3 Methods with Covariates

We now introduce methods that use information from exogenously measured covariates to improve mortality forecasts. Information coded in covariates such as tobacco consumption, gross domestic product, and education rates are ignored in the methods of chapter 2. To the extent that we understand mortality and can measure its causes, therefore, the methods described herein have the potential to improve forecasts substantially.

Methods using covariates that have appeared in the literature, however, often do not directly model the age-mortality relationship and exclude the considerable information we have about it. As such, these methods should not be considered generalizations of those in chapter 2 and will not necessarily produce better forecasts. They will improve forecasts only when the additional information from the covariates outweighs the loss of information about the age profile of mortality. The methods we describe in part II combine the insights from the approaches in both chapters.

3.1 Equation-by-Equation Maximum Likelihood

The idea of equation-by-equation maximum likelihood (ML) is to analyze each time series (i.e., for one age, sex, country, and cause) with a separate regression. Few scholars would forecast mortality 30 years ahead using a single maximum likelihood regression applied equation by equation to a short time series, but the method is an obvious building block for most other methods described here. We explicate it here for expository purposes, to provide a baseline comparison with other approaches, to introduce some important issues in forecasting we use elsewhere, and as a starting point to introduce our notation. We consider two equation-by-equation approaches, based on the exponential-Poisson and linear-normal (or least-squares) specifications, with the latter an approximation to the former.

3.1.1 Poisson Regression

We observe d_{it}, the number of people who die in year t in cross section i. Because this variable is an event count, a reasonable starting point is to assume that d_{it} is a Poisson process, with unknown mean λ_{it}. (Obviously, more sophisticated event count models may be chosen, such as one based on the negative binomial or generalized event count

distributions [Cameron and Trivedi, 1998; King, 1989a; King and Signorino, 1996], but we do not need them for the expository purpose of this chapter.) We summarize this information as follows:

$$d_{it} \sim \text{Poisson}(\lambda_{it}), \quad \text{E}[d_{it}] = \lambda_{it}, \quad \text{Var}[d_{it}] = \lambda_{it}. \tag{3.1}$$

At the core of an equation-by-equation Poisson regression lies some specification for the expected value of the dependent variable as a function of some covariates. In our case, we are interested in the mortality rate $M_{it} \equiv d_{it}/p_{it}$, where p_{it} is the population of cross section i at time t. To define this formally, denote by $\mathbf{Z}_{it}$ a $1 \times k_i$ row vector of covariates and by $\boldsymbol{\beta}_i$ a $k_i \times 1$ column vector of coefficients. (Throughout this book, we use **bold** to indicate vectors or matrices with at least one dimension equal to k_i and Greek to denote unknown parameters; see appendix A.) A common choice is the following log-linear specification:

$$\text{E}[M_{it}] \equiv \exp(\mathbf{Z}_{it}\boldsymbol{\beta}_i). \tag{3.2}$$

Combining specification 3.2 with equation 3.1, we summarize the equation-by-equation Poisson regression model as follows:

$$p_{it}M_{it} \sim \text{Poisson}\left(p_{it}\exp(\mathbf{Z}_{it}\boldsymbol{\beta}_i)\right). \tag{3.3}$$

The log-likelihood function for model 3.3 is then:

$$\ln \mathcal{P}(M \mid \boldsymbol{\beta}) \propto \sum_i \sum_t p_{it}\left[M_{it}\mathbf{Z}_{it}\boldsymbol{\beta}_i - \exp(\mathbf{Z}_{it}\boldsymbol{\beta}_i)\right]. \tag{3.4}$$

Here, as in the rest of the book, we have used the convention that, when we drop an index of a variable, we mean that we are forming a matrix by taking the union over all the dropped indices.

The ML estimator of this model, which we denote by $\hat{\boldsymbol{\beta}}_i^{\text{PML}}$, satisfies the following first-order conditions:

$$0 = \sum_t p_{it}\mathbf{Z}_{it}\left[M_{it} - \exp\left(\mathbf{Z}_{it}\hat{\boldsymbol{\beta}}_i^{\text{PML}}\right)\right]. \tag{3.5}$$

3.1.2 Least Squares

Although the Poisson regression outlined in the previous section is appealing because it explicitly recognizes the event count nature of the data, it has a computational disadvantage in that the associated ML estimator requires optimizing nonlinear equations for which no analytical solution exists.

It is possible to modify the approach of the previous section in such a way that the resulting ML estimator is a least-square estimator and therefore can be computed as a solution of a linear system. The idea is to replace the Poisson distribution of equation 3.1 with the log-normal distribution. These two distributions are obviously very different near

the origin, but if the expected value of the number of deaths is not too small, then they are fairly similar. The assumption of log-normality implies that the logarithm of the number of deaths is normally distributed. Hence, we define the log-mortality rate as

$$m_{it} = \ln M_{it} = \ln \frac{d_{it}}{p_{it}}, \tag{3.6}$$

which would then be normally distributed (for $d_{it} \neq 0$). Because the logarithmic function has unbounded range, a linear specification for log-mortality can be used, suggesting the following model:

$$\begin{aligned} m_{it} &\sim \mathcal{N}\left(\mu_{it}, \sigma_i^2\right) \\ \mu_{it} &= \mathbf{Z}_{it}\boldsymbol{\beta}_i, \end{aligned} \tag{3.7}$$

where $\mu_{it} \equiv \mathrm{E}[m_{it}]$ is the expected log-mortality rate, and m_{it} is assumed independent over time after conditioning on $\mathbf{Z}$ (where $\mathbf{Z}$ may include lags of m_{it}). This model is also more flexible than Poisson regression, because the variance is not restricted to equal the mean.

Equation 3.7 posits a linear specification for the expected value of log-mortality, $\mathrm{E}[\ln M_{it}] \equiv \mathrm{E}[m_{it}] \equiv \mu_{it}$, while the previous section posits a linear specification for the log of the expected value of mortality, $\ln \mathrm{E}[M_{it}] \equiv \lambda_{it}$. The two models are obviously close but not identical. One key difference between the models is that, in equation 3.7, log-mortality is not defined when the realized value of mortality is zero. This need not be a problem when the expected value of the number of deaths is large enough, because in this case we will never observe zero deaths in a given year. However, when the expected value of deaths is small (say less than 10), which can happen because the cause of death is rare or the population group is small, some adjustment must be made.

A common and easy solution is to assign to each cross section an extra 0.5 deaths every year before taking the log (Plackett, 1981, p. 5). The constant added should not be arbitrary, because one can generate almost any coefficient values by tinkering with this constant (King, 1988). Plackett's justification for 0.5 comes from his approximation of the log of expected mortality by the expectation of the log of the number of deaths plus a constant. He shows that the approximation is optimized when the constant is 0.5. This issue is strictly related to the fact that the log-normal approximation to the Poisson distribution is not appropriate when the expected value of deaths is small, and so a somewhat better approach, short of moving to a full event count model, would be to replace the zeros with a number obtained by some imputation technique. We analyze this issue formally in section 6.5 (page 112). For simplicity, we assume in this chapter that the data happen to contain no observed zeros or that adding 0.5 deaths is a satisfactory solution.

Under these assumptions, the likelihood function for the model in equation 3.7 is simply:

$$\mathcal{P}(m \mid \boldsymbol{\beta}_i, \sigma_i^2) \propto \prod_i \sigma_i^{-T} \exp\left(-\frac{1}{2\sigma_i^2}\sum_t (m_{it} - \mathbf{Z}_{it}\boldsymbol{\beta}_i)^2\right). \tag{3.8}$$

A more general formulation can be obtained by allowing each observation m_{it} in the preceding equation to be weighted with some exogenous weight b_{it}. This is formally done

by replacing the specification in equation 3.7 with

$$m_{it} \sim \mathcal{N}\left(\mu_{it}, \frac{\sigma_i^2}{b_{it}}\right) \tag{3.9}$$
$$\mu_{it} = \mathbf{Z}_{it}\boldsymbol{\beta}_i.$$

The weight b_{it} may reflect prior knowledge, such as knowing that an observation in a particular year is very noisy, or knowing that the variance of log-mortality is, under certain conditions, inversely proportional to the expected value of the number of deaths. We explore some of these possibilities in section 6.5. The parameter σ_i in equation 3.9 has a different meaning from the parameter σ_i in equation 3.7.

If the approach of equation 3.9 is taken, it is convenient to make a change of variable so that the weights b_{it} disappear from the likelihood. Defining $y_{it} \equiv \sqrt{b_{it}} m_{it}$ and $\mathbf{X}_{it} \equiv \sqrt{b_{it}}\mathbf{Z}_{it}$, we rewrite the likelihood as follows:

$$\mathcal{P}(y \mid \boldsymbol{\beta}_i, \sigma_i^2) \propto \prod_i \sigma_i^{-T} \exp\left(-\frac{1}{2\sigma_i^2}\sum_t (y_{it} - \mathbf{X}_{it}\boldsymbol{\beta}_i)^2\right). \tag{3.10}$$

The maximum likelihood estimator is therefore the following computationally fast, weighted least-squares estimator:

$$\boldsymbol{\beta}_i^{\text{wls}} = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i y_i. \tag{3.11}$$

3.1.3 Computing Forecasts

Forecasts can be computed from a given model with covariates in at least three ways. These computations apply to the equation-by-equation model as well as some other models (such as the pooled model described in section 3.2). They include forecasted covariates, no covariates, or lagged covariates, which we now explain.

Forecasting Covariates The first possibility is to fit the model to the observed data, forecast the covariates using separate models for each, assume that the relationship between m and $\mathbf{Z}$ remains constant between the in-sample and out-of-sample periods, and use the forecasted values of $\mathbf{Z}$ to create fitted values for m. This approach can work well if indeed the relationship remains constant into the future and if high-quality covariate forecasts are available. In practice, the uncertainty in the covariate forecasts propagates into mortality forecasts, greatly increasing the forecast variance. In addition, a valid application of a two-stage procedure like this should ideally be accompanied by forecasts that simultaneously model the full set of covariates without zero correlation restrictions. This method is especially important for computing appropriate uncertainty estimates and also for sensitivity tests based on using ranges of plausible values for each covariate. Unfortunately, generally accepted multivariate forecasts like these are not available.

Autoregressive Models A second method by which one can use a model like this to forecast mortality is to drop all variables from $\mathbf{Z}$ except functions of lags of m and to use a standard multi-step-ahead forecasting method. For example, if $\mathbf{Z}$ includes only a lag of mortality, we forecast for year t by computing $\hat{\mu}_t = \mathbf{Z}_{i,t-1}\boldsymbol{\beta}_i$. Subsequent steps involve

using $\hat{\mu}_t$ in place of lagged mortality. This procedure can obviously be generalized to include full autoregressive integrated moving average (ARIMA) models by changing the functional form in equation 3.7. This procedure works if the model closely approximates the data generation process, but it is heavily model-dependent, and so minor deviations from the model get amplified through each step, sometimes driving forecasts far off the mark when many steps are required. Because the uncertainty in multistep forecasts amplifies especially quickly, ARIMA models tend to be used mostly for only much longer observed time series than are available for this problem. Of course, a potentially more serious problem with this procedure is that information from other covariates is ignored. In experiments with our data, we found that even when a time series was highly autocorrelated, an autoregressive model produced inferior out-of-sample forecasts due to high levels of model dependence.

Lagged Covariates A final procedure forecasts t periods ahead by using the one-step-ahead forecasting algorithm applied to data where the covariates are lagged by t periods. The disadvantage of this procedure is that either data are lost at the start of the period (where the values of $\mathbf{Z}_{it}$ for $t < 1$ are required) or earlier values of the covariates must be collected or "backcast." Backcast covariates can be created via a combination of statistical methods and expert judgment and are uncertain but will normally be less uncertain than covariate forecasts. The key disadvantage of this method is that, to forecast t periods ahead, it posits a relationship between two sets of variables lagged t periods, which may not be reasonable for some applications.

Recommendations We experimented with all three methods, but found that covariates were often informative and that expert backcasts seemed far more reliable than forecasts. Smoking rates, to take one example, can in principle go in any direction henceforth, but prior to 1950 we have considerable evidence that people smoked less, which at a minimum bounds the estimates between zero and the 1950 value. In practice, even more information is available, all of which should be used. Although any of the three methods can be used, we find that lagging covariates is the most scientific approach. Unlike autoregressive approaches, the information in the covariates is used. Moreover, although we had hoped to gather the best covariate forecasts from the appropriate experts in fields corresponding to each covariate, we found in practice no less uncertainty in forecasted covariates than in forecasts of mortality, and so they could not be used to reliably improve forecasts.

3.1.4 Summary Evaluation

We note that equation-by-equation weighted least squares forecasts well in some instances, such as with high-quality data for causes with many deaths (e.g., U.S. forecasts of cardiovascular disease among older men). But for most countries, which have shorter time series, or for younger age groups or other diseases, this method forecasts quite poorly. The time series on which it is based is normally short, and the variance estimates are large. We find that this approach typically overfits the data—resulting in a model that matches the in-sample data well but forecasts out-of-sample poorly. The result is a lot of noise: averaged over cross-sectional units, this method may even be nearly unbiased, but the variance is so large for any one as to make it unusable in most circumstances. This method also treats all cross-sectional units as if they are independent, whereas the data and observable prior knowledge strongly suggest otherwise.

This method falls at the low-bias, high-variance end of the continuum representing bias-variance trade-offs. Almost all other methods combine different cross sections in attempts to trade some bias for larger reductions in variance.

3.2 Time-Series, Cross-Sectional Pooling

Whereas equation-by-equation weighted least squares is a low-bias, high-variance approach, Murray and Lopez's (1996) pooling approach is just the opposite, producing low variance but potentially high-bias estimates. The advantages of one are thus the weaknesses of the other. Neither is optimal, but understanding the contributions of each is helpful in building better models.

3.2.1 The Model

Whereas the equation-by-equation approach pools all time periods for a single age, sex, cause, and country, the Murray-Lopez model pools all time periods and countries for a given age, sex, and cause. They use a linear-normal model, and hence:

$$\begin{aligned} m_{ict} &\sim \mathcal{N}\left(\mu_{ict}, \sigma_i^2\right) \\ \mu_{ict} &= \mathbf{Z}_{ict}\boldsymbol{\beta}_i, \end{aligned} \tag{3.12}$$

where quantities are indexed by country c ($c = 1, \ldots, C$), time period t ($t = 1, \ldots, T$), and the fixed cross-sectional unit i ($i = 1, \ldots, n$) that now stands for age, sex, and cause. Thus, the variables in any one regression vary over c and t, and the coefficients are allowed to vary only over i (age, sex, and cause).

Because this model allows each coefficient to be based on many more observations than for equation-by-equation least squares ($C \times T$ at most compared to T at most, respectively), the variance of the quantities estimated is substantially smaller. However, the model makes the implausible assumption that $\boldsymbol{\beta}_i$ is the same for every country. Murray and Lopez did not pool age groups, because their public health perspective emphasizes that different age groups die of different causes and at different rates. From a political science perspective, however, the constancy assumptions are not plausible: the model assumes the direct effect of an extra year of education is the same in the United States as it is in Tajikistan, and the effect of a given increase in GDP is the same in Benin as it is in Germany. Although we do not have the same strong prior information about the exact levels of these coefficients across countries as we have across age groups, especially because they are not necessarily causal effects, we have no good reason to think that they are the same across all national boarders or even for all countries within a region, as Murray and Lopez do. (From a political science perspective, pooling over age groups makes more sense than pooling over countries, but this practice would be biologically absurd. Researchers, it seems, are always most comfortable pooling over variables outside their own discipline.)

What is the consequence of pooling cross sections with coefficients that are not the same? At a minimum, parameters that vary more than indicated in the model lead to standard errors that are too small, but if the variation in the parameters is related to the

variables in the equation, then a direct application of the model in equation 3.12 will lead to bias. Indeed, every empirical test we run confirms that the coefficients vary considerably across countries and pooling often induces high levels of bias.

3.2.2 Postestimation Intercept Correction

Murray and Lopez (1996) were, of course, aware of the problem of pooling coefficients that were not completely the same. The technology to partially pool in an appropriate fashion did not exist, and so they needed to pool something. Pooling countries was the most obvious choice, and probably the best, given the constraints available at the time. But because they were aware of the problems, they found an approach (though somewhat unconventional) to the problem that partially corrects for the difficulties while still allowing for pooling.

Instead of adding fixed effects for countries, which after all would address only unmodeled parameter variation in the intercept, they use a technique known as *intercept correction* (Clements and Hendry, 1998), which does this and a bit more. The idea is to fit the model in equation 3.12, compute a predicted value to create a "forecast" for the last in-sample observation, calculate the error in fitting this last data point, and assume the same method will generate the same error for all subsequent out-of-sample forecasts. Because they use covariate forecasting to construct their forecasts, this method amounts to computing a k-year-ahead forecast as the usual regression-predicted value minus the last observed residual:

$$\mathbf{Z}_{ic,T+k}\hat{\boldsymbol{\beta}}_i - e_{icT},$$

where $e_{icT} = (m_{icT} - \mathbf{Z}_{icT}\hat{\boldsymbol{\beta}}_i)$ is the residual in country c and the last observed year T. This procedure may seem flawed, but because, under the model, $\mathrm{E}[e_{icT}] = 0$, subtracting e_{icT} from the usual model forecasts does not change the expected value of the forecast and so introduces no bias. However, intercept correction increases the variance of the model-based forecast (by $\mathrm{V}[e_T] = \sigma_i^2$), and so if the model is correct, the procedure will be suboptimal. This makes sense, of course, because if the series is noisy, the forecast depends on *one* data point that could be far from any underlying systematic pattern.

On the other hand, if the model is wrong due to a structural shift in the underlying parameters near the end of the in-sample period, or to large differences across countries, then intercept correction can bring the forecasts back in line by reducing bias. This method does not correct for bias due to variation in the slope parameters, but it adds some robustness to model violation in short-term forecasts (and in long-term forecasts biased by a constant-shift factor) at the cost of some efficiency. Moreover, the higher variance is not a terrible cost under this approach, given how many more observations are available in this pooled method to reduce the variance.

3.2.3 Summary Evaluation

We find that without intercept correction, out-of-sample forecast errors from the Murray-Lopez model are immense: the coefficients are noisily estimated and often average to approximately zero, and the levels are often far off. With intercept correction, the forecasts are vastly improved because at least the levels are closer to correct, and when mortality

changes slowly over time, near-zero coefficients on the covariates and a level correction is not a bad combination. However, the approach still leaves considerable room for improvement: the age profiles are not constrained to be reasonable, most dynamics are usually missed, and the forecasts are typically not close to the mark. Although intercept correction no doubt improves a pooling model with constant coefficients, the large changes it introduces make it hard to imagine a justification for a model so biased in the first place. For small to moderate changes in the forecasts, intercept correction is sometimes a reasonable and practical procedure. But when changes due to intercept correction are large and frequent, as for log-mortality, researchers should probably pursue a different approach when available.

3.3 Partially Pooling Cross Sections via Disturbance Correlations

One (non-Bayesian) way to allow cross sections to borrow strength partially, without pooling, is to postulate a correlation among the set of disturbances for the separate cross-sectional regressions—as in seemingly unrelated regression models (SURM) (Zellner, 1962). Because the correlation in these models is among scalars (the disturbances), SURM can be made to work even if different covariates are available for different cross sections, which addresses one of the problems of pooled cross-sectional time series.

SURM, however, was not intended to, and does not, resolve other important difficulties. If the explanatory variables are the same for each cross section, then SURM reduces to equation-by-equation least squares. In this situation, the technique offers no efficiency advantage. When explanatory variables are highly correlated, even if not the same, SURM gives results that are very similar to least squares. This outcome suggests that a stronger notion of smoothing is required for our problem, and indeed for most applied time-series, cross-sectional problems.

Finally, to make SURM work, one needs knowledge of the disturbances and their correlations with the covariates, information that, if available, should be used. In dealing with complex multivariate problems with numerous cross sections, however, it is not clear how confident we can be about knowledge regarding the behavior of these unobservable quantities. For SURM to work well, we must be in the odd situation where the things we *know* about the cross sections (the covariates) are minimally correlated, but at the same time the things we *do not know* (the disturbances) are maximally correlated. This situation may occur sometimes, such as when different uncorrelated covariates are available for different cross sections, and when powerful covariates common to all the cross sections can be clearly identified but not measured (a situation that may also lead to omitted variable bias). However, the SURM framework clearly does not provide the kind of assumptions researchers would want to rely on to build a general method of borrowing strength between related linear regressions.

More generally, using SURM just because it is a convenient mathematical framework with which to introduce correlations among the cross sections does not seem wise without a good substantive reason. In practice, empirical results using SURM often do not differ

much from least squares. Instead, our guiding principle is that we should use the prior knowledge we truly have, rather than knowledge it would be convenient to have but we do not possess. In the next part, we follow this principle and show how certain types of prior knowledge are easily incorporated in the analysis of cross-sectional time-series models using a Bayesian framework.

3.4 Cause-Specific Methods with Microlevel Information

We now discuss several approaches to forecasting mortality that include external sources of information in different ways. All share the feature of decomposing mortality into separate components, each of which is forecast separately to arrive at a mortality forecast.

3.4.1 Direct Decomposition Methods

We illustrate direct decomposition methods and their assumptions by summarizing the PIAMOD (prevalence, incidence, analysis model) approach of forecasting cancer mortality rates (Verdecchia, De Angelis, and Capocaccia, 2002). This method uses cancer registry data that summarize the histories of individual patients to forecast cancer population mortality rates. Cancer registries are available in subsets of highly developed countries but not in most other parts of the world, and so the technique is not as widely applicable as others.

Modeling

PIAMOD mortality forecasts are based on a deterministic relationship that decomposes mortality at one time into the prevalence and incidence of, and relative survival from, cancer, along with the death rate. Denote the fraction of people who die in age group a from cancer as M_a, and decompose it as

$$M_a = \sum_{a'=0}^{a} (1 - v_{a'}) \pi_{a'} \tau_{aa'} \theta_{aa'}, \tag{3.13}$$

where the sum is over all cohorts from birth up to age a, and where $(1 - v_{a'})$ is the fraction of healthy individuals at time a' (one minus the prevalence), $\pi_{a'}$ is the incidence rate (the probability of contracting cancer between ages a' and $a' + 1$), $\tau_{aa'}$ is the relative survival probability at age a for a diagnosis of cancer at age a', and $\theta_{aa'}$ is the crude death rate from cancer from age a' to $a' + 1$ among those diagnosed with the disease.

Then each of the component parts is itself forecast, and plugged into equation 3.13 to produce a forecast of M_a. The healthy population is estimated by using equation 3.13, setting $\theta_{aa'} = 1$, and summed up to year $a - 1$. Incidence and relative survival are then estimated to give an estimate of the healthy population as well as for use directly in equation 3.13.

Incidence $\pi_{a'}$ is estimated as a logistic or exponential regression, as polynomial functions of age and cohort, respectively:

$$g(\pi_{at}) = \alpha + \sum_{j=1}^{j_a} \eta_j a^j + \sum_{j=1}^{j_c} \omega_j (t-a)^j, \tag{3.14}$$

where $g(\cdot)$ is the log or logit link, α, $\eta_1, \ldots, \eta_{j_c}$, and $\omega_1, \ldots, \omega_{j_c}$ are estimable parameters. Verdecchia, De Angelis, and Capocaccia (2002) decide on the number of polynomial terms, j_a and j_c, via a stepwise procedure that involves a function of likelihood ratios. They report an application where the polynomial for age is of degree eight and for cohort is degree two, which seems to them to be both reasonable and consistent with previous applications. To forecast incidence, the age and cohort effects are assumed to be constant into the future. Sometimes the cohort linear term (which, of course, is indistinguishable from a period effect) is let to drift (and in which case the logit instead of log link is used to avoid exponential growth).

The relative survival probability, $\tau_{aa'}$, is estimated by a mixture model for age-sex-period stratum i for the time since diagnosis d:

$$\tau_a(d) = \alpha_i + (1-\alpha_i)\exp\left(-(\lambda_i d)^{\gamma_i}\right), \tag{3.15}$$

with the weight α_i modeled as a logistic function of time in order to help control for right censoring:

$$\alpha_i(t) = \frac{1}{1+\pi_0 \exp(\pi t)}. \tag{3.16}$$

Survival is forecast by merely choosing different scenarios and computing the mortality forecast for each. The pessimistic scenario of Verdecchia, De Angelis, and Capocaccia (2002) assumes that survival improvements do not occur. Their optimistic scenario assumes that survival rates continue to improve exactly as in previous years. They do not consider a scenario where improvements would occur at an increasing rate.

Finally, PIAMOD models the crude death rate due to cancer, $\theta_{aa'}$, as

$$\theta_{aa'} = \left(1 - \frac{\tau_{aa'}}{\tau_{a,a'+1}}\right)(1-q_a^*),$$

where q_a^* is the probability of death from competing causes (i.e., other than cancer) at age a for a birth cohort surviving to age a. Because q_a^* is normally unobserved, the population crude death rate q_a is often substituted instead.

3.4.2 Microsimulation Methods

Another approach to forecasting based on external, exogenous sources of information is microsimulation. The idea here is to develop a computational (as distinct from formal or statistical) model of the life course of people with or without particular illnesses and to cull information from the academic literature to set parameter values.

The canonical microsimulation approach to disease-specific mortality forecasting is Weinstein et al.'s (1987; see also Salomon et al., 2002) model of death from (and incidence and cost of) coronary heart disease. The idea is to set up a flow chart of life paths, including onset of coronary heart disease, paths through treatment, recurrence, and therapeutic interventions, and death outcomes. The boxes with these (and other) states, and arrows that represent transitions among them, are articulated with as much detail as information warrants. The key fact about this approach is that no estimation is carried out by the authors. Instead, parameter values, such as the transition probabilities, are gleaned from prior academic studies.

Although Weinstein's specific model has been used for forecasting and compared to actual out-of-sample mortality data, this is not the typical approach taken in this literature. Most of these models are capable of producing forecasts, conditional on the assumptions, but they are not really empirical forecasting models and have not been validated on out-of-sample data. Instead, they are closer to formal models, informed by the empirical literature, or in some sense even abstract representations of literature reviews, and demonstrations of what the literature, taken as a whole, says about a particular cause of death.

3.4.3 Interpretation

The PIAMOD and microsimulation approaches include many reasonable modeling decisions. Each equation in PIAMOD or box and arrow in a microsimulation approach may lead to or summarize insights about a component of mortality. The approaches include additional information, and they use data closer to individual human beings, about which we have more real biological knowledge than we do about population aggregates.

Counterbalancing these advantages are the detailed data requirements. In part because of these substantial data demands, the approaches have not been widely used in practical forecasting applications with multiple causes of death. In addition, the detailed individual-level modeling means that more modeling decisions need to be made, so the result is a fairly large number of decisions about which there is little prior knowledge. This outcome can be seen in PIAMOD, from the high-dimensional polynomial modeling of incidence, to the mixture model for relative survival, to the assumptions of constant age profiles over time. Mortality forecasts are necessarily highly dependent on many of these modeling decisions.

Similarly, many ways of laying out the boxes and arrows of the microsimulation approach can be seen as equally plausible, and there exist few ways of validating parts of the model. Although this approach remains a creative way to summarize vast tracks of literature, it is difficult to use to produce reliable forecasts for many causes of death.

3.5 Summary

The methods of forecasting mortality described in this chapter fall on a dimension from low bias, high variance, for equation-by-equation analyses, to higher bias and lower variance, such as the Murray-Lopez forecasts. Both of these methods include covariates, which are

intended to code some of what we know about the patterns of mortality from numerous biological and aggregate empirical studies. Parametric curve-fitting also falls at the higher-bias, lower-variance end of the continuum, but it excludes knowledge we may have in the form of measured covariates (although it would not be difficult to add covariates to the forecasts of the parameters in equation 2.3.)

Better methods of forecasting will generally come from more information, and so the methods we develop in subsequent chapters will all allow the use of covariates, when available. However, the key intuition provided by each method in prior literature reviewed here will also be retained. That is, we begin with equation-by-equation analyses and, as with pooling approaches, we give up some bias for larger reductions in variance. We also seek to improve forecasts by finding and incorporating additional information wherever possible.