Demographic Forecasting

Demographic Forecasting

Federico Girosi and Gary King

with contributions from Kevin Quinn and Gregory Wawro

PRINCETON UNIVERSITY PRESS

Princeton and Oxford

Published by Princeton University Press, 41 William Street, Princeton, New Jersey 08540

In the United Kingdom: Princeton University Press, 6 Oxford Street,
Woodstock, Oxfordshire OX20 1TW

Library of Congress Cataloging-in-Publication Data

Girosi, Federico.
Demographic forecasting/Gary King and Federico Girosi.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-691-13094-1 (alk. paper)—978-0-691-13095-8 (pbk.: alk. paper)
1. Mortality—Forecasting—Methodology. 2. Mortality—Statistical methods.
3. Demography. I. King, Gary, 1958-. II. Title.
HB1321.G56 2008
304.6/40112 22 2008062102

British Library Cataloging-in-Publication Data is available

This book has been composed in Times

Printed on acid-free paper. ∞

press.princeton.edu

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Figures

Tables

Preface

We introduce a framework for forecasting age-sex-country-cause-specific mortality rates that can incorporate more information, and thus has the potential to forecast better, than existing approaches. Mortality forecasts are used in a wide variety of academic fields and for global and national health policy making, medical and pharmaceutical research, and social security and retirement planning.

As it turns out, the tools we developed in pursuit of this goal also have broader statistical implications, in addition to their use for forecasting mortality or other variables with similar statistical properties. First, our methods make it possible to include different explanatory variables in a time-series regression for each crosssection, while still borrowing strength from one regression to improve the estimation of all. Second, we show that many existing Bayesian (hierarchical and spatial) models with explanatory variables use prior densities that incorrectly formalize prior knowledge. Many demographers and public health researchers have fortuitously avoided this problem so prevalent in other fields by using prior knowledge only as an ex post check on empirical results, but this approach excludes considerable information from their models. We show how to incorporate this demographic knowledge into a model in a statistically appropriate way that also turns out to have the advantage of requiring many fewer adjustable parameters than classic Bayesian approaches. Finally, we develop a set of tools useful for developing models with Bayesian priors in the presence of partial prior ignorance. This approach also provides many of the attractive features claimed by the empirical Bayes approach but does so fully within the standard Bayesian theory of inference.

Software and Data

Accompanying this book is a free and open source software package that implements all our suggestions (see http://GKing.Harvard.edu/yourcast/ for a copy). The software is entitled "YourCast: Time Series Cross-Sectional Forecasting with Your Assumptions" to emphasize a key intended contribution of our approach: that the assumptions made by the statistical model you run are governed entirely by your choices and your assumptions, and the sophistication of those assumptions and the degree to which they match empirical reality are limited primarily by what you may know or are willing to assume rather than by arbitrary choices hidden behind or hard-coded into a complicated mathematical model. Although some of the tools we introduce require technical sophistication to implement, the ideas are conceptually straightforward. As such, the software and methods should be usable even by those who decide not to digest all of our detailed mathematical arguments.

YourCast is distributed as an R package and is part of the R Project for Statistical Computing (R Development Core Team, 2007). Included in that package are demonstrations that include all the data and code necessary to replicate all the empirical analyses in this book (King, 1995). The complete data set from which examples are drawn is available in Girosi and King (2006).

Background

The methods developed in this book rely on fields of statistics and mathematics, at least some of which are likely to be unfamiliar to many interested in mortality forecasting. Yet, given the highly important public policy issues at stake, the advantage to scholars and citizens of any forecasting improvement, even when achieved via unfamiliar mathematical techniques, should, in our view, outweigh higher costs to researchers in learning the methods. We have thus not shied away from introducing new methods but have tried to reduce the associated costs to researchers in a variety of ways. Most importantly, we explain our methodology in a way that should make all of our results accessible to those who are familiar only with linear regression analysis and a course in Bayesian inference. We also include in appendix A a detailed glossary of notation. In addition, because different aspects of the necessary mathematical background are likely unfamiliar to different audiences, as they were even to us when we started, we offer an extensive mathematical refresher in appendix B that should be relatively complete as is.

Although we have attempted to keep the book as readable as possible, we have also included, for more mathematically sophisticated readers, all necessary proofs and evidence so that the work would be relatively self-contained.

Publications

In some of the fields with which the content of this book intersects, almost all new results appear first in articles. Book presses are left to print texts that only summarize prior research. In this project, we resisted the temptation to send preliminary or partial results to scholarly journals because we felt the whole of our book would be greater than the sum of the parts, sliced into smaller articles, and because our goal was to produce a relatively complete and usable forecasting method in practice. Thus, although we have presented preliminary results in talks over the years, and shared earlier versions of this manuscript, this book is the first complete account of our approach.

Acknowledgments

We are especially grateful for the contributions of Kevin Quinn and Greg Wawro. Although they should not be held responsible for our errors or arguments, we could not have written this book without what we learned from them. Wawro and Quinn were postdoctoral fellows at what is now the Institute for Quantitative Social Science (in the 1998–1999 and 1999–2000 academic years, respectively), and each contributed a great deal. Greg Wawro replicated Murray and Lopez's (1996) forecasts, uncovered some surprising errors (traced to a remarkably basic error in the SAS Statistics package), and managed to improve substantially the quality of the World Health Organization's mortality database. He also demonstrated the failure of a whole range of classic econometric tests and techniques when applied to our data: paradoxically, we found that standard tests frequently confirmed the presence of problems such as cointegration, nonstationarity, parameter heterogeneity, nonrandom missingness, dynamic processes, and omitted variable bias, but following the textbook econometric correction for each using a variety of off-the-shelf approaches degraded or only slightly improved the forecasts. We now accept new model features only if they demonstrate their worth by improving out-of-sample forecasting performance, specification tests for the various econometric problems being largely beside the point.

Kevin Quinn had the vision to see the connection between what we were trying to accomplish and the literature on Markov random fields. He worked out how to extend the Markov random field approach, designed for modeling physical space, to processes that varied over conceptual space (such as age). He also designed clever ways to implement these ideas via Gibbs sampling in Gauss code. These contributions were invaluable.

We also thank Chris Murray, who originally challenged us to develop a method of forecasting mortality that outperformed existing approaches—including his own—and his offices at the World Health Organization supplied us with data and research support. Chris, along with David Evans, Majid Ezzati, Emmanuela Gakidou, Alan Lopez, Colin Mathers, and Josh Salomon, taught us a great deal about mortality data and patterns. They were all especially helpful during the sessions we had pouring over thousands of forecasts from successive versions of our model. To say they were good sports every time we showed up in Geneva lugging two linear feet of printouts in a tiny font is no small understatement.

This project is also an outgrowth of a National Institute of Aging grant on the "Global Burden of Disease in Aging Populations" (P01 AG17625-01). The frequent meetings of this distinguished group of scholars have provided invaluable feedback on our work and, through site visits around the world, have greatly informed our analyses. Our thanks to the other project leaders on this grant, David Cutler, Ken Hill, Alan Lopez, and Chris Murray, to Richard Suzman at NIA, and to the many other participants in our meetings.

Our thanks to Emmanuela Gakidou, James Honaker, Catherine Michaud, Ken Scheve, Ajay Tandon, and Neils Tomijiman for logistical help with and insightful comments on our work. Thanks also to Sam Abrams, Chris Adolph, Marcus Augustine, Anders Corr, Alexis Diamond, Suzanne Globetti, Josh Goldstein, Mie Inoue, Ethan Katz, Ryan Moore, Claudia Pedroza, Nirmala Ravishankar, Heather Stoll, and Liz Stuart for superb research assistance, and Anita Goldpergel and Elena Villalon for expert programming assistance. Our appreciation goes to Gadi Geiger for his mechanical engineering contributions. For helpful discussions on previous drafts or talks, we thank Marc Alexander, Barbara Anderson, Jim Berger, Heather Booth, Brad Carlin, Majid Ezzati, Andrew Gelman, Ed George, Ken Hill, John Maindonald, Xiao-Li Meng, Jamie Robins, Don Rubin, Samir Soneji, Herbert Smith, Len Smith, Dalene Stangl, Liz Stuart, Leonie Tickle, John Wilmoth, and Chris Zorn. Our special thanks to Ron Lee and Nan Li for their generous help with the demographic literature and insights into their models and perspectives, and Sander Greenland for patiently teaching us about the epidemiology literature. Micah Altman, Shawn Bunn, Matt Cox, and William Wei at the Harvard-MIT Data Center provided a series of first rate computational environments that always matched our quickly evolving needs. Thanks also to Jaronica Fuller for consistently cheering up everyone within miles of our offices, and Marie Cole, Beverly MacMillen, and Kim Schader for organizing the team we assembled and making difficult administrative issues transparent. For research support, in addition to the National Institutes of Aging, we are grateful to the National Science Foundation (SES-0112072, IIS-9874747) and the World Health Organization and, at Harvard, the Weatherhead Initiative and the Institute for Quantitative Social Science.

Demographic Forecasting