posterior is a simple but tedious matter of selecting only the terms that contain the parameter of interest, discarding all other multiplicative terms as proportionality constants, and simplifying/rearranging what's left to determine the resulting distribution. If we begin with the parameter α, the relevant terms in the posterior are:

$$\begin{aligned} p(\alpha|.) &\propto p(\alpha_i|\alpha,\tau^2)p(\alpha) \\ &\propto \left(\prod_{i=1}^{n} \frac{1}{\sqrt{\tau^2}} \exp\left\{-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right\}\right) \frac{1}{\sqrt{s^2}} \exp\left\{-\frac{(\alpha-m)^2}{2s^2}\right\}. \end{aligned}$$

From this expression, the leading fractions involving the variances can be removed as normalizing constants (they do not depend on α), and the exponential expressions can be combined to obtain:

$$p(\alpha|.) \propto \exp\left\{\left(-\frac{1}{2}\right)\left(\frac{\tau^2(\alpha-m)^2 + s^2\sum_{i=1}^{n}(\alpha_i-\alpha)^2}{\tau^2 s^2}\right)\right\}.$$

Next, we can expand the numerator of the exponential, extract terms not involving α as constants, and we have:

$$p(\alpha|.) \propto \exp\left\{\left(-\frac{1}{2}\right)\left(\frac{\tau^2\alpha^2 - 2\tau^2\alpha m - 2s^2\alpha\sum\alpha_i + ns^2\alpha^2}{\tau^2 s^2}\right)\right\}.$$

Rearranging terms, we obtain:

$$p(\alpha|.) \propto \exp\left\{\left(-\frac{1}{2}\right)\left(\frac{(\tau^2+ns^2)\alpha^2 - 2\alpha(\tau^2 m + s^2\sum\alpha_i)}{\tau^2 s^2}\right)\right\}.$$

As we did in Chapter 3, we can complete the square in α, and we find that the conditional posterior for α is:

$$p(\alpha|.) \propto N\left(\frac{\tau^2 m + s^2\sum\alpha_i}{\tau^2+ns^2} \quad , \quad \frac{\tau^2 s^2}{\tau^2+ns^2}\right) \tag{9.2}$$

The conditional posterior distribution for each α_i is even easier to obtain. Once again, we begin with terms involving only α_i. We should realize, however, that, for each individual i, the only relevant terms in the product are those involving that particular individual. Thus, the conditional posterior for person i ($\forall i$) is:

$$\begin{aligned} p(\alpha_i|.) &\propto p(Y|\alpha_i,\sigma^2)p(\alpha_i|\alpha,\tau^2) \\ &\propto \left(\prod_{t=1}^{2}\frac{1}{\sqrt{\sigma^2}}\exp\left\{-\frac{(y_{it}-\alpha_i)^2}{2\sigma^2}\right\}\right)\left(\frac{1}{\sqrt{\tau^2}}\exp\left\{-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right\}\right). \end{aligned}$$

We can follow the same steps as for α, and we obtain:

$$p(\alpha_i|.) \propto \exp\left\{\left(-\frac{1}{2}\right)\left(\frac{(2\tau^2+\sigma^2)\alpha_i^2 - 2\alpha_i(\tau^2\sum y_{it} + \sigma^2\alpha)}{\tau^2\sigma^2}\right)\right\}.$$

If we complete the square in α_i, we find that:

$$p(\alpha_i|.) \propto N\left(\frac{\tau^2\sum y_{it} + \sigma^2\alpha}{2\tau^2+\sigma^2} \quad , \quad \frac{\tau^2\sigma^2}{2\tau^2+\sigma^2}\right). \tag{9.3}$$

The variance parameters σ^2 and τ^2 can be derived following the same strategy. The conditional posterior for σ^2 is:

$$p(\sigma^2|.) \propto p(Y|\alpha_i,\sigma^2)p(\sigma^2|a,b).$$

After substitution we obtain:

$$p(\sigma^2|.) \propto \left(\prod_{i=1}^{n}\prod_{t=1}^{2}\frac{1}{\sqrt{\sigma^2}}\exp\left\{-\frac{(y_{it}-\alpha_i)^2)}{2\sigma^2}\right\}\right)\frac{1}{(\sigma^2)^{c+1}}\exp\left\{-\frac{d}{\sigma^2}\right\},$$

and after some simplification, we get:

$$p(\sigma^2|.) \propto \left(\sigma^2\right)^{-(n+c+1)} \exp\left\{\frac{-\left(\sum_{i=1}^{n}\sum_{t=1}^{2}(y_{it}-\alpha_i)^2+2d\right)}{2\sigma^2}\right\}.$$

This result shows that the conditional posterior for σ^2 is an inverse gamma distribution:

$$p(\sigma^2|.) \propto IG\left(n+c\,,\ \frac{\sum_{i=1}^{n}\sum_{t=1}^{2}(y_{it}-\alpha_i)^2+2d}{2}\right). \tag{9.4}$$

The conditional posterior for τ can be derived similarly. The posterior is:

$$\begin{aligned} p(\tau^2|.) &\propto p(\alpha_i|\alpha,\tau^2)p(\tau^2) \\ &\propto \left(\prod_{i=1}^{n}\frac{1}{\sqrt{\tau^2}}\exp\left\{-\frac{(\alpha_i-\alpha)^2}{2\tau^2}\right\}\right)\frac{1}{(\tau^2)^{a+1}}\exp\left\{-\frac{b}{\tau^2}\right\}. \end{aligned}$$

After simplification, we obtain:

$$p(\tau^2|.) \propto IG\left(n/2+a+1\,,\ \frac{\sum_{i=1}^{n}(\alpha_i-\alpha)^2+2b}{2}\right) \tag{9.5}$$

(see Exercises).

Given a complete set of conditional posterior distributions, we can implement a Gibbs sampler for the model by sequentially drawing from these conditionals. Below is an R program that conducts the Gibbs sampling:

```
#R program for simple random effects model
#read data
y=as.matrix(read.table("c:\\internet_examp.dat")[,3:4])

m=0; s2=10000; a=c=.001; b=d=.001; tau2=1; sigma2=1; malpha=0
n=nrow(y)

for(i in 1:20000){

#draw alpha_i
alpha= rnorm(n,
mean=(((tau2*(y[,1]+y[,2]))+sigma2*malpha)/(2*tau2+sigma2)),
sd=sqrt((tau2*sigma2)/(2*tau2+sigma2)))

#draw malpha
malpha=rnorm(1,
mean=(tau2*m+s2*sum(alpha))/((tau2+n*s2)),
sd=sqrt((tau2*s2)/((tau2+n*s2))))

#draw tau2
tau2=rgamma(1, shape=(n/2+a), rate=(sum((alpha-malpha)^2)+2*b)/2)
```

```
tau2=1/tau2

#draw sigma2
sigma2=rgamma(1, shape=n+c, rate=(sum((y-alpha)^2) +2*d)/2)
sigma2=1/sigma2

#write results to file
if(i%%10==0 | i==1)
 {print(c(i,alpha[1],malpha,tau2,sigma2))
  write(c(i,alpha[1],malpha,tau2,sigma2),
        file="c:\\bart2.out",append=T,ncol=5)}
}
```

As with previous programs, the first block reads in the data and establishes starting (and fixed) values for the parameters. The hyperparameters associated with the hyperpriors for α, τ^2, and σ^2 are fixed to 0, 10,000, .001, .001, .001, and .001, respectively, in order to ensure that the hyperparameters have little influence on the results (see Exercises). The starting values for the population/grand mean (α) as well as for τ^2 and σ^2 are arbitrarily set to benign values.

Subsequent sections of the program constitute nothing more than iteratively sampling from the conditional posterior distributions derived above.

Although this R program is relatively short, the derivation of the conditional distributions was a tedious process. Fortunately, however, a software package exists that allows us to simulate values from the posterior distributions for the parameters of this model more directly: WinBugs. WinBugs is a freely available software package that simplifies Gibbs sampling for a variety of models. The syntax for WinBugs is substantially similar to that of R, but many of the conditional posterior distribution derivations are done for us by WinBugs, reducing the need to derive the conditional posterior distributions manually. For example, a WinBugs program for the same example involves nothing more than specifying the likelihood, prior, and hyperprior distributions and parameter as follows:

```
#Winbugs program for simple random effects model
model
{
 for(i in 1:9249)
  {
   for(j in 1:2)
    {
     y[i,j]~dnorm(alpha[i],sigma2inv)
    }
  alpha[i]~dnorm(malpha,tau2inv)
  }
malpha~dnorm(0,1.0E-4)

tau2inv~dgamma(.01,.01)
```

```
tau2<-1/sqrt(tau2inv)

sigma2inv~dgamma(.01,.01)
sigma2<-1/sqrt(sigma2inv)
}
```

The syntax in this program is similar to that of R with a few exceptions. First, the tilde is used to simulate from distributions. Second, "$<-$" is used to assign values to variables.[2] Third, the parameterization of the normal distribution in WinBugs involves a precision parameter rather than a variance parameter. The precision is simply the inverse of the variance, and so, we can recover the variance parameter simply by inverting the draw from the gamma distribution for the precision parameters.

The key results from the R program, the equivalent WinBugs program, and the equivalent maximum likelihood results obtained from STATA (versions 8 and 9 were used throughout) using the `xtreg` procedure are presented in Table 9.2. As the results show, all three approaches yielded virtually the same results and therefore lead to the same conclusions. The Bayesian results, however, whether from R or WinBugs, yield more information by default than the STATA results, because the Bayesian approach yields distributions for all parameters/quantities of interest, including the variance parameters.

Table 9.2. Results of hierarchical model for two-wave panel of income and Internet use data.

Variable	R	WinBugs	STATA `xtreg`
Population Mean (α)	2.103(.005)	2.103(.005)	NA
Intercept	NA	NA	2.103(.005)
$\sqrt{\tau^2}$	0.434(.004)	0.434(.004)	0.434
$\sqrt{\sigma^2}$	0.311(.002)	0.311(.002)	0.311
$\tau^2/(\tau^2+\sigma^2)$	0.661(.006)	0.660(.006)	0.660

Note: Posterior means (and posterior standard deviations) are reported for R and WinBugs algorithms. Generalized least squares estimates (and standard errors) are reported for STATA.

Overall, these results indicate that mean log wages are 2.103 log-dollars per hour with a standard deviation of .434 log-dollars. Within individuals, the standard deviation of wages was .311 log-dollars, and the ratio of the between-individual to total variance is about 66%. This result suggests that much of the variation we observe in log-wages—as we might expect—is due

[2] This syntax can also be used in R, but I have generally not done so throughout the text.

to differences between individuals and not within individuals across the two-year period. As a side note, the total variance in hourly wages is equal to $\tau^2 + \sigma^2$. Because we obtain estimates for both of these variances—the "between-individual" and "within-individual" variances—hiearchical models like this one are sometimes called "variance components" models.

The next step in our hierarchical modeling approach is to allow variation in the group level parameters to be functions of group level variables and to let the individual level (here, time-specific level) random error term to be a function of individual level variables. First, for example, we could include group level characteristics in our model by decomposing the random intercept into a regression on group level variables. For example, suppose we now wish to determine whether sex influences respondents' wages. In that case, we can specify the model as:

$$\begin{aligned} y_{it} &\sim N(\alpha_i + \alpha_{(1)}\text{sex}_i,\ \sigma^2) \\ \alpha_i &\sim N(\alpha_{(0)}, \tau^2) \\ \alpha_{(0)} &\sim N(m0, s0) \\ \sigma^2 &\sim IG(a, b) \\ \alpha_{(1)} &\sim N(m1, s1) \\ \tau^2 &\sim IG(c, d). \end{aligned}$$

Essentially the only substantial difference between this and the previous model is that the individual-specific intercept has now been decomposed into a population intercept and an effect of sex. A WinBugs program for this model is simple to specify from these distributions:

```
model
{
 for(i in 1:9249)
  {
   for(t in 1:2)
    {
     y[i,t]~dnorm(alpha[i],sigma2inv)
    }
  alpha[i]~dnorm(mu[i],tau2inv)
  mu[i]<-alpha0+alpha1*sex[i]
  }
alpha0~dnorm(0,1.0E-4)
alpha1~dnorm(0,1.0E-4)

sigma2inv~dgamma(.01,.01)
sigma2<-1/sqrt(sigma2inv)

tau2inv~dgamma(.01,.01)
tau2<-1/sqrt(tau2inv)
```

```
}
```

In this program, I have specified independent (univariate) normal distribution priors for the population mean and the parameter representing the influence of sex. The fact that I have specified independent priors, however, does not imply that the two parameters are necessarily uncorrelated in the posterior. In fact, the two parameters are highly negatively correlated, as Figure 9.1 shows.

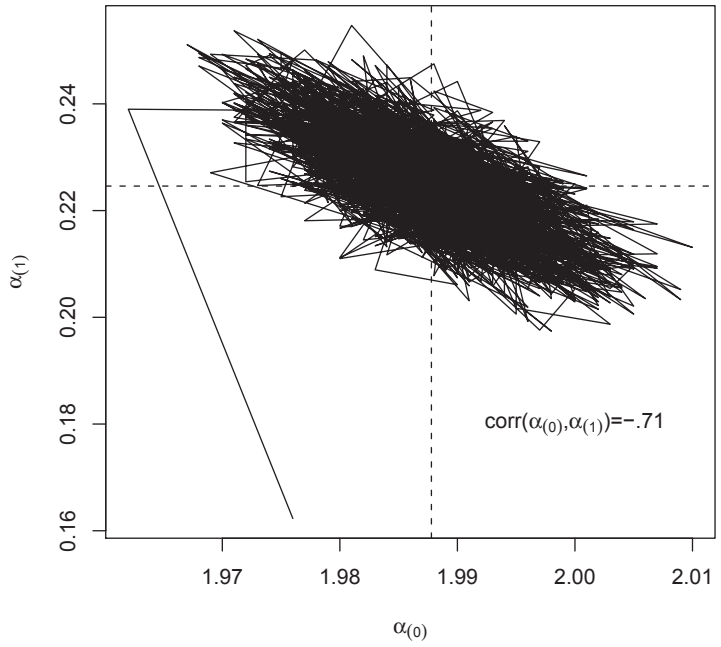

Fig. 9.1. Two-dimensional trace plot of $\alpha_{(0)}$ and $\alpha_{(1)}$ parameters (dashed lines at posterior means for each parameter).

The posterior mean for the adjusted population mean ($\alpha_{(0)}$) was 1.99 (s.d. = .007), and the mean for the influence of sex ($\alpha_{(1)}$) was .225 (s.d. = .0098), indicating that males have higher log wages. The only additional change between this and the previous model is the magnitude of τ^2. Recall that τ^2 reflects unexplained between-individual variation in the random intercept for log-wages. With the inclusion of sex as an explantory variable differentiating individuals' wages, τ^2 has been reduced. Its posterior mean is now .419 (s.d. of .004), which is a reduction of 3.5% over the mean value obtained under the

previous model. This reduction can be viewed as an R^2 term; put another way, sex differences account for 3.5% of the between-individual variance in wages.

Additional time-invariant variables can be easily included to further account for between-individual variation in wages. But what if we would like to consider the influence of time-varying covariates? For example, suppose we are interested in examining the extent to which Internet usage at a given point in time influences wages at the same point in time. Our data include time-specific measures of Internet usage, measured at the same points in time that wages are measured. There are two ways we can accomplish this goal. First, we can allow such covariates to influence the time-specific outcomes directly:

$$\begin{aligned} y_{it} &\sim N(\alpha_i + \alpha_{(1)}\text{sex}_i + \alpha_{(2)}\text{Internet}_{it},\ \sigma^2) \\ \alpha_i &\sim N(\alpha_{(0)}, \tau^2) \\ \alpha_{(0)} &\sim N(m0, s0) \\ \alpha_{(1)} &\sim N(m1, s1) \\ \alpha_{(2)} &\sim N(m2, s2) \\ \sigma^2 &\sim IG(a, b) \\ \tau^2 &\sim IG(c, d) \end{aligned}$$

In this model, time-specific wages are considered a function of individual random intercepts and time-specific Internet usage indicators, and the random intercepts are considered a function of a grand mean and an indicator for sex.[3] A WinBugs program to implement this model is as follows:

```
model
{
 for(i in 1:9249)
  {
   for(t in 1:2)
    {
     y[i,t]~dnorm(mu[i,t],sigma2inv)
     mu[i,t]<-(alpha[i]+alpha1*sex[i])+alpha2*internet[i,t]
    }
   alpha[i]~dnorm(alpha0,tau2inv)
  }
alpha0~dnorm(0,1.0E-4)
alpha1~dnorm(0,1.0E-4)
alpha2~dnorm(0,1.0E-4)

sigma2inv~dgamma(.01,.01)
sigma2<-1/sqrt(sigma2inv)
```

[3] An equivalent way of specifying this model is: $y_{it} \sim N(\alpha_i + \alpha_{(2)}\text{Internet}_{it},\ \sigma^2)$, with $\alpha_i \sim N(\alpha_0 + \alpha_{(1)}\text{sex}_i,\ \tau^2)$.

```
tau2inv~dgamma(.01,.01)
tau2<-1/sqrt(tau2inv)
}
```

This program is only slightly more complicated than the previous programs. The only substantial differences are that (1) we have included the new parameter ($\alpha_{(2)}$) within the double loop (i, t), and (2) we have incorporated a prior distribution for it. The results of this model suggest that Internet usage does, in fact, influence income. The posterior mean for the influence of Internet usage is .18 (s.d. = .0075), and the intercept ($\alpha_{(0)}$) falls to 1.86 (s.d. = .009).

9.2.2 Random effects: The random coefficient model

As written, the last model in the previous section forces the effect of Internet usage to be constant across time: There was only a single parameter representing the effect of Internet usage on wages. This constraint may introduce error into the model if, in fact, the influence of Internet usage on wages varies across time. Thus, a second way we can include this time-varying variable is to allow the influence of Internet usage to vary across time. This model is:

$$\begin{aligned}
y_{it} &\sim N(\alpha_i + \alpha_{(1)}\text{sex}_i + \alpha_{(2t)}\text{Internet}_{it},\ \sigma^2) \\
\alpha_i &\sim N(\alpha_{(0)}, \tau^2) \\
\alpha_{(0)} &\sim N(m0, s0) \\
\alpha_{(1)} &\sim N(m1, s1) \\
\alpha_{(21)} &\sim N(m2, s2) \\
\alpha_{(22)} &\sim N(m3, s3) \\
\sigma^2 &\sim IG(a, b) \\
\tau^2 &\sim IG(c, d)
\end{aligned}$$

The alterations of the WinBugs program to accommodate this new parameter are very slight: The `alpha2` parameter must be subscripted appropriately (i.e., `alpha2[t]`), and an additional hyperprior distribution must be incorporated. By some terminologies, we can now call the model a *random coefficient model,* because a *slope*—and not simply an intercept—is now considered a function of other variables.[4]

[4] It may be easier to recognize that allowing `alpha2` to vary across time implies that `alpha2` is now a slope, and not simply an intercept, if we consider that our current representation is equivalent to specifying α_2 to be a function of a dummy variable reflecting time of measurement: $\alpha_2 = \beta_0 + \beta_1 I(t = 2)$, where β_1 is a regression slope.

The results of this model do not vary substantially from those obtained when the effect of Internet usage was treated as constant. However, the influence of Internet usage at time 1 was found to be .167 (s.d. = .009), while the effect of Internet usage at time 2 was .188 (s.d. = .008). A distribution for a new variable representing the difference between these parameters was constructed in order to determine whether this difference is greater than 0; 99.9% of the mass of the resulting distribution was above 0 (posterior mean of .02; s.d. = .006), which indicates that Internet usage indeed influenced wages to a greater extent at the second wave of the study than at the first wave.

From a substantive perspective, this result seems to be more consistent with the view that Internet usage influences income than the view that wages influence Internet usage. That is, Internet availability has become less dependent on income over time as the hardware for accessing the Internet (i.e., computers and modems), as well as Internet service, has become cheaper. If wages influenced Internet usage, on the other hand, we might expect the influence of wages on Internet use to decrease rather than increase over the period of observation. Thus, the result we obtained may be explained such that Internet usage builds social capital, allowing individuals to find or acquire better, higher paying jobs.

One could still argue, however, that higher paying jobs have become increasingly dependent on Internet usage/access, and that a polarization of the labor market is occurring. Thus, higher paid workers have increasingly come to use the Internet, while lower paid jobs continue to not require Internet access/use.

The relationship between Internet usage and income may not just vary across time; it may vary across individuals. For example, individuals in low-income, low-skill occupations may get less of a return to their income from using the Internet. In contrast, individuals in high-skilled occupations may get a large return to their income from using the Internet. In order to examine this possibility, we can alter the model so that the $\alpha_{(2)}$ parameter varies by individual (i) rather than by time (t). Thus, the model becomes:

$$
\begin{aligned}
y_{it} &\sim N(\alpha_i + \alpha_{(1)}\text{sex}_i + \alpha_{(2i)}\text{Internet}_{it},\ \sigma^2) \\
\alpha_i &\sim N(\alpha_{(0)}, \tau^2) \\
\alpha_{(2i)} &\sim N(\alpha_{(20)}, \tau_2^2) \\
\alpha_{(0)} &\sim N(m0, s0) \\
\alpha_{(1)} &\sim N(m1, s1) \\
\alpha_{(20)} &\sim N(m2, s2) \\
\sigma^2 &\sim IG(a, b) \\
\tau^2 &\sim IG(c, d) \\
\tau_2^2 &\sim IG(e, f)
\end{aligned}
$$

This model is easily implemented in WinBugs with only minor changes to our previous programs:

```
model
{
 for(i in 1:9249)
  {
   for(t in 1:2)
    {
     y[i,t]~dnorm(mu[i,t],sigma2inv)
     mu[i,t]<-alpha[i]+alpha1*sex[i]+alpha2[i]*internet[i,t]
    }
  alpha[i]~dnorm(alpha0,tau2inv)
  alpha2[i]~dnorm(alpha20,tau20inv)
  }
 alpha0~dnorm(0,1.0E-4)
 alpha1~dnorm(0,1.0E-4)
 alpha20~dnorm(0,1.0E-4)

 sigma2inv~dgamma(.01,.01)
 sigma2<-1/sqrt(sigma2inv)

 tau2inv~dgamma(.01,.01)
 tau2<-1/sqrt(tau2inv)

 tau20inv~dgamma(.01,.01)
 tau20<-1/sqrt(tau20inv)
}
```

The results of this model suggest that there is considerable variation in the relationship between Internet usage and income across individuals. The estimated mean effect of Internet usage ($\alpha_{(2i)}$) was .205, and the estimated standard deviation for this effect (τ_2) was .224. This result yields (under the assumption that the random effect $\alpha_{(2)}$ is normally distributed) a 95% probability interval for the influence of Internet usage of [-.234, .644], which indicates that Internet usage may be, in some cases, harmful to wages (playing games at the office, lowering productivity?!).

What factors determine the influence of Internet usage on wages? In other words, why do some people appear to benefit from using the Internet, whereas others do not? We have previously decomposed the individual-specific random intercepts into an adjusted intercept and an effect of respondent's sex. When we begin to allow regression parameters (like the the one capturing the influence of Internet usage) to vary across individuals, we can also decompose it into a regression on higher level factors. For example, suppose we assumed that sex not only influenced the random intercept for wages, but also that it influences the extent to which Internet usage affects income. We can easily incorporate this idea into our model as follows. I switch notation slightly to avoid confusion:

$$
\begin{aligned}
y_{it} &\sim N(\alpha_i + \beta_i \text{Internet}_{it},\ \sigma^2) \\
\alpha_i &\sim N(\alpha_{(0)} + \alpha_{(1)}\text{sex}_i,\ \tau_\alpha^2) \\
\beta_i &\sim N(\beta_{(0)} + \beta_{(1)}\text{sex}_i,\ \tau_\beta^2) \\
\alpha_{(0)} &\sim N(m1, s1) \\
\alpha_{(1)} &\sim N(m2, s2) \\
\beta_{(0)} &\sim N(m3, s3) \\
\beta_{(1)} &\sim N(m4, s4) \\
\tau_\alpha^2 &\sim IG(a, b) \\
\tau_\beta^2 &\sim IG(c, d) \\
\sigma^2 &\sim IG(e, f)
\end{aligned}
$$

This model clarifies the hierarchical structuring of the data and parameters. Each individual's income is a function of his/her own intercept and slope, and these individual-level intercepts and slopes are determined, in part, by sex—a characteristic that differentiates individuals. The model consists of seven vague hyperprior distributions, one for each of the parameters that are not themselves endogenous within the model.

This model is sometimes called a multilevel or hierarchical model with cross-level interactions. The cross-level interactions, although not immediately apparent in the above specification, can be observed if we revert to the equation-based, more classical representation of the model. Under that approach:

$$
\begin{aligned}
y_{it} &= \alpha_i + \beta_i \text{Internet}_{it} + e_{it} \\
\alpha_i &= \alpha_{(0)} + \alpha_{(1)}\text{sex}_i + u_i \\
\beta_i &= \beta_{(0)} + \beta_{(1)}\text{sex}_i + v_i,
\end{aligned}
$$

with appropriate specifications for the variances of the errors at each level. If we then substitute the expressions for α_i and β_i into the first equation, we obtain:

$$
y_{it} =
$$

$$
\alpha_{(0)} + \alpha_{(1)}\text{sex}_i + u_i + \beta_{(0)}\text{Internet}_{it} + \beta_{(1)}\text{sex}_i \times \text{Internet}_{it} + v_i\text{Internet}_{it} + e_{it}.
$$

In this representation, we have a grand mean ($\alpha_{(0)}$) and an individual adjustment to it (u_i), a main effect of sex ($\alpha_{(1)}$), a time-constant main effect of Internet usage (β_0) and an individual adjustment to it (v_i), an interaction effect between sex and Internet usage ($\beta_{(1)}$), and an error term (e_{it}). The interaction term is considered a cross-level interaction, because sex is measured

at the individual level (the "group" in this context), whereas Internet usage is measured at the within-individual level. Historically, prior to the widespread use of hierarchical modeling, this model was estimated simply using OLS regression with the relevant interaction. However, as we have discussed, and as this equation shows, the OLS approach is not optimal, because it absorbs the various random quantities (i.e., u_i, $v_i$internet$_{it}$, and e_{it}) into a single error term for each individual. These error terms are assumed to be independent across time-specific observations, but, as the single subscripting for u_i and v_i suggest, they are not truly independent.

Returning to the Bayesian specification, the model can be implemented very easily in WinBugs with the following code:

```
model
{
 for(i in 1:9249)
  {
   for(t in 1:2)
    {
     y[i,t]~dnorm(mu[i,t],sigma2inv)
    mu[i,t]<-alpha[i]+beta[i]*internet[i,t]
    }
alpha[i]~dnorm(ma[i],tauinv.alpha)
beta[i]~dnorm(mb[i],tauinv.beta)
ma[i]<-alpha0 + alpha1*sex[i]
mb[i]<-beta0 + beta1*sex[i]
  }
alpha0~dnorm(0,1.0E-4)
alpha1~dnorm(0,1.0E-4)
beta0~dnorm(0,1.0E-4)
beta1~dnorm(0,1.0E-4)

sigma2inv~dgamma(.01,.01)
sigma2<-1/sqrt(sigma2inv)

tauinv.alpha~dgamma(.01,.01)
tau.alpha<-1/sqrt(tauinv.alpha)

tauinv.beta~dgamma(.01,.01)
tau.beta<-1/sqrt(tauinv.beta)
}
```

The key results of this model are not only that men make higher wages than women ($\alpha_{(0)} = 1.86$; $\alpha_{(1)} = .20$), but also that Internet usage has substantially higher returns for men than for women ($\beta_{(0)} = .18$; $\beta_{(1)} = .05$). In fact, based on these point estimates, the return to income of Internet usage for men is 28% greater than it is for women. A 95% interval estimate of this percentage is [11%, 48%].

9.2.3 Growth models

Often, we may wish to include time as one of our variables affecting an outcome. For example, in the previous model, we allowed the effect of Internet usage on wages to vary across individuals, but we could also consider that wages grow at differential rates for individuals. Similarly, we found earlier that the influence of Internet usage on wages varied across time. We may therefore consider specifying a model in which wages are expected to grow at differential rates for individuals, with Internet usage influencing the rate of growth. This type of model is often called a "growth model," or "latent growth model," because we are modeling the time-specific outcomes as realizations of an underlying growth process that unfolds across age/time at the individual level. Such a model may look like:

$$
\begin{aligned}
y_{it} &\sim N(\alpha_i + \beta_i t_{it},\, \sigma^2) \\
\alpha_i &\sim N(\alpha_{(0)} + \alpha_{(1)}\text{sex}_i + \alpha_{(2)}\text{Internet}_i,\, \tau_\alpha^2) \\
\beta_i &\sim N(\beta_{(0)} + \beta_{(1)}\text{sex}_i + \beta_{(2)}\text{internet}_i,\, \tau_\beta^2) \\
\alpha_{(0)} &\sim N(m1, s1) \\
\alpha_{(1)} &\sim N(m2, s2) \\
\alpha_{(2)} &\sim N(m3, s3) \\
\beta_{(0)} &\sim N(m4, s4) \\
\beta_{(1)} &\sim N(m5, s5) \\
\beta_{(2)} &\sim N(m6, s6) \\
\tau_\alpha^2 &\sim IG(a, b) \\
\tau_\beta^2 &\sim IG(c, d) \\
\sigma^2 &\sim IG(e, f).
\end{aligned}
$$

Although this model has a lengthy specification, it is has a fairly straightforward interpretation. Individual wages are expected to start and grow at individual-specific levels and rates (α_i and β_i, respectively). An individual's specific level and rate is then seen as depending on his/her sex and Internet usage. The remaining lines of the model specification are simply hyperpriors for the various parameters.

A couple of notes are in order regarding the growth model presented above. First, I have included Internet usage measured only at the first point in time. The reason for this is that the model is underidentified if we attempt to estimate it with Internet usage treated as a time-varying covariate influencing individual-specific effects of time (see Exercises). Second, given that this model only consists of two waves of data, the model is only measuring the extent to which sex and Internet usage influence change in wages over a single time interval, making the model nothing more than a slightly different parameterization of a change score regression model. Third, because of the limited

number of waves, some additional constraints must be enforced. One is that the error variance σ^2 must be constrained to be time invariant. Often, growth models allow this parameter to vary across time (see Bollen and Curran 2006), but here we simply cannot allow that, given our limitation of having only two time-specific measures per person. The results of this model can be found in Table 9.3.

Table 9.3. Results of "growth" model for two-wave panel of income and Internet use data.

Parameter Meaning	Parameter	Posterior Mean(s.d.)
Adjusted intercept for time-1 wages	α_0	1.74(.016)
Influence of sex on wages	α_1	0.259(.015)
Influence of Internet on wages	α_2	0.296(.016)
Adjusted intercept for change in wages	β_0	0.033(.009)
Influence of sex on change in wages	β_1	−0.013(.009)
Influence of Internet on change in wages	β_2	0.006(.009)
Residual variance in wages	σ^2	0.308(.002)
Residual variance in time-1 wages	τ_α^2	0.383(.004)
Residual variance in change in wages	τ_β^2	0.061(.006)

Note: Posterior means (and posterior standard deviations) are reported.

These results indicate that sex and Internet usage each influence baseline wages, with men earning more than women (see $\alpha_{(1)}$) and Internet users earning more than nonusers (see α_2). Indeed, the Internet effect is roughly 20% larger than the sex effect. The results also indicate that wages grew slightly across the one-year time period (see $\beta_{(0)}$). Wages grew less for men (see $\beta_{(1)}$), but more for Internet users (see $\beta_{(2)}$), although this effect was slight at best (observe the posterior standard deviation for $\beta_{(2)}$ compared with its mean). These results may also be written in equation form to clarify their interpretation:

$$\begin{aligned} E(\text{wages}_{it}) &= \alpha_i + \beta_i \\ E(\alpha_i) &= 1.74 + .259\text{male}_i + .296\text{Internet}_{i1} \\ E(\beta_i) &= .033 - .013\text{male}_i + .006\text{internet}_{i1}. \end{aligned}$$

For a fuller, more detailed example involving more time points of measurement, I examine health trajectories of individuals across a 20-year span. My assumption is that health tends to decline across the life course of individuals, and that baseline health and the rate of decline in health are a function of age, sex, race, area and type of residence, and education. My primary interest is in examining how socioeconomic status (measured by education) influences the

health differential across time. One hypothesis in the literature—the cumulative advantage hypothesis—argues that the health gap between high and low SES groups widens across age as a function of the cumulative disadvantage that low SES generates across the life course (see Lynch 2003). At young ages, risk factors like smoking and lack of health care access matter little, because most young adults are quite healthy. However, across age, exposure to risk factors accumulates and produces a larger health differential. An alternate hypothesis is the age-as-leveler hypothesis. This hypothesis argues that the health gap narrows across age because age overwhelms all risk factors—the biological effect of aging supercedes any socially based risk factor (see House et al. 1994). Often a selective mortality argument is also advanced to support this hypothesis: that the observed health gap at a particular age is ultimately a between-individual measure, and only the health of survivors is observed. Thus, those with the poorest health have been eliminated from the observed population, and the gap is simply a comparison of a robust subset of lower SES individuals with higher SES individuals. In other words, there are different populations being compared at young and older ages (see Lynch 2003 for extensive discussion).

A life course perspective suggests that we should examine trajectories of health for individuals, and that selective mortality should be "controlled out." One way to do this is to allow decedents to be included in the model, rather than to exclude them, as cross-sectional analyses must do (because only survivors can be observed in a cross-section). A Bayesian growth model can easily handle the unbalanced data that result from mortality, and health trajectories can even be estimated for individuals for whom we only observe a single measure. Their trajectories become a compromise between their observed measures and those of persons with similar covariate profiles who do survive. Ultimately, this approach underestimates the rate of decline in health, because surviving low-SES individuals drive the estimate of the mean growth rate, and surely decedents have/had steeper—but unobserved—rates of health decline. However, this argument implies that the finding with regard to the cumulative advantage hypothesis are conservative.

For this example, I again use the data from the National Health and Nutrition Examination Survey (NHANES) and its follow-ups, the National Health Epidemiologic Follow-up Surveys (NHEFS) (see Chapter 8 for a description). After eliminating individuals who were missing on one or another variable in the analyses and individuals whose final status in 1992 was unknown, the analytic sample consisted of 6,403 persons.

In this example, I include only individuals who were between 30 and 34 years of age at baseline, because age presents a problem in these analyses: The variable "age" represents both age and cohort. Research has shown that a common pattern for the health gap between individuals with low versus high SES across age is divergent until midlife and then convergent after (see House et al. 1994). This pattern is a function of two things: selective mortality and cohort change in the importance of education in affecting health (see Lynch

2003). Thus, for the sake of simplicity in this example, I restrict the analyses to the 608 individuals who fall in this age range, eliminating cohort effects.

I include age (mean = 32.0, s.d. = 1.4), sex (male = 1, 41.6%), race (nonwhite = 1, 12.3%), region (south = 1, 28.1%), urban residence (urban = 1, 23.2%), and education (in years, mean = 12.6, s.d. = 2.6, minimum = 0, maximum = 17) as second-level covariates that may influence the random intercept and slope factors. The outcome measure is self-rated health measured on a 5-point Likert scale ranging from excellent health (5) to poor health (1). Health measured on a 5-point scale is known to be a reliable and valid indicator of health (especially at younger ages), and the data are fairly symmetric, with a slight skew toward excellent health. I expect that individuals random intercepts are relatively high, and that in general, health declines between 30 and 55—the age range covered by the study. Furthermore, I expect that education differentiates health at baseline, with higher educated individuals having better health than lower educated ones. Finally, if the cumulative advantage hypothesis is true at least prior to age 55, education serves to reduce the rate of decline in health. This hypothesis implies that the growth rate in health is negative in general, but that education's influence on the growth rate is positive.

Below is the WinBugs program specifying the growth model:

```
model
{
 for(i in 1:608)
  {
   for(t in 1:pyrs[i])
    {
     h[i,t]~dnorm(mu[i,t],sigma2inv)
    mu[i,t]<-alpha[i]+beta[i]*yr[i,t]
    }
alpha[i]~dnorm(ma[i],tauinv.alpha)
beta[i]~dnorm(mb[i],tauinv.beta)
ma[i]<-alpha0 + alpha1*age[i] + alpha2*male[i] + alpha3*nonw[i] +
       alpha4*south[i] + alpha5*urban[i] + alpha6*educ[i]
mb[i]<-beta0                  + beta2*male[i]  + beta3*nonw[i]  +
       beta4*south[i]  + beta5*urban[i]  + beta6*educ[i]
  }
alpha0~dnorm(0,1.0E-4)
alpha1~dnorm(0,1.0E-4)
alpha2~dnorm(0,1.0E-4)
alpha3~dnorm(0,1.0E-4)
alpha4~dnorm(0,1.0E-4)
alpha5~dnorm(0,1.0E-4)
alpha6~dnorm(0,1.0E-4)
beta0~dnorm(0,1.0E-4)
beta2~dnorm(0,1.0E-4)
beta3~dnorm(0,1.0E-4)
beta4~dnorm(0,1.0E-4)
```

```
beta5~dnorm(0,1.0E-4)
beta6~dnorm(0,1.0E-4)

sigma2inv~dgamma(.01,.01)
sigma2<-1/sqrt(sigma2inv)

tauinv.alpha~dgamma(.01,.01)
tau.alpha<-1/sqrt(tauinv.alpha)

tauinv.beta~dgamma(.01,.01)
tau.beta<-1/sqrt(tauinv.beta)
}
```

This program, although longer than our previous growth model program because of the inclusion of additional level 2 covariates, is only slightly different from it. In order for WinBugs to handle unbalanced data—that is, data collected at different times and on different numbers of occasions for different respondents—I include a variable called `pyrs`, which tells the program at how many occasions the respondent was interviewed, and time of measurement is treated as a time-specific, individual-level variable. Individuals who die—or are lost—before the first follow-up (after baseline) contribute only a single person-year record and single measure of time. Persons who die—or are lost—before the second follow-up contribute two person-year records, etc. In these data, there are 16 persons who contribute one person-year record, 7 persons who contribute two records, 6 who contribute three, and 579 who contribute the maximum of four. These data provide some initial indication that there is some education-based selective mortality: The mean for education among persons who contribute 4 person-records is 12.7, whereas the mean for those who contribute fewer records is 10.9. In other words, the less-educated die earlier than the more-educated.

The remainder of the model is virtually identical to the one presented earlier, only with more covariates and therefore more hyperprior distributions. One note is in order: I do not include the effect of respondent's age on growth. The reason for this is that for age to influence the growth rate, either (1) the underlying latent health trajectories must be assumed to be nonlinear or (2) there are cohort differences in growth rates (see Mehta and West 2000).

I ran the program for 10,000 iterations and retained the last 1,000 samples for inference. Figure 9.2 shows 200 sampled values for the random intercepts and random slopes for four individuals. Person 1 only survived through the first wave of the study; person 17 survived through two waves; person 24 survived through three waves; and person 35 survived through all four waves. As the figure shows, the scatter of points is widest for person 1, reflecting the lack of certainty about this individual's true random intercept and slope values due to the existence of only one observed measure for his health. As the number of time points observed increases, the variance in the random intercept and slope for each individual decreases. For example, in the bottom

right plot, the random intercept and slope scatter is centered very narrowly over approximately (3, −.08), which indicates that we are fairly certain that this individual's latent trajectory starts around 3 health units at baseline and declines about .08 units per year.

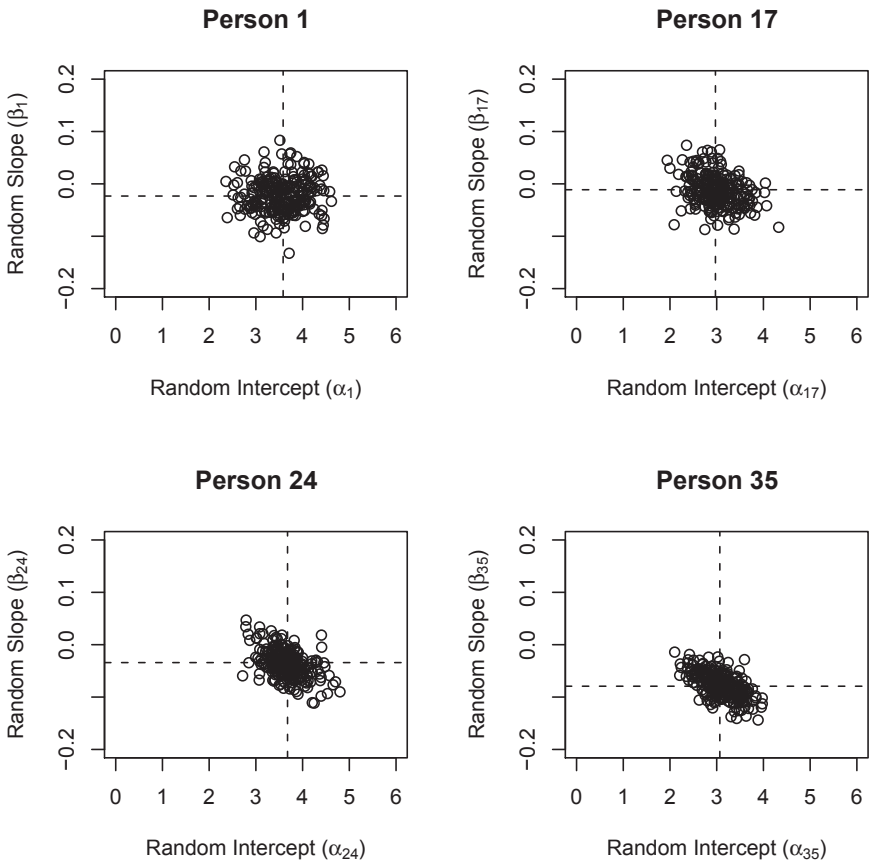

Fig. 9.2. Scatterplots of four persons' random intercepts and slopes from growth curve model of health (posterior means superimposed as horizontal and vertical dashed lines).

Table 9.4 presents the posterior means and standard deviations for the model parameters. The columns in the table report the influence of each covariate on the random intercept and random slope. The intercept for the random intercept term was 4.52. Older persons (recall the age range was only 30-34) reported worse health than younger persons at baseline (−.06). Men

reported better health at baseline than women (.05), and persons from the South reported worse health (−.05), but these effects were not substantially different from 0, based on posterior probabilities that the parameters were greater than (or less than) 0, truncated to the p-value ranges used by classical statistics (i.e., $p < .05$). Nonwhites and persons living in urban areas reported worse health than whites and persons living in other areas. Finally, education had a strong, positive effect on baseline health.

Almost none of the covariates influenced the random slope. The intercept for the random slope was negative, implying that the tendency was for health to decline slightly across the 20-year period. Males and nonwhites had a slightly steeper decline in health, although these effects would not be statistically significant by classical standards. Persons from the South and from urban areas had shallower declines in health than persons from other areas, although, again, these effects would not be statistically significant by classical standards. Finally, education had the expected positive effect (.001, $p < .1$), indicating that health trajectories do diverge across age (for the range from age 30 to age 55), such that persons with more education experience a shallower decline in health across age than persons with less education. Indeed, although the coefficient's magnitude appears small, the results indicate that a person with 17 years of schooling (the maximum) would experience a rate of health decline only 43% as great as a person with 0 years of schooling and only 76% as great as a person with 12 years of schooling.

Table 9.4. Results of growth curve model of health across time.

Variable	Random Intercept	Random Slope
Intercept	4.52(.70)***	−0.03(.01)**
Age	−0.06(.02)**	
Male	0.05(.08)	−0.006(.005)
Nonwhite	−0.54(.11)***	−0.006(.008)
South	−0.05(.08)	0.003(.005)
Urban	−0.23(.08)***	0.003(.005)
Education	0.11(.01)***	0.001(.0009)#
Variance	0.37(.03)	0.001(.0001)
Within-ind. Variance	0.42(.02)	

Note: The Bayesian estimates are posterior means. The p-values are the probabilities that the parameter exceeds 0 (either positively or negatively), truncated to the classical cutpoints of $\#p < .1$, $*p < .05$, $**p < .01$, $***p < .001$.

The results can be used in two ways to predict health trajectories. First, we may directly use the simulated latent intercepts and slopes for individuals *in the sample* (as shown in Figure 9.2). For example, we could use the poste-

rior means for these simulated intercepts and slopes to construct an expected trajectory: $y_{it} = \mu_{\alpha_i} + \mu_{\beta_i} \times t$. Second, we may use the posterior distributions for the model parameters—the covariate effects—to compute predicted latent intercepts and slopes for persons with particular covariate profiles. This approach allows us to predict trajectories for individuals *out of the sample*, in addition to those in the sample. Person 1 shown in Figure 9.2 was a 31-year-old nonwhite male living in a non-southern, urban area with 11 years of schooling. Based on the posterior means for the effects of the covariates, this individual would have a predicted intercept of 3.22 for his health trajectory and a predicted slope of $-.022$.

Figure 9.3 shows these two types of predicted trajectories for the four individuals shown in the previous figure, along with their observed health measures. The solid line in each graph shows the predicted trajectory based on the posterior means of the simulated, individual-specific random intercepts and slopes (i.e., the simulated values from Figure 9.2). The dashed line in each graph shows the model predicted trajectory based on the posterior means of the parameters applied to each individual's covariate profile.

There is a substantial difference between these two trajectories, as well as between either trajectory and the observed health measures. This variation reflects the different types of (error) variance captured by the model. The discrepancies between the solid-line trajectories and the observed health measures are captured by the within-individual error variance parameter σ^2. In brief, we do not expect each individual's health measure to fall exactly on the solid line, because a number of unobserved factors may "bump" an individual off of his/her expected, latent health trajectory at any point in time. Instead, what the model has attempted to capture is the best fitting line for the observed health measures. This error may be reduced by including time-specific measures into the model as we did in the previous section in the model in which we included Internet usage as a time-varying covariate.

The discrepancies between the solid and dashed-line trajectories, on the other hand, reflect the extent of between-individual variation captured (or not!) by the covariates in the model. Put another way, if the covariates *perfectly* explained all differences between individuals' health trajectories, the solid and dashed lines would perfectly coincide. The fact that these lines are not overlapping suggests that our covariates do a poor job differentiating individuals in the sample. This conclusion is foretold by the lack of strong results in Table 9.4, especially with respect to the general lack of effect of covariates on the latent growth rate. Indeed, if we consider the estimated rate of decline in health for each individual in Figure 9.3, all four individuals are expected to have similar, shallow rates of health decline that obviously do not match the observed health declines (or those predicted by the simulated individual-specific random effects). In contrast, the estimated intercepts for these trajectories show greater variability, reflecting the stronger effects of the covariates in predicting baseline health. In an additional model (not shown), I re-estimated this growth model with no covariates to obtain estimates of

the variance of the mean latent intercept and slope. An R^2 for the effects of the covariates on the estimated latent intercept was found by computing $1 - \tau^2_{\alpha,\text{cov}}/\tau^2_{\alpha,\text{nocov}}$, using the posterior means for these variance parameters from the two models. A similar calculation was performed for the variance of the latent slope (τ^2_β). The results indicated that the covariates reduced the between-individual variance in the latent intercept by 29% (i.e., $R^2 = .29$), but the covariates reduced the between-individual variance in the latent slope by only 10%. These results confirm that our covariates have little effect on the latent slope, and therefore, it is no surprise that our two types of predicted trajectories differ substantially.

As a final note on growth modeling, the use of growth models has been rapidly expanding in psychology and sociology over the last decade, in part because the growing availability of longitudinal (panel) data has enabled the investigation of life course processes for which growth modeling is well suited. Additionally, growth models have become increasingly popular, because they can be estimated via a variety of software packages, including HLM and various structural equation modeling packages (see Willett and Sayer 1994; see also McArdle and Epstein 1987, Meredith and Tisak 1990 and Rogosa and Willett 1985). The HLM approach closely resembles the modeling strategy developed in this section. The structural equation modeling approach, on the other hand, is in some ways more intuitive, although it is mathematically equivalent to the Bayesian and HLM approaches.[5] However, that approach typically requires balanced data—that is, data that have been collected at the same time and at all times for all individuals in the sample. This latter requirement can be relaxed by assuming that individuals who are missing at one or more occasions are missing at random and estimating the model using a full information maximum likelihood (FIML) estimator. The former restriction, however, is not easily relaxed. However, estimating the model using a Bayesian approach or using other hierarchical modeling packages offer a straightforward way to handling unbalanced data. For more details on latent growth modeling within a structural equation modeling framework, I highly recommend Bollen and Curran (2006).

9.3 A note on fixed versus random effects models and other terminology

One issue that makes understanding hierarchical models difficult is the terminology that different disciplines and statistical paradigms use to describe various features of the models. In this section, I hope to clarify some of the terminology, although there is certain to be some disagreement regarding my

[5] In fact, for each growth model example presented here, I estimated the equivalent model using a structural equation approach. The results were nearly identical.

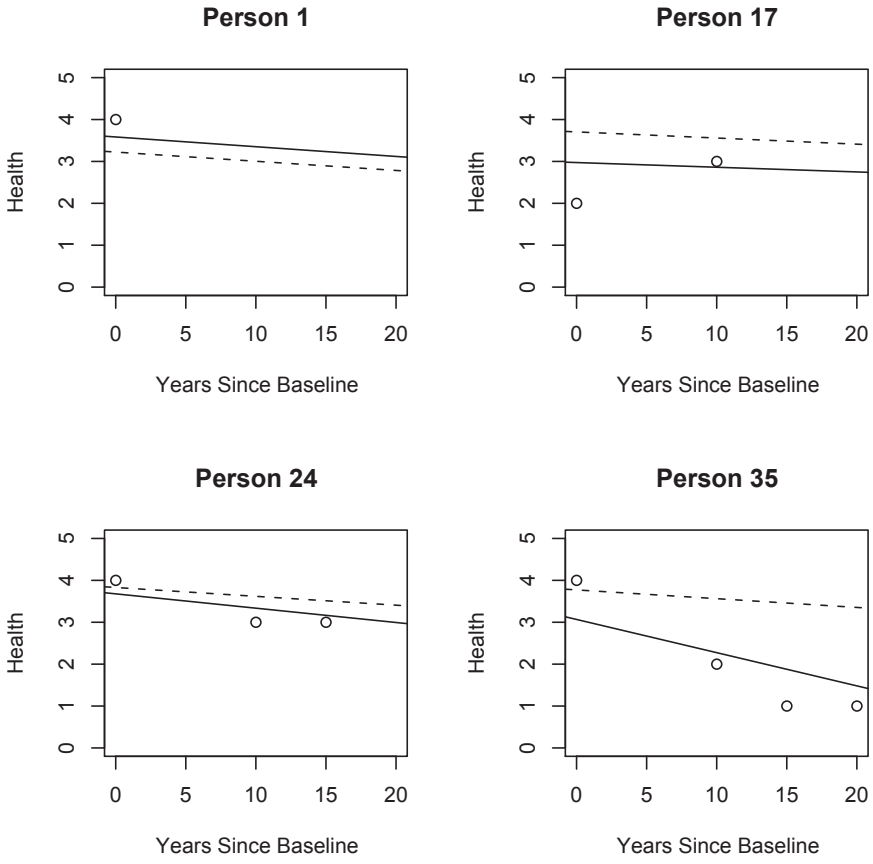

Fig. 9.3. Predicted trajectories and observed health for four persons: The solid lines are the predicted trajectories based on the posterior means of the random intercepts and slopes from Figure 9.2; and the dashed lines are the predicted trajectories based on the individuals' covariate profiles and posterior means of the parameters in Table 9.4

use of terms. To be sure, many of the terms used in discussions of hierarchical modeling have not had static definitions over time, adding to the confusion.

First, the terms "fixed effects" and "random effects" are frequently tossed about in discussions of hierarchical modeling. From a Bayesian perspective, controversy over these terms is often much ado about nothing, because from a Bayesian view (1) parameters are seen as random quantities arising from proper probability distributions, making all effects "random"; and (2) fixed effects models generally contain "random" effects, making the distinction between fixed and random effects models somewhat dubious. Consider the OLS

regression model $Y = X\beta + e$, which is often considered to be a fixed effects regression model. In this model, X is considered a *fixed* variable matrix, and β is considered a fixed regression parameter vector—i.e., "fixed effects." From a classical statistical standpoint, the only random quantity in this model is the vector e, which is generally portrayed as random by the expression $e \sim N(0, \sigma_e^2 I_n)$. In other words, in the classical representation, e is a random effect because it comes from a *specified* probability distribution. The β vector, on the other hand, is considered fixed—these parameters are what they are in the population and do not stem from a probability distribution. From a Bayesian view, however, β *may* be considered as a vector of random effects, because we can produce a posterior probability distribution for the vector. The only difference between the Bayesian approach to this model and the classical approach is that the classical approach implicitly assumes uniform prior distributions on β, whereas a Bayesian approach makes this assumption explicit in the formulation of the prior. Whether we consider β fixed or random, nonetheless, one could argue that the model is a random effects model with some fixed effects (β) if the priors are left unspecified.

Next, consider the basic random effects model considered in this chapter in which individuals have their "own" intercepts or means:

$$y_{it} \sim N(\alpha_i,\, \sigma^2),$$

with $\alpha_i \sim N(\alpha_0, \tau^2)$. From a Bayesian perspective, this model is considered a random effects model, because the α_i are treated as arising from a normal distribution with parameters α_0 and τ^2. A classical statistician, on the other hand, might introduce a dummy variable for each observation, coupled with a β for each dummy variable, and call this model a fixed effects model, because the β vector could be considered a fixed parameter vector. In other words, the classical statistician may specify the model as an OLS regression model, $Y = X\beta + e$, again with X being a matrix of dummy variables, β being a vector of effects of these dummy variables, and $e \sim N(0, \sigma_e^2)$ being considered the only random quantity. The data structure in this specification would be a person-year matrix, with each individual contributing t rows, with X having dummy variables for each person-record corresponding to each person. From a Bayesian view, this is a random effects model, but from a classical view, this is still a fixed effects model. The Bayesian, however, recognizes that, again, the only difference between these models is the explicit statement that each α_i (intercept/mean) has a proper prior distribution; the classical statistician again implicitly assumes uniform priors distributions on these "fixed" effects.

The next step in our modeling process in this chapter was to incorporate additional individual-level (level 2) variables via essentially the decomposition of the intercept term into a regression on individual-level factors. Specifically, we allowed individuals' α_i to be a function of their sex. One representation of this model is:

$$y_{it} \sim N(\alpha_i, \sigma^2)$$
$$\alpha_i \sim N(\alpha_{(0)} + \alpha_{(1)}\text{sex}_i,\ \tau^2),$$

along with appropriate (vague) hyperpriors for the hyperparameters $\alpha_{(0)}$, $\alpha_{(1)}$, σ^2, and τ^2. Alternatively, but equivalently, the model may be specified as we did earlier:

$$y_{it} \sim N(\alpha_i + \alpha_{(1)}\text{sex}_i, \sigma^2)$$
$$\alpha_i \sim N(\alpha_0, \tau^2),$$

again with appropriate priors for $\alpha_{(0)}$, $\alpha_{(1)}$, σ^2, and τ^2. A Bayesian then would call this a random intercept model. The classical statistician, on the other hand, would write this model as:

$$y_{it} = \alpha_i + \alpha_1\text{sex}_i + e_{it}$$
$$e_{it} \sim N(0, \sigma^2)$$
$$\alpha_i = \alpha_0 + u_i$$
$$u_{it} \sim N(0, \tau^2).$$

After substituting the third equation into the first, we would obtain:

$$y_{it} = \alpha_0 + u_i + \alpha_1\text{sex}_i + e_{it}.$$

Under this representation, the classical statistician would claim that α_0 and α_1 are fixed effects, and that the only random effects are u_i and e_{it}. If u_i is considered a component of α_0, then the model *could* be called a random intercept model with fixed effects. Once again, however, the Bayesian would argue that the explicit assignment of proper priors for α_0 and α_1 makes the model a random effects model: The classical approach is implicitly assuming uniform priors on these parameters.

In subsequent steps of our modeling building process, we included Internet usage as a time-varying (level 1) variable, and we eventually allowed the influence of Internet usage on wages to vary across individuals and we allowed the individual-specific influence of Internet usage to be a function of individuals' sex:

$$y_{it} \sim N(\alpha_i + \beta_i\text{Internet}_{it},\ \sigma^2)$$
$$\alpha_i \sim N(\alpha_0 + \alpha_1\text{sex}_i,\ \tau_\alpha^2)$$
$$\beta_i \sim N(\beta_0 + \beta_1\text{sex}_i,\ \tau_\beta^2),$$

once again with appropriate hyperprior distributions for the higher level hyperparameters. Using Bayesian terminology, this model is a "random coefficients" model, because the regression coefficient β_i is allowed to vary across individuals. The classical approach, however, would find

$$y_{it} = \alpha_0 + \alpha_1 \text{sex}_i + u_i + \beta_0 \text{Internet}_{it} + \beta_1 \text{sex}_i \text{Internet}_{it} + v_i \text{Internet}_{it} + e_{it}$$

after subsitution and might call the model a fixed effects model with random intercepts, random coefficients, and cross-level interactions.

To make a long story short, all of these models are considered hierarchical models because there is a hierarchical structure to the parameters. They may also be called multilevel models because the variables in the models are measured at different levels (time-specific measures and individual-level measures). Additionally, all the models contain random effects and may therefore be called random effects models, despite the fact that the classical statistician may prefer to include the term "fixed effects" in describing them. When the regression parameters—and not simply the intercepts—are allowed to vary across individuals, they may be called "random coefficient models." When time is included as a variable and its influence—a random coefficient—is allowed to vary across individuals, the model may be called a "(latent) growth (curve) model." Finally, all of these models are sometimes called "mixed models," because they generally include both fixed and random effects when the terms "fixed" and "random" are applied to distinguish between effects that have implicit versus explicit prior distributions.

9.4 Conclusions

In this chapter, we have covered considerable ground. We began by discussing how we can use the conditional probability rule to produce hierarchical structure in the parameters and obtain a posterior distribution for all parameters in a simple model without covariates. We then discussed how hierarchical regression models can easily be constructed to capture hierarchical structure in both the parameters and the data with variables measured at different levels. Finally, we showed how the general hiearchical linear regression model can be specialized to examine growth in the outcome over time by including time in the model as a covariate. As the chapter demonstrated, the Bayesian approach is naturally suited to hierarchical modeling. Indeed, the Bayesian approach handles hierarchicality so easily that virtually no text on Bayesian statistics omits hierarchical modeling, and I can find no Bayesian text that *only* covers hierarchical modeling. For further reading on Bayesian hierarchical modeling, as I said at the beginning of the chapter, I recommend Gelman et al. (1995). I also recommend Gill (2002) for an introductory exposition, and I suggest Spiegelhalter et al. (1996) for an illustrative example of growth modeling.

9.5 Exercises

1. Show the steps in the derivation of the conditional posterior distribution for τ in Equation 9.5.
2. Explain why the values chosen to complete the hyperprior specification in Section 9.2.1 are noninformative.
3. Explain in your own words why the first growth model presented in Section 9.2.3 cannot allow Internet usage to be a time-varying variable in the model as it is specified.
4. Using your own data, write an R routine to estimate a growth model. Then write a WinBugs routine to estimate the same model. Are the results similar?