1 Qualitative Overview

1.1 Introduction

This book introduces a set of methodological techniques that borrows from and is designed to build on parts of the scholarly disciplines of demography, statistics, political science, macroepidemiology, public health, actuarial science, regularization theory, spatial analysis, and Bayesian hierarchical modeling. Our approach would also seem applicable in a variety of substantive research applications in these and other disciplines. In this chapter, we describe the purpose of this book in four ways, the last three being consequences of the first.

The narrowest view of this work is an attempt to address the goal we originally set for ourselves: to create *a class of statistical methods for forecasting population death rates* that outperforms existing alternatives—by producing forecasts that are usually closer to out-of-sample mortality figures in large-scale comparisons; by more reliably fitting well-known patterns in mortality data as they vary over age groups, geographical regions, and time; and generally by incorporating more available quantitative and qualitative information.

Mortality analyses are of widespread interest among academics, policymakers, industry researchers, and citizens worldwide. They are used for retirement fund planning, directing pharmaceutical research, and planning public health and medical research interventions. They constitute our running example and source of data for empirical validation. The World Health Organization (WHO), which has made use of earlier versions of our approach, relies on worldwide mortality forecasts to estimate morbidity (by using the relationship between death rates and the prevalence of certain diseases), to make ongoing public health recommendations to specific countries and regions, and as a source of information for the health ministries in member countries.

The class of methods we introduce is also applicable to problems other than mortality forecasting. In particular, any research that uses a method like linear regression to apply to time-series analyses in more than one cross section may benefit from our approach. Although the only data we present are about mortality, the methods would seem to be directly applicable to forecasting variables from classical demography and macroepidemiology, such as fertility rates and population totals; from economics, such as income and trade; from political science, such as electoral results; and from sociology, such as regional crime rates. Section 1.2 elaborates.

In order to accomplish our original goal of forecasting mortality rates, we found it necessary to develop a series of new tools that turn out to have wider implications. These new tools also suggest alternative ways of understanding the content of this book.

Thus, a second and broader view of this book is as *a way to specify and run a set of linear regression models that builds on existing Bayesian approaches and automates what would otherwise be time-consuming decisions*. Our key methodological result is demonstrating how to borrow strength by partially, or probabilistically, pooling separate cross sections based on similarities across only the expected values of the dependent variables instead of the traditional practice of pooling based only on similar coefficients. This result, which to our knowledge has not been attempted before in hierarchical models, also enables us to pool data from regressions that have different explanatory variables in each cross section, or the same explanatory variables with different meanings. Because in applications like ours researchers can directly observe the dependent variable but never know the coefficient values, this approach makes it easier to base prior distributions on known information rather than optimistic speculation. We also offer several new ways of building priors that better represent the types of knowledge substantive experts possess and new ways of modeling ignorance and indifference, and understanding when they are appropriate in Bayesian models. These methods incorporate and extend the key attractive features of empirical Bayes models but without having to resort to a theory of inference outside the standard Bayesian framework.

A third way of understanding this work is an implication of the second. We show that *the most common Bayesian method of partially pooling multiple coefficients in cross sections thought to be similar is often inappropriate as it frequently misrepresents prior qualitative knowledge*. The idea of pooling can be powerful, but many hierarchical and multilevel Bayesian models with covariates in the literature are flawed in this way. This claim affects several scholarly literatures in statistics and related fields, such as at least some work in hierarchical modeling, spatial smoothing, and, for relevant applications, the social and natural sciences. Recognizing this problem was what led us to smoothing on the expected value of the dependent variable rather than the coefficients. This approach helped in our goal of forecasting, but it also led to a huge reduction in the number of required (but often unintuitive) parameters users were required to set. We provide the mathematical and statistical tools to accomplish this in theory and practice.

A final way to view this book is as a small step in *reconciling the open warfare between cross-national comparativists in political science, sociology, public health, and some other fields with the area studies specialists that focus on one country or region separately from the rest of the world*. In political science, for example, the current animosity between quantitative cross-national comparativists and area studies scholars originated in the expanding geographic scope of data collection in the 1960s. As quantitative scholars sought to include more countries in their regressions, the measures they were able to find for all observations became less comparable, and those that were available (or appropriate) for fewer than the full set were excluded. Area studies scholars appropriately complain about the violence these procedures do in oversimplifying the reality they find from their in-depth (and usually qualitative) analyses of individual countries but, as quantitative comparativists continue to seek systematic comparisons, the conflict continues. By developing models that enable comparativists to include different explanatory variables, or the same variables with different meanings, in the time-series regression in each country, we hope to eliminate a small piece of the basis of this conflict. The result should permit statistical analyses and data collection strategies that are more sensitive to local context and that include more of the expertise of area studies specialists. Indeed, even if the area studies specialist in each country would prefer a unique set of explanatory variables, our methods enable a scholar to estimate all these regressions together, marshaling the efforts of many

scholars and enabling them to work together without sacrificing what they each bring to the table.

We now discuss each of these four perspectives on our work in turn, saving technical discussions for subsequent chapters.

1.2 Forecasting Mortality

Almost all countries—democracies and authoritarian regimes, rich countries and poor countries, nations in the North and those in the South—make efforts to improve the health of their populations. Though not at the same rates or with the same success, most attempt to reduce mortality and increase health. Indeed, more than 9% of the world's economy (and 15% of the U.S. economy) is devoted to health care spending. As in other areas of public policy, information about the problem can help tremendously in amelioration efforts. For this reason, WHO has regularly forecast mortality and morbidity for the entire world. These forecasts are used by WHO, other international institutions, donor countries, and the health ministries and public health bureaucracies within each country to direct the flow of funds in the most effective way possible to the population groups in most need or which can be helped the most. Mortality forecasts are also used to assess the future security of retirement and social security plans, public and private insurance schemes, and other public policies that depend on specific population and mortality counts.

In recognition of the value of this information, but also for other unrelated reasons, enormous quantities of money are spent on vital registration systems in many countries. For example, in the United States, laws in each of the 50 states require death certificates be completed for every person when he or she dies. Federal law then mandates the central compilation and publication of these data and other vital statistics. Vital registration data are also collected in many countries around the world, some with excellent data, some with insufficient coverage, and some with either nonexistent or estimated rates. Each of the 779,799,281 deaths recorded in our database was (in principle) coded from an official death certificate, but in some cases the annual mortality rates were estimated by WHO or local officials.[1]

1.2.1 The Data

Our mortality data (Girosi and King, 2006) have the following structure. For 191 countries, 24 causes of death,[2] 17 age groups, and 2 sexes, we observe an annual time series of the number of people who have died and the population in that subgroup. For our purposes, the death rate is of interest: the number of people who have died divided by the number

[1] Survey methods exist for estimating all cause and cause-specific mortality in countries without vital registration. See Gakidou and King (2006) and King and Lu (2008) and the citations therein.

[2] The 24 categories of death are called "clusters" in the international classification of diseases. These include all-causes (which is the sum of all other causes), malaria, AIDS, tuberculosis, other infectious diseases (i.e., other than malaria, AIDS, and tuberculosis), lung cancer, cancer of the mouth and esophagus, liver cancer, stomach

of people at risk of dying (the population). The time series of death rates for each of these 155,856 cross-sectional units usually ends in 2000 and is between 2 and 50 observations long. Time-series, cross-sectional data are commonly used in the social sciences. Less common are data like these that have 4 cross-classified cross sections for each time period. The methods we develop here usually help more with larger numbers of cross sections, but they can be advantageous with as few as 2.

For the time series representing a single year, sex, age, country, and cause, we also observe a subset of the following covariates (explanatory variables): gross domestic product (adjusted for purchasing power parity), tobacco consumption (in some cases based on direct information and in others inferred based on present lung cancer rates), human capital, total fertility, fat consumption, a time trend as a rough measure of technology.

Enormous effort worldwide went into producing and collecting these data, but much still remains missing. Every country in the world has at least 2 observations, but only 34 countries have at least 35 time-series observations; 17 have 20–34 observations; 33 have 2–19; and 107 have only 2 observations. Many of the countries with only 2 "observations" were imputed by WHO experts.

In figure 1.1 we provide a graphical representation of the distribution of the number of observations for all-cause mortality. The red points in the graph represent the percentage of countries for which the number of observations is larger than a given amount, read on the horizontal axis. In green we show the percentage of the world population living in those countries. For example, the figure shows that 50–60% of the world's population lives in countries whose age-specific mortality time series has more than 10 observations. The key message is that data are sparse, and thus any forecasting method that is not applicable to problems with fewer than 10 observations will fail to make forecasts of any kind for 40–50% of the world's population.

Africa, AIDS, and malaria are the areas with the most missing data. Data are usually good for member countries in the organization for Economic Co-operation and Development (OECD) but not for most of the rest of the world. If mortality is observed for any age, it is (almost) always observed for all ages. All-cause mortality (the number of people who die regardless of cause) is often observed for more countries than cause-specific mortality, although counts for some diseases are observed when all-cause is not; if any cause is observed, all causes, ages, and sexes are normally observed. We treat the covariates as fully observed, if they exist in a cross section at all, although parts are exogenously imputed based on statistical techniques in combination with expert judgment.

The noise in death rates appears to be mainly a function of the size of the cross-sectional unit, so that small countries, younger age groups, and rarer causes of death are associated with more noise. Measurement error, which is one contributor to noise, is somewhat more prevalent in smaller cross sections, but many exceptions exist; the real cause would appear to be underfunded data collection facilities in less-developed countries. Data outside the OECD also tend to be noisier. Our judgment, based on interviews and some statistical tests, is that fraud does not seem to be a major problem, but many types of data errors exist.

---

cancer, breast cancer, cervix cancer, other malignant neoplasms, infectious respiratory disease, chronic respiratory disease, cardiovascular disease, digestive disease, maternal conditions, perinatal conditions, all other diseases, transportation accidents, other unintentional injuries, suicide, homicide, and war.

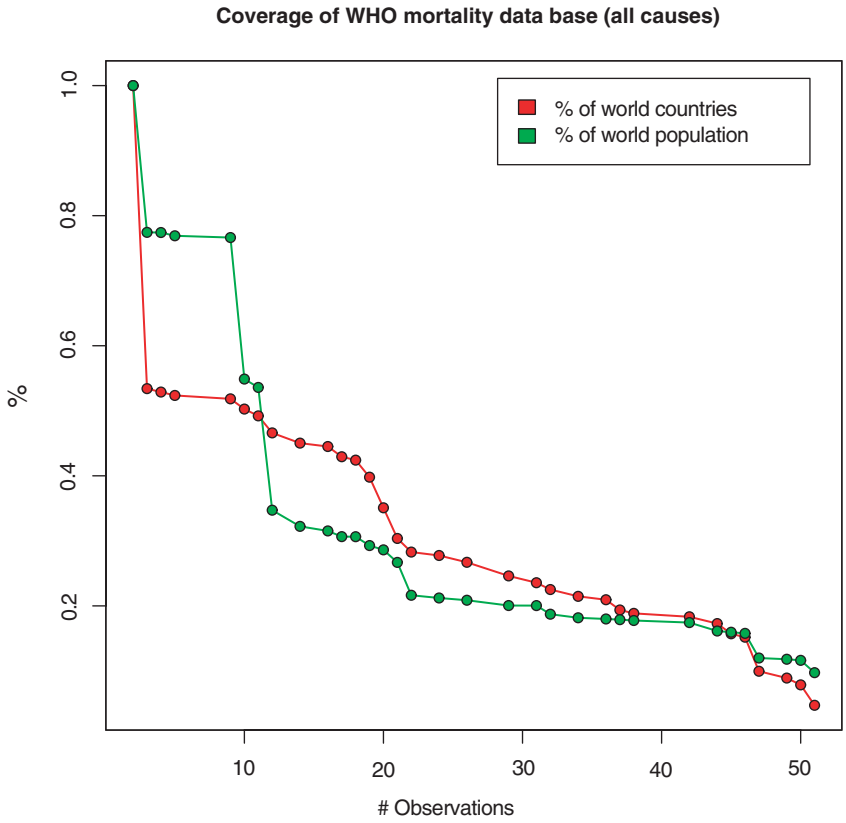

FIGURE 1.1. Distribution of number of observations for all-cause mortality: the vertical axis represents the percentage of countries (and the percentage of the world population living in those countries) whose time series have a number of observations larger than a given amount, read on the horizontal axis.

1.2.2 The Patterns

Health policymakers and public health, medical, and pharmaceutical researchers are primarily interested in cause-specific rather than total mortality. They need cause-specific information to direct appropriate treatments to population subgroups and to aid in have a chance at understanding the mechanisms that give rise to observed mortality patterns. Researchers also use relationships between cause-specific mortality and some (ultimately fatal) illnesses to estimate the prevalence of these illnesses.

Others, such as economists, sociologists, actuarial scientists, insurance companies, and public and private retirement planners, are primarily interested in total or "all-cause" mortality. The cost to a retirement plan, for example, does not change with changes in the causes of death unless the compositions of cause-specific mortality add up to different totals. However, those interested in forecasting only all-cause mortality are still well advised to examine closely, and attempt to forecast, cause-specific mortality. For example, the U.S. Social Security Administration takes into account forecasts of the leading causes of death in its all-cause forecasts, albeit in an informal, qualitative way.

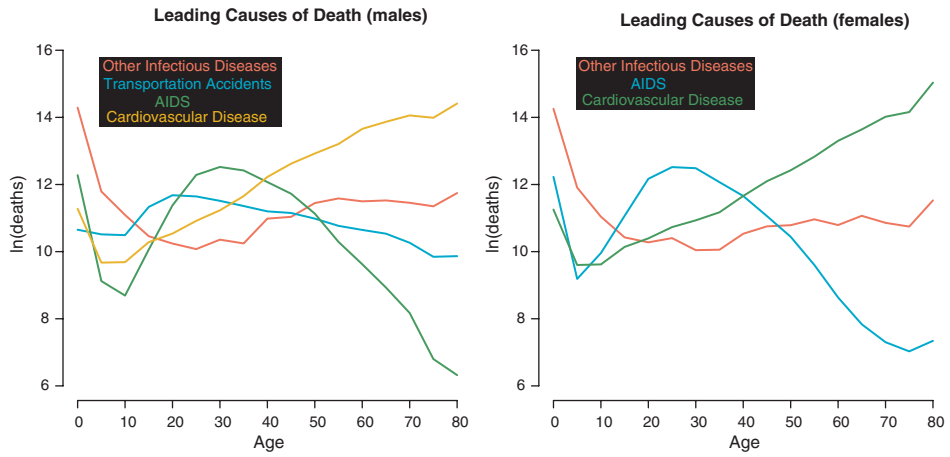

FIGURE 1.2. Leading causes of deaths worldwide by sex. Logarithm of the number of deaths worldwide by leading cause for males and females. A cause of death is included in each graph if more members of the respective sex died from it than from any other cause for at least one age group.

We offer in figure 1.2 an overview of the leading causes of death worldwide in the year 2000. In each graph, we have included each cause of death for all age groups that is the leading cause for at least one age group. The graph on the left is for males and shows that, worldwide, the leading cause of death for males aged 10 and younger is infectious diseases other than AIDS, malaria, and tuberculosis. For those older than age 10, death due to this cause declines and then increases over age groups, but it is not the leading cause of death for any other age group. From 10 to 20, the leading cause of death is transportation accidents; these increase before declining, but AIDS takes over until about age 35 as the leading cause of death among males. For older-aged men, cardiovascular disease is the leading killer. The patterns are similar for females, although the numbers are slightly lower than for males. In addition, transportation accidents never rise to the leading cause of death among females and so do not appear.

Figure 1.2 also demonstrates that different causes of death have widely divergent age profiles. Some have mortality rates that start high for infants and rise as they age. Some drop and then rise, and some follow still other patterns. However, through all this diversity, a clear pattern emerges: *the age profiles are all relatively smooth.* Whereas 5-year-olds and 80-year-olds die at very different rates, people in any pair of adjacent age groups die at similar rates from any cause.

We now illustrate the same point about the diversity and smoothness of log-mortality age profiles by offering a broader picture of different causes of death. For this figure, we change from the log of deaths to the log of the deaths per capita, otherwise known as the *log-mortality rate*. We average the age profile of the log-mortality rate over all available years and all 191 countries in our database. Figure 1.3 presents these average age profiles for the 23 causes of death in females, 1 in each plot. Figure 1.4 offers a parallel view for the 20 causes of death in males. (We exclude perinatal conditions in both because most of the deaths are in the first age group.) Age groups are 5-year increments from 0 to 75 and 80+ and are labeled by the lower bound.

These figures provide a baseline for understanding the units of log-mortality for our subsequent analyses. They also clearly demonstrate that an appropriate method of

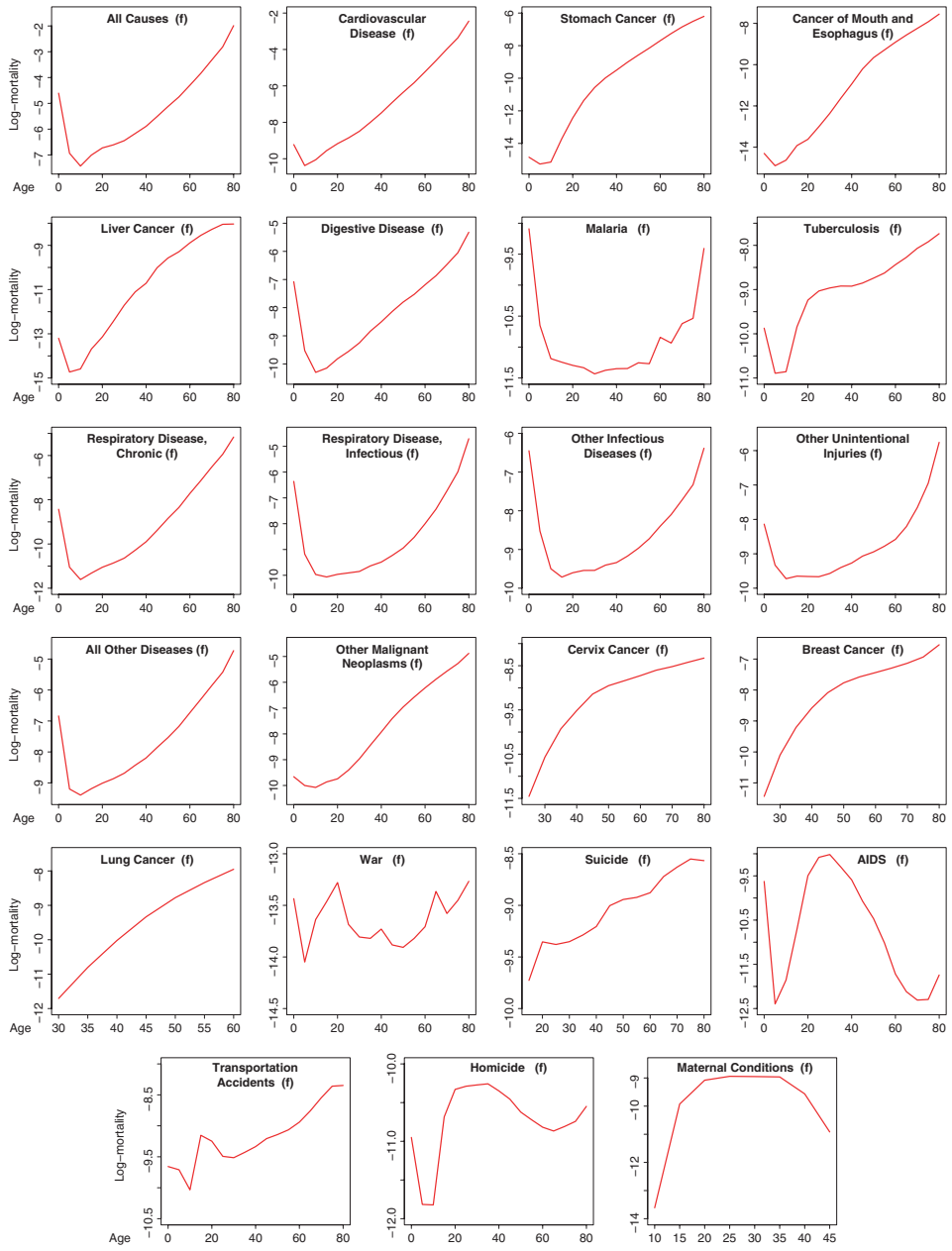

FIGURE 1.3. World age profiles for 23 causes of death in females. The age profiles have been averaged over all 191 countries and over all available years.

forecasting mortality must be capable of modeling great diversity in the age profile across diseases (and sex and country) while at the same time guaranteeing that log-mortality remains smooth over the age profile. All 43 age profile graphs are fairly smooth, with some exceptions for younger age groups, but they follow a large diversity of specific patterns.

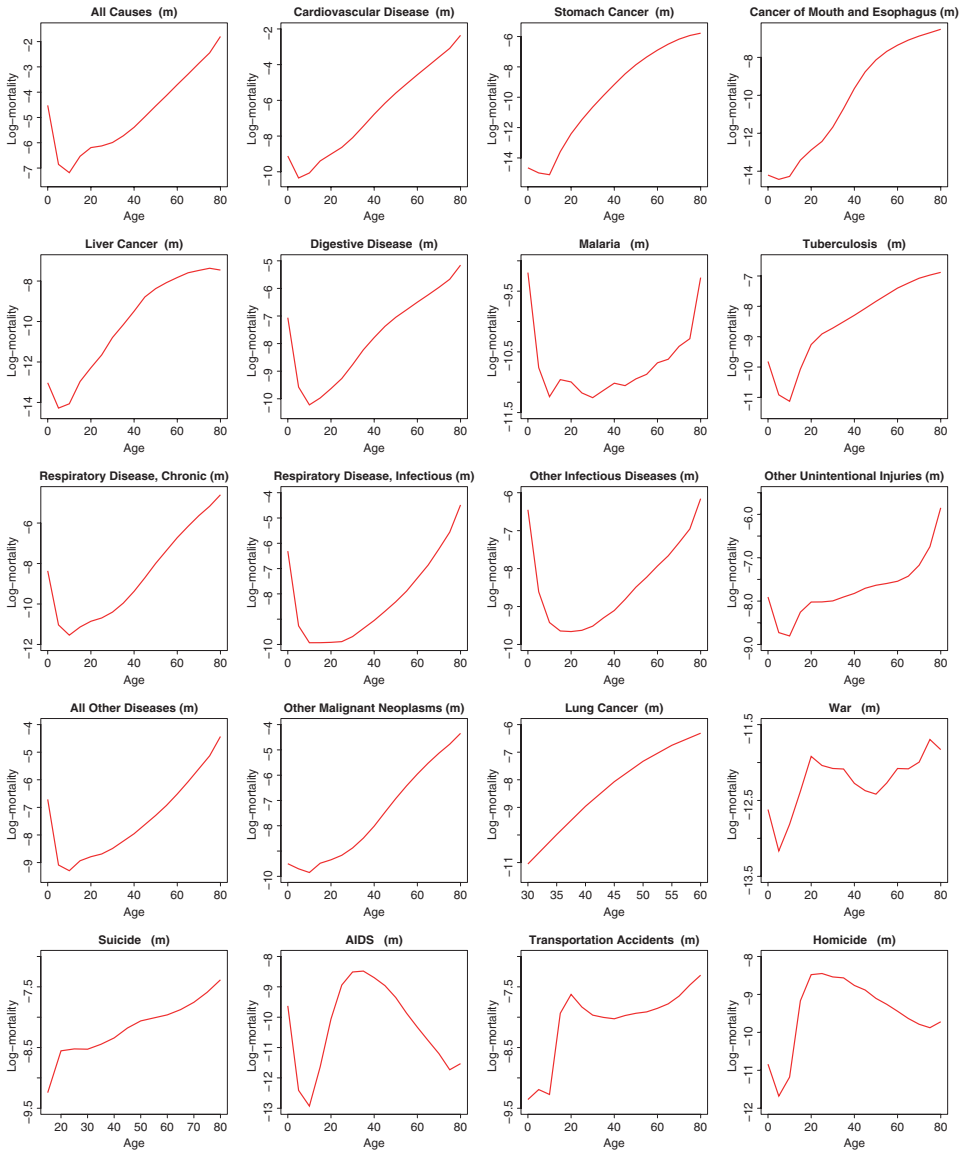

FIGURE 1.4. World age Profiles for 20 causes of death in males. The age profiles have been averaged over all 191 countries and over all available years.

1.2.3 Scientific versus Optimistic Forecasting Goals

A list of all valid methods of learning about the future developed thus far is as follows: waiting. As a methodology, waiting requires no special expertise, is easy to apply (for some!), and produces highly certain results. Unfortunately, whether researchers make forecasts or not, public policy makers will not wait because getting the future wrong will often produce enormous costs that can be denominated in dollars, deaths, and disease.

Forecasting may be impossible, and policymakers may get it wrong, but they will try. As such, and despite the uncertainties, researchers could hardly have more motivation to make forecasts better and to try to formalize and improve what is often informal qualitative guesswork. Indeed, almost any advance in learning about future mortality rates has the potential to inform and rationalize health care spending, to mobilize resources to improve care where it is needed, and ultimately to reduce mortality.

Because modern scientific forecasting is not soothsaying, we need to be clear about what it can and cannot accomplish, even at best. So here is our definition:

> Forecasting is the (1) systematic distillation and summary of relevant information about the past and present that (2) by some specific assumption may have something to do with the future.

The first point is the forecasters' only real accomplishable scientific goal, while the second is pure assumption that is by definition beyond the range of, and not fully defensible on the basis of, the data. Even genuine repeated out-of-sample forecasts that turn out close to the observed truth have no necessary relationship to subsequent forecasting success.

Put differently, *science* is about collecting the facts and the deeper understanding that explains the facts or what the facts imply. In contrast, *forecasting* involves more than science. It also involves unverifiable assumptions, justified by argument, analogy, the success of previous efforts, and reasoned speculation that by definition goes beyond direct evidence. The best forecasts you can make, then, are those that provide the best systematic summaries of all that is known and also your personal assumptions (which is why we call our software "YourCast"). The purpose of this book is to provide methods that make it easier to include more existing information, in as many forms as it comes, in your forecasting model so that we can better summarize the present and quantify and understand your assumptions about the future.

Our public policy goal is to provide forecasting methods for the public health establishment that can incorporate more information than those now used and that are more systematic. The death rate forecasts of interest are those up to 30 years in the future for each time series, with the needed forecast horizon often differing by cause of death. In most applications, time-series analysts would not normally be willing to make forecasts that predict 20 to 30 years into the future on the basis of even 50 observations, and we often have fewer. The difference here is that death rates usually move slowly and smoothly over time so that forecasting with low error rates at least a few years out is usually easy. Moreover, when considered as a set, and combining it with qualitative knowledge that experts in the field have gathered from analyses of numerous similar data sets, the information content available for making forecasts is considerably greater than when treating each time series separately and without qualitative information.

But the uncertainties should not be overlooked, and valid 30-year forecasts are not likely to be accurate with any known degree of certainty. New infectious diseases can spring up, mutate, and kill millions, as occurred with the HIV pandemic and the 1918 influenza, or affect few, as in the severe acute respiratory syndrome (SARS) outbreak in 2003. Unexpected developments in medical technology can drastically reduce the effects of a disease, as occurred with HIV treatments in the developed countries in the past decade or with the earlier eradication of smallpox and the reductions in polio and shistosomiasis throughout the world. Military conflict, terrorism, and natural disasters have demonstrated the potential to wreak havoc without advance notice. Changes can even occur without a cause that we can unambiguously identify, as with the dramatic increase in cardiovascular disease in Eastern Europe in the late 1980s. Physicians cannot always produce accurate

time-of-death predictions for individuals with serious illnesses, and no law of nature guarantees that errors will always cancel out when individuals are aggregated. As Lee and Carter (1992, p. 675) write, almost tongue in cheek, "Perhaps we should add the disclaimer that our [confidence] intervals do not reflect the possibilities of nuclear war or global environmental catastrophe, for indeed they do not." In fact, model-based confidence intervals can never be accurately calibrated because the model can never be counted on to reliably include all relevant future events. Forecasts, which are by definition extrapolations outside the range of available data, are necessarily model-dependent to a degree (King and Zeng, 2006, 2007).

These qualifications put all forecasting methods in context. In practice, mortality forecasts are always at best conditional on the state of knowledge we have today. In this book, we go farther than existing methods primarily by incorporating a larger fraction of (known) information, but we obviously cannot include information that is unknown at the time of the forecasts. Thus, to make the same point in yet another way, the forecasts produced by the methods in this book, and all other methods in the literature, can hope to tell us what will happen in the future only if current conditions and patterns hold. If future mortality is driven by factors we have not included, or if the relationship between the factors we include and mortality changes in unexpected ways, our forecasts will be wrong. Moreover, because conditions do change, and will likely change, we do not expect forecasts made today to be accurate. Indeed, much of the stated purpose of policymakers and funding agencies in the health arena is to change these conditions to reduce future mortality rates. To put it even more starkly, a *central* goal of the global public health, medical, research, governmental, and intergovernmental infrastructures is to make our forecasts wrong. We hope they succeed.

The fact that even we expect that our forecasts will be proved incorrect does not mean that our methods and forecasts should be considered free from scientific evaluation. Far from it. The only issue is finding the right standard. True out-of-sample forecasts are infeasible because it would take several years for data to come in and mortality forecasts are needed immediately and continuously. And even if we waited, had our forecasts confirmed, and decided to use those forecasts for policymaking, nothing could stop the world from changing at that point to make our subsequent forecasts inaccurate. Because policymakers are effectively trying to do this anyway, actual out-of-sample forecasts, which we use frequently, are no panacea.

The optimal scientific standard for us would be to make actual out-of-sample forecasts in an (infeasible and unethical) experiment where some country is held to have the same standards, conditions, types of medical care, and the like in a future period as during the period from which our data come. But although this hypothetical experiment would validate the model, it is obviously of no practical use. To approximate this situation, we set aside some number of years of data (the "test set" or "out-of-sample period"), fit the model in question to the rest of the data (the "training set" or "in-sample period"), and then compare out-of-sample forecasts to the data set aside. This standard makes us somewhat less vulnerable to being proved wrong than we would with real out-of-sample forecasts, which is a perhaps necessary flaw in our approach. Nevertheless, the enormous number of forecasts we need to compute would make any attempt to stack the deck extraordinarily difficult, even if we had set out to do so intentionally. Although we have run hundreds of thousands of such out-of-sample comparisons between our approach and previous methods offered in the literature, and although earlier drafts of this manuscript recorded summaries

of many of these runs, we exclude most here (we summarize the rest in chapter 12). The reason is that our method does better because it includes more information and so these tests merely demonstrate the predictive value of available information. Comparisons are either an unfair fight—because we use more information than relevant competitors—or merely an evaluation of the quality of available information, neither of which tells us much about our methods. In addition, as explained previously, forecasts always require choices about assumptions regarding the future, and we wished to evaluate our methods, not particular assumptions about the future, which we will leave to users of our methods. Thus, with this book we are releasing our full original data set and our software program that makes it easy for readers to choose their own assumptions, to make their own forecasts, and, if they wish, to make their own assessments of their forecasting accuracy when using our methods with their assumptions.

1.3 Statistical Modeling

We offer here a qualitative overview of the statistical modeling issues we tackle in this book. We discuss this modeling in the context of our mortality example, although everything in this section would also apply to any outcome variable that a researcher would be willing to specify as a linear function of covariates. The methods introduced are an improvement over linear regression (and seemingly unrelated linear regressions and other related procedures) for any analysis with more than one regression over at least some of the same time periods. They can be extended to nonlinear models with the same priors we introduce, but we have not done so here.

In our application, the more than 150,000 individual time series created substantial difficulties in data management, computational speed, and software architecture. To put the problem in context, if we spent as much as a single minute of computational and human time on each forecast, we would need more than three months (working 24 hours a day, 7 days a week) to complete a single run through all the data using only one of the many possible specifications. We thus quickly found that if a feature of a method was not automated (i.e., if it required human judgment to be applied to the time series in each cross section), it became almost impossible to use. An important goal of the project thus became automating as much as possible. Fortunately, this goal is entirely consistent with, and indeed almost equivalent to, the goals of good Bayesian modeling, because automation is merely another way of saying that we need to include as much available a priori information in the model as possible. So this has indeed been our goal. Throughout our work, we have unearthed and then incorporated in our models successive types of prior knowledge, which is what accounts for most of our forecasting improvements.

Demographers, in contrast, often use the vast majority of their prior knowledge as a way to evaluate rather than to improve their methods. Although this procedure is an essential part of most research, including ours, it is not optimal from a scientific perspective because the method is not formalized in a way others can apply and so is less vulnerable to being proved wrong. At its worst, their procedure merely adjusts a model specification until it produces forecasts consistent with one's prior opinion—in which case the forecast is nothing more than an expert judgment and the statistical model is reduced to an irrelevant scientific-looking decoration. In our view, the future of demographic modeling and

forecasting lies primarily in identifying and conditioning on more and more information in statistical models. Classical demographers have gleaned enormous information from decades of careful data analyses; the key now is to marshal this for making inferences.

To explain our alternative approach, we now begin with a single least-squares regression as a basic building block, and then explain how we add in other sources of information. Consider a long annual series of the log-mortality rate for a single cross-sectional unit (one age × sex × cause × country, such as log-mortality in 45-year-old South African males who die of cardiovascular disease), and a set of covariates (explanatory variables) that code its systematic causes or predictors. If this cross section has few missing data or measurement errors, informative covariates, and a high signal-to-noise ratio, then least-squares regression might do reasonably well. Unfortunately, for the vast majority of the time series in our data, measurement error, missing data, unexpected changes in the data, and unusual patterns conspire to make a simple least-squares, time-series regression model produce incorrect or even absurd forecasts. Normally, the problem is not bias but enormous variance. This problem becomes clear when one thinks of the fanciful goal of forecasting thirty years ahead on a single time series with say twenty observations or with more observations but high rates of measurement error, etc.

Thus, our key goal in analyzing these data is to identify and find ways of incorporating additional information in order to increase efficiency. The most common way to do this would be to pool. For example, we could pool all countries within a region, effectively assuming that the coefficients on the covariates for all those countries are identical. This method would increase efficiency, but—if, in fact, the coefficients varied—it could then introduce bias.[3] Strict pooling in this way is also contrary to our prior information, which does not normally indicate that coefficients from neighboring countries would be identical. Instead of having to decide by hand which countries (or cross sections) to pool, which is infeasible, we could think about automating this decision by relaxing the requirements for pooling. Instead of requiring neighboring countries to have identical coefficients, we could allow them to have similar coefficients. This kind of "partial pooling," or "smoothing," or "borrowing strength" from neighboring cross-sectional units to improve the estimation (and efficiency) of each one, is common in modern Bayesian analysis, and we were therefore able to build on a considerable body of prior work (which we properly cite in subsequent chapters).

Partially pooling countries by assuming similarity of coefficients in neighboring countries is an improvement and serves to automate some aspects of the analysis. (To better reflect knowledge of public health, we allowed "neighboring" to refer to similar countries, not always strictly adhering to geographic contiguity.) We extend this logic to combine partial pooling of neighboring countries and consecutive time periods, simultaneously with partial pooling of adjacent age groups. The fact that 5-year-olds and 80-year-olds die of different causes and at very different rates would normally prevent pooling these groups. However, we also know that 5-year-olds and 10-year-olds die at similar rates, as do 10-year-olds and 15-year-olds, and 15-year-olds and 20-year-olds. Thus, we simultaneously pool over neighboring countries, adjacent age groups, and time (and we allow smoothing of interactions, such as trends in neighboring age groups), to result in a

[3] Assuming a regression coefficient is constant when it in fact varies can cause one to underestimate standard errors and confidence intervals. If, in addition, the actual coefficients assumed to be constant are correlated with one of the measured variables, then the estimated coefficient will be a biased estimate of the average of the true coefficients.

form of multidimensional, nonspatial smoothing. This step also provides a more powerful approach to reducing the dimensionality of mortality data than the 180-year tradition of parametric modeling in classical demography (Gompertz, 1825; Keyfitz, 1982).

Each of these steps generated an improvement in efficiency, fit, and forecasting stability. Yet, it was still very difficult to use with so many forecasts. We had to spend inordinate amounts of time tuning the priors for different data sets, and finding what seemed like the "right" parameters was often impossible. Eventually, we realized that the problem was with the most fundamental assumption we were making—that the coefficients in neighboring units were similar. In fact, the scale of most of the coefficients in our application is not particularly meaningful, and so although our model had been saying they should be similar, and large literatures on Bayesian hierarchical modeling and spatial modeling smooth in this way, experts in public health and demography do not really have prior beliefs about most of these coefficients. For one, most of these coefficients are not causal effects and so are not even of interest to researchers or the subject of any serious analysis. They are at best partial or direct effects—for example, the effect of gross domestic product on mortality, after controlling for some of the consequences of GDP, such as human capital and total fertility. This coefficient is not the subject of study, it is not directly observable, and, because some of the control variables are both causally prior and causally consequent, it has no natural causal interpretation. Moreover, even for variables about which these scholars possess considerable knowledge, such as tobacco consumption, the real biological knowledge is at the individual level, not at the aggregated national level. Because of aggregation, finding a negative coefficient on tobacco in predicting mortality would neither contradict the obvious individual-level relationship nor be particularly surprising. Finally, even when some knowledge about the coefficients exists, the prior must include a normalization factor that translates the effect of tobacco consumption, say, into the effect of GDP. Unfortunately, no information exists in the data to estimate this normalization, and prior knowledge about it almost never exists.

Our methods allow for partially pooling coefficients also, for the situations for which it may be useful. But we also added a new feature that alleviates many of the remaining problems in our application and that we believe will work in some others. Thus, instead of (partially) pooling coefficients, about which we had little real prior knowledge, we developed a way to partially pool based on expected mortality. Scholars have been studying mortality rates for almost two centuries and know a great deal about the subject. Although we do not observe expected mortality, for which our priors are constructed, every observed mortality rate is a direct and usually fairly good estimate of expected mortality. Priors formulated in this way correspond closely to the nature of the variables we are studying. This procedure also makes it substantially easier to elicit information from subject matter experts. It also directly satisfies the goal of Bayesian analysis by incorporating more prior information appropriately. And it serves our goal of automating the analysis, because far less cross-section-specific tuning is required (and many fewer hyperparameter values need be set) when using this formulation.

Because the mapping from the vector of expected mortality rates, on the scale of our prior, to the coefficients, on the scale of estimation, is a many-to-few transformation, it may seem less than obvious how to accomplish this. We have, however, developed a relatively straightforward procedure to do this (that we describe in chapter 4). An especially interesting and useful result is that, because the variance matrix and its hyperparameters drop out, the prior turns out to require many fewer adjustable parameters than partially pooling coefficients, most of which can be set on the basis of known demographic

information about known mortality rather than on unknown coefficients. The resulting analysis generates the metric space necessary even to compare regressions with entirely different covariate specifications, including different numbers of covariates and different meanings of the covariates included in each cross section.

We also found that our method of putting priors on the expected outcome variable, rather than on the coefficients, turned out to solve a different problem that affects many time-series, cross-sectional data collection efforts. The issue in these efforts is that interesting variables are available in some countries or cross-sectional units but not in others. The problem is that existing methods require the same variables to be available for all the units. This is easy to see when trying to pool coefficients, because omitting a variable from the time series in one cross-sectional unit will make all the coefficients take on a different meaning, and so they become impossible to pool either partially or completely (at least without adding untenable assumptions). The result is that in order to use existing methods, scholars routinely make what would otherwise seem like bizarre data analysis decisions. The uncomfortable choice is normally one among:

1. omitting any variables not observed for all units, which risks attributing differences to biases from omitted variables;
2. excluding cross-sectional units for which some variables are not available, which risks selection bias; or
3. running each least-squares analysis separately, equation by equation, which risks large inefficiencies.

Researchers have, of course, been infinitely creative in choosing ad hoc data analysis strategies to limit the effects of these problems, but the lack of a better method clearly hinders research in numerous fields.

Our method of pooling on expected mortality avoids this choice by allowing researchers to estimate whatever regression in each cross-sectional unit they desire and to borrow strength statistically from all similar or neighboring units, even with different specifications (i.e., different covariates), by smoothing the expected value of the dependent variable instead of each of the coefficients. Borrowing strength statistically in this way greatly increases the power of the analyses compared to simple equation-by-equation regressions. Making choices about priors is also much simpler. As a result, with these methods, scholars can collect and use whatever data are most appropriate in each country or cross-sectional unit, so long as they have a dependent variable with the same meaning across countries. The data to which our methods are most useful have many cross sections and a relatively short time series in each.[4]

Another situation where smoothing the expected outcome variable can be useful is when the same explanatory variables are in the specification for each cross-sectional unit, but they are measured differently in each. For example, measures of the gross national product, and other economic variables, are denominated in the currency of each country. The methods we provide to smooth the expected outcome variable enable scholars to use variables in whatever denominations are most meaningful in each country.

What follows in this book contains much technical material; however when viewed from this perspective, the result is simply a better way of running least-squares regressions.

[4] A different approach to this problem might be to impute multiply entire variables when they are missing in a cross-sectional unit, but this would require further methodological work because methods to do this have been developed primarily for independent cross sections such as survey data (Gelman, King, and Liu, 1999).

The logic is along the lines of seemingly unrelated regressions—such that if you have several regressions to run that are related in some way, estimating them jointly results in more efficient estimation than separate equation-by-equation analyses. Of course, the result is very different because, for example, even identical explanatory variables produce more efficient results and the precise information assumed and modeled is very different. The technology to produce our estimates may seem complicated, but the logic and the end result are quite simple. Indeed, the end result is still a set of regression coefficients, predicted values, and any other quantities of interest that normally come from a linear regression. The only difference is that the new estimates can include considerably more information and will have generally superior statistical properties to the old ones. In particular, they have lower mean-square error and can cope far better with measurement error, short time series, noisy data, and model dependence. In our application, this added efficiency also produces much better out-of-sample forecasts and more-accurate regression coefficient estimates.

1.4 Implications for the Bayesian Modeling Literature

Our work has an implication for the Bayesian modeling literature and applies to many hierarchical or multilevel models with clusters of exchangeable units, or spatial models imposing smoothness across neighboring areas, with multiple covariates. The implication when applied to applications like ours is that *many of the prior densities commonly put on coefficients to represent qualitative knowledge about clustering or smoothness are misspecified*. We summarize our argument here and elaborate it in subsequent chapters.

As described in the previous section, some coefficients on explanatory variables are causal effects about which we might have some real prior knowledge. However, most variables included in regression-type analyses are controls (i.e., "confounders," "antecedent covariates," or "pretreatment variables"), and the coefficients on these controls are usually viewed as ancillary. The problem is that the claim that researchers have prior knowledge about coefficients that have never been observed directly and most consider a nuisance is dubious in some contexts (although not all, as we explain in section 4.3). These coefficients may have even been estimated, but they have rarely been the subject of study, and so in many situations, such as our application, there exists little sound scientific basis for claiming that we possess any substantial degree of prior knowledge about them.

But, one may ask, don't we often have strong knowledge that neighboring cross sections are similar? Of course, but the issue is what "similar" means in the usual qualitative judgment about prior knowledge, and how we should go about formalizing such similarities. If we are predicting mortality, we might imagine that New Hampshire and neighboring Vermont have similar cause-specific mortality rates, but that does not necessarily imply that the coefficients on the variables that predict mortality in these two American states are similar. In fact, if the covariates differ between the two states, then the *only* way mortality can be similar is if the coefficients are different. As such, imposing a prior that the coefficients are similar in this situation would directly violate our qualitative knowledge.

Put differently, what many researchers in many applications appear to mean when they say that "neighboring states are similar" is that the dependent variable (or the expected

value of the dependent variable) takes on similar values across these states—not that the coefficients are necessarily similar. Moreover, similarity in one does not necessarily imply similarity in the other. In this area, like so many in mathematics and statistics, formalizing qualitative intuitions and knowledge with some precision often leads to counterintuitive conclusions.

We develop tools that enable researchers to put a prior directly on the expected value of the dependent variable, allowing these to be smooth across neighboring areas or all shrunk together in the case of an exchangeable group within a cluster. The result is a different prior, and thus a different model, in most applications, even when we translate this prior into its implications for a prior on the coefficients. In our mathematical analyses and empirical experiments, these priors outperform priors put directly on the coefficients: they fit the data better, they forecast better, and they better reflect our qualitative knowledge. They also require fewer adjustable parameters and a natural space within which to make all relevant comparisons, even among coefficients and no matter the scale of the covariates.

1.5 Incorporating Area Studies in Cross-National Comparative Research

In a variety of disciplines, and often independent of the disciplines, area studies scholars have explored individual countries and regions as separate areas of study. These scholars have contributed a large fraction of the basic descriptive information we have about numerous countries, but the advantages of the incredible depth and detailed analyses they perform are counterbalanced by the absence of comparison with other areas. Those focusing on different countries work in parallel but without much interaction and without systematic comparison. In the middle of the past century, coincident with the rise of behavioralism and quantification in the social sciences, scholars began to analyze some of the same questions as area studies scholars by systematic quantitative country comparisons.

Although these trends also affected anthropology, sociology, public health, and other areas, we tell the story from the perspective of political science where the conflict is particularly pronounced. Political science is also among the most diverse of scholarly disciplines, and it includes scholars from all the other disciplines affected.

Political scientists began to be comparative on a global scale in the 1960s, vastly expanding the scope of their data collection efforts. Over the rest of the century, they traveled to every corner of the planet to observe, measure, and compare governments, political systems, economies, conflicts, and cultures. Venturing out by oneself to a foreign (which meant primarily non-American) land to collect data became a right of passage for graduate students in the subfield called comparative politics. Other scholars built large cross-national data sets that spanned ever increasing sets of countries. Whereas "comparative study [was once] comparative in name only" (Macridis, 1955), the comparative politics subfield and political science more generally emerged in this period as a more modern, international discipline.

As this period also marked the height of the behavioral movement in the discipline, many researchers became enthusiastic quantifiers and sought to measure concepts across as many nations as possible. Political science made important strides during this period but, in its efforts to expand comparisons across diverse cultures, researchers also created a variety

of strained quantifications and mismeasured concepts, which often led to implausible conclusions, or at least to conclusions without a demonstrable connection to empirical reality.

The reaction from traditional area studies scholars and others was fierce. The data gathered were widely recognized as sometimes reliable but rarely valid: "The question with our standard variables on literacy, urbanization, occupation, industrialization, and the like, is whether they really measure common underlying phenomena. It is pretty obvious that, across the world, they do not; and this quite aside from the reliability of the data gathering agencies" (Sartori, 1970; p. 1039). Sartori (see also Collier and Mahon, 1993) talked about "conceptual stretching" and the "traveling problem". Research practices were characterized as "indiscriminate fishing expeditions for data" (LaPalombara, 1968). Generally, relations between the two camps resembled some of the wars we have studied more than a staid scholarly debate: "no branch of political science has been in more extreme ferment than comparative politics during the last fifteen years" (LaPalombara, 1968, p. 52).

In the past three to four decades, comparative politics researchers have improved their data collection techniques. Their procedures are more professional, more replicable, and better documented, and the results are often even permanently archived (in the Inter-University Consortium for Political and Social Research, which was formed during this period by political scientists). Political scientists have also developed better theories to help guide data collection efforts, and as a result of all this work, the concepts underlying our quantifications and the measures themselves have improved. Data sets have moved from small cross-sectional snapshots to large time-series, cross-sectional collections. And methods for analyzing data like these have also become more sensitive and adapted to the problems at hand (Beck and Katz, 1995, 1996; Beck, Katz, and Tucker, 1998; Stimson, 1985; Western, 1998; Zorn, 2001).

As a result of these improvements, the respect for quantitative work among those who know the details of individual countries has improved (as has the use of standards of scientific inference in qualitative research; King, Keohane and Verba 1994), but we still have a long way to go. Indeed, the field has not resolved the key *comparative* problem of quantitative comparative politics. The problem remains in part because it may be inherently unresolvable. Political scientists want broad cross-national and comparable knowledge and simultaneously need detailed context-specific information. They want unified concepts that apply to places that are so different that the concepts very well may not be comparable. Is bartering for a goat in Ghanzi (a town in land-locked Botswana) and buying a GPS-equipped luxury yacht in Fort Lauderdale really comparable after translating pula into U.S. dollars, adjusting for purchasing power, and dividing by the implied cost of the goat? And that's money. What about ideas without natural units and without international organizations devoted to making them comparable—concepts like support for the government, partisan identification, social capital, postindustrialism, political freedom, human security, and many of the other rich concepts that political scientists study?

Quantitative scholars are, of course, painfully aware of these problems even when not explicitly working to solve them. The pain surfaces most obviously during data collection when scholars are torn between the desire to have one large comparable data set that they can stack up to run regressions on—thus needing the same variables measured over every country and time period—and the desire to respond to the area studies critique by collecting contextually and culturally sensitive measures. The issue is that contextually sensitive measures almost by definition involve collecting different variables, or the same variables with different meanings, in each country. The methods that have existed through this entire period, however, required the identical variables with the same meanings in all countries.

Area studies scholars thus seem to have at least two remaining complaints about the quantitative cross-national comparative literature: some oppose quantification in principle, and others do not like the particular types of cross-national quantifications and comparisons that have been conducted. In our view, a key methodological constraint—requiring scholars to have the same variables with the same meanings across countries—has helped to conflate these two distinct complaints in scholarly debates held over the years. The methods offered here relax this underlying methodological constraint and enable scholars to use different explanatory variables, or those with different meanings, in the time-series regressions in different countries. In that way a time-series, cross-sectional data analysis can still be done but with more meaningful measures. We hope this development will eliminate some of the basis of the second area studies complaint, making the resulting analyses more appropriate to each area studied. As importantly, methods along these lines may encourage scholars to collect data in different ways. These methods will not alleviate metaphysical objections to quantification, but if our methods enable or encourage scholars to collect more contextually sensitive, country-specific data, perhaps some of the reasons for the first complaint will also be reduced.

Thus, we can express a formalization and simplification of the preferences of area studies scholars (i.e., those who would allow some quantification) by saying that they would want to run a separate time-series regression within each country, using whatever explanatory variables are appropriate to use within that country and without the constraints imposed by having to build comparable measures across countries. Of course, they want more than this, but they surely want at least this. Probably the leading area study, the field of U.S. politics, has employed this strategy increasingly, and with increasing success, over the past several decades. The problem is that in many other countries, long time series are not available, and so the results of any such analyses would be highly inefficient, making this strategy mostly infeasible. Most quantitative cross-national comparativists would probably also like to adopt this strategy, if it were feasible, because it would enable them to incorporate the detailed information and insights of the area studies scholars. Our methods enable quantitative area studies scholars to collect whatever explanatory variables they believe are appropriate and nevertheless to enable experts from different regions to work together—by borrowing strength across the different regions to improve estimation for each one individually—without any constraint other than a dependent variable with a common meaning. This approach hardly solves all problems of cross-national comparison, but it should make it possible for those with different perspectives to work together more productively.

1.6 Summary

This book can be viewed as a set of methods for forecasting mortality, a new approach to statistical modeling, a critique and method for improving an aspect of the Bayesian modeling literature, or a step toward resolving some of the disputes between area studies experts and cross-national comparativists. The methods introduced should also have substantive applications well beyond these areas, some of which we discuss in the chapters to come.