Part II.
Statistical Modeling

In part II, we introduce a class of statistical models that generalize linear regression for time-series, cross-sectional analyses. We also provide new methods for identifying, formalizing, and incorporating prior information in these and other models. Chapter 4 introduces our model and a new framework for generating priors. Chapter 5 extends the framework to grouped continuous variables, like age groups. Chapter 6 explains how to connect modeling choices to known substantive information. Then chapter 7 implements our framework for a variety of other variables, like those which vary over geographic space, and various types of interactions. Chapter 8 provides more detailed comparisons between our model and spatial models for priors on coefficients and extends our key results and conclusions to Bayesian hierarchical modeling.

4 The Model

4.1 Overview

The specific models introduced in this chapter all depend on the specification of priors, which we introduce here and then detail in the rest of part II. Details of how one can compute estimates using this model appear in part III. Our strategy is to begin with the models in chapter 3 that use covariates and add in information about the mortality age profile in a different way than in the models in chapter 2. After putting all the information from both approaches in a single model, we then add other information not in either, such as the similarity of results from neighboring countries or time periods.

In this section, we use a general Bayesian hierarchical modeling approach to information pooling.[1] Although developing models within the Bayesian theory of inference is entirely natural from the perspective of many fields, in several respects it is a departure for demography. It contrasts most strikingly with the string of scholarship extending over most of the past two centuries that seeks to find a low-dimensional parametric form for the mortality age profile (see section 2.4 and the remarkable list in Tabeau 2001). It has more in common with principle components approaches like Lee-Carter, in that we also do not attempt to parameterize the age profile with a fixed functional form, but our approach is more flexible and capable of including covariates as well as modeling patterns known from prior research in demography or any other patterns chosen by the researcher. Our approach also contrasts with the tendency of demographers to use their detailed knowledge only as an ex post check on their results. We instead try to incorporate as much of this information as possible into the model. Our methods tend to work better only when we incorporate information demographers have about observed data or future patterns.

The opposite, of course, applies too. Researchers forecasting variables for which no prior quantitative or qualitative analyses or knowledge exists will not benefit from the use of our methods. And those who use incorrect information may degrade their forecasts by adding priors.

Although our approach can work in principle with any relevant probability density for log-mortality, including those based on event count models discussed in section 3.1.1, we fix ideas by developing our model by building on the equation-by-equation least-squares

[1] For other approaches to Bayesian hierarchical modeling, and for related ideas, see Blattberg and George (1991), Gelman et al. (2003), Gelman and Hill (2007), Gill (2002), and Western (1998).

model, described in section 3.1.2. This model is

$$m_{it} \sim \mathcal{N}\left(\mu_{it}, \frac{\sigma_i^2}{b_{it}}\right) \quad i = 1, \ldots, N, \quad t = 1, \ldots T \tag{4.1}$$
$$\mu_{it} = \mathbf{Z}_{it}\boldsymbol{\beta}_i,$$

where, as before, m_{it} is the log-mortality rate (or a generic dependent variable) with mean μ_{it} and variance σ_i^2/b_{it}, b_{it} is some exogenous weight, and $\mathbf{Z}_{it}$ is a vector of exogenous covariates. We are not concerned with the choices of the weights b_{it} and the covariates $\mathbf{Z}_{it}$ here (we discuss these specification decisions in chapter 6). Although they are crucial, their specifics have no effect on the overall structure of our model.

Specification 4.1 forms the basic building block of our hierarchical Bayesian approach, and so we now interpret the coefficients $\boldsymbol{\beta}_i$ and the standard deviations σ_i as random variables, with their own prior distributions. We denote the prior for the variables σ_i generically as $\mathcal{P}(\sigma)$. The prior for the coefficients $\boldsymbol{\beta}$, which usually depends on one or more hyperparameters θ, we denote by $\mathcal{P}(\boldsymbol{\beta}|\theta)$. The hyperparameters θ also have a prior distribution $\mathcal{P}(\theta)$. (By our notation conventions, $\mathcal{P}(\theta)$ and $\mathcal{P}(\sigma)$ are different mathematical expressions; see appendix A.)

We choose the specific functional form of the priors $\mathcal{P}(\sigma)$ and $\mathcal{P}(\theta)$ to make the computations simple (usually a Gamma or inverse-Gamma density), with the mean and variance set to be diffuse so they do not have an important effect on our results. In contrast, our central arguments in this chapter, and most of the rest of this book, are about the specification for the prior for the coefficients, $\mathcal{P}(\boldsymbol{\beta}|\theta)$. This prior will be taken as highly informative, reflecting considerable prior knowledge. The issue at hand is deciding precisely how to formalize this prior knowledge in this density.

Using the likelihood function $\mathcal{P}(m|\boldsymbol{\beta}, \sigma)$ from equation 3.8 (page 45), and assuming that σ is a priori independent of $\boldsymbol{\beta}$ and θ, we express the posterior distribution of $\boldsymbol{\beta}$, σ and θ conditional on the data m as

$$\mathcal{P}(\boldsymbol{\beta}, \sigma, \theta|m) \propto \mathcal{P}(m|\boldsymbol{\beta}, \sigma)\left[\mathcal{P}(\boldsymbol{\beta}|\theta)\mathcal{P}(\theta)\mathcal{P}(\sigma)\right], \tag{4.2}$$

where $\mathcal{P}(\boldsymbol{\beta}, \theta, \sigma) \equiv \mathcal{P}(\boldsymbol{\beta}|\theta)\mathcal{P}(\theta)\mathcal{P}(\sigma)$ is the prior. Once the prior densities have been specified, we usually summarize the posterior density of $\boldsymbol{\beta}$ with its mean,

$$\boldsymbol{\beta}^{\text{Bayes}} \equiv \int \boldsymbol{\beta}\mathcal{P}(\boldsymbol{\beta}, \sigma, \theta|m)\, d\boldsymbol{\beta} d\theta d\sigma, \tag{4.3}$$

(or the mode) and can then easily compute forecasts using one of the three methods described in section 3.1.3.

This section provides a framework for the information pooling problem. By choosing a suitable prior density for $\boldsymbol{\beta}$, we summarize and formalize prior qualitative knowledge about how the coefficients $\boldsymbol{\beta}$ are related to each other, so that information is shared among cross sections. If the prior for $\boldsymbol{\beta}$ is specified appropriately, the information content of our estimates of $\boldsymbol{\beta}$ will increase considerably. This, in turn, can result in more informative and more accurate forecasts.

4.2 Priors on Coefficients

As we have described, a common way to derive a prior for $\boldsymbol{\beta}$ is to use the following kind of prior knowledge: "similar" cross sections should have "similar" coefficients. The most common approach is to use a class of Markov random field priors, which are an example of an intrinsic autoregressive prior. These models are closely related to the autoregressive priors pioneered by Besag and his colleagues (Besag, 1974, 1975; Besag and Kooperberg, 1995; Iversen, 2001) in that they allow spatial smoothing for units like age groups that vary over nongeographical space. The priors formalize this knowledge by introducing the following density:

$$\mathcal{P}(\boldsymbol{\beta}|\Phi) \propto \exp\left(-\frac{1}{2}H^{\beta}[\boldsymbol{\beta}, \Phi]\right), \tag{4.4}$$

where

$$H^{\beta}[\boldsymbol{\beta}, \Phi] \equiv \frac{1}{2}\sum_{ij} s_{ij}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|^2_{\Phi}, \tag{4.5}$$

where we use the notation $\|\mathbf{x}\|^2_{\Phi}$ to refer to the weighted Euclidean norm $\mathbf{x}'\Phi\mathbf{x}$ and where the symmetric matrix s is known as the *adjacency matrix*, and its elements can be thought of as the inverse of the "distance," or the proximity, between cross section i and cross section j.[2] It is useful, for future reference, to write equation 4.5 in an alternative way:

$$H^{\beta}[\boldsymbol{\beta}, \Phi] = \sum_{ij} W_{ij}\boldsymbol{\beta}_i'\Phi\boldsymbol{\beta}_j, \tag{4.6}$$

where $W = s^{+} - s$ is a positive semidefinite symmetric matrix whose rows sum to 1 (see appendix B.2.6, page 237). The matrix Φ is a generic symmetric, positive definite matrix of parameters, which help summarize the distance between vectors of coefficients. Because it is usually unknown, the matrix is considered to be a set of hyperparameters to be estimated, with its own prior distribution. In practice this matrix is likely to be taken to be diagonal in order to limit the number of the unknowns in the model, although this specification implies the strong assumption that elements of the coefficient differences $(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j)$ are a priori independent of each other. It also implies that Φ is constant over i, which we show later is highly improbable in many applications.

The function $H^{\beta}[\boldsymbol{\beta}, \Phi]$ assumes large values when similar cross sections (e.g., s_{ij} "large") have coefficients far apart (i.e., $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_{\Phi}$ is also "large"). Therefore, equation 4.4 simply says that, a priori, the most likely configurations of coefficients $\boldsymbol{\beta}$ are those in which similar cross sections have similar coefficients or, in other words, those in which the coefficients $\boldsymbol{\beta}$ vary smoothly across the cross sections.

A key point is that the prior defined by equations 4.4 and 4.5 is *improper* (which means that the probability density in equation 4.4 integrates to infinity; see appendix C).

[2] Although constraining the elements of this matrix to be positive is consistent with their interpretation as proximities, the constraint is not necessary mathematically. The only constraint on s is that the quadratic form defined by equation 4.5 be positive definite.

The improperness stems from the fact that the function $H^{\beta}[\boldsymbol{\beta}, \Phi]$ is constant and equal to 0 whenever $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ are equal, regardless of the levels at which the equality occurs (i.e., in the subspace $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j, \forall i, j = 1, \ldots, N$). This causes no statistical difficulties: because the likelihood is proper (and normal), the posterior is always proper. Indeed, an improper prior is a highly desirable feature in most applications, because it constrains the regression coefficients not to be close to any particular value (which would normally be too hard to specify from prior knowledge) but only to be similar to each other. In other words, the prior density in equation 4.4 is sensitive not to the absolute values or levels of the coefficients, only to their relative values.

Example 1 Suppose the cross sections are labeled only by countries (with no age group subclassification). Then s could be a symmetric matrix of zeros and ones, where $s_{ij} = 1$ indicates that country i and country j are "neighbors," in the sense that we believe they should have similar regression coefficients. Neighbors could also be coded on the basis of physical contiguity, proximity of major population areas, or frequency of travel or trade between the countries. In practice, the matrix s would need to be constructed by hand by a group of experts on the basis of their expectations of which countries should have similar coefficients, which, of course, requires the experts to understand the meaning of all the regression coefficients and how they are supposed to vary across countries. ☒

Example 2 Suppose cross sections are labeled by age groups, or by a similar variable with a natural ordering (with no country-level subclassification). Then s could be a tridiagonal matrix of ones, so that every age group has as neighbors its two adjacent age groups. A more general choice is a band matrix, with the size of elements decaying as a function of the distance from the diagonal. Although the general pattern desired may be clear from prior knowledge, choosing the particular values of the elements of s in this situation would be difficult, because they do not directly relate to known facts or observed quantities. ☒

4.3 Problems with Priors on Coefficients

For any Bayesian approach to model 4.1, we will ultimately need a prior for $\boldsymbol{\beta}$. The issue is how to turn qualitative and impressionistic knowledge into a specific mathematical form. But herein lies a well-known disconnect in Bayesian theory: because the prior knowledge we have is typically in a very different form than the probability density we ultimately need, the task of choosing a density often requires as much artistic choice as scientific analysis. In many Bayesian models, this disconnect is spanned with a density that forms a reasonable approximation to prior knowledge.

Unfortunately, in some applications of Bayesian models with covariates, the disconnect can be massive and the resulting density chosen is often inappropriate. Our critique applies to many Bayesian spatial or hierarchical models that put a prior on a vector of coefficients and where the prior is intended to convey information. The problem here is that the jump from qualitative knowledge to prior density is too large, and some steps are skipped or intuited incorrectly. This argument applies to many models with spatial smoothing, like that in equation 4.4, and more generally to hierarchical models with clusters of units that include covariates. We describe these problems here with spatial smoothing and make the extension to hierarchical models, featuring clusters of exchangeable units, in section 8.2.

4.3.1 Little Direct Prior Knowledge Exists about Coefficients

To put a prior on the vector $\boldsymbol{\beta}$, we need to be in possession of nonsample knowledge about it. When an element of $\boldsymbol{\beta}$ coincides with a specific causal effect, the claim to nonsample knowledge is complicated, but sometimes plausible. For example, we know that 25 years of tobacco consumption causes a large increase in the probability of death (from lung cancer, heart disease, and other causes) in humans. However, the $\boldsymbol{\beta}$ in our models are at the population level, and so they are not necessarily causal effects. For example, if we observe that tobacco consumption is positively related to lung cancer mortality across countries, it may be that smokers are getting lung cancer, but it could also be true—on the basis of the same patterns in the aggregate data—that it is the nonsmokers who happen to live in countries with high tobacco consumption who are dying at increasing rates. Whether we can imagine the reason for such a pattern is unimportant. It can occur, and if it does, the connection from the aggregate level relationship to the individual level at which the biological causal effect is known may be severed. This, of course, is an example of the well-known ecological inference problem (Goodman, 1953; King, 1997), the point being that without special models to deal with the problem, $\boldsymbol{\beta}$ may not contain causal effects even for a covariate as apparently obvious as tobacco consumption.

A better case for the claim of prior knowledge about $\boldsymbol{\beta}$ may be variables where the causal effect operates at the country level. For example, a democratic electoral system or a comprehensive health care system may lead to lower mortality from a variety of causes. Although these effects would also operate at the individual level, the causal effect could plausibly occur at the societal level. In that situation, no ecological inference problem exists, and the case that we may really possess prior knowledge about at least this coefficient is more plausible.

Even, however, when one coefficient is truly causal, its interpretation is clear, and much prior knowledge exists about its likely direction and magnitude, it is typically not the only coefficient. Normally, we have a set of control variables with corresponding coefficients. The problem is that coefficients on control variables are treated as nuisance parameters, are typically not the subject of study, and are rarely of any direct interest. As such, even for regressions that include well-specified causal effects, we will often have little direct prior knowledge about most of the coefficients.

In some areas, of course, whole literatures have built up around repeated analyses of similar specifications that lead to much knowledge about coefficients. For example, in epidemiological studies, age and sex are often the only important confounders and are typically used to stratify the data sets prior to analyses. Many of the statistical specifications then include a treatment variable only weakly correlated with the pretreatment control variables included in the analysis, each of which has only a small effect on the outcome. As Greenland (2001, p. 667–668) points out, sufficient knowledge does exist to put priors on the coefficients in situations like these. Indeed, many of these studies are based on case-control designs where estimating base probabilities was not done and long thought impossible (although it has now been shown to be straightforward; see King and Zeng 2002), and so in these areas putting priors on the expected outcome is more difficult and perhaps less natural.

A final point is that $\boldsymbol{\beta}$ is not scale invariant with respect to $\mathbf{Z}$: if we double $\mathbf{Z}$, we are also halving $\boldsymbol{\beta}$. This is not a problem if we truly understand the coefficients, because we would merely scale everything appropriately and set the prior to suit. When the coefficients'

values are not fully understood, however, several problems can ensue. The main problem here is that the whole model requires that $\boldsymbol{\beta}$ take on the same meaning for all cross-sectional units. If the meaning or scale of $\mathbf{Z}$ changes at all, however, then the prior should change. Yet, the parameters Φ in equation 4.4 have no subscript and are assumed constant over all units. In some situations, this is plausible but, even for variables like GDP, we expect some changes in scale over the units, even after attempts to convert currencies and costs of living.

This problem is sometimes addressed by standardizing $\mathbf{Z}$ in some way, such as by subtracting its sample mean and dividing by its sample standard deviation. This undoubtedly helps in some situations, but it just as certainly does not solve the problem. For a simple example, suppose one covariate is GDP per capita and another is real disposable income per 100 population. Suppose that the right normalization here is to multiply GDP by 100 (even though this assumes away a host of other potential problems). Now suppose that, for whatever reason, GDP per capita varies very little over countries in some data set, but real disposable income varies enormously. In that situation, standardization would exacerbate the problem rather than solve it. The general problem here is that the sample does not necessarily contain sufficient information with which to normalize the covariates. Some exogenous information is typically needed.

4.3.2 Normalization Factors Cannot Be Estimated

Whether the coefficients are meaningful or not, the prior in equation 4.4 contains the expression $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_\Phi$, which implies that the coefficients can all be made comparable. In particular, it assumes that we can translate the coefficient on one variable in a single cross-sectional regression to the scale of a coefficient on another variable in that cross section or some other cross section. Indeed, this prior specifies a particular metric for translation, governed by the hyperprior parameter matrix Φ.

To be more specific, we denote individual explanatory variables by the index v, and rewrite equation 4.5 (page 59) as

$$\begin{aligned} H^\beta[\boldsymbol{\beta}, \Phi] &\equiv \frac{1}{2}\sum_{ij} s_{ij}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_\Phi^2 \\ &= \sum_{ijv} W_{ij}\boldsymbol{\beta}_i^v b_j^v, \end{aligned} \tag{4.7}$$

where W is a function of s defined in appendix B.2.6 (page 237), and, most importantly,

$$b_j^v \equiv \sum_{v'} \Phi_{vv'}\boldsymbol{\beta}_j^{v'} \tag{4.8}$$

is the translation of coefficient v' in cross section j, into the same scale as that for coefficient v in cross section i.

As is more obvious in this formulation, Φ serves the critical role of normalization constants, making it possible to translate from one scale to another. For example, if we multiply degrees Celsius by 9/5 and add 32, we get degrees Fahrenheit, where the numbers 9/5 and 32 are the normalization constants. The translations that Φ must be able to perform include normalizing to the same scale (1) coefficients from different

covariates in the same cross-sectional regression, (2) coefficients from the same covariate in different cross-sectional regressions, and (3) coefficients from different covariates in different cross-sectional regressions. Each of these three cases must be made equivalent via normalization, and all this prior knowledge about normalization must be known ex ante and coded in Φ.

The role of the normalization can be seen even more clearly by simplifying the problem to one where Φ is diagonal. In this situation, the normalization factor is especially simple:

$$b_j^v = \Phi_{vv}\boldsymbol{\beta}_j^{v'},$$

and so Φ_{vv} simply provides the weights to multiply into the coefficient vector in one cross section to get the coefficient vector in another cross section.

This alternative formulation is appropriate only if we have prior knowledge that different components of $\boldsymbol{\beta}$ are independent. In other words, although equation 4.8 contains the correct normalization, regardless of independence assumptions, the prior in equation 4.7 allows us to use only those parts of the normalization that are relevant to forming the posterior, and independence among components of $\boldsymbol{\beta}$ means that the cross-product terms (i.e., when $v \neq v'$) in equation 4.8 would not be needed. Although assuming that elements of a prior are independent is common in Bayesian modeling, the assumption of independence is far from innocuous here, because the result can greatly affect the comparability of coefficients from different cross sections or variables and thus can enormously influence the final result.

The key to putting priors on coefficients is knowing Φ. Without this knowledge, the translation from one scale to another will be wrong, the prior will not accurately convey knowledge, and our estimates and forecasts would suffer. Unfortunately, because we often know little about many of the $\boldsymbol{\beta}$ coefficients, researchers usually know even less about the values in Φ. Any attempt within the Bayesian theory of inference to bring the data to bear on the prior parameter values will fail, which is easy to see by trying to estimate Φ by maximizing the posterior: because Φ does not appear in the likelihood, the entire likelihood becomes an arbitrary constant and can be dropped. As such, under Bayes, the data play no role in helping us learn about Φ; all information about it must come from prior knowledge, which, of course, is the problem.

Some scholars try to respond to the lack of knowledge of Φ as good Bayesians by adding an extra layer to the modeling hierarchy and putting a proper hyperprior on Φ. Ultimately, however, we always need to choose a mean for the distribution of Φ. And that deeply substantive choice will be critical. Adding variance around the mean does not help much in this situation because it merely records the degree to which the smoothing prior on $\boldsymbol{\beta}$ (and our knowledge of the normalization factor) is irrelevant in forming the model posterior: if a prior on the coefficients is to do any good, one must know the normalization factor, Φ, or choose a sufficiently narrow variance for the prior on it. Otherwise, no Bayesian shrinkage occurs, and the original motivation for using the model vanishes.

The fact is that Φ is inestimable from the given data and must be imposed a priori with exogenous knowledge. Adding a prior so that Φ is identified does not help unless that prior is also meaningful, because the estimates will be the results of prior specification rather than empirical information. If prior knowledge about the normalization factor does not exist, then the model cannot be meaningfully specified.

4.3.3 We Know about the Dependent Variable, Not the Coefficients

When experts say that neighboring countries, adjacent age groups, or a set of regressions are all "similar," they are not usually talking about the similarity of the coefficients. It is true that in Bayesian analysis, we need a prior on coefficients, and so it may seem reasonable to attach the qualitative notion of similarity to the formal Bayesian prior density in equation 4.4. But reasonable it is not always. In most situations, it seems that "similarity" refers to the dependent variable or the expected value of the dependent variable, not the coefficients, and assuming that similarity in the expected value of the dependent variable applies to similarity in the coefficients turns out to be a serious flaw. Except when most prior evidence comes from case-control studies, and so the expected value of the dependent variable is not typically estimated, similarity would mostly seem to refer to the dependent variable rather than coefficients.

Even if experts from public health and demography are willing to accept the linear functional form we typically specify, $\mu_{it} \equiv \mathbf{Z}_{it}\boldsymbol{\beta}_i$, they do not normally observe the coefficients, $\boldsymbol{\beta}$, or even any direct implications of them. Many of them are not quantities of interest in their research, because most do not directly coincide with causal effects. Instead, the only outcome of the data generation process that researchers get to observe is the log-mortality rate, m_t, and it, or at least the average of multiple observations of it, serves as an excellent estimate of the expected log-mortality rate. As such, it is reasonable to think that analysts might have sufficient knowledge with which to form priors about the expected mortality rate, even if most of the coefficients are noncausal and on different scales.

Indeed, we find that when asking substantive experts for their opinion about what countries (or age groups, etc.) are alike, they are much more comfortable offering opinions about the similarity of expected mortality than regression coefficients. In fact, on detailed questioning, they have few real opinions on the coefficients even considered separately. This point thus follows the spirit of Kadane's focus on prior elicitation methods that are "predictive" (focusing on the dependent variable) rather than "structural" (focusing on the coefficients) (Kadane et al., 1980; Kadane, 1980).

To see why priors on μ do not translate automatically into priors on $\boldsymbol{\beta}$ without further analysis, consider a simple version of the cross-sectional variation in the expected value of the dependent variable, $\mu_{it} \equiv \mathbf{Z}_{it}\boldsymbol{\beta}_i$, at one point in time t. This version is merely the difference between two cross sections i and j, if we assume that the covariates in cross sections i and j are of the same type:

$$\begin{aligned} \mu_{it} - \mu_{jt} &= \mathbf{Z}_{it}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j) + (\mathbf{Z}_{it} - \mathbf{Z}_{jt})\boldsymbol{\beta}_j \\ &= \text{Coefficient variation} + \text{Covariate variation.} \end{aligned} \tag{4.9}$$

This expression decomposes the difference (or variation in) the expected value of the dependent variable into coefficient variation and covariate variation. A prior on variation in the expected value does not translate directly into coefficient variation because it ignores covariate variation. In other words, this expression demonstrates that having $\boldsymbol{\beta}_i \approx \boldsymbol{\beta}_j$ does *not* guarantee that the expected value of the dependent variable assumes similar values in cross sections i and j, because of the term $(\mathbf{Z}_{it} - \mathbf{Z}_{jt})\boldsymbol{\beta}_j$, which is not necessarily small. Obviously, the more similar $\mathbf{Z}_{it}$ is to $\mathbf{Z}_{jt}$, the smaller is this term. However, there is no

reason, a priori, for which two cross sections with similar patterns of mortality should have similar patterns of the *observed* covariates: some of the similarity may arise from patterns of the unobservables, or, when some of the covariates are "substitutes" of each other, from a different mix. For example, two countries might achieve similar patterns of mortality due to cardiovascular disease by different means: one could have first-class surgical and pharmaceutical interventions that keep people alive but very poor public health and education facilities that might prevent illness in the first place, and the other could have the opposite pattern. In this situation, we would observe differences in covariates and their coefficients but similar mortality rates.

4.3.4 Difficulties with Incomparable Covariates

But even when the covariates behave in such a way that this extra source of variation is not an issue, another problem may surface. In the previous section, we implicitly assumed that all cross sections share the same "type" of covariates and the same specification. However, the dependent variable may have different determinants in different cross sections, and some covariates may be relevant in some cross sections but not in others. For example, in forecasting mortality rates, we know that fat and cigarette consumption are important determinants of mortality, but these covariates are observed only in some subset of countries. Similarly, we would not expect the availability of clean water to explain much variation in mortality rates in most of the developed world. In this situation, we could pool the corresponding coefficients only in the cross sections for which these covariates are observed, but then we might introduce unpredictable levels of pooling bias. In general, pooling coefficients is not a viable option when we have different covariates in different cross sections.

Moreover, even when we have the same type of covariates in all cross sections, pooling coefficients makes sense only if the covariates are directly comparable. A simple example is the case of GDP: if we want to pool the coefficients on GDP, this covariate will not only have to be expressed in the same currency (say U.S. 1990 dollars) but also be subjected to further adjustments, such as purchasing power parity, which are not trivial matters. Having a variable with the same name in different countries does not guarantee that it means the same thing. If it does not, substance matter experts would have no particular reason to believe that the coefficients from the regression of a time series in one cross section would be similar to that in another, because the coefficients themselves would mean entirely different things.

4.4 Priors on the Expected Value of the Dependent Variable

In this section, we show how to address the issues from section 4.3, using the simple idea of focusing attention on the expected value of the dependent variable, rather than on the coefficients. Researchers may know fairly precisely how the expected value is supposed to vary across cross sections, or something about its behavior over time, or interactions among these or other variables.

To get the usual Bayesian modeling technology to work, however, we ultimately need priors expressed in terms of the coefficients because they are the parameters to be estimated. We therefore propose the following two-step strategy, aimed at deriving a prior density on the regression coefficients, but constructed from knowledge of priors specified on the expected value of the dependent variable. First, we specify the prior in terms of the expected value, and then we add information about the functional form and translate it into a prior on the coefficients.

The idea of putting a prior on the expected value of an outcome variable and then deducing (either mathematically or qualitatively) the implied prior on the coefficients has appeared in several literatures for single-equation models. See, for example, work on prior elicitation (Ibrahim and Chen, 1997; Kadane et al., 1980; Kadane, 1980; Laud and Ibrahim, 1996, 1995; Weiss, Wang, and Ibrahim, 1997), the covariance structure of wavelet coefficients (Jefferys et al., 2001), predictive inference (West, Harrison, and Migon, 1985; Tsutakawa and Lin, 1986; Tsutakawa, 1992), and logistic (Clogg et al., 1991), and other generalized linear models (Oman, 1985; O'Hagan, Woodward, and Moodaley, 1990; Bedrick, Christensen, and Johnson, 1996; Greenland and Christensen, 2001; Greenland, 2001). Our approach builds on these literatures by generalizing these ideas to hierarchical models, where it turns out the benefits of the approach are even greater. To our knowledge, these generalizations and their additional benefits have not been noted before in the literature.

4.4.1 Step 1: Specify a Prior for the Dependent Variable

We begin by thinking nonparametrically (i.e., qualitatively, before entertaining a specific parametric functional form) about the expected value of the dependent variable as a function μ_{it} of the cross-sectional index i $(i = 1, \ldots, N)$ and time t $(t = 1, \ldots, T)$. Although μ is naturally thought of as an $N \times T$ matrix, it will be more convenient to think of it as the column vector in $\mathbb{R}^{T \times N}$ obtained by concatenating the N time series corresponding to the different cross sections, one after the other. The experts' knowledge can be seen as a set of L statements about properties of μ, and we assume that it is possible to translate them into L formulas of the form:

$$H_l[\mu] \quad \text{should be small} \quad l = 1, \ldots L, \tag{4.10}$$

where H_l are functionals of μ (a functional is a map from a set of functions into the set of real numbers).[3]

[3] The idea that prior knowledge can be represented in statements like that in equation 4.10 is very old. In its simplest form, we see it in the method of graduation (Whittaker, 1923; Henderson, 1924), but its broad and full formalization was given by Tikhonov in the framework of regularization theory (Tikhonov, 1963; Tikhonov and Arsenin, 1977; Morozov, 1984), where the functionals H_l are usually called *stabilizers* or *regularization functionals*. Similar ideas appear in the theory of splines, starting with the seminal work of Schoenberg (Schoenberg, 1946; de Boor, 1978; Wahba, 1990) where the functionals H_l are usually called *smoothness functionals*. The fact that we build a prior density starting from a stabilizer is not by chance: there is a deep connection between regularization theory and Bayesian estimation, which was first unveiled by Kimeldorf and Wahba (1970; see also Wahba, 1990), as well as one between the method of graduation and Bayesian estimation (Taylor, 1992; Verrall, 1993).

Example 1 A simple example is the case where we believe μ varies very little over time in each cross section but the different cross sections have no necessary relationships. In this situation, we could have $L = N$, and $H_i[\mu]$ could be the average rate of temporal variation of μ in cross section i. If we think that this set of constraints is too restrictive, we may want to enforce our statements only on average and then replace the N functionals H_i with the single functional $\sum_i H_i[\mu]$. ⊠

Example 2 Suppose we know that, for any given year t, the cross-sectional profile of μ_{it} should be not too far from some specified profile g_i over cross sections, but we have no prior beliefs about time-series patterns. Then, we could set $L = T$ and $H_t[\mu] = \sum_i (\mu_{it} - g_i)^2$, for all t. Alternatively, we may want to enforce our statements only on average and then replace the T functionals H_t with the single functional $\sum_t H_t[\mu]$. ⊠

We now put these statements in a probabilistic form. Think of μ as a random variable, and define a *normal* probability density on μ as

$$\mathcal{P}(\mu \mid \theta) \propto \exp\left(-\frac{1}{2}\sum_l \theta_l H_l[\mu]\right) \equiv \exp\left(-\frac{1}{2}H[\mu, \theta]\right),\ \mu \in \mathbb{R}^{T\times N}, \tag{4.11}$$

where $\theta = (\theta_1, \ldots, \theta_l)$ is a set of positive parameters (often called hyperparameters, regularization parameters, or smoothing parameters). The choice of the exponential function in the preceding equation is a matter of convenience at this point and is in line with practical applications, while the factor $\frac{1}{2}$ is there only to simplify future calculations. The probability density in equation 4.11 assigns high probability only to those configurations such that $\theta_l H_l[\mu]$ is small for all l, which is precisely what we want in a formal version of equation 4.10. The parameters θ_l control how small we want $H_l[\mu]$ to be. A simple but important observation is that they control all the moments of the prior density. Therefore, if additional information is available about some moments (e.g., we may have an idea of what the variance of μ might be), then these parameters could be determined from prior knowledge. In general, they will be known with some uncertainty, and therefore they are usually taken to be random variables with known prior distributions. Because their precise value is not relevant in this section, we take them as user-specified for the moment and will consider the problems of their choice in chapter 6 and estimation in chapter 9.

4.4.2 Step 2: Translate to a Prior on the Coefficients

Equation 4.11 is a convenient and flexible way to summarize prior knowledge, but it tells only half of the story, the other half being told by the covariates. In fact, the density in equation 4.11 is defined over the entire space $\mathbb{R}^{T\times N}$ and assigns positive probability to *any* realization of the vector μ. However, this does not take into account the linear specification in equation 4.1, which says that only the values of μ explained by the covariates $\mathbf{Z}$ can be realized—that is, μ must lie in some subspace $\mathbb{S}_{\mathbf{Z}} \subset \mathbb{R}^{T\times N}$. Therefore, the prior 4.11 is valid only in the subspace $\mathbb{S}_{\mathbf{Z}}$, and it should be set to 0 outside it.

To formalize this result, we rewrite the specification in equation 4.1 in matrix form as $\mu = \mathbf{Z}\boldsymbol{\beta}$, where $\boldsymbol{\beta} \in \mathbb{R}^{\sum_i k_i}$ is the column vector obtained by concatenating the N vectors $\boldsymbol{\beta}_i$; k_i is the number of covariates in cross section i; and $\mathbf{Z}$ is block diagonal, with the

data matrices $\mathbf{Z}_i$ forming the blocks. Then the set $\mathbb{S}_{\mathbf{Z}}$ is formally defined as $\mathbb{S}_{\mathbf{Z}} \equiv \{\mu \in \mathbb{R}^{T\times N} | \mu = \mathbf{Z}\boldsymbol{\beta} \text{ for some } \boldsymbol{\beta} \in \mathbb{R}^{\sum_i k_i}\}$. We summarize this information by writing:

$$\mathcal{P}(\mu|\theta) \propto \begin{cases} \exp\left(-\frac{1}{2}H[\mu,\theta]\right) & \text{if } \mu \in \mathbb{S}_{\mathbf{Z}} \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

It is now clear that *on the subspace $\mathbb{S}_{\mathbf{Z}}$, that is, on the support of the prior, the relationship $\mu = \mathbf{Z}\boldsymbol{\beta}$ is invertible* by the usual formula $\boldsymbol{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mu$ (if we assume only that $\mathbf{Z}$ is of full rank). This result implies that we can use equation 4.12 to derive a probability distribution for $\boldsymbol{\beta}$. Because the transformation $\mu = \mathbf{Z}\boldsymbol{\beta}$ is linear, its Jacobian is an irrelevant constant (the tools to compute it are presented in appendix C), and we obtain the probability density for $\boldsymbol{\beta}$ by simply plugging $\mu = \mathbf{Z}\boldsymbol{\beta}$ in equation 4.12. Therefore, we write equation 4.12 in terms of the coefficients $\boldsymbol{\beta}$ as

$$\mathcal{P}(\boldsymbol{\beta}|\theta) \propto \begin{cases} \exp\left(-\frac{1}{2}H^{\mu}[\boldsymbol{\beta},\theta]\right) & \text{if } \mu = \mathbf{Z}\boldsymbol{\beta} \in \mathbb{S}_{\mathbf{Z}} \\ 0 & \text{otherwise,} \end{cases} \tag{4.13}$$

where

$$H^{\mu}[\boldsymbol{\beta},\theta] \equiv H[\mathbf{Z}\boldsymbol{\beta},\theta]. \tag{4.14}$$

Equation 4.13 contains exactly the same amount of information as contained in equation 4.12. The fact that $\mathcal{P}(\mu \mid \theta)$ is 0 outside of $\mathbb{S}_{\mathbf{Z}}$ is given by the expression $\mu = \mathbf{Z}\boldsymbol{\beta}$, and the density of μ on $\mathbb{S}_{\mathbf{Z}}$ implied by the prior on the coefficients in equation 4.13 is, by construction, equation 4.12.

For clarity of notation, we use the superscript μ in $H^{\mu}[\boldsymbol{\beta},\theta]$ to remind us that, unlike the version in equation 4.4 used to smooth directly based on the coefficients, this density has been derived using knowledge about μ. (Because in our formulation the prior densities are always of the form 4.13 for some appropriate choice of the function in the exponent, we will often refer to the function in the exponent as "the prior," without the risk of confusion.)

4.4.3 Interpretation

To make meaningful comparisons between any two units, they both need to be measured on a common scale. The difficulty in our problem is that the vectors of $\boldsymbol{\beta}$s from different equations, each with a different set of covariates, have different lengths and meanings, and so each is associated with a distinct "metric space" (see section B.1.2, page 219). Our approach takes advantage of the fact that μ from all cross sections form a natural (and single) metric space, and functionals defined on μ dictate relationships among the coefficients $\boldsymbol{\beta}$ due to the specification $\mu = \mathbf{Z}\boldsymbol{\beta}$. Although we still cannot map $\boldsymbol{\beta}$ into a common metric space, μ_i controls $\boldsymbol{\beta}_i$ for all i, and so we can make comparisons between μ_i and μ_j and, given the deterministic relationship between μ and $\boldsymbol{\beta}$, come up with an expression for how close $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ are.

We convey the math involved in this procedure via a simple analogy. Suppose $\theta = \beta_1 - \beta_2$ and $P(\theta) = N(\theta|0,\sigma^2)$. Then, the density $P(\beta_1,\beta_2) \equiv N(\beta_1 - \beta_2|0,\sigma^2)$ is singular bivariate Normal defined over β_1, β_2 and constant in all directions but $(\beta_1 - \beta_2)$. In our more general case, we start with the one-dimensional $P(\mu_{it})$, and treat it as the multidimensional $P(\boldsymbol{\beta}_i)$, constant in all directions except for $\mathbf{Z}_{it}\boldsymbol{\beta}_i$.

The final result, expressed in equation 4.13, is important because it has the desired property: it assigns high probability only to configurations of the coefficients $\boldsymbol{\beta}$ such that the corresponding predicted values of the dependent variable $\mu = \mathbf{Z}\boldsymbol{\beta}$ have high probability, which conforms to our prior knowledge. Notice that the specification $\mu = \mathbf{Z}\boldsymbol{\beta}$ holds for the years for which we have observations as well as for the years for which we need to make a forecast; this implies that the covariates in the preceding expressions have a temporal range that extends into the future and that prior knowledge is enforced on both the past and the future of the expected value of the dependent variable.[4]

After these steps have been performed, the rest is standard Bayesian analysis: We plug the prior of equation 4.13 into the expression for the posterior distribution of equation 4.2, leaving to be solved only the computational problem of calculating equation 4.3, which we address in chapter 9.

Chapters 5 and 7 are devoted to showing how this approach works in practice by deriving explicit expressions for priors on $\boldsymbol{\beta}$ in a diverse variety of important cases. The pleasant surprise is that, for a wide choice of smoothness functionals $H[\mu, \theta]$, the implied prior for $\boldsymbol{\beta}$ turns out to have a mathematical form very similar to the one in equation 4.5, which is well understood, but without the shortcomings discussed in section 4.3.

4.5 A Basic Prior for Smoothing over Age Groups

For expository purposes, we begin by detailing a very simple prior on μ, which, although not sophisticated enough to be used in most applications, generates a prior for the coefficients with all the relevant characteristics useful for understanding our approach.

4.5.1 Step 1: A Prior for μ

We assume for the moment that there is only one country and A age groups, so that the expected value of the dependent variable in age group a at time t is μ_{at}, with $a = 1, \ldots A$ and $t = 1, \ldots, T$. We consider a very simple form of prior knowledge: at any point in time, *nearby age groups have similar values of* μ. A simple way to represent this kind of knowledge is based on the average squared difference of expected log-mortality in adjacent age groups (averaged over time periods):

$$H[\mu, \theta] \equiv \frac{\theta}{T} \sum_{t} \sum_{a=1}^{A-1} (\mu_{at} - \mu_{a+1,t})^2, \quad \text{should be small.}$$

This smoothness functional has two important properties:

1. It takes on small values only when nearby age groups have similar values of μ.

[4] This implies that the notation $\sum_t$ has different meanings when it appears in the likelihood and in the prior, but for notational simplicity, we do not distinguish between the two.

2. It is indifferent to arbitrary, time-dependent, shifts constant across the age profiles. More precisely, it is invariant with respect to the transformation:

$$\mu_{at} \rightsquigarrow \mu_{at} + f_t \qquad \forall f_t \in \mathbb{R}.$$

In other words, for any given year, the prior associated with the preceding functional suggests we are ignorant with respect to the level of the dependent variable. (We formalize and generalize this notion of prior indifference in section 5.1.)

We now rewrite the preceding functional in a form more amenable to generalization. First we define the matrix s such that $s_{aa'} = 1$ if and only if $|a - a'| = 1$, and 0 otherwise. If we use this notation, the preceding functional can be written as

$$H[\mu, \theta] \equiv \frac{\theta}{2T} \sum_t \sum_{aa'} s_{aa'}(\mu_{at} - \mu_{a't})^2 = \frac{\theta}{T} \sum_t \sum_{aa'} W_{aa'} \mu_{at} \mu_{a't}, \tag{4.15}$$

where we have defined the matrix $W = s^+ - s$ (see appendix B, page 237). Therefore the prior density for μ has the form:

$$\mathcal{P}(\mu|\theta) \propto \exp\left(-\frac{\theta}{2T} \sum_t \sum_{aa'} W_{aa'} \mu_{at} \mu_{a't}\right). \tag{4.16}$$

One feature of the prior density in equation 4.16 is that it has zero mean and is symmetric around the origin, so that the probability of an age profile and its negative are the same. Depending on the application this may or may not be appropriate. For example, this is not realistic when analyzing logged mortality rates: in this case, we know that, in any given year, the age profiles will look, on average, like some cause-specific "typical" age profile $\bar{\mu} \in \mathbb{R}^A$. Fortunately, it is easy to modify the prior to take this information into account, by letting the prior have mean $\bar{\mu}$ in any year:

$$\mathcal{P}(\mu|\theta) \propto \exp\left(-\frac{\theta}{2T} \sum_t \sum_{aa'} W_{aa'}(\mu_{at} - \bar{\mu}_a)(\mu_{a't} - \bar{\mu}_{a'})\right). \tag{4.17}$$

One issue is where $\bar{\mu}$ comes from. One productive procedure is to have an expert draw pictures of his or her expectations for average log-mortality age profile, $\bar{\mu}$. Alternatively, a reasonable typical age profile $\bar{\mu}$ could be synthesized from the data, or borrowed from other countries not in the analysis, possibly after some preprocessing and smoothing, and subject to the "approval" of some expert. In this case it looks like we have a data-dependent prior, which, of course, would not seem "prior" and so would not appear appropriate. However, this problem is easily solved by noticing that using a prior that is not mean zero is equivalent to using a mean-zero prior in which we have replaced the dependent variable m_{at} with $m_{at} - \bar{\mu}_a$. In the following, therefore, we keep using equation 4.16 rather than the more cumbersome equation 4.17, where we keep in mind that, depending on the application, μ may be either the expected value of the dependent variable or its deviation from the typical age profile $\bar{\mu}$.

4.5.2 Step 2: From the Prior on μ to the Prior on β

We now proceed to the second step and substitute our specification in the functional in equation 4.15. In this case, the specification is simply $\mu_{at} = \mathbf{Z}_{at}\boldsymbol{\beta}_a$, and substituting it into equation 4.15, we obtain

$$\begin{aligned} H^{\mu}[\boldsymbol{\beta}, \theta] &\equiv \frac{\theta}{T} \sum_{aa't} W_{aa'}(\mathbf{Z}_{at}\boldsymbol{\beta}_a)(\mathbf{Z}_{a't}\boldsymbol{\beta}_{a'}) \\ &= \theta \sum_{aa'} W_{aa'}\boldsymbol{\beta}'_a \mathbf{C}_{aa'}\boldsymbol{\beta}_{a'}, \end{aligned} \tag{4.18}$$

where the second line uses the fact that the coefficients $\boldsymbol{\beta}$ do not depend on time and so the sum over time can be performed once for all, and where we have defined the matrix:

$$\mathbf{C}_{aa'} \equiv \frac{1}{T}\mathbf{Z}'_a\mathbf{Z}_{a'},$$

so that $\mathbf{Z}_a$ is the usual data matrix of the covariates in cross section a, which has $\mathbf{Z}_{at}$ for each row. Hence, the prior for $\boldsymbol{\beta}$, conditional on the parameter θ, is now simply:

$$\mathcal{P}(\boldsymbol{\beta}|\theta) \propto \exp\left(-\frac{1}{2}\theta \sum_{aa'} W_{aa'}\boldsymbol{\beta}'_a \mathbf{C}_{aa'}\boldsymbol{\beta}_{a'}\right). \tag{4.19}$$

This gives the desired prior over the coefficients, which we have built using prior knowledge only on the expected value of the dependent variable μ. Its key characteristics are the facts that: (1) because the covariates $\mathbf{Z}_{at}$ and $\mathbf{Z}_{a't}$ are of dimensions k_a and $k_{a'}$, respectively, then $\mathbf{C}_{aa'}$ is a rectangular $k_a \times k_{a'}$ matrix, and *it does not matter whether we have the same number or type of covariates in the two cross sections a and a′;* in addition, (2) *the entire matrix of normalization coefficients* Φ *drops out and so need not be specified or estimated.*

4.5.3 Interpretation

While we postpone to chapter 8 a thorough comparison between this prior and the prior we would have obtained by imposing smoothness of the coefficients over age groups, as described in section 4.2, it is useful to write them side by side as follows:

$$\mathcal{P}(\boldsymbol{\beta}|\theta) \propto \exp\left(-\frac{1}{2}\theta \sum_{aa'} W_{aa'}\boldsymbol{\beta}'_a \mathbf{C}_{aa'}\boldsymbol{\beta}_{a'}\right) \quad \Leftrightarrow \quad \mathcal{P}(\boldsymbol{\beta}|\Phi) \propto \exp\left(-\frac{1}{2} \sum_{aa'} W_{aa'}\boldsymbol{\beta}'_a \Phi\boldsymbol{\beta}_{a'}\right).$$

This comparison helps us emphasize that focusing on the expected value of the dependent variable allows us to solve two problems at the same time: (1) we replaced an entire unknown matrix Φ with the quantities $\theta\mathbf{C}_{aa'}$, which are known up to a single scalar parameter; and (2) our formulation allows each cross section to have its own specification and therefore for different covariates in different cross sections. In addition, the new prior, although conceptually very different from the usual prior over the coefficients, is

computationally similar and does not imply any additional difficulties from the point of view of the implementation.

In addition, while having real prior knowledge about Φ is extremely rare, the one remaining parameter in our formulation, θ, is directly linked to quantities that we can directly interpret and on which we are likely to have prior information. Although we put off a detailed exploration until section 6.2, we report here a key result. If we let $\mu_{at} = \mathbf{Z}_{at}\boldsymbol{\beta}_a$, the quantity

$$\frac{1}{T}\sum_t \sum_{aa'} s_{aa'}(\mu_{at} - \mu_{a't})^2 = \sum_{aa'} W_{aa'}\boldsymbol{\beta}'_a \mathbf{C}_{aa'}\boldsymbol{\beta}_{a'},$$

which appears in the exponent of our prior, represents an average over time and in a mean square sense *how much the expected value of the dependent variable varies from one age group to the next*. Postponing some technicalities related to the fact that the prior in equation 4.19 is improper, the expected value of the preceding quantity under the prior in equation 4.19 is:

$$E\left[\sum_{aa'} W_{aa'}\boldsymbol{\beta}'_a \mathbf{C}_{aa'}\boldsymbol{\beta}_{a'}\right] = \frac{K}{\theta},$$

where K is a number that depends on W and the matrices $\mathbf{C}_{aa'}$ and that can be easily computed. In most applications, we will have an idea of how much the expected log-mortality rate varies between adjacent age groups, and so we could easily specify a range of values for the left side of this equation. This then immediately leads to a range of values for θ, and therefore to a prior density $\mathcal{P}(\theta)$. This is a stark contrast to putting priors on coefficients where the normalization matrix Φ is unknown but still would need to be specified.

Before we proceed to analyze more sophisticated priors, two key remarks are in order. First, as pointed out, this prior is invariant with respect to the transformation $\mu_{at} \rightsquigarrow \mu_{at} + f_t,\ \forall f_t \in \mathbb{R}$. This implies that the prior is constant over an entire subspace of its domain, and therefore its integral over the domain is infinite: in other words, the prior is improper. This causes no difficulties because our likelihood is normal and proper, and therefore our posterior is proper too. It is also a highly desirable feature of a prior, smoothing expected values of the dependent variable toward each other but without requiring one to specify at what specific level they smooth toward. Note that this feature is distinct from (proper) priors that are merely diffuse with respect to a parameter. This prior will have *no* effect on constant shifts in the posterior, no matter how much weight is put on it (or, correspondingly, how small we make its variance).

Second, we mentioned that the prior presented is not necessarily what we recommend in applications. The reason for this can be seen by rewriting it as

$$\mathcal{P}(\mu|\theta) \propto \exp\left(-\frac{\theta}{2T}\sum_t \sum_{a=1}^{A-1}(\mu_{at} - \mu_{a+1,t})^2\right). \tag{4.20}$$

This version illuminates the feature of this prior that increments in log-mortality between adjacent age groups are independent, and therefore samples from this prior will tend to look like a random walk (as a function of *age*, not time), and therefore it will not be

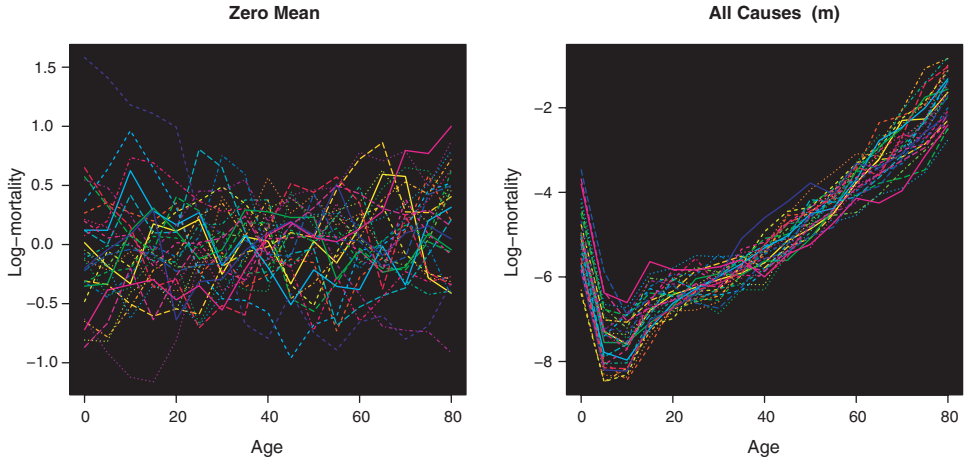

FIGURE 4.1. Samples from the prior in equation 4.20 with zero mean (*left graph*) and nonzero mean (*right graph*). Depicted are 17 age groups, at 5-year intervals. For the right graph, the mean $\bar{\mu}$ is the average age profile for all cause mortality in males. The average has been computed over all the countries with more than 20 observations and over all the available years. The value of θ has been chosen so that the standard deviation of μ_a is 0.5, on average over the age groups. Notice the jagged behavior of each line in the two graphs, making these priors undesirable for most applications.

particularly smooth. In figure 4.1, we present samples from the preceding prior (with each draw an $A \times 1$ vector represented by one line) for a fixed time t. There are 17 age groups, at 5-year intervals, so that $A = 17$. The left graph gives the case of a zero mean, while in the right graph, we add a mean to the prior, where the mean is a typical age profile for all-cause male mortality. Each age profile is quite jagged over age groups, indicating that it is not a very good representation of our prior knowledge. The reason for the lack of smoothness is that the increments between nearby age groups are specified to be independent. We will see in chapter 5 that by introducing correlations between the increments, much smoother and hence much more appropriate age profiles can be obtained.

4.6 Concluding Remark

In the rest of this book, we develop specific models within the class of models defined in this chapter. In particular, chapters 5 and 7 repeatedly use the two steps offered in section 4.4 for each of a variety of different data types to derive new models under the framework of this chapter.

Although we use the simple linear-normal likelihood throughout our work, nothing in our approach would necessarily constrain one to that model. Our methods for specifying the priors on the expected value of the dependent variable need not be changed for other likelihood models, such as nonlinear functional forms or nonnormal densities. There would be different and more difficult computational considerations, of course, but the general approach offered here still applies.