Forecasting cohort childlessness: Bayesian modeling based on historical patterns in the Human Fertility Database

Carl Schmertmann *Florida State University*
Emilio Zagheni *Queens College, City University of New York*
Joshua Goldstein *Max Planck Inst. for Demographic Research*

EXTENDED ABSTRACT

Introduction

Childlessness among US women at age 45 had been rising steadily until very recently, from a low of 6% among women born in 1935 to near 17% for women in the 1954 birth cohort. There are indications that the trend may have reversed (Dykstra 2009; Human Fertility Database [HFD] 2011), with lower childlessness among women born in the late 1950s and early 1960s. Are we at a turning point? Will recent increases in period fertility eventually translate into lower childlessness among currently-young cohorts? In this article, we present a new forecasting method that attempts to answer such questions, and to assess the associated uncertainty.

Childlessness is interesting in part for what it reveals about the forces driving fertility change. Theoretical explanations for increases in childless include two important strands. On the one hand, the second demographic transition literature emphasizes changes in social pressures and norms, which in some sense liberate people from the social obligations of having children. In this perspective, increased childlessness represents a closer alignment of desires and behavior, a kind of previously unmet demand for child-free life. On the other hand, the literature on work-family balance (the second shift, the incomplete revolution, etc.) emphasizes that the level of childbearing achieved in different societies reflects the degree of compatibility of family and work life. In this way of thinking, higher childlessness is a symptom of incompatible institutions and expectations.

Childlessness also has important social and individual consequences. Childbearing decisions affect educational attainment and labor force participation. Absence of children may have consequences for care received at old age, may result in new forms of social support, and may transform the role and functions of traditional kinship and friendship networks.

Forecasting childlessness for cohorts of women that are still in childbearing ages is a major challenge for demographers (Morgan and Chen 1992, Sobotka 2004), and all forecasts are of course imperfect. In this paper, we propose a Bayesian approach to forecasting age-period-cohort surfaces of fertility rates, using the Human Fertility Database (HFD 2011) as a source of *a priori* knowledge about patterns and variability. Our approach takes into account both what is happening to each cohort, and secular change between cohorts over time. The Bayesian model allows us to complete fertility histories for young cohorts in a way that is consistent with historical regularities observed across space and time. Moreover, our method allows improved evaluation of uncertainty about future trends in childlessness.

A Bayesian model with priors based on historical first-birth rates

We model childlessness by first considering the Lexis surface of unconditional first-birth rates

$$\theta_{ac} = \frac{\text{first births to women age } a \text{ in cohort } c}{\text{number of women age } a \text{ in cohort } c}$$

over 30 integer ages a=15…44 and 43 calendar-year birth cohorts c=1950…1992. Note that these are *demographic rates of the second kind*, because the denominator does not exclude women who have had a first birth. For cohort c, the proportion childless at age 45 is

$$Z_c = 1 - \sum_{a=15}^{44} \theta_{ac}$$

Some of the (age,cohort) rates belong to the past and can be estimated very accurately from vital statistics, while others lie in the future and must be forecast. For example, from the HFD we can estimate that for US women born in 1960, $\theta_{30,1960} \approx .0324$ and final childlessness was $Z_{1960} \approx .158$. In contrast, past data provide only a partial picture of first-birth incidence and childlessness for US women born in 1985: estimates of first-birth rates for this cohort at young ages (15-22) are already available in the HFD, but rates at higher ages – and thus Z_{1985} – must be forecast.

We propose a Bayesian model in which all first-birth rates – both past and future – are unknown quantities about which we can express uncertainty. We stack all 1290 (age,cohort) combinations of interest into a vector θ, sorted by age within cohort. In loose but intuitive notation the model for US women is

$$\underbrace{P(\,\theta \mid Data, History)}_{\text{posterior}} \propto \underbrace{L(Data \mid \theta\,)}_{\text{likelihood}} \cdot \underbrace{f\left(\theta \middle| History\right)}_{\text{prior}}$$

where

- *Data* represents available HFD estimates for the US (age,cohort) cells of interest, such as $\theta_{30,1960} \approx .0324$ above, stacked into a single vector
- *History* represents HFD data for cohorts born (anywhere, US or not) 1900-1949.

The *L*( ) function is a standard likelihood. It answers the question *How likely are the HFD estimates if the true rates are* θ*?*, and assigns higher probabilities to rate surfaces that match published estimates closely.

The *f*( ) function encodes prior knowledge gained from surfaces of first-birth rates for pre-1950 cohorts in the HFD, as described below. It answers the question *How likely are different contemporary rate surfaces, given the patterns and variability in earlier HFD data?*, and assigns higher probabilities to sets of rates that better match historical patterns.

Because the posterior distribution on the left is a product of *L*( ) and *f*( ), it assigns high probabilities to θ surfaces that fit the available data well *and* have historically plausible patterns over age and cohort. The explicit compromise between fitting past estimates and maintaining plausible patterns allow us to estimate both the most likely levels of future childlessness for US cohorts, and, equally importantly, to estimate our uncertainty about those levels.

We use a normal approximation to the likelihood, specifically

$$\ln \underbrace{L(Data \mid \theta\,)}_{\text{likelihood}} = const - (Data - \mathbf{V}\theta)' \, \mathbf{\Psi}^{-1} \left(Data - \mathbf{V}\theta\right)$$

where **V** is a matrix of zeroes and ones such that **V**θ is the subset of rates for which there are already existing HFD estimates, *i* indexes that subset, W_i is the number of women from whom the HFD rate in cell *i* was estimated, and **Ψ** is a diagonal matrix of estimated variances Var($Data_i$)≈ $Data_i/W_i$. This approximation ignores the non-zero covariances between estimated rates of the second kind within each cohort, but in practice that subtlety does not matter for our estimates. Because HFD data come from national populations and sample sizes W_i are so large, any reasonable model of the sampling process yields extremely low likelihoods for proposed surfaces θ on which rates in past cells *i* are not almost identical to published estimates $Data_i$. In short, the likelihood portion of the model practically insists that the surface θ must match HFD estimates very closely for (age,cohort) cells belonging to the past.

The main novelty in our forecast method is the construction of the prior distribution *f*( ) from a large historical dataset. We only outline the basic ideas in this Extended Abstract, leaving many details for the final paper and presentation.

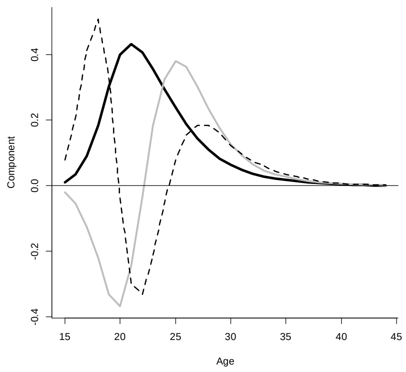
We want to identify patterns in the HFD estimates of the 1900-1949 cohorts (*History*), for both

- age schedules within cohorts, and
- time series at each age 15…44

We deal with these in turn in the next two subsections.

Age Patterns/Cohort Schedule Shapes

In order to identify age patterns, we first aggregated available HFD first-birth rates for cohorts born in any country prior to 1950. Discarding cohorts with incomplete rate data at any age left us with 152 complete historical schedules (32 from the US, 18 from Bulgaria, 21 from Canada, 16 from the Czech Republic, and so on). We performed a singular value decomposition on the 152x30 matrix of schedules to derive 3 principal components (illustrated in the adjacent graph). Over 95% of the variation in age-specific first-birth rates can be explained by modeling the historical rates as weighted sums of these components, which essentially correspond to overall level of first births (solid black), postponement to higher ages (solid grey), and variance over age (dotted black).



Our prior for the cohort shapes in a contemporary surface θ is that they *also* should be well approximated by the three SVD components, with approximation errors that behave similarly to those in the historical HFD. More specifically, one can always decompose rates for each cohort *c* into a least-squares projection and an error vector

$$\theta_{\bullet c} = \underset{\substack{\text{projection onto} \\ \text{SVD components}}}{\mathbf{X}\beta_c} + \underset{\substack{\text{approximation} \\ \text{error}}}{e_c}$$

Our prior is that approximation errors for cohorts in a contemporary surface θ will have an approximately normal distribution with mean E(e_c)=**0** and a covariance matrix **Ω** = E($e_c e_c$□) = average outer product of approximation errors in the historical data. Thus, contemporary surfaces are more likely when they mimic cohort patterns and approximation errors in the historical HFD. In matrix terms this yields a “shape prior”

$$\underset{\text{prior}}{\ln f_{shape}(\ \theta \mid History)} = const\ -\ \sum_{c} e_c' \, \mathbf{\Omega}^{+} \, e_c$$

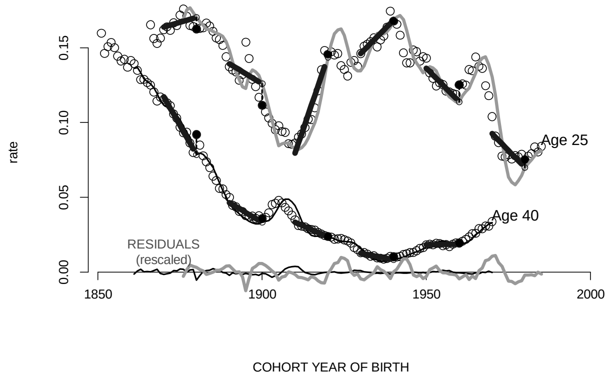
that penalizes θ surfaces with implausible cohort shapes. (We use the generalized matrix inverse, $\mathbf{\Omega}^{+}$, because this shape prior is improper and therefore by construction the historical HFD estimate of **Ω** is not full rank – the full paper will have more details.)

Time Patterns/Age-Specific Rate Series

Our second prior is that the time series of rates at each age 15…44 should have smoothness properties similar to historical HFD rates. To operationalize this we assume that time series at each age *a* are approximately linear over 10-year periods, such that residuals *u* in the expression

$$\begin{aligned}\theta_{ac} &= \left\{ \text{OLS forecast from } \theta_{a,c-1} \cdots \theta_{a,c-10} \right\} + u_{ac} \\ &= \left\{ \sum_{j=1}^{10} w_j \, \theta_{a,c-j} \right\} + u_{ac} \qquad c = 1960...1992;\ w_j = (7-j)/15\end{aligned}$$

are small and behave like their historical HFD equivalents. The adjacent figure shows some examples of this model for historical HFD data for Swedish age-specific birth rates (this illustration shows births at all parities, at maternal ages 25 and 40). Local linearity is generally a reasonable assumption, but one-cohort ahead forecast residuals are notably larger in the more volatile time series for 25-year olds.



We stack over ages within cohorts to produce a vector expression for the OLS forecasts of cohort schedules for women born between 1960 and 1992:

$$\theta_{\bullet c} = \left\{ \sum_{j=1}^{10} w_j \ \theta_{\bullet, c-j} \right\} + u_c \qquad c = 1960...1992$$

Our prior for time series smoothness is that vectors of approximation errors in this expression are normal with mean E(u_c)=**0** and covariance E($u_c u_c$□)=**Γ**=average outer product of these approximation errors in historical HFD data. . Thus, contemporary surfaces are more likely when they mimic time series approximation errors in the historical HFD. In matrix terms this yields a “time prior”

$$\ln \underbrace{f_{time}(\ \theta\ \mid History)}_{\text{prior}}\ = const\ -\ \sum_c u_c' \, \Gamma^{-1} u_c$$

that penalizes θ surfaces with implausible time series patterns.

Combining the Shape and Time Priors

Because shape residuals e_c and time series residuals u_c are not independent, we developed a reweighting procedure that appropriately calibrates the joint distribution. We omit the details in this Extended Abstract, but the end result is a 1290x1290 penalty matrix **P**, and an improper prior distribution for the surface that penalizes both implausible cohort shapes and implausible time series of rates:

$$\ln \underbrace{f_{combined}(\ \theta\ \mid History)}_{\text{prior}}\ = const\ -\theta' \mathbf{P}\,\theta$$

Posterior Distribution

With normal priors and a normal likelihood function, the posterior distribution is also normal, with quadratic penalties both for lack of fit to the available data and for implausible shapes and time series:

$$\ln \underbrace{P(\theta\ \mid Data\ )}_{\text{posterior}}\ = const\ -\ \underbrace{(Data - \mathbf{V}\theta)'\ \boldsymbol{\Psi}^{-1}\ (Data - \mathbf{V}\theta)}_{\text{fitting penalty}}\ -\ \underbrace{\theta' \mathbf{P}\,\theta}_{\text{historical plausibility penaly}}$$

The mean and variance of this posterior have closed-form matrix solutions, so that

$$\theta\ \mid Data\quad \sim N\left(\mu_{post}\ ,\ \Sigma_{post}\right)$$

with

$$\mu_{post} = \left[\mathbf{V}'\ \boldsymbol{\Psi}^{-1}\mathbf{V} + \mathbf{P}\right]^{-1} \mathbf{V}'\ \boldsymbol{\Psi}^{-1}\, Data$$
$$\boldsymbol{\Sigma}_{post} = \left[\mathbf{V}'\ \boldsymbol{\Psi}^{-1}\mathbf{V} + \mathbf{P}\right]^{-1}$$

Childless in any one cohort *c* is $Z_c\ = 1 - \gamma_c'\,\theta$, where γ_c is a 1290-vector of dummy indicators for membership in cohort *c*, so that its posterior distribution is

$$Z_c\ \mid Data\quad \sim N\left(1 - \gamma_c'\,\mu_{post}\ ,\ \ \gamma_c'\,\boldsymbol{\Sigma}_{post}\,\gamma_c\right)\qquad \text{c=1950…1992}$$

For cohorts that have already reached their 45th birthdays, these posterior distributions are extremely precise, because the only uncertainty in first-birth rates comes from sampling variance. For example,

$$Z_{1960}\ \mid Data\quad \sim N\left(.1575\ ,\ \ \mathrm{var} = .00000038\right)$$

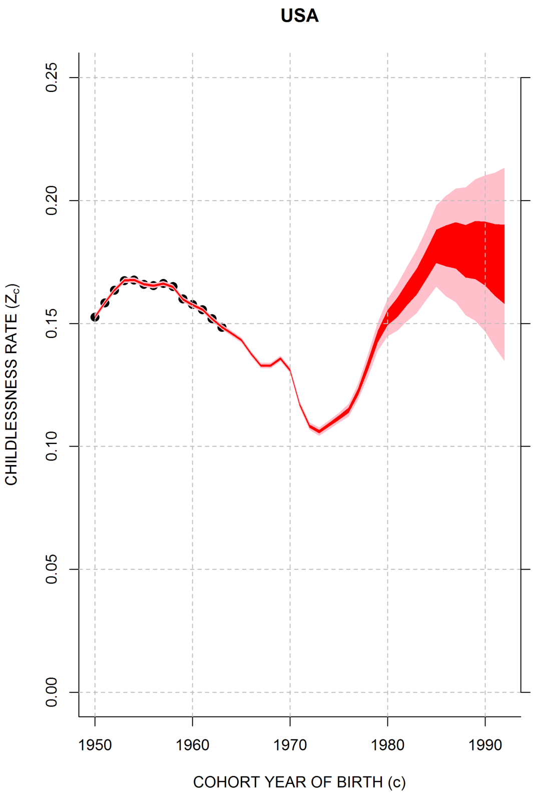
corresponding to a 90% posterior probability interval [.157,.159]. In contrast, for cohorts with incomplete histories in the HFD the model produces forecasts whose uncertainty is calibrated to historical volatility in cohort schedule shapes and time trends, such as

$$Z_{1985} \mid Data \sim N\left(.1814\ ,\ \text{var} = .00010108\right)$$

with a 90% probability interval of [.165,.198].

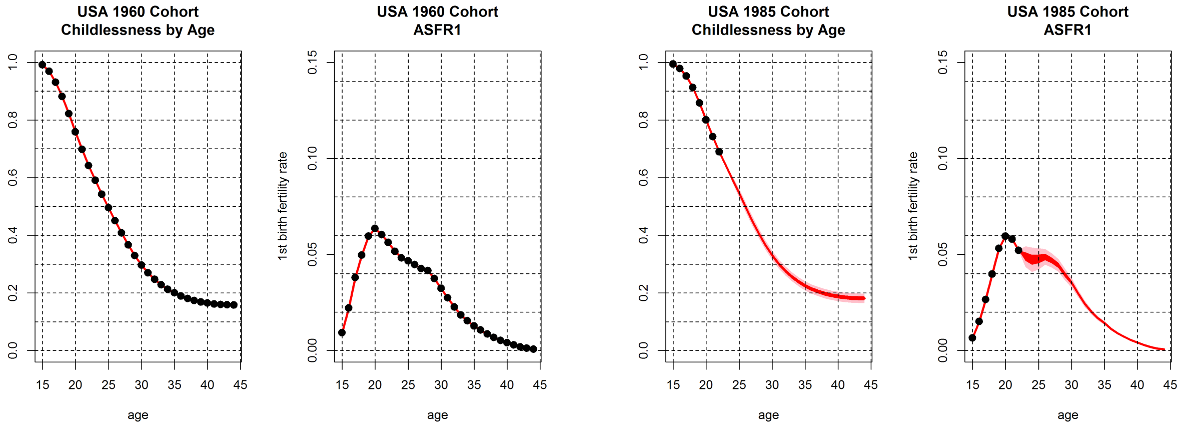
Example Forecasts of US Childlessness

We briefly illustrate the output of the forecast model for US cohorts. The adjacent plot shows the time series of forecast final childlessness levels Z_c. Solid dots represent cohorts that had already reached their 45th birthday at the time of data collection. Dark and light bands illustrate 50% and 90% posterior probability intervals for the estimates, respectively.



One can see from the plot that our model suggests very little uncertainty about the ultimate levels of childlessness among woman born before 1980, even for those who are still at childbearing ages. The high precision of these pre-1980 estimates arises mainly from two sources related directly to our shape and time priors: (1) given a cohort schedule of first-birth rates through the mid- to late-20s, the shape of the remaining schedule is easy to predict, and (2) for cohorts nearing the end of childbearing, we have recent data on slightly-older cohorts at the remaining ages, so that only short temporal extrapolations of age specific rates are necessary to “fill in” the remaining schedule. These are not really new insights. However, a Bayesian model allows automatic estimation of the associated forecast uncertainty, and suggests that for women already past the modal age of first birth rates that uncertainty is very small.

The forecast model provides age-specific first-birth rates as well as final cohort childlessness. The two small figures below use the same graphical conventions as the previous plot, this time to illustrate model results for the 1960 and 1985 birth cohorts by age. Downward-sloping curves show the proportion childless at each age, and the hump-shaped curves show first-birth rates from the forecast Lexis surface.



The model produces a rich set of similar results, including rates for other parities, which we will also analyze for the final paper and presentation.

References & Resources

P Dykstra, 2009. Childless old age, pp. 671-690 in Uhlenberg (ed.), *International Handbook of Population Aging*. Springer.

F Girosi and G King, 2008. *Demographic Forecasting*. Princeton University Press, Princeton NJ.

Human Fertility Database (HFD), 2011. Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at www.humanfertility.org (data downloaded on 2 Nov 2011)

SM Lynch, 2007. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, New York.

SP Morgan and R Chen, 1992. Predicting childlessness for recent cohorts of American women. *International Journal of Forecasting* 8:477-493.

National Center for Health Statistics (NCHS), 2011. Cohort fertility tables. Available at http://www.cdc.gov/nchs/nvss/cohort_fertility_tables.htm#documentation.

T Sobotka, 2004. *Postponement of childbearing and low fertility in Europe*. Doctoral thesis, University of Groningen. Dutch University Press, Amsterdam.