6 Model Selection

Like any statistical method, and especially any Bayesian method, our approach comes with a variety of adjustable settings and choices. As discussed in chapter 1, however, we needed to make so many forecasts for our application that we had to limit these choices as much as possible. Of course, limiting choices is almost the same process as ensuring that the choices made were based on readily available knowledge. This is often not the case in Bayesian modeling, where hyperparameter values and other choices are based on guesses, default settings, “reference priors,” or trial and error.

Thus, in this chapter we try to connect every adjustable setting to some specific piece of knowledge that demographers and others have readily available from their empirical analyses. For example, at no point should users need to set the value of an important hyperparameter that has no obvious meaning or connection to empirical reality. Our job in developing these methods, as we see it, is to bring new information to bear on the problem of demographic forecasting, and so we put considerable effort into taking existing qualitative and quantitative knowledge bases in demography and the related social sciences and providing easy and direct connections to the choices necessary in using our methods.

In this chapter, we discuss choices involving the degree of smoothness (section 6.1), the prior for the smoothing parameter (section 6.2), where in the function to smooth (section 6.3), covariate specification (section 6.4), and variance function specification (section 6.5). The results of this chapter made it possible for us to design easy-to-use software that implements our methods, because we were able to translate apparently arcane choices into substantively meaningful decisions about which much is known.

6.1 Choosing the Smoothness Functional

In chapter 5, we considered a family of smoothness functionals based on the derivative of order n. The parameter n plays multiple roles in the smoothness functionals of the type of equation 5.10:

1. It determines the *local* behavior of the function, that is, how the functions looks on a small interval of the domain. Increasing values of n correspond to samples that are locally more and more smooth.

2. It determines the size of the null space of the functional, which consists of the $\mathbb{n}$-dimensional space of polynomials of degree $\mathbb{n} - 1$.
3. It also determines the *global* shape of the samples, that is, how many global "bumps" (or changes of direction) it has, and also how many zeros. This is a side effect of controlling the size of the null space: when $\mathbb{n}$ increases, the polynomials in the null space have more and more bumps. Therefore, we can have functions with many bumps, similar to polynomials, with very small values of the smoothness functional: these functions will appear with high probability, which explains why samples from the prior with large $\mathbb{n}$ display, over the whole domain, a high degree of oscillation even if they are locally very smooth.

Reducing an entire proximity matrix to one parameter $\mathbb{n}$ is tremendously convenient, but the fact that only one parameter controls different characteristics of the samples drawn can sometimes be a disadvantage in practical applications. We show how to avoid this disadvantage now. Suppose, for example, that we wish our samples to be locally very smooth, with the kind of local smoothness associated with $\mathbb{n} = 4$. However, we may not want the global behavior associated with $\mathbb{n} = 4$ (see figure 5.3, page 89), because it has many bumps, and we may not want as a null space the space of polynomials of degree three, because it is too large (i.e., it may exclude constraints we wish to impose). Suppose, instead, we want the null space to consist of the space of linear functions. As it turns out, we can have the best of both worlds by considering a larger class of smoothness functionals that includes mixtures of those we have considered until now:

$$H[\mu, \theta] \equiv \sum_{i=1}^{K} \theta_i \int_0^T dw^{\mathrm{time}}(t) \int_0^A dw^{\mathrm{age}}(a) \left(\frac{d^{\mathbb{n}_i} \mu(a, t)}{da^{\mathbb{n}_i}} \right)^2 , \tag{6.1}$$

where $\theta_i \geq 0$, and we use the convention that the numbers $\mathbb{n}_i$ are listed in ascending order. We refer to smoothness functionals of this form as *mixed smoothness functionals*, whereas we refer to a smoothness functional of the form 5.10 as a *standard smoothness functional*.

The reason mixed smoothness functionals are useful is that they enable separate control over the size of the null space and the degree of local smoothness. The size of the null space is controlled by $\mathbb{n}_1$, the lowest order of derivative in the functional, because, in order for the smoothness functional to be zero, all the terms of the sum in equation 6.1 must be zero. The degree of local smoothness is controlled by $\mathbb{n}_K$, the highest order of derivative in the functional. In order for the smoothness functional to have small values, all the individual smoothness functionals in the sum must assume small values, and if the term with $\mathbb{n}_K$ does not assign a small value, the smoothness functional will not assume a small value. This makes clear that what is really important in the mixed smoothness functional is the choice of $\mathbb{n}_1$ and $\mathbb{n}_K$, which suggests that we can probably limit ourselves in most applications to the case $K = 2$.

Example We now return to the example at the beginning of this section where we desire a smoothness functional with samples that look locally very smooth, do not have many global bumps, and have a null space consisting of the set of linear functions. The local smoothness could be obtained from a standard smoothness functional with $\mathbb{n} = 4$, but this will have a null space that is too large and will also have samples with many bumps.

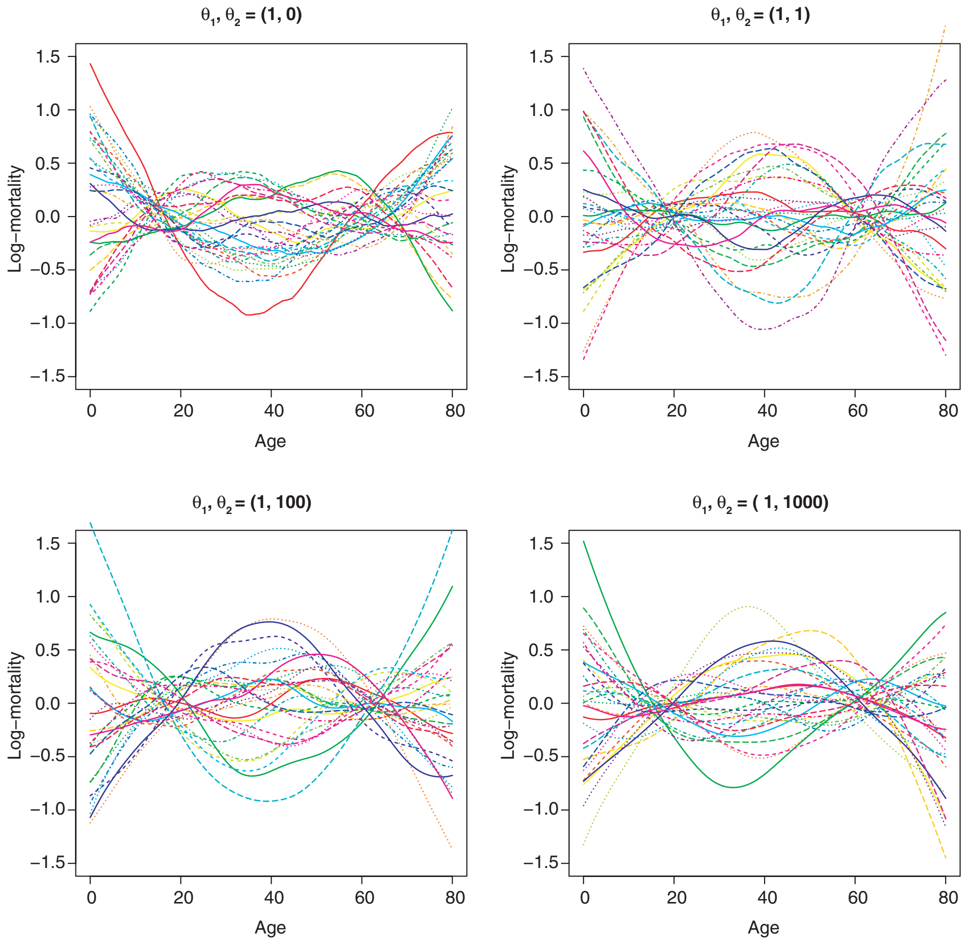

FIGURE 6.1. Age profile samples from "mixed" smoothness priors with $K = 2$, for different values of θ_2. Here $A = 80$, and there are 81 age groups, from 0 to 80. The graphs are on the same scale. In order to make the graphs comparable, the values of θ_1 and θ_2 have been scaled, in each graph, by a common factor, so that the standard deviation of μ_a is 0.3, on average over the age groups.

If we use a standard smoothness functional with $\mathbb{n} = 2$, we get the right null space, but the samples might not be smooth enough for our purposes.

Therefore, we use a mixed smoothness functional with $K = 2$. The fact that the null space must be the set of linear functions immediately determines that $\mathbb{n}_1 = 2$. Because we want samples that look locally very smooth, we choose $\mathbb{n}_2 = 4$. The last thing we need to choose is the size of the parameters θ_1 and θ_2. For the purpose of this illustration, all that matters is the relative size of these two numbers, so we fix θ_1 to an arbitrary number, say 1. Obviously, if we want the samples to look very smooth, we should give much more importance to the prior with the highest derivative. Figure 6.1 gives samples from the "mixed" prior for $\theta_1 = 1$ and for four different values of θ_2: 0, 1, 100, 1,000.

Notice how increasing the value of θ_2 leaves the "global" shape of the samples and the number of bumps unchanged (because they do not depend on the part of the prior with higher derivative), while the local smoothness steadily increases. ⊠

Thus, in a mixed smoothness functional, what determines the qualitative behavior of the samples from the prior are the lowest and highest degrees of the derivative, $\mathbb{n}_1$ and $\mathbb{n}_K$. Thus, in practice we usually limit ourselves to $K = 2$. Although results are sensitive to the choice of the prior, it is unlikely that they are that sensitive. If we also added to figure 6.1 a smoothness functional with $\mathbb{n} = 3$, we would not see major changes in the samples from the prior.

In practice many users will never need a mixed smoothness functional, and in fact a standard smoothness functional with $\mathbb{n} = 2$ *will give reasonable results in most cases.* This is consistent with experience from the literature on nonparametric regression, where cubic splines, or thin-plate splines (which are related to our prior with $\mathbb{n} = 2$), are used most of the time. The point of this section is that if we need something more sophisticated, it is readily available, and no additional concepts are required. From a computational point of view, because mixed functionals are sums of standard functionals, and because a linear combination of quadratic forms is also a quadratic form, the priors determined by both standard and mixed smoothness all have exactly the same general mathematical form as equation 5.12.

It is true that mixed smoothness functionals have more parameters than standard smoothness functionals, but the relative size between the parameters can be determined in advance and kept fixed, leaving only one global hyperparameter. For example, with $K = 2$, we suggest parameterizing the prior as follows:

$$H[\mu, \theta] \equiv \theta \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left[\left(\frac{d^{\mathbb{n}_1} \mu(a, t)}{da^{\mathbb{n}_1}} \right)^2 + \lambda \left(\frac{d^{\mathbb{n}_2} \mu(a, t)}{da^{\mathbb{n}_2}} \right)^2 \right],$$

where the parameter λ is easily chosen by drawing from the mixed prior and selecting the value that leads to the “best”- looking samples (i.e., by making graphs like the ones in figure 6.1. With this parametrization, everything we say about the smoothing parameter for standard smoothness functionals in the next section also applies to θ in the preceding mixed functional.

6.2 Choosing a Prior for the Smoothing Parameter

All priors introduced thus far come with a smoothness parameter, which we denote by θ. If we assume θ is known, then its effect on the forecast is qualitatively clear: larger values of θ correspond to smoother predictions, which pay correspondingly less attention to the data. In the limit, with θ going to infinity, our Bayesian estimate leads to a projection that lies in the null space of the prior—the specific element of which is chosen by the likelihood—because the only configuration of coefficients with nonzero probability are those for which the smoothness functional is zero. (This contrasts with proper priors that return a single point, such as the prior mean, when θ goes to infinity.) If we treat θ as a random variable, then this point is still valid when we replace θ with its expected value.

The parameter θ can be viewed in two ways, each leading to a different type of procedure for choosing it. The first is to consider θ as a free parameter of the theory, for which we happen to have no direct information. In this situation, we could use relatively automated algorithms to choose θ's optimal value. In our case, the optimal value is the one that minimizes an estimate of the forecast error. Usually these algorithms rely on the idea of *cross validation*: one or more data points are left out of the data set and used as a "test set" to estimate the accuracy of the forecast on new, unseen data. By repeating this procedure many times over different definitions of the test set, one can construct reasonable estimates of the forecast error and choose the value of θ that minimizes it. One method based on this idea is generalized cross validation, pioneered by Grace Wahba and her associates (see especially Golub, Heath, and Wahba, 1979; and Wahba, 1980, 1990). Another set of techniques based on a similar idea goes under the generic name of "bootstrapping" (see Efron, 1979, 1982, and the lengthy review in Efron and Tibshirani, 1993). A third approach to the choice of the optimal smoothness parameter is structural risk minimization, which is a very general approach to statistical inference and model selection (Vapnik, 1998; Hastie, Tibshirani, and Friedman, 2001).

A second way to look at the smoothness parameter is to consider it at the same level of other quantities, such as $\boldsymbol{\beta}$ and σ, and to treat it as a random variable with its own *proper* prior distribution $\mathcal{P}(\theta)$. (However, unlike $\boldsymbol{\beta}$ and σ, the prior for θ must be proper because the likelihood contains no information about it.) This implies that we must have an idea of what the mean and the variance of $\mathcal{P}(\theta)$ should be. While such information is often not available in many applications where smoothness functionals are typically used, as in pattern recognition, it *is* usually available for the applications described in this book. The main observation is that, although demographers do not have direct prior knowledge about θ in those terms, they typically do have knowledge about quantities determined by θ. Therefore, if the relationship between these quantities and θ can be inverted, knowledge about these quantities translates into knowledge about θ.

We formalize this idea in two stages. First, we consider a nonparametric prior for the age profiles, of the type discussed in the previous chapters, ignoring the covariates and the time dimension. Then, we introduce covariates and show how the nonparametric approach can be modified in order to be used in practical applications.

6.2.1 Smoothness Parameter for a Nonparametric Prior

In this section, we disregard our linear specification and the time dimensions and simply use the following smoothness prior for the age profiles:

$$\mathcal{P}(\mu) \propto \exp\left(-\frac{1}{2}\theta(\mu - \bar{\mu})' W^{\text{age},2}(\mu - \bar{\mu})\right), \tag{6.2}$$

where μ is an $A \times 1$ age profile vector, $W^{\text{age},2}$ is the matrix that corresponds to a smoothness functional of the form in equation 5.8 (page 81), with a derivative of order $\mathbb{n} = 2$, and $\bar{\mu}$ is a "typical" age profile for other infectious diseases in males. The question we address here is, what is a reasonable value for θ? The answer to this question is, simply put, a value that produces reasonable sample age profiles. An example of reasonable and unreasonable age profiles is shown in figure 6.2, where we plot two sets of 20 age profiles, each sampled from the prior in equation 6.2. The only difference between the left and right graphs (other than the randomness of the sampling process) is the value of the smoothness parameter θ, which is larger in the right than left graph. Even to the untrained eye, the graph on the right would seem to correspond to an unlikely choice for the smoothness parameter, because each age

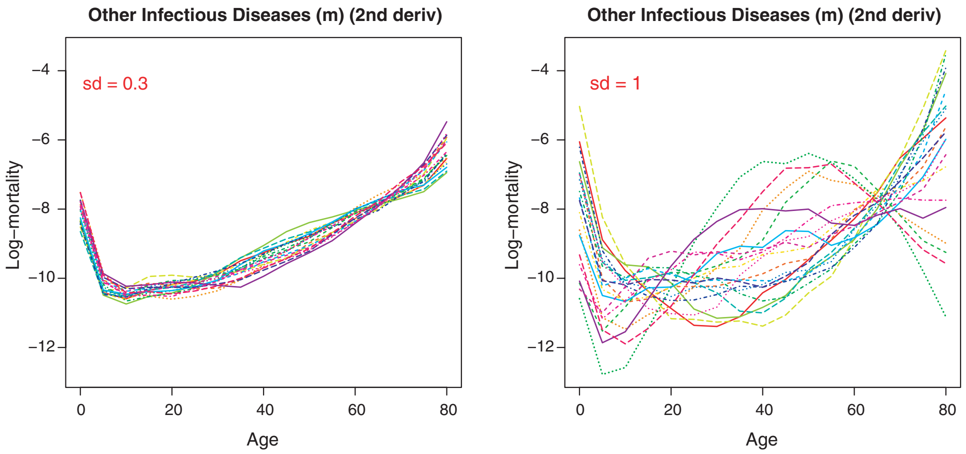

FIGURE 6.2. Samples from the prior in equation 6.2 corresponding to different values of θ. In the text, we show how the standard deviation ("sd" in the figure) of log-mortality across samples at each age determines θ.

group, except maybe age group 70, exhibits an unacceptably large variance. Those who study mortality age profiles typically have exactly this type of information. Therefore, if we had to choose between the two figures, we could easily choose the one on the left. The fact that we are not indifferent between these two figures shows that we are not indifferent to different values of θ and, therefore, that prior knowledge about θ exists and could be used to define at least a range in which θ should lie.

In principle, we could apply this strategy. For many values of $\theta > 0$, sample from the prior and draw graphs like the ones in figure 6.2. Then let subject matter experts choose a prior representative of their prior knowledge. The range of smoothness parameters θ associated with the selected figures will give us a range of acceptable values of θ, which we use to build a reasonable probability distribution $\mathcal{P}(\theta)$.

Of course, this approach in practice is too time-consuming, and fortunately easier ways to achieve the same result exist. As it turns out, the only fact we need to know is the basis of the experts' judgment about the samples. For example, we may postulate that experts judge the different figures according to the overall degree of variability of the age profiles, measured by the average of the age-specific variance of the age profiles. Alternatively, it is possible that experts are confident that, in a certain age range, say 20 to 75, log-mortality does not increase more (or less) than a certain amount going from one age group to the next, and therefore they judge the age profiles accordingly.

We now formalize this approach. We assume that experts' opinions can be summarized by a statement of the following form: "on average, the value of $F(\mu)$ is $\bar{F}$," where $F(\mu)$ is the function of the age profiles that experts implicitly use as a basis for their judgment. For example, if experts judge samples from the priors by the overall degree of variability of the age profiles, we could set $F(\mu)$ to be the average (over age groups) of the age-specific variance of the prior:

$$F(\mu) \equiv \frac{1}{A}\sum_{a=1}^{A}(\mu_a - \bar{\mu}_a)^2. \tag{6.3}$$

The expected value of $F(\mu)$ is clearly a function of the parameter θ, and therefore setting the expected value $F(\mu)$ to the experts' value $\bar{F}$ uniquely determines θ, by the following equation:

$$\mathrm{E}_{\perp}[F(\mu)|\theta] = \bar{F}, \tag{6.4}$$

where the subscript $\perp$ reminds us that the prior in equation 6.2 is improper and all expected values must be taken with respect to the subspace orthogonal to the null space of the prior (see appendix C). Equation 6.4 can be solved for θ either numerically or analytically, depending on the choice of F, and therefore used to set the mean of the prior $\mathcal{P}(\theta)$. Because the value $\bar{F}$ will always be provided with an uncertainty interval, the uncertainty on $\bar{F}$ can be easily translated in uncertainty on θ, and therefore used to estimate the variance of $\mathcal{P}(\theta)$.

The relationship between $\bar{F}$ and θ is easily seen for the choice of $F(\mu)$ shown in equation 6.3, which corresponds to the average of the age-specific variance of the prior. Experts are more likely to think in terms of standard deviations, rather than variance, and therefore it is convenient to define $\bar{F}$ in this case as σ_{age} and refer to σ_{age} as the *standard deviation of the prior*. Using this definition, and using the formulas of appendix C to perform the calculation in equation 6.4, we obtain the following:

$$\mathrm{E}_{\perp}[F(\mu)|\theta] \equiv \frac{1}{A}\sum_{a=1}^{A}\mathrm{E}_{\perp}[(\mu_a - \bar{\mu}_a)^2] = \frac{1}{A\theta}\mathrm{tr}\left(W^{\text{age},2}\right)^{+} = \bar{F} = \sigma^2_{\text{age}}, \tag{6.5}$$

where the superscript + stands for the generalized inverse (appendix B.2.5, page 235), and the matrix $W^{\text{age},2}$ is known (see section 5.2.5). Equation 6.5 can now be solved for θ as

$$\theta = \frac{\mathrm{tr}\left(W^{\text{age},2}\right)^{+}}{A\sigma^2_{\text{age}}}. \tag{6.6}$$

In fact, we used this expression to produce figure 6.2: the graph on the left used a value of $\sigma_{\text{age}} = 0.3$, which seems, a priori, an empirically reasonable number (i.e., the quantity "sd" reported in the top left corner of the figure). Plugging this number in the preceding equation, we obtained a value of θ to use in our simulation. Similarly, for the graph on the right, we choose θ such that $\sigma_{\text{age}} = 1$, which, based on our experience with age profiles, seems unrealistically large.

6.2.2 Smoothness Parameter for the Prior over the Coefficients

The formulas reported in the previous section are not what we use for forecasting (although they would be appropriate for simple univariate smoothing). In fact, in practical applications our age profiles can lie only in the span of the covariates, and the priors we use are defined over the set of coefficients $\boldsymbol{\beta}$. In order to simplify some of the formulas, it is convenient to rewrite our linear specification as $\mu = \bar{\mu} + \mathbf{Z}\boldsymbol{\beta}$, where we have defined

$$\boldsymbol{\beta} \equiv \begin{vmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_A \end{vmatrix}, \quad \mathbf{Z} \equiv \begin{vmatrix} \mathbf{Z}_1 & 0 & \dots & 0 \\ 0 & \mathbf{Z}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathbf{Z}_A \end{vmatrix}, \quad \tilde{\mu}_a \equiv \mathbf{Z}_a \boldsymbol{\beta}_a, \quad \tilde{\mu} \equiv \mathbf{Z}\boldsymbol{\beta} = \begin{vmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \\ \vdots \\ \tilde{\mu}_A \end{vmatrix}. \tag{6.7}$$

In other words, in this section the specification $\mathbf{Z}\boldsymbol{\beta}$ refers to the *centered* age profiles, $\tilde{\mu} = \mu - \bar{\mu}$. Under this definition, the prior over the coefficients $\boldsymbol{\beta}$ corresponding to the nonparametric prior in equation 6.2 has zero mean and can be written as

$$\mathcal{P}(\boldsymbol{\beta} \mid \theta) \propto \exp\left(-\frac{1}{2}\theta \sum_{aa'} W^{\text{age},2}_{aa'} \boldsymbol{\beta}'_a \mathbf{C}_{aa'} \boldsymbol{\beta}_{a'}\right). \tag{6.8}$$

By introducing the matrix D_{age} as

$$D_{\text{age}} \equiv \begin{vmatrix} W^{\text{age}}_{1,1}\mathbf{C}_{1,1} & W^{\text{age}}_{1,2}\mathbf{C}_{1,2} & \dots & W^{\text{age}}_{1,A}\mathbf{C}_{1,A} \\ W^{\text{age}}_{2,1}\mathbf{C}_{2,1} & W^{\text{age}}_{2,2}\mathbf{C}_{2,2} & \dots & W^{\text{age}}_{2,A}\mathbf{C}_{2,A} \\ \dots & \dots & \dots & \dots \\ W^{\text{age}}_{A,1}\mathbf{C}_{A,1} & W^{\text{age}}_{A,2}\mathbf{C}_{A,2} & \dots & W^{\text{age}}_{A,A}\mathbf{C}_{A,A} \end{vmatrix}, \tag{6.9}$$

the prior for the coefficients takes the form:

$$\mathcal{P}(\boldsymbol{\beta} \mid \theta) \propto \exp\left(-\frac{1}{2}\theta \boldsymbol{\beta}' D_{\text{age}} \boldsymbol{\beta}\right). \tag{6.10}$$

For a given set of covariates $\mathbf{Z}$, a sample of log-mortality age profiles is obtained by sampling the prior in equation 6.10, obtaining a random set of coefficients $\boldsymbol{\beta}$, and plugging $\boldsymbol{\beta}$ in the specification $\mu = \bar{\mu} + \mathbf{Z}\boldsymbol{\beta}$. Notice that this procedure generates T age profiles, which are linked over the time dimensions by the time variation of the covariates $\mathbf{Z}$.

The discussion of section 6.2.1 still applies, except for the fact that the functions $F(\mu)$ will contain an average over time, in addition to the average over age groups. For example, if we think that experts have knowledge about the standard deviation of the prior, we set

$$F(\mu) \equiv \frac{1}{AT} \sum_{a=1}^{A} \sum_{t=1}^{T} (\mu_{at} - \bar{\mu}_a)^2, \tag{6.11}$$

where $\mu = \bar{\mu} + \mathbf{Z}\beta$. Equation 6.4 is therefore replaced by

$$\mathrm{E}_{\perp}[F(\mu)|\theta] = \mathrm{E}_{\perp}[F(\bar{\mu} + \mathbf{Z}\beta)|\theta] = \bar{F}, \tag{6.12}$$

where the expected value is now taken over $\boldsymbol{\beta}$ using the prior 6.8. In order to see how this can be used in practice, we explicitly perform the calculation in equation 6.12 with $F(\mu)$

given by equation 6.11. Using the formulas of appendix C, we show that

$$\mathrm{E}_{\perp}[F(\mu)](\theta) \equiv = \frac{1}{AT}\sum_{a=1}^{A}\sum_{t=1}^{T}\mathrm{E}_{\perp}[(\mathbf{Z}_{at}\boldsymbol{\beta}_a)^2] = \frac{1}{AT}\mathrm{E}_{\perp}[\boldsymbol{\beta}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\beta}] = \frac{1}{AT\theta}\mathrm{Tr}\left(\mathbf{Z}D^{+}\mathbf{Z}'\right).$$

Therefore, equation 6.12 now reads

$$\frac{1}{AT\theta}\mathrm{Tr}\left(\mathbf{Z}D^{+}\mathbf{Z}'\right) = \bar{F} = \sigma^2_{\text{age}}, \tag{6.13}$$

and solving for θ, we obtain

$$\theta = \frac{\mathrm{Tr}\left(\mathbf{Z}D^{+}\mathbf{Z}'\right)}{AT\sigma^2_{\text{age}}}. \tag{6.14}$$

Equation 6.14 is ready to be used in practical applications. It says that all we need to know in order to set a reasonable value for θ is an estimate of the standard deviation of the prior σ_{age}, a quantity that is easily interpretable and, in our experience, is easy to elicit from subject matter experts.

But what specific numbers for σ_{age} should one choose? In our application, the scale of log-mortality means that a standard deviation of about 0.1 is usually a reasonable starting point, and it implies excursions of about plus or minus 0.3 (three standard deviations) around the prior mean. Obviously, each case is different, and there is no substitute for good judgment. Therefore, our general recommendation at least in our mortality data is to start with 0.1 and try other values near 0.1. For example, because the effect of the smoothing parameter operates on a logarithmic scale, we usually also try values 0.05 and 0.2 (or sometimes as high as 0.3). If the highest value gives reasonable results, there is no reason to move to lower values, with the risk of introducing unnecessary bias by restricting likelihood too much. In many examples, a range of values for σ gives similar results.

A key point is that setting θ, or equivalently the standard deviation of the prior σ_{age}, sets *all* the other "summary measures" of the prior. Therefore, if we have knowledge of several summary measures, a good strategy is usually to study the behavior of all of them as a function of the smoothness parameter of the prior. If each summary measure suggests different values for θ, then our prior "knowledge" may well be logically inconsistent and should be reconsidered.

Consider, for example, the case in which experts know that, in a certain age range $\mathcal{A}$, the change in log-mortality from one age group to the next is expected to remain around a certain level, or that, on average over time, is not expected to exceed a certain level. This information could be captured respectively by the following two summary measures:

$$F_1(\mu) \equiv \frac{1}{T\#\mathcal{A}}\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}|\mu_{at} - \mu_{a-1,t}| \tag{6.15}$$

$$F_2(\mu) \equiv \frac{1}{T}\sum_{t=1}^{T}\max_{a\in\mathcal{A}}|\mu_{at} - \mu_{a-1,t}|, \tag{6.16}$$

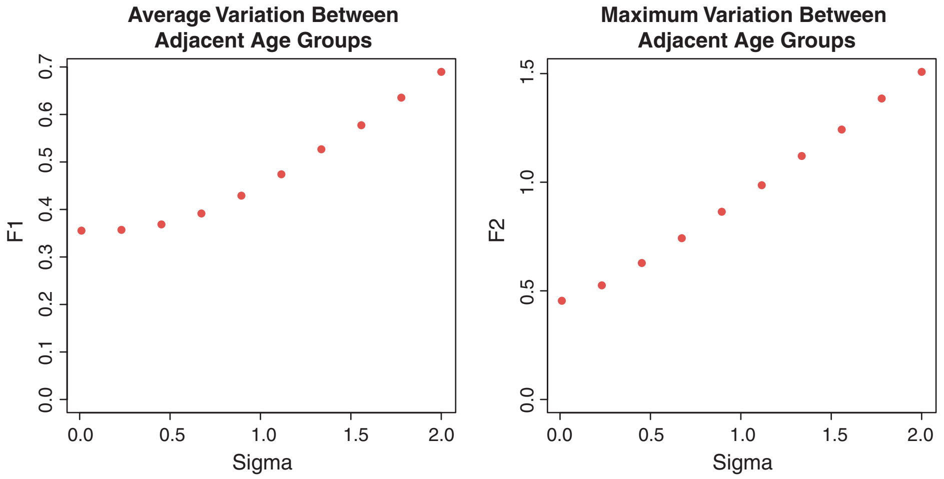

FIGURE 6.3. The expected value of the summary measures F_1 and F_2 defined in equation 6.16 as a function of σ_{age}. This example refers to mortality from all causes in males, and the typical age profile μ was obtained by averaging the age profile of log-mortality of all countries in our database with more than 15 observations. The country of interest is the United States, and the covariates are a linear trend and GDP.

where $\mathcal{A}$ could be, for example, age range 20 to 80. The reason for restricting age groups to the range $\mathcal{A}$ could be that, for many causes of death, log-mortality varies much more in younger age groups, and deviates from the common, almost linear, pattern observed in older age groups.

The expected values of F_1 and F_2 depend on θ, and therefore on σ_{age}. In figure 6.3 we report the expected values of F_1 and F_2 as a function of σ_{age} for a wide range of values of σ_{age}.

For very small values of σ_{age}, the prior is concentrated around the typical age profile $\bar{\mu}$, and therefore the expected values of F_1 and F_2 reflect the properties of μ. As seemed likely, the expected value of both F_1 and F_2 increase with σ_{age}. Especially interesting in figure 6.3 is that, for both summary measures, the expected value does not depend strongly on σ_{age} for small values of σ_{age} (say $\sigma_{\text{age}} < 0.5$): in this region the properties of the samples from the prior, measured by F_1 and F_2, reflect quite closely the properties of the typical age profile $\bar{\mu}$, and are therefore within very reasonable limits. Only after a threshold is reached do increases in σ_{age} begin to produce noticeable increases in the summary measures.

The shape of the curve corresponding to the summary measure F_1 is quite typical, in our experience. In fact, this summary measure is closely related to another summary measure:

$$F_3(\mu) \equiv \frac{1}{T\#\mathcal{A}} \sum_{t=1}^{T} \sum_{a\in\mathcal{A}} (\mu_{at} - \mu_{a-1,t})^2,$$

and we expect that

$$\sqrt{\mathrm{E}_{\perp}[F_3(\mu)|\theta]} \approx \mathrm{E}_{\perp}[F_1(\mu)|\theta].$$

Note that F_3 is a quadratic in μ, and therefore in $\boldsymbol{\beta}$. Furthermore, when we use the formulas of appendix C, the expected value of any quadratic form in $\boldsymbol{\beta}$ is a linear function of $\frac{1}{\theta}$, and therefore a linear function of σ^2_{age}. This implies the following dependency of $\mathrm{E}_\perp[F_1(\mu)|\theta]$ on σ_{age}:

$$\mathrm{E}_\perp[F_1(\mu)|\theta] \approx \sqrt{k_1 + k_2\sigma^2_{\text{age}}},$$

which is precisely the behavior shown in the left panel of figure 6.3.

The fact that a single parameter, σ_{age}, determines all the properties of the samples of the prior is both an advantage and a disadvantage. It certainly simplifies our task of choosing θ, but it may also create problems: In the preceding example, the summary measures F_1 and F_2 assumed reasonable values, but what if the samples from the prior have the desired standard deviation but not the desired values of other summary measures? This is possible, especially if the summary measures involve dimensions on which the prior does not have much control, such as the time behavior. Therefore, in these cases additional priors must be used, increasing the number of smoothness parameters and therefore the number of summary measures over which we have control. This issue is discussed in chapter 7.

6.3 Choosing Where to Smooth

In chapter 5, we considered smoothness functionals of the form:

$$H[\mu,\theta] \equiv \theta \int_0^T dw^{\text{time}}(t) \int_0^A dw^{\text{age}}(a) \left(\frac{d^{\mathsf{n}}\mu(a,t)}{da^{\mathsf{n}}}\right)^2, \tag{6.17}$$

where the measure $dw^{\text{age}}(a)$ allows us to enforce a varying degree of smoothness for different age groups. In this section, we elaborate on this point and offer some practical examples of the effect of making different choices for $dw^{\text{age}}(a)$.

The choice of $dw^{\text{age}}(a)$ is related to the meaning of the dependent variable μ. As mentioned in chapter 5, μ could be the expected value of log-mortality or its deviation from some “typical” age profile $\bar{\mu}$. To clarify, we use μ to refer to the expected value of log-mortality and introduce $\bar{\mu}$ explicitly for prior’s mean for expected mortality. Because we are interested in the behavior over age groups at any fixed point in time, we drop the time variable, so that μ is only a function of age, and the smoothness functional becomes

$$H[\mu,\theta] \equiv \theta \int_0^A dw^{\text{age}}(a) \left(\frac{d^{\mathsf{n}}}{da^{\mathsf{n}}}(\mu(a) - \bar{\mu}(a))\right)^2. \tag{6.18}$$

If we choose $\bar{\mu} = 0$, and therefore a prior with zero mean, a nonuniform measure $dw^{\text{age}}(a)$ penalizes the variation in log-mortality from one age group to the next more in certain age groups and less in others. For example, suppose $dw^{\text{age}}(a)$ is such that it weights older age groups more than younger ones. Samples from such a prior will oscillate more and exhibit more variation at younger age groups, whereas they will be “stiffer” at older age groups.

We illustrate this idea in figure 6.4, where each graph displays 100 samples from the prior associated to the smoothness functional in equation 6.18. For these figures, we have

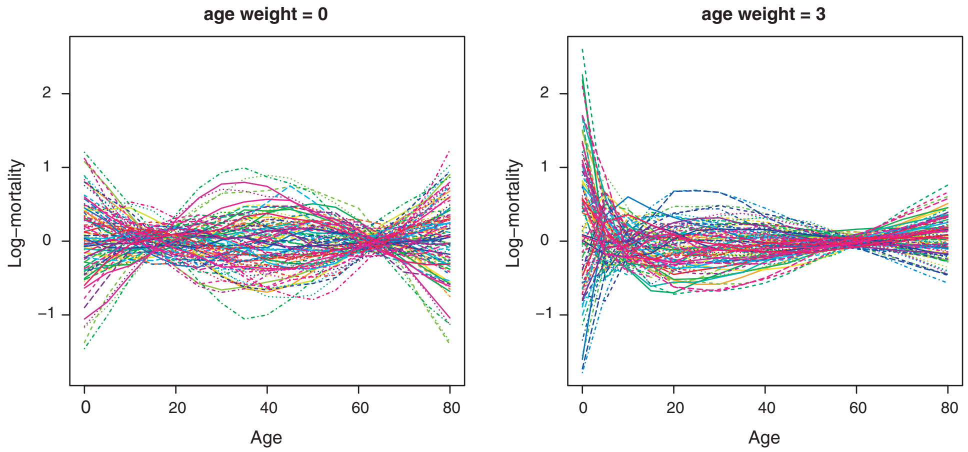

FIGURE 6.4. One hundred random draws from the smoothness functional in equation 6.18, with $\mu = 0$, $\mathbb{n} = 2$, and measure $dw^{\text{age}}(a) = a^l da$. For the graph on the left, $l = 0$ (a uniform measure), and for the graph on the right, $l = 3$. The standard deviation, averaged over age groups is the same in both graphs and equal to 0.3.

chosen $\bar{\mu} = 0$, $\mathbb{n} = 2$, and a measure of the form $dw^{\text{age}}(a) = a^l da$, $l \geq 0$. The graph on the left corresponds to $l = 0$, that is, to a uniform measure, and the one on the right, to $l = 3$. Notice that the graphs differ in two respects. First, when $l = 3$, the variation in log-mortality from one age group to the next, in each sample, is much larger for younger age groups than for older ones, as expected. Second, within each age group, the variance of log-mortality is, on average, higher in younger age groups (compare the huge variation observed at age 0 with the smaller variation observed at age 80).

In order to get an idea of how the parameter l can affect the result, we introduce a different way of looking at a smoothness functional. Instead of looking at samples from the prior, we use it to smooth the data in a classic nonparametric framework and plot the results for different values of θ, which indices how much effect the prior has. A large class of nonparametric smoothers is the one defined by the following minimization problem:

$$\min_{\mu} \ \|\mu - m\|^2 + \theta \mu' W \mu, \tag{6.19}$$

where m is an observed age profile $m = (m_1, \ldots, m_A)$ and $\mu' W \mu$ is the discretized version of the smoothness functional 6.18. Smoothers of this type are common in regularization and spline theory, and can be interpreted as Bayes estimates, as shown by Kimeldorf and Wahba (1970). For our purposes, the only thing we need to know about the problem 6.19 is that its solution is given by

$$\mu(\theta) = (I + \theta W)^{-1} m \tag{6.20}$$

and that it has the following special cases:

$$\mu(0) = m \ , \qquad \lim_{\theta \to \infty} \mu(\theta) = P_{\circ} m,$$

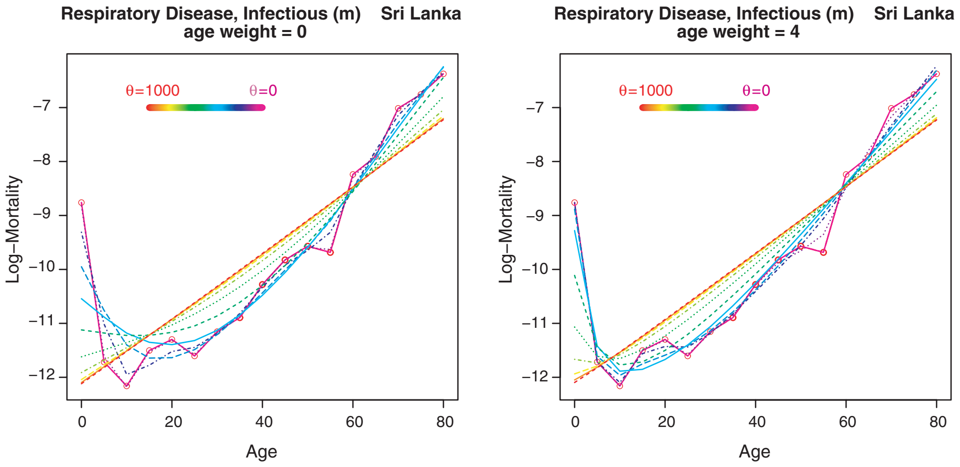

FIGURE 6.5. Smoothed versions of the age profiles of respiratory infectious disease in Sri Lankan males (data from year 2000). Here we use prior 6.18 with zero mean and report in each figure the results for θ from 0 to 1,000. For the graph on the left, $l = 0$ (uniform measure), and for the graph on the right, $l = 4$.

where $P_\circ$ is the projector onto the null space of the smoothness functional in equation 6.18. The last identity is easily derived using the eigenvector-eigenvalue decomposition of W and partitioning the eigenvectors into a basis for the null space and a basis for its orthogonal complement. Its interpretation is simple: when θ goes to infinity, the smoothed version of m is its best approximation from the null space of W. For example, if we choose $\mathbb{n} = 2$, so that the null space consists of the linear functions, when θ goes to infinity, the smoothed version of the data is the straight line that best fits m.

In figure 6.5 we report the result of the smoother in equation 6.20 for different values of θ and for different choices of l. The data in the two graphs are the same (the red dots) and correspond to the age profile of log-mortality for respiratory infectious disease in Sri Lankan males in year 2000. In each graph we have plotted the smoothed version of the data for 12 values of θ from 1,000 to 0 (the value 1,000 is, for all practical purposes, equal to infinity). The smoothed curves are color-coded along the rainbow colors: the lines in red to yellow correspond to very large values of θ, while those in blue-violet to very small values of θ. The only difference between the graphs is the value of l, which is 0 in the left graph and $l = 4$ in the right graph. Notice that when $l = 0$, too much smoothing occurs at younger age groups, and it is difficult to find a value of θ that smoothes the data well. When $l = 4$, instead, the smoothed curves are allowed to remain fairly steep and to "bend" a considerable amount at young ages, even for relatively large values of θ, producing a better range of smoothing curves.

Notice also how the smoothed curves become a straight line when θ becomes very large. This is an undesirable feature of this smoothness functional, and a consequence of having chosen $\bar{\mu} = 0$ and a null space consisting of straight lines. If a "typical" age profile $\bar{\mu}$ is available, much better results can be obtained, as we will now show. For the purpose of this experiment, we have synthesized a profile $\bar{\mu}$ by averaging the age profiles of the 50 countries (not including Sri Lanka) for which at least 20 observations are available (this criterion aims to select countries with "good" time series). Alternatively, we could have chosen to average only the countries that are "neighbors" of Sri Lanka.

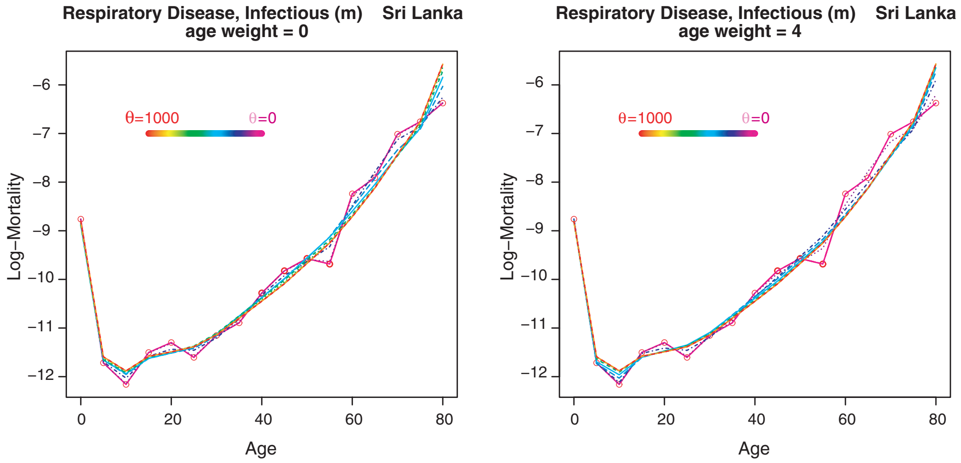

FIGURE 6.6. Smoothed versions of the age profiles of respiratory infectious disease in Sri Lankan males (data from year 2000). Here we use prior 6.18 with mean $\bar{\mu}$ different from 0, and report in each figure the results for θ from 0 to 1,000. For the graph on the left, $l = 0$ (uniform measure), and for the graph on the right, $l = 4$.

We now show in figure 6.6 the same kind of graphs we showed in figure 6.5, with the only difference being that $\bar{\mu}$ is no longer zero. Now there is much less difference between the two graphs, because the mean of the prior is responsible for explaining most of the large variation between age groups at younger ages. This could be expected: using a prior with nonzero mean is, in fact, equivalent to using a prior with zero mean where the dependent variable is the deviation of log-mortality from the typical age profile. While for log-mortality there is a strong argument for penalizing nonsmooth behavior less at younger age groups, a similar argument is less clear if the dependent variable is the deviation of log-mortality from the typical age profile. In this case, the reason for having $l \neq 0$ is slightly different: in some cases, knowledge about the shape of the age profiles could be more accurate in older age groups than in younger age groups, and therefore we would like to have a prior whose variance is higher in younger age groups. By smoothing less at younger age groups, we are also allowing the variance within each of the young age groups to be higher. Therefore, even if the prior has nonzero mean, we may still want to use a value of l different from 0. We illustrate this effect in figure 6.7, which is the counterpart of figure 6.4, but with $\bar{\mu}$ chosen as in figure 6.6. Now the samples from the prior are quite similar: the main difference between the two graphs is that, within each group, the variance of the samples is higher at younger age groups. Because the graphs in figure 6.6 correspond to the priors whose samples are represented in figure 6.7, it is not surprising that the results with $l = 0$ and $l = 3$ are fairly similar.

Obviously other forms of prior knowledge on the shape of age profiles could be available, which may not be represented by the simple choice $dw^{\text{age}}(a) = a^l da$ (e.g., one may want to allow more variation both in young and old age groups, but not in the middle ones). We suggest that, in any case, researchers study graphs of the kind we have produced here in order to understand what is the prior that best represents their knowledge.

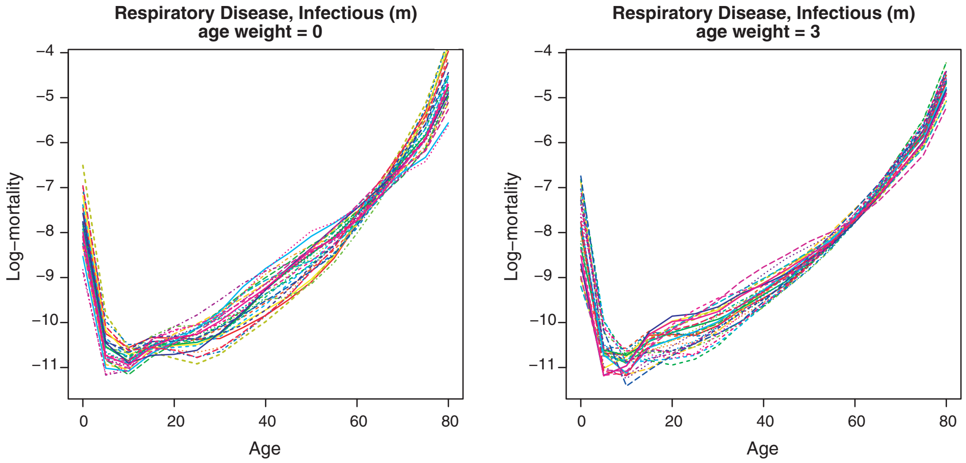

FIGURE 6.7. One hundred random draws from the smoothness functional in equation 6.18, with a nonzero mean $\bar{\mu}$, $\mathbb{n} = 2$, and measure $dw^{\text{age}}(a) = a^l da$. For the graph on the left, $l = 0$ (a uniform measure), and for the graph on the right, $l = 3$. The standard deviation, averaged over age groups is the same in both graphs and equal to 0.3.

6.4 Choosing Covariates

The choice of covariates in regression models is normally a major decision, or more important than most of the other statistical issues that often arise. Indeed, the same rules apply in forecasting with our models as with any other use of regression for forecasting: choose covariates that pick up on systematic patterns that are likely to persist, rather than idiosyncratic features likely to overfit in-sample data only. Reduce the chances of overfitting by using priors to reduce the effective sample space or, if necessary, drop covariates. Et cetera. The importance of these usual cautions is hard to overestimate, because even well-designed priors will not always avoid the bias induced by misspecifying covariates. But our procedure involves an additional factor that is implied by everything that has come before in this book and that we now make explicit.

The second step of our two-step procedure in section 5.2 is to project the prior specified in terms of the expected value of the dependent variable μ on the subspace spanned by the covariates $\mathbf{Z}_{at}$ into the lower-dimensional vector of coefficients $\boldsymbol{\beta}$. Effectively, we are able to invert what would be a noninvertible (many-to-one) relationship by restricting the full prior on μ to the subspace that spans $\mathbf{Z}$, for which the equation $\mu_{at} = \mathbf{Z}_{at}\boldsymbol{\beta}_a$ is invertible. The key to the whole procedure, however, is having a set of covariates that makes it possible to express the relationships of interest specified under the prior for μ. The danger is that a rich prior for μ could be matched with an impoverished set of covariates such that the resulting prior restricted to the subspace spanned by the covariates is not able to reflect most of the interesting patterns allowed under the original unrestricted prior for μ.

Thus, we now provide tools with which one can check to see that important characteristics of the prior are not lost when we take the projection. We focus in particular on the null space, because when we project the nonparametric prior on the space spanned by the covariates, we also project its null space, and in principle it is even possible for this operation to cause the null space to disappear or at least to be greatly reduced. For example, if we impose a prior that says μ is smooth over age groups but include covariates that are not smooth, then it could be that the only way the prior could produce smoothness is to set the slope coefficients to zero. For other examples, no values of the $\boldsymbol{\beta}$ parameters can produce forecasts consistent with the qualitative prior on μ. Because this would be an undesirable feature, we need to understand the conditions under which this could happen and how to avoid it.

6.4.1 Size of the Null Space

We denote by $\mu_t \in \mathbb{R}^A$ an age profile in year t, and write the nonparametric prior on μ as

$$H[\mu, \theta] = \frac{\theta}{T} \sum_t \mu_t' W \mu_t, \tag{6.21}$$

where $W = W^{\text{age},n}$ (the squared discretized derivative of μ with respect to age) in the rest of this section. We assumed a zero mean for this prior, because the mean is irrelevant for the computation of the dimension of null space. This prior is defined over μ in $H[\mu, \theta]$, which represents the $A \times T$ dimensional space of the T age profiles. Let $\mathfrak{N}(W)$ denote the null space of W (which, as per section 5.1, determines what patterns of μ the prior is indifferent to), and let $\dim(\mathfrak{N}) \equiv \text{nullity}(W)$ denote its dimensionality.

The null space of the nonparametric prior in equation 6.21 is obtained by allowing the age profile of each year to vary *independently* from the other years, in $\mathfrak{N}$. This implies that the dimensionality of the null space of the prior 6.21 is $T \times \text{nullity}(W)$. When we project the prior on the space spanned by the covariates, we obtain the following:

$$H[\boldsymbol{\beta}, \theta] = \frac{\theta}{T} \sum_{aa'} W_{aa'} \boldsymbol{\beta}_a' \mathbf{C}_{aa'} \boldsymbol{\beta}_{a'}. \tag{6.22}$$

This prior is defined on a much smaller space than equation 6.21. Or, in other words, the prior in equation 6.21 on μ, restricted to the space where $\mu = \mathbf{Z}\boldsymbol{\beta}$ holds, has a much smaller null space than before imposing the restriction.

In order to fix ideas, assume 17 age groups and 60 years of observations for a set of 7 covariates, with a nonparametric prior involving the second derivative only, so that $\text{nullity}(W) = 2$. The prior for μ is defined over $\mathbb{R}^{1020}$ ($1{,}020 = 60 \times 17$), and its null space has dimension 120 ($120 = 60 \times 2$). The prior on the coefficients in equation 6.22 is defined over $\mathbb{R}^{119}$ ($119 = 7 \times 17$), which is less than the dimensionality of the whole null space of the nonparametric prior!

In order to study the dimensionality of the null space of the prior on the coefficients in equation 6.22, it is convenient to start from the simple case in which the covariates do not vary across age groups (e.g., like GDP). Therefore the number of covariates in age groups 1 to A is the same, and we denote it by k. In this case, we have

$$\mathbf{C}_{aa'} \equiv \mathbf{C} \;\; \forall a, a',$$

where $\mathbf{C}$ is a symmetric $k \times k$ matrix and the prior has the form:

$$H[\boldsymbol{\beta}, \theta] = \frac{\theta}{T} \sum_{aa'} W_{aa'} \boldsymbol{\beta}'_a \mathbf{C} \boldsymbol{\beta}_{a'}. \tag{6.23}$$

Because the dimensionality of the null space obviously does not depend on the particular coordinate system we use to compute it, we perform a convenient change of variables. We assume, without loss of generality, that the covariates are orthogonal. Therefore the substitution

$$\sqrt{\mathbf{C}} \boldsymbol{\beta}_a \to \boldsymbol{\beta}_a \tag{6.24}$$

is an invertible transformation (i.e., C^{-1} will exist because of the absence of collinearity among the covariates), and we can study the prior in equation 6.23 in the new system of coordinates:

$$H[\boldsymbol{\beta}, \theta] = \frac{\theta}{T} \sum_{aa'} W_{aa'} \boldsymbol{\beta}'_a \boldsymbol{\beta}_{a'}.$$

Now denote by β_a^q the q-th component of the vector $\boldsymbol{\beta}_a$, so that $q = 1, \ldots, k$, and the $A \times 1$ vector $\boldsymbol{\beta}^q$ whose elements are β_a^q. Then we rewrite the preceding expression, which sums over age groups, as one which sums over covariates:

$$H[\boldsymbol{\beta}, \theta] = \frac{\theta}{T} \sum_{q=1}^{k} (\boldsymbol{\beta}^q)' W \boldsymbol{\beta}^q. \tag{6.25}$$

Recall that $k = 7$ in our running numerical example. In order for $H[\boldsymbol{\beta}, \theta]$ to be zero, each vector $\boldsymbol{\beta}^q$, with $q = 1, \ldots, k$, must be in the null space of W, which has dimension nullity(W). Therefore, *the null space of the prior in equation 6.23 has dimension* nullity$(W) \times k$. Using the numbers in the preceding example, we would have that the prior over the coefficients, which is defined over $\mathbb{R}^{119}$, has a null space of dimension 14 $(14 = 2 \times 7)$.

6.4.2 Content of the Null Space

Expression 6.25 also allows us to identify the exact content of the null space. Let us choose all the $\boldsymbol{\beta}^q$ to be 0 except for $q = q^*$, and let us assume that the prior over age groups is a standard smoothness prior with derivative of order $\mathfrak{n}$ (for the mixed smoothness, similar reasoning applies). Then the null space of W is the set of polynomials of degree $\mathfrak{n} - 1$. Therefore, $\boldsymbol{\beta}^{q^*}$ is in the null space of W if it can be written as

$$\beta_a^{q^*} = \sum_{j=0}^{\mathfrak{n}-1} v_i^{q^*} a^j \ , \quad \text{for any } v_i \in \mathbb{R} \ , \quad i = 0, \ldots, \mathfrak{n} - 1.$$

For this choice of coefficients, the patterns of log-mortality that belong to the null space of the prior are described as

$$\mu_{at} = z_t^{q^*} \beta_a^{q^*} = z_t^{q^*} \sum_{j=0}^{\mathfrak{n}-1} v_j^{q^*} a^j \ , \quad \text{for any } v_j \in \mathbb{R} \ , \quad j = 0, \ldots, \mathfrak{n} - 1.$$

These are patterns that at any point in time have an age profile that looks like a polynomial of degree $\mathbb{n}-1$, but whose coefficients evolve over time as the covariate $z_t^{q^*}$. Suppose, for example, that $\mathbb{n}=2$ and that q^* corresponds to the covariate GDP, so that $z_t^{q^*}=\text{GDP}_t$. Then a pattern in the null space of the prior can be written as

$$\mu_{at}=\text{GDP}_t(v_1+v_2 a) \quad \text{for any } v_1, v_2 \in \mathbb{R}.$$

This reasoning can be used for any given $q^*=1,\ldots,k$, and taking a linear combination of the corresponding $\boldsymbol{\beta}^q$, we can obviously span the null space of the prior. Therefore, the null space of the prior consists of patterns of log-mortality of the following general form:

$$\mu_{at}=\sum_{q=1}^{k} z_t^q \sum_{j=0}^{\mathbb{n}-1} v_j^q a^j, \quad \text{for any } v_j^q \in \mathbb{R},$$

where the coefficients v_j^q are arbitrary numbers (notice that there are exactly nullity$(W)\times k$ of them).

Covariates That Vary over Age Groups So far we have discussed the restrictive case in which the covariates are the same for all the age groups. The main observation necessary to understand the general case is that *the more the covariates differ across the age groups, the smaller the dimension of the null space of the prior*. The reason for this is that in order for a cross-sectional time series to be in the null space of the prior, the age profile for each year must be in the null space of W. That means that the coefficients have to satisfy, for each year, a complicated condition involving the covariates and the matrix W. In other words, *if the covariates have no regularity across the age groups, it may be impossible for the prior to find a set of coefficients that satisfies a requirement of regularity for every year.*

We reinforce this intuition with the following example. Suppose the covariates z_{at}^r are zero mean, unit standard deviation, independent and identically distributed (i.i.d.) random variables. If T is large enough, we will have

$$\mathbf{C}_{aa'}^{qr}=\frac{1}{T}\sum_{t=1}^{T} z_{at}^q z_{a't}^r \approx \delta_{aa'}\delta_{qr},$$

where δ is Kronecker's delta. In this case the prior 6.22 becomes

$$H[\boldsymbol{\beta},\theta]=\frac{\theta}{T}\sum_a W_{aa}\|\boldsymbol{\beta}_a\|^2.$$

Because W is semipositive definite, its diagonal elements W_{aa} are positive, and therefore the null space of this functional is $\boldsymbol{\beta}_a=0$ for all $a=1,\ldots,A$, and the prior is proper: the lack of correlation of the covariates across age groups has shrunk the null space of the prior to zero.

In order to quantify this intuition, let us consider the case in which there are k covariates, but some of them may be missing in some age groups (by making the number of unique covariates large enough, any case can be seen as a special case of this). Suppose, for example, we have seven covariates, one of which is missing below a certain age. The prior in equation 6.22 is now defined over a space of dimensionality smaller than $A\times k$.

However, we can rewrite it as a prior defined over $\mathbb{R}^{A\times k}$, but with a constraint: we can "fill in" the missing covariates with arbitrary values, while constraining the corresponding coefficients to be zero. This constraint is formalized by saying that the coefficients $\boldsymbol{\beta}$ belong to a subspace $\mathbb{S}$ of $\mathbb{R}^{A\times k}$. Because the unconstrained prior has the same covariates in each age group, the dimensionality of its null space is $\dim(\mathfrak{N}) \times k$. It follows that the dimensionality of the null space of the constrained prior has to be lower than that, and therefore $\dim(\mathfrak{N}) \times k$ provides a convenient upper bound.

A lower bound is available as well. In fact, let us drop the covariate that is missing in some age groups altogether or, equivalently, let us set to zero the coefficients corresponding to this covariate for all age groups. The resulting prior also has the same covariate in all the age groups, and therefore the dimensionality of its null space is $\dim(\mathfrak{N}) \times (k-1)$. Because this is a projection of the prior in equation 6.22 over a lower dimensional subspace, this is a lower bound for the dimension of the null space of the prior 6.22. This result is comforting: it implies that *as long as there is at least one covariate (e.g., the constant) that is the same across all age groups, the prior 6.22 is improper*. More precisely, if there are l covariates that are the same across the age groups, the dimension of the null space is at least $\dim(\mathfrak{N}) \times l$. We conjecture that this number is actually the correct dimensionality of the null space, but we have not proved it yet.

The discussion in this section applies to the prior defined over age groups, but its logic also applies to any of the priors that we describe in the next chapter, where, instead of smoothing the expected value of the dependent variable over age groups, we smooth it, for example, over time or countries. A detailed example of how the null space depends on the covariates is presented in section 7.1.

6.5 Choosing a Likelihood and Variance Function

Through most of this book, and in our software implementation, we specify the logarithm of the mortality rate to be normally distributed, a choice we refer to as "the normal specification." We also model the mean of the normal density as a linear function of the covariates, and we have let its variance be an unknown parameter σ_i^2, indexed by the cross-sectional index. Because our approach is Bayesian, we model the variances as random variables with a probability density $\mathcal{P}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$. This density must satisfy two constraints: it must reflect knowledge we have about the variances, and it must lead to a computationally feasible model. The problem of finding a suitable model for $\mathcal{P}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$ cannot be discussed without discussing the motivations behind our choice of the normal specification and the approximations and sources of errors associated with it. Therefore, we start this section by reviewing the usual rationale for the normal specification.

6.5.1 Deriving the Normal Specification

The raw variable we observe is d_{it}, the number of people who die in year t in crossection i. Because of the "count" nature of this variable, a reasonable starting point is to assume that d_{it} can be described by a Poisson process, with unknown mean λ_{it}. We summarize this information as follows:

$$d_{it} \sim \text{ Poisson}(\lambda_{it}), \quad \mathrm{E}[d_{it}] = \lambda_{it}, \quad \mathrm{Var}[d_{it}] = \lambda_{it}. \tag{6.26}$$

This model is highly constrained, because the mean and variance of this density are not independent. A more flexible model would be given by a Pólya process, rather than a Poisson process, where the Poisson density would be replaced by a negative binomial, in which the variance can be any number larger than the mean, or the generalized event count model that allows the variance to be greater than or less than the mean (King, 1989a; King and Signorino, 1996). We do not consider alternative processes here, since that would considerably lengthen our exposition without leading us in the end to different practical choices. From a conceptual point of view the various count models are appealing, and there is nothing in our model that prevents us from using any of them. Because they simply correspond to different choices of the likelihood in the expression for the Bayesian estimator in equation 4.3 (page 58), the same priors still apply. However, because they would lead to fairly complicated implementations, we look for a computationally simpler alternative.

The key observation at this point is that, if we think of the Poisson density as a function of a continuous random variable, then it can be well approximated, under certain conditions and for appropriate choices of the parameters, by a log-normal density. The log-normal is a density with two free parameters, ν and ϱ, whose functional form is reported in appendix B.3.3 (page 239). In the following, if a random variable d has a log-normal density, we write $d \sim \log \mathcal{N}(\nu, \varrho^2)$. If we want to approximate a Poisson density with a log-normal density, we must choose the parameters ν and ϱ in such a way that the mean and variance of the log-normal match the mean and variance of the Poisson density. By using the formulas in appendix B.3.3 (page 239), it is easy to see that the log-normal approximation to the Poisson density of equation 6.26 is

$$d_{it} \sim \log \mathcal{N}\left(\log \lambda_{it} + \frac{1}{2} \log\left(\frac{\lambda_{it}}{1 + \lambda_{it}}\right), \log\left(1 + \frac{1}{\lambda_{it}}\right)\right). \tag{6.27}$$

The advantage of the approximation of the Poisson density by a log-normal is that

$$x \sim \log \mathcal{N}(\nu, \varrho^2) \Longleftrightarrow \log x \sim \mathcal{N}(\nu, \varrho^2).$$

Thus, if we could model the observed number of deaths by equation 6.27, it would follow immediately that log-mortality would be modeled with a normal distribution. In fact, dividing d_{it} in equation 6.27 by population p_{it} and using the preceding property, we obtain

$$m_{it} \sim \mathcal{N}\left(\log \frac{\lambda_{it}}{p_{it}} + \frac{1}{2} \log\left(\frac{\lambda_{it}}{1 + \lambda_{it}}\right), \log\left(1 + \frac{1}{\lambda_{it}}\right)\right). \tag{6.28}$$

Because by definition $\lambda_{it}/p_{it} = \mathrm{E}[M_{it}]$, where $M_{it} = d_{it}/p_i t$, the preceding expression implies that

$$\mu_{it} \equiv \mathrm{E}[m_{it}] = \mathrm{E}[\log M_{it}] = \log \mathrm{E}[M_{it}] + \frac{1}{2} \log\left(\frac{\lambda_{it}}{1 + \lambda_{it}}\right) \geq \log \mathrm{E}[M_{it}].$$

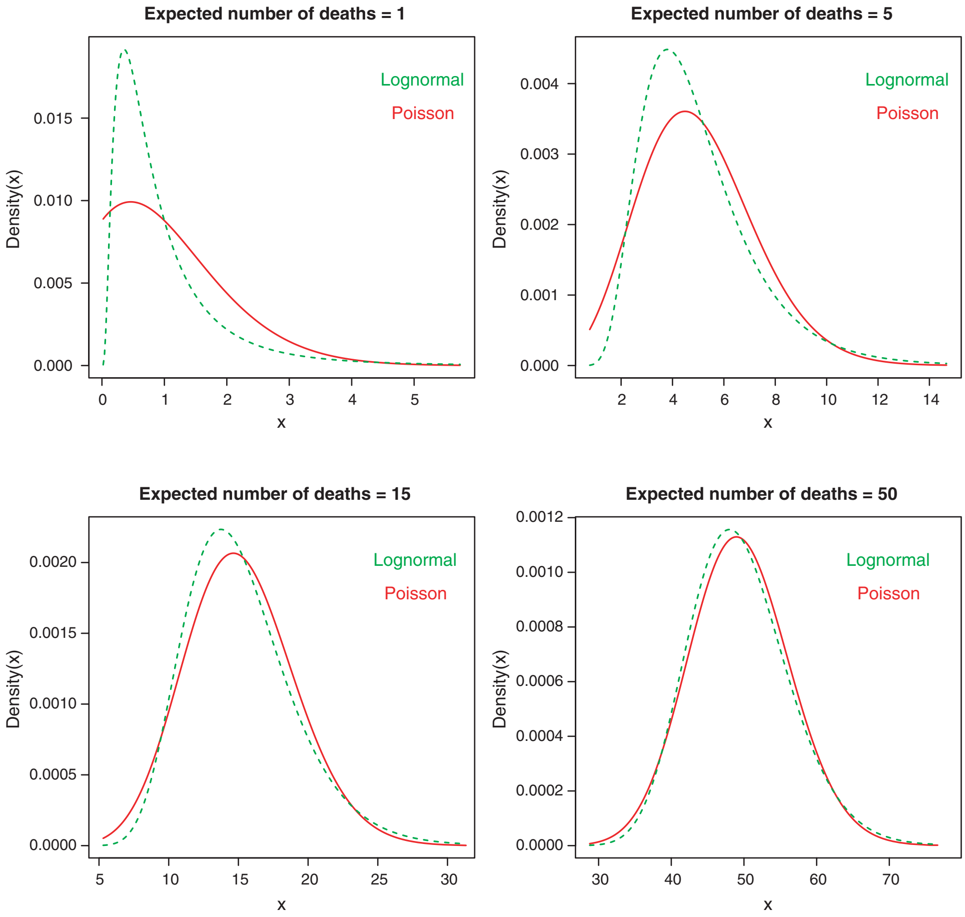

FIGURE 6.8. The log-normal approximation (in green) to the Poisson density (in red), as a function of number of deaths, for different expected numbers of deaths.

As expected (from Jensen's inequality), the expected value of log mortality is larger than the log of the expected value of mortality, with the difference decreasing as the expected number of deaths increases.

Before discussing the implications of the preceding expression, we first study precisely when the approximation that led to it is appropriate.

6.5.2 Accuracy of the Log-Normal Approximation to the Poisson

Here we compare the Poisson density, seen as a function of a continuous variable, with its log-normal approximation. Figure 6.8 shows both densities for four different values of the mean.

As illustrated in figure 6.8, the approximation improves as the expected value of the number of deaths, λ_{it}, increases, with large errors occurring when $\lambda_{it} < 5$. The crucial difference between the log-normal and the Poisson density is the behavior at the origin: while the log-normal density is 0 at the origin (because it contains a negative exponential in $(\log d_{it})^2$), the Poisson is not. Therefore, a sample from the log-normal may generate d_{it} close to zero, but never zero, whereas a sampling from the Poisson (as a discrete distribution) can certainly generate $d_{it} = 0$. This implies that any attempt to use the log-normal density in a likelihood where the data are generated by a Poisson distribution will result in an attempt to compute the logarithm of zero.

To give an idea of the numbers involved, with a Poisson density when $\lambda_{it} = 1$, the probability of observing a 0 is 37%, whereas when $\lambda_{it} = 5$, this probability drops to 0.7%, dropping to a negligible value of 4.5×10^{-5} when $\lambda_{it} = 10$.

These considerations suggest that there are three "regimes" in which we may need to operate: a large value, a small value, and a very small value of λ_{it}.

Large Value of $\boldsymbol{\lambda_{it}}$ In this case, the observed number of deaths contain no zeros. This is likely to happen when the expected number of deaths λ_{it} is always larger than 10 or 15. In this situation, equation 6.27 can be simplified:

$$d_{it} \sim \log\mathcal{N}\left(\log\lambda_{it} + \frac{1}{2}\log\left(\frac{\lambda_{it}}{1+\lambda_{it}}\right), \log\left(1 + \frac{1}{\lambda_{it}}\right)\right) \approx \log\mathcal{N}\left(\log\lambda_{it}, \frac{1}{\lambda_{it}}\right).$$

To see the amount of error involved in these approximations, let us take $\lambda_{it} = 15$. In this case, the term $\frac{1}{2}\log(\frac{\lambda_{it}}{1+\lambda_{it}})$ in the mean is -0.032, which is negligible when compared to $\log\lambda_{it} = 2.7$ (of a factor 100). For the term in the variance, we have that $\log(1 + \frac{1}{\lambda_{it}}) = 0.064$, which is well approximated by $\frac{1}{\lambda_{it}} = 0.066$.

The corresponding simplified specification for log-mortality is then:

$$m_{it} \sim \mathcal{N}\left(\log\frac{\lambda_{it}}{p_{it}}, \frac{1}{\lambda_{it}}\right). \tag{6.29}$$

In this regime, the expected value of log-mortality and the log of the expected value of mortality essentially coincide, and the variance of log-mortality is inversely proportional to the expected value of the number of deaths. This situation is very common when dealing with all-cause mortality or for the leading causes of death in countries that are not too small and, in most cases, for other than very young ages. The pattern is less common as we move to rarer causes of death, small countries, or younger age groups.

To provide some specificity, consider a common cause of death, cardiovascular disease, in a hypothetical country, similar to the United States, with a total population 280 million, for age group 70–74 among males. Reasonable values in this case are $\lambda_{it} = 60{,}000$ and $p_{it} = 4{,}000{,}000$, which correspond to $\mathrm{E}[M_{it}] = 1.5\%$. Under these conditions, the problem of zeros is nonexistent, and the log-normal and Poisson density are virtually identical.

Now "scale" this hypothetical country down by a factor of 600, keeping the mortality rate constant. We would obtain an expected number of deaths of 100 in a population of 6,666 people, in a country with a total population of about 470,000. For this country, we would still not expect any zeros in the observed number of deaths, and the log-normal approximation would still be appropriate. If we scaled our initial country down by a factor 36,000, then we could run into problems: the expected number of deaths would be only 1.6 in a population of 111 people, and the total population of the country would be only 7,777. Under these conditions, we should expect a nonnegligible number of zeros in the observed number of deaths, which would make the application of the model in equation 6.29 practically impossible (because we would need to take the logarithm of zero) and imprecise (even if by luck we do not have zeros, the log-normal density does not approximate the Poisson density very well in this circumstance).

Let us make this discussion more empirical. In year 2000, two countries whose population in age group 70–74 was around 111 were the Cook Islands and Palau. The observed number of deaths by cardiovascular disease for males in this age group for the two countries was 3 and 5, respectively, probably reflecting higher mortality rates than in the United States. In younger age groups, these countries exhibit nonnegligible numbers of zeros in observed deaths.

Small Value of $\boldsymbol{\lambda_{it}}$ In this case, the data will have some observed zeros, although most of the data will not contain zeros. We can expect this situation whenever λ_{it} is somewhere between 2 and 10 (for $\lambda_{it} = 2$, a Poisson density will generate data that are 0 about 13% of the time). We face two problems in this case : (1) in this range of λ_{it} the log-normal is not a very good approximation of the Poisson distribution, and (2) we do not know what to do when $d_{it} = 0$ because its logarithm is not defined. A common "fix" to the second problem consists of assigning to each cross section an extra 0.5 deaths every year Plackett (1981).

Because there are great computational advantages in retaining a normal specification for log-mortality, like the one in equation 6.29, we now study the size of errors associated with this procedure. To begin, suppose λ_{it} is small, but we add 0.5 deaths to each observation and proceed as if λ_{it} were large. How large is the error we make if we estimate λ, assuming that equation 6.29 still holds?

To study this question, we take 500,000 draws from a Poisson distribution with mean λ, for $0.5 \leq \lambda \leq 10$ (we omit the indices it for simplicity, as if we were considering one specific cross section in one specific year). To these points, we add a value of 0.5 and then we take their logarithm. We consider the result our sample of log-mortality, which we analyze as if it were normally distributed according to equation 6.29. Let $\hat{\mu}$ be the empirical average of log-mortality in our sample. If equation 6.29 holds, then we can estimate λ as $\hat{\lambda} = e^{\hat{\mu}}$. We repeat this procedure for many different values of λ, and for each value we compute the percentage error $\frac{|\hat{\lambda}-\lambda|}{\lambda}$. We report our results in figure 6.9.

The approximation error we obtain with this procedure is surprisingly small, dropping below 2% for λ greater than 2. Although this is reassuring, it does not mean that the density of log mortality with the "fix" is well represented by equation 6.29: it merely says that its expected value is well approximated by the expected value of the density 6.29 (although a similar phenomenon holds for the variance, too). In order to perform a more stringent test, we perform a different simulation.

Thus, we generate a sample of 500,000 points from a Poisson distribution with mean λ for $0.5 \leq \lambda \leq 10$, add 0.5, and take their logarithm as before. We consider the resulting sample our data for log-mortality, which we now analyze as if it were normally distributed according to $\mathcal{N}(\mu, \sigma^2)$, where estimates $\hat{\mu}$ and $\hat{\sigma}$ of μ and σ are obtained in a standard way. We do not make any assumption about how μ is related to λ. In order to estimate λ, we sample from $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ to obtain a new sample for log-mortality, which we then convert to a sample for the number of deaths by simple exponentiation. This step is crucial, because this sample will not look like the sample obtained from the Poisson density, especially when λ is small (it will have a log-normal distribution). As a final step we compute the empirical average of the mortality values from the new sample, which is an estimate of λ that we denote by $\hat{\lambda}$. We repeat this procedure for many different values of λ, and for each value we compute the percentage error $\frac{|\hat{\lambda}-\lambda|}{\lambda}$. We report our results in figure 6.10, which now displays larger errors.

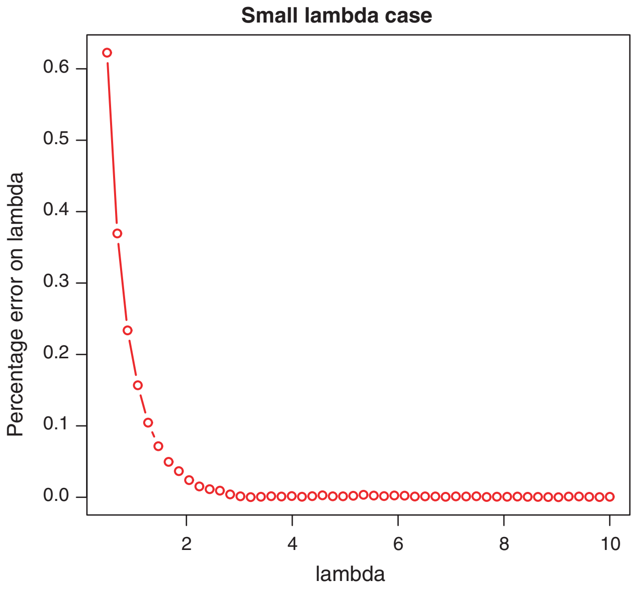

FIGURE 6.9. The error in estimating the expected number of deaths from log-mortality with zeros in observed deaths and 0.5 added to each observation. The expected number of deaths is estimated under the assumption that equation 6.29 is still valid.

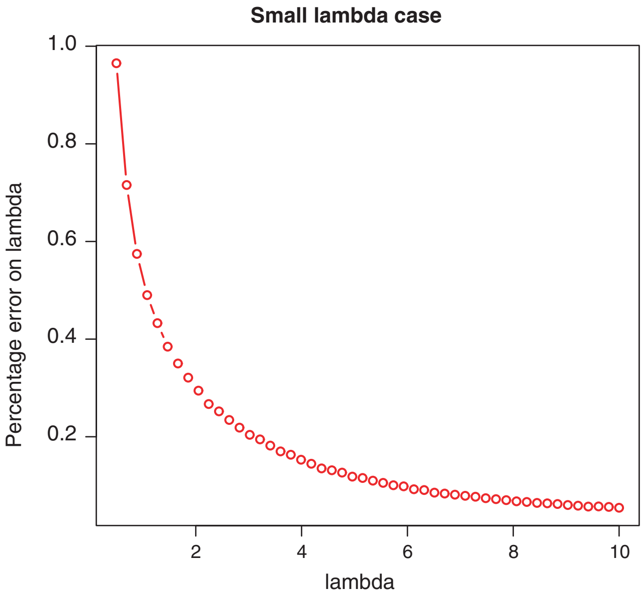

FIGURE 6.10. Error in estimating expected deaths from log-mortality with observed zeros in the number of deaths and with 0.5 deaths are added to each observation. We have not assumed that equation 6.29 holds.

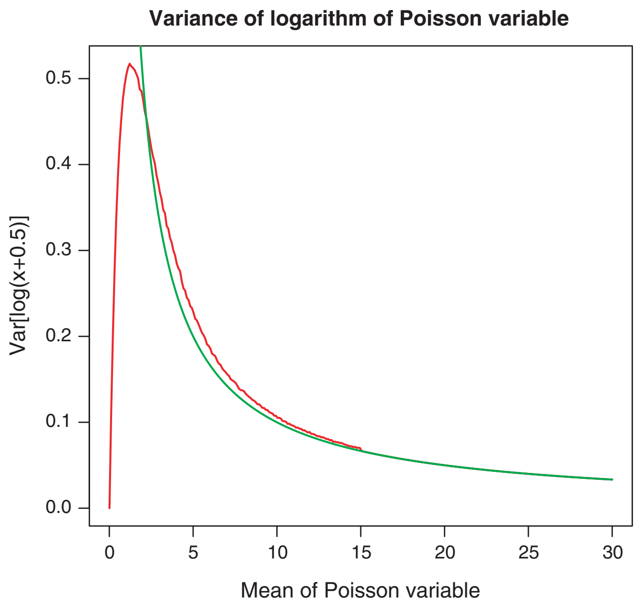

FIGURE 6.11. Approximating the variance of the logarithm of a Poisson variable. In red we report $V[\log(d_{it} + 0.5)]$ for $d_{it} \sim \text{Poisson}(\lambda_{it})$, as a function of λ_{it}. In green we report the function $\log(1 + \frac{1}{\lambda_{it}})$, the variance of the logarithm of the number of deaths, as a function of λ_{it}, when the Poisson and the log-normal density are close to each other and there is no 0.5 additional term.

But how large are these errors? When λ is small, say three, the standard deviation of the Poisson density is quite large (for $\lambda = 3$, it is $\sqrt{3} = 1.73$), and therefore there is a lot of variation built in the data. Thus, expecting great accuracy in estimating λ is unreasonable, even if we use the correct assumptions. Examining figure 6.10 with this in mind, and noting that a percentage error on λ of even 20% is not large on this scale, we note that the "fix" of *adding 0.5 deaths to each observation could probably be used for values of λ as small as 3 without serious consequence*. This assessment takes explicitly into account the difference between the log-normal and the Poisson distribution, and therefore it is more informative than the one of figure 6.9.

The figures shown so far suggest a range of values of λ_{it} such that adding 0.5 to the number of deaths does not noticeably destroy the information about the expected values of death, while still allowing one to use a normal specification for log-mortality. It remains to be shown that this procedure does not alter the structure for the variance. To this end it is instructive to perform a simple simulation: to compute the variance of $\log(d_{it} + 0.5)$ when $d_{it} \sim \text{Poisson}(\lambda_{it})$ for several values of λ_{it}. Given the preceding results, we would expect, in the regime in which the Poisson and the log-normal density are not too far apart, that this variance is equal to $\log(1 + \frac{1}{\lambda_{it}})$ (see equation 6.27). We test this hypothesis in figure 6.11, where on the horizontal axis we have λ_{it}, and the curve in green is $\log(1 + \frac{1}{\lambda_{it}})$ as a function of λ_{it}, while the one in red is $V[\log(d_{it} + 0.5)]$ for $d_{it} \sim \text{Poisson}(\lambda_{it})$. Again, *serious deviations between these two curves occur only for $\lambda_{it} < 3$.*

Figure 6.11 underlines an important point about the variance of log-mortality. In general it is not possible to define the random variable log-mortality $m_{it} = \log d_{it}$ (assuming population $p_{it} = 1$), unless we know that d_{it} never assume zero values, which is certainly not the case when λ_{it} is small and d_{it} is a Poisson process. Therefore, it does not make sense to talk of the variance of log-mortality when λ_{it} is very small. It does makes sense to define the random variable $\log(d_{it} + \alpha)$, where α is any number larger than 0. It is tempting therefore to interpret the variance of log-mortality as the variance of $\log(d_{it} + \alpha)$ and then let α go to 0. Unfortunately, figure 6.11 suggests that this cannot be done. The figure, which corresponds to $\alpha = 0.5$, is representative of the behavior of the variance of $\log(d_{it} + \alpha)$: for $\lambda_{it} = 0$, the variance is 0 (because the density is concentrated at the origin), and in a neighbor of $\lambda_{it} = 0$, the variance increases with λ_{it} before starting to decrease. For values of α smaller than 0.5, the location of the maximum will shift to the left and the curve will become steeper, but the shape of the curve remains the same. As a result, the limit of this curve for α going to 0 does not exist at the origin (the curve becomes discontinuous: it is 0 at the origin and then becomes a finite, large number as soon as we leave the origin). Therefore, writing $\mathrm{V}[\log m_{it}] \approx \frac{1}{\lambda_{it}}$ for λ_{it} going to 0 cannot be correct.

In order to understand what situations correspond to "small" expected values of deaths, we perform an exercise similar to the one we have done for λ large. We start with the same large country (total population 280,000,000), and consider a cause of death not as common as cardiovascular disease, for example, homicide, in the same age groups as before, that is 70–74. If we consider the male population, a reasonable value for λ_{it} is 135, and for p_{it} is 4,000,000, which corresponds to $\mathrm{E}[M_{it}] = 3.4 \times 10^{-5}$. Let us now scale this country down by a factor of 45, keeping mortality the same: this makes the total population approximately 6,200,000, with $\lambda_{it} = 3$ and $p_{it} \approx 88,000$. For such a country, we would expect to see a number of zero observed deaths, and we would have to add 0.5 deaths to each observation if we wanted to retain the normal specification for log-mortality.

Again, we compare these calculations to real data: in the year 2000, two countries with population in the age group 70–74 were close to 88,000—Denmark and Finland—and both report some zeros for the number of deaths. The total population of both countries is around 5.1 million. For Finland the average number of deaths over the past 10 years has been 2.1, while for Denmark it has been 0.7, and in both cases the number of zeros observed in the time series is consistent with expected deaths of that size (ignoring the downward time trend).

Very Small Value of $\boldsymbol{\lambda_{it}}$ This case corresponds to situations where many observations have $d_{it} = 0$, which is likely to happen when λ_{it} is smaller than 2 or 3. The data in these cases contain very little information, and assigning 0.5 deaths to each observation could be highly distortive. The problem is not so much the correct specification of the density but the paucity of data. Absent prior information, it is not clear that any sort of meaningful statistical inference can be performed on the data. Because we do have prior information, we use it to deal with these cases with an appropriate preprocessing imputation stage. Whenever zeros are found in the data, we fill them in with values borrowed from nearby cross sections and nearby points in time. An alternative to this preprocessing is simply to consider the zero values as missing and to let the prior take over, although this risks selection bias. We use the preprocessing approach mostly because of an implementation issue: mortality data have the convenient feature that if the number of deaths is observed in one age group, it is normally observed in all age groups.

This regime is common when studying very small countries, such as islands in the Pacific. In these cases it can easily happen that an entire age profile is 0, even for causes of death that are not very rare. Obviously, adding 0.5 deaths would be wrong in these cases, and the distinction between an entirely zero age profile and entirely missing mortality rate is very small. Less dramatic cases occur in small countries such as Honduras or Nicaragua for causes of death such as breast cancer. For example, in Nicaragua, in the period before 1980, in every year the age profiles had only 3 or 4 nonzero observations, with the typical value hovering around $d_{it} = 3$. Because the shape of the age profile for breast cancer is fairly well known, it is not difficult to use the nonzero observations to fit a reasonable age profile, and therefore impute the observations corresponding to the zero values.

In our applications, we have not found reasons to give up the computational advantages of the normal specification to use a Poisson or negative binomial specification, especially because our primary interest is in the forecast point estimate. We have seen in this section, and confirmed in our experiments, that when the expected number of deaths λ_{it} is large, the Poisson specification does not help, and when λ_{it} is small but not very small, adding 0.5 deaths to each observation does not cause enough error to justify changing the specification. In short, once 0.5 deaths are added to each observation, equation 6.29 is a reasonable choice.

The results of this section leave open the possibility that, instead of a Poisson density, a different density should be used as starting point for evaluating our approximation. In particular, a different density could allow the variance of the number of deaths to be less tightly tied to the expected value λ_{it}. We address this issue by using a specification for the variance slightly more general than the one suggested by equation 6.29, a topic that we discuss in the next section.

6.5.3 Variance Specification

If the normal specification in equation 6.29 is correct, we would expect to see the variance of log-mortality to be inversely proportional to the expected number of deaths. Because for every year and cross section we have only one observation, we cannot test this hypothesis directly. The value of λ_{it} could be approximated with the observed value d_{it}, but we cannot do the same to get the variance of m_{it}, for which we need at least two observations. The problem obviously is that the random variables involved are not stationary. A quick way around that is to assume that they are "temporarily" stationary, so that we can assume m_{it} and $m_{i,t+1}$ are drawn from the same distribution. In this case, we take their average absolute differences as an estimate of the standard deviation of m_{it}, which according to the model should be $\frac{1}{\sqrt{\lambda_{it}}} \approx \frac{1}{\sqrt{d_{it}}}$. Therefore, to check how well the following relationship holds, we make this comparison:

$$\frac{1}{T}\sum_t |m_{i,t+1} - m_{it}| \approx \frac{1}{T}\sum_t \frac{1}{\sqrt{d_{it}}}, \tag{6.30}$$

where we are averaging over time in order to reduce the estimation variance in computing the standard deviation.

We offer some illustrative examples for cardiovascular disease in men in figure 6.12 and breast cancer in women in figure 6.13. Each of the four graphs in each figure is specific to a country, cause of death, and gender. Each graph plots the left side of equation 6.30,

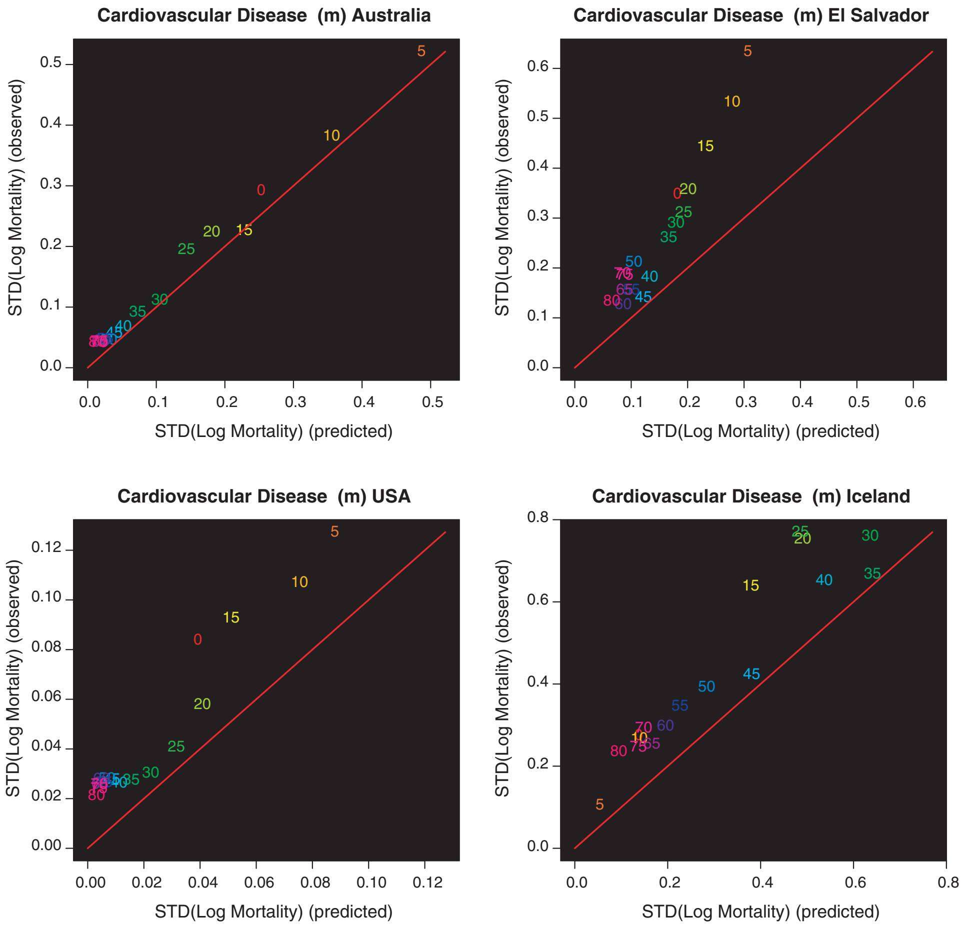

FIGURE 6.12. The log-normal variance approximation for cardiovascular disease in men. The left and right sides of equation 6.30 are plotted against each other for 17 age groups. Deviation from the 45° line indicates countries and diseases where the approximation holds less well.

on the vertical axis, against its right side, where the cross-sectional index i varies over 17 age groups. A red line is drawn at 45°, where the points should be if the relationship in equation 6.30 held exactly (and without measurement error). Note that the vertical axis, which determines the meaning of specific deviations from the 45° line, is different in each graph.

In some of these graphs, such as for cardiovascular disease in Australian males and for breast cancer in females in Italy and Malta, the relationship holds quite well, especially considering the approximation involved. In other cases, for example, cardiovascular disease in males in the United States and Iceland, the relationship holds qualitatively, in the sense that the pattern is correct. In cases such as breast cancer in Kuwait, it is hard to say, since the points are so dispersed relative to the narrow range in which mortality varies over ages. Examples like El Salvador are clearly violations, but in practice, we let our variance approximation be a scalar multiple of the true variance, which means that deviations from the 45° line that fall around a line through the origin, such as in El Salvador, fit well. We

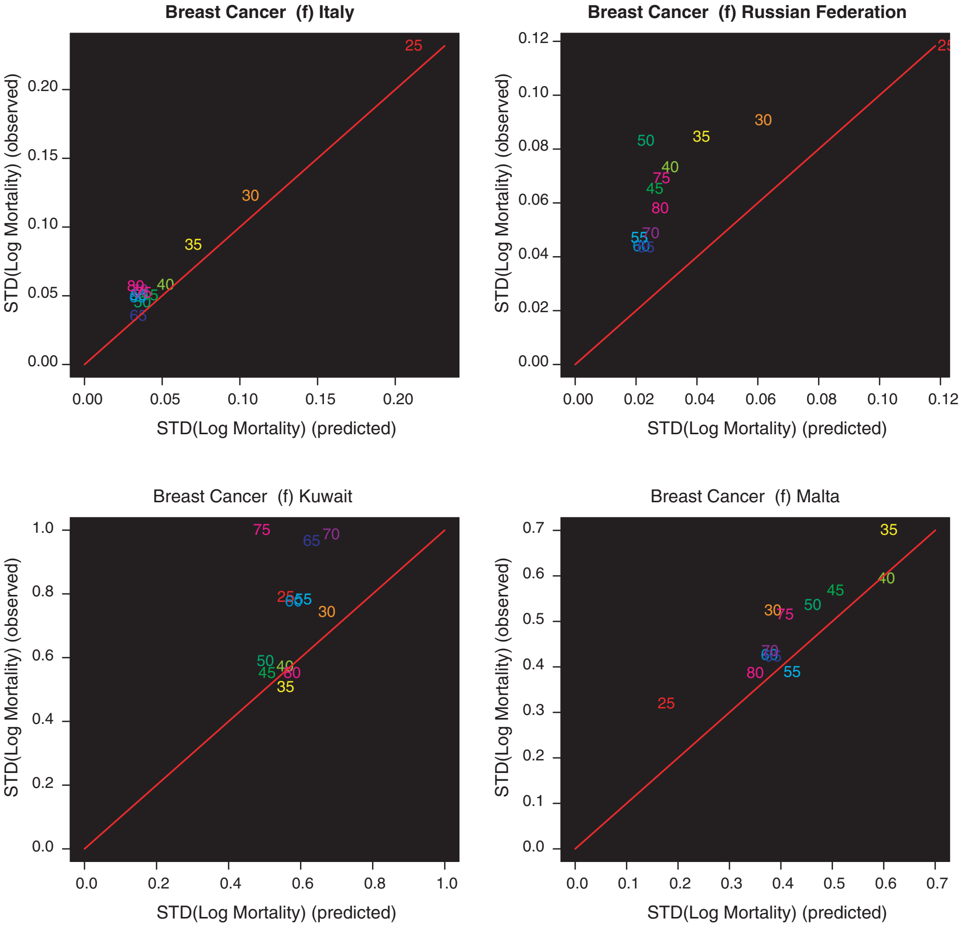

FIGURE 6.13. The log-normal variance approximation for breast cancer in women. The left and right sides of equation 6.30 are plotted against each other for 17 age groups. Deviation from the 45° line indicates countries and diseases where the approximation holds less well.

could also introduce an additive scalar correction, to allow for linear deviations that do not pass through the origin, but we have not found this necessary.

After having looked at many different countries and many different causes of death, the tentative conclusion we have drawn is that, when the number of deaths is not too small, it is qualitatively true that the variance of log-mortality decreases with the expected number of deaths, although the exact functional form may vary. To return to the question posed at the beginning of this section, How do we translate this knowledge into a prior for the standard deviations?

Begin by assuming the usual linear specification for the expected value of log-mortality $\mu_{it} = \mathbf{Z}_{it}\boldsymbol{\beta}_i$ and that λ_{it} is large enough so that equation 6.29 is valid. Then we can identify the expected value of log-mortality with the logarithm of the expected value of mortality and therefore set $\lambda_{it} = p_{it}\exp(\mathbf{Z}_{it}\boldsymbol{\beta}_i)$. If we truly believed in this model, we would then have a different standard deviation σ_{it} for each observation. A way to capture the inverse proportionality between σ_{it} and λ_{it} would be to make the σ_{it} correlated with the regression coefficients $\boldsymbol{\beta}_i$, and write $\mathcal{P}(\sigma^2 \mid \boldsymbol{\beta}) = \prod_{it} \mathcal{P}(\sigma_{it}^2 \mid \boldsymbol{\beta}_i)$, and set $\mathcal{P}(\sigma_{it}^2 \mid \boldsymbol{\beta}_i)$ to some density whose mean value is $\frac{1}{\lambda_{it}} = \frac{1}{p_{it}} \exp(-\mathbf{Z}_{it}\boldsymbol{\beta}_i)$. An obvious choice is to set $\sigma_{it} = \frac{\sigma_{it}^*}{\sqrt{\lambda_{it}}}$ where σ_{it}^* is a random variable whose expected value is 1. An extreme case of this model would be to set $\sigma_{it}^2 = \frac{1}{p_{it}} \exp(-\mathbf{Z}_{it}\boldsymbol{\beta}_i)$.

There are at least two problems with this approach. First, having σ and $\boldsymbol{\beta}$ correlated is computationally complicated and would lead to a fairly slow implementation in terms of Markov Chain Monte Carlo. Second, having one standard deviation for each observation leads to an unreasonable number of parameters.

Thus, we first modify this approach by removing the dependency of the standard deviations on time and hence writing σ_i instead of σ_{it}. This step is reasonable because the variation over time is small relative to the variation over cross sections (age groups in particular). Second, we model σ_i as inversely proportional to some average historical level of the number of deaths for cross section i, which we assume to be known a priori and denote by $\bar{\lambda}_i$. If we think of $\bar{\lambda}_i$ as a number that sets the order of magnitude of σ_i, then we model the uncertainty around σ_i by writing $\sigma_i = \frac{\sigma_i^*}{\sqrt{\bar{\lambda}_i}}$ and taking σ_i^* to be a random variable with mean value 1. This model still leads us to introduce $C \times A$ random variables, which is large in some applications. We have also found it useful to reduce further the number of parameters by allowing σ_i^* to vary only by age, although this latter choice is grounded in our experience and not in any theory that guarantees that it will hold in other applications.

To summarize, our variance specification has the form:

$$\sigma_{ca} = \frac{\sigma_a^*}{\sqrt{\bar{\lambda}_{ca}}}, \quad \mathrm{E}[\sigma_a^*] = 1. \tag{6.31}$$

The presence of $\bar{\lambda}_{ca}$ in the variance has a flavor similar to empirical Bayes. In fact, it is not reasonable to expect that we can elicit estimates of these quantities directly from experts, and some method that uses data needs to be used. A straightforward implementation would use the average historical value of number of deaths as an estimate of $\bar{\lambda}_{ca}$, or the average of the predicted values of a least-squares regression. This we find too data-dependent. In order to remain close in spirit to the rest of the book, we choose to borrow the value of $\bar{\lambda}_{ca}$ from neighboring cross sections, using the same weights we use in the prior for the regression coefficients. Although not entirely satisfactory, this approach seems to be a good compromise between practicality and statistical theory.