A Stochastic Predictive Model for Life Expectancy

Sam Dawley
Department of Chemistry
Johns Hopkins University
Baltimore, MD 21218
sdawley1@jhu.edu

Oliver Wolff
Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218
owolff1@jhu.edu

Abstract

Among the most salient determinants of a country's condition, life expectancy provides a snapshot of the economic status, societal status, and even geographical location at a glance. Using data on national GDP per capita, population, and continent, we utilized a hierarchical Bayesian model to estimate the life expectancy of countries since 1950 with a double logistic function. Results of the model and a comparison of predicted life expectancy for geographically similar countries versus their actual life expectancy is presented and discussed. Potential sources of bias in the model which may cause performance issues in certain parts of the globe are analyzed: a reminder that cognizance of limitations in the predictive power of data science increase its validity and reliability.

The GitHub repository containing code associated with this project can be found here:
`https://github.com/ofw24/friendly-potato`

1 Introduction

Among the most critical pitfalls of the frequentist paradigm is its reliance on the long-run probability: the assumption that the chance of an event happening is the limit of its relative frequency within the entire sample space. This shortcoming is especially dangerous in applications of statistical and machine learning methods to more far-reaching and global problems, e.g., facial recognition for identification, supervised and unsupervised learning for advanced artificial intelligence, and decision-making algorithms for evaluating interesting problems. Whether the applications be private, industrial, or international, allowing inferential models to account for stochastic uncertainty within real-world problems helps to build statistical frameworks which can more accurately capture complexity that is otherwise ignored within frequentist statistics.

Addressing practical observation and consequent prediction amidst uncertainty is done most effectively within the Bayesian paradigm; hierarchical Bayesian models may account for not only potential randomness but also sampling variability within and between different experiments. Herein we present a Bayesian hierarchical model for predicting life expectancy within a particular country based on easily attainable factors such as population size, continent, and gross domestic product (GDP). Unlike deterministic methods which model life expectancy as a function of the previous years' average time of death, stochastic methods under the Bayesian framework allow greater flexibility in assigning parameters to our model as well as allowing for within- and between-country variability that naturally arise with geographic data sets.

2 Motivation

Life expectancy is a critical factor in determining a nation's ability to provide its citizens with their most basic needs. Socioeconomic stability, predicted population size, and governmental fragility

are some examples of statistics which influence average life expectancy and to some extent can be predicted by this metric. Creating an accurate predictive model for worldwide, continental, and national life expectancy is therefore integral for international governing bodies, such as the United Nations, among other more local administrations and governments. A Bayesian predictive model for life expectancy allows for more accurate comparisons of expected national prosperity against the observed. Moreover, such methods ultimately afford a technique of assessing the impact of historically significant events on the average life expectancy both spatially and temporally.

3 Literature Review

Population dynamics are the subject of many introductory courses on differential equations; fitting population to a logistic curve as a function of time is an elementary exercise in nearly all of them. As such, this technique remains invaluable for estimating life expectancy, as improvements in nutrition, education, and healthcare have pushed lifetimes to a point where environmental factors are less frequently the primary cause of death.

There exist numerous published papers regarding Bayesian estimation of life expectancy reliant on a variety of population parameters, many of which make use of the logistic function and its related logit distribution. As discussed by Raftery et al., the predicted increase in life expectancy can be modeled as a double logistic function, denoted $g(\ell_{c,t} \mid \boldsymbol{\theta}^{\boldsymbol{c}})$. Under this predictive model, predicted life expectancy is given by

$$\ell_{c,t+1} = \ell_{c,t} + g(\ell_{c,t}) + \varepsilon_{c,t+1} \tag{1}$$

where $\ell_{c,t}$ is the current life expectancy, $\ell_{c,t+1}$ is the future life expectancy, and $\varepsilon_{c,t+1}$ represents randomness in the model. The results of Figure 1a support the shape of a double logistic function $g(\ell_{c,t+1})$, and is consistent with five-year gain among the literature.

Bongaarts further discusses the use of logistic functions in modeling the "force of mortality", i.e., the expected age of death for a given population (Boongart, 2005), and in particular, making use of the shifting logistic model $\mu(x,t)$:

$$\mu(x,t) = \frac{\alpha(t)e^{\beta x}}{1+\alpha(t)e^{\beta x}} + \gamma(t) \tag{2}$$

The rate parameter β is constant with the population and time whereas $\alpha(t)$ and $\gamma(t)$ vary with level and rate of mortality, respectively. Taking $\gamma(t)$ to be randomness associated with statistical uncertainty in life expectancy within a population, (2) becomes the sum of a logistic function and noise: functionally equivalent to the predicted life expectancy in Equation (1).

Ultimately, one of the best known approaches to stochastic projection of life expectancy and mortality is that of Lee & Carter, modeling age-specific mortality using a random walk model with constant drift (Lee, 1992). Improvements on this idea include smoothing age-specific mortality rates over age and time, though, both this technique and Lee & Carter's original idea suffer from the assumption that the rate of change of age-specific mortality remains constant with time (Raftery, 2013).

Clearly, this need not always be the case. Hence, a model which accounts for changes in age-specific mortality both spatially and temporally is necessary to more effectively predict future life expectancies. Building off of Lee & Carter, a random walk model using *nonconstant* drift provides an outlet for testing a stochastic model for predicting global average life expectancies.

4 Methodology

Future life expectancy may be projected deterministically from the life expectancy of the country c in the current five-year time period t, denoted $\ell_{c,t}$, plus the expected gain in life expectancy $g(\ell_{c,t})$:

$$\ell_{c,t+1} = \ell_{c,t} + g(\ell_{c,t}) \tag{3}$$

The gain $g(\ell_{c,t})$ is a double logistic function of the current life expectancy:

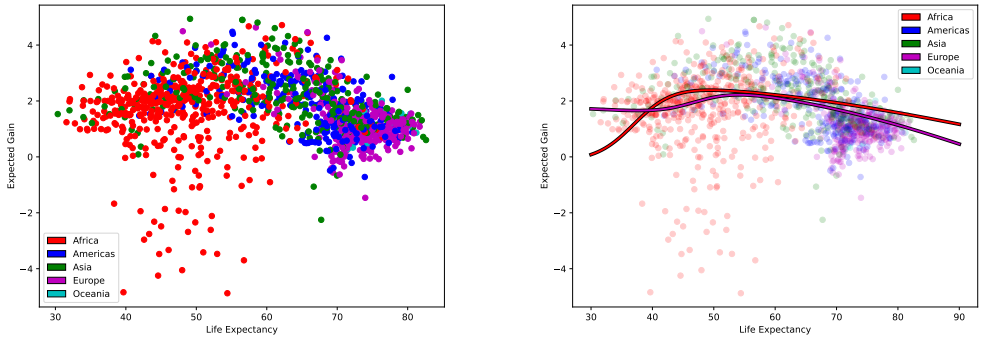
$$\begin{aligned} g(\ell_{c,t}|\boldsymbol{\theta}^c) = &\frac{k^c}{1+\exp\left(-2\ln(9)\Delta_1^c[\ell_{c,t} - \sum_{j=1}^4 \Delta_j + 0.5\Delta_1^c]\right)} \\ &+ \frac{z^c - k^c}{1+\exp\left(-2\ln(9)\Delta_3^c\left[\ell_{c,t} - \Delta_4^c - 0.5\Delta_3^c\right]\right)} \end{aligned} \tag{4}$$

for a vector of parameters $\boldsymbol{\theta}^c = (\Delta_1^c, \Delta_2^c, \Delta_3^c, \Delta_4^c, k^c, z^c)$. The constant factors are chosen so that the parameters of $\boldsymbol{\theta}^c$ remain on an interpretable scale (Alkema, 2011). Most importantly, the parameters Δ_i^c are *decline parameters* which characterize rates of life expectancy gain across different portions of the double logistic curve and the parameter z^c is the asymptotic average rate of increase in life expectancy as described in Oeppen and Vaupel (Oeppen, 2002). This deterministic model may be naïvely fit to the raw data, as shown in Figure 1b, to provide some intuition for the shapes of these types of curves and best fit.

Figure 1: Raw data.

(a) Gain in average life expectancy across all countries across all continents.

(b) Naïve fit of the double logistic function (4) to the average gain in life expectancy within a continent.



Amending this model to account for uncertainty involves two changes: allowing stochastic movement within each country and allowing the parameters of $g(\ell_{c,t})$ to vary between countries over a continuous rather than discrete range (Raftery, 2013). The resultant hierarchical model enjoys variability among all parameters of $\boldsymbol{\theta}^c$ and includes error from uncertainty. In particular, allowing $\ell_{c,t+1} = \ell_{c,t} + g(\ell_{c,t}|\boldsymbol{\theta}^c)$ with

$$\Delta_{i,t+1}^c|\sigma_{\Delta_{i,t}} \sim N\left(\Delta_{i,t}, \sigma_{\Delta_i}^2\right), \quad k_{t+1}^c|\sigma_k \sim N\left(k_t, \sigma_k^2\right), \quad \text{and} \quad z_{t+1}^c|\sigma_z \sim N\left(z_t, \sigma_z^2\right) \tag{5}$$

supports the parameters of $\boldsymbol{\theta}^c$ within each country over a continuous range. Finally, we add a random perturbation $\varepsilon_{c,t+1}$ to account for uncertainty within the model:

$$\ell_{c,t+1} = \ell_{c,t} + g(\ell_{c,t}|\boldsymbol{\theta}^c) + \varepsilon_{c,t+1} \tag{6}$$

Choosing Gaussian noise to model the uncertainty suffers from assuming health is independent of age; as the observed average life expectancy increases the variability in time of death most certainly decreases due to health issues that arise as one gets older. This idea is supported in Figure 2 which plots the residuals of predicted life expectancy using simple ordinary least squares estimation of the parameters $\boldsymbol{\theta}^c$. Hence, the distribution of $\varepsilon_{c,t+1}$ was chosen such that

$$\varepsilon_{c,t+1} \sim N\left(0, \omega^2 S^2(\ell_{c,t})\right) \tag{7}$$

where S is the spline fit to a plot of the predicted residual versus observed life expectancy and $\omega \sim \text{Uniform}\,(0, 10)$.

4.1 Parameter Estimation

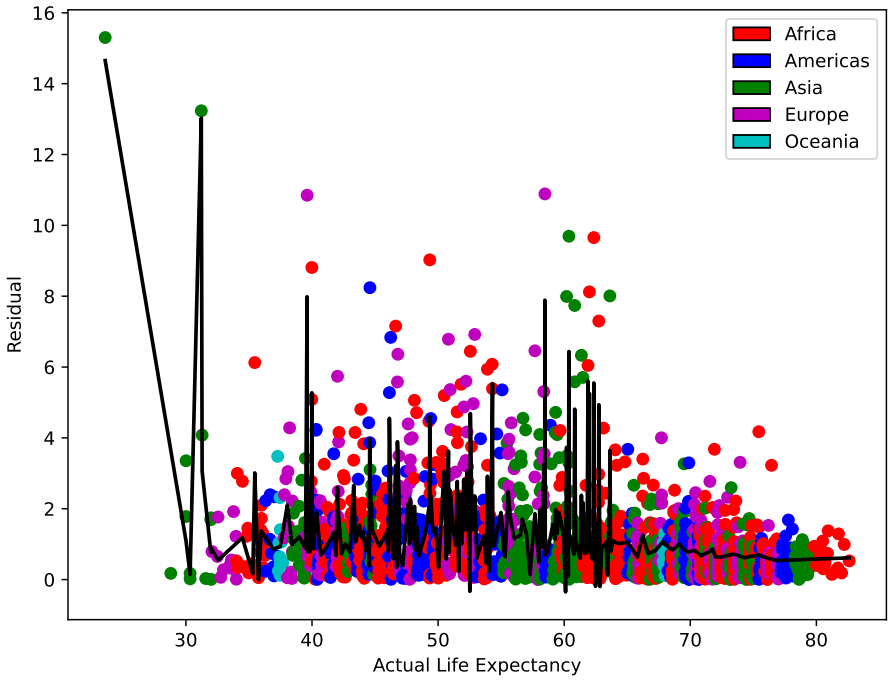
The Bayesian hierarchical model adopted relies on prior distributions for the variances of each country-specific parameter $\{\Delta_i^c, k^c, z^c\}$:

$$\sigma_j^2 \sim IG\left(\nu, \delta_j^2\right) \quad \text{with} \quad j = \Delta_1^c, \Delta_2^c, \Delta_3^c, \Delta_4^c, k^c, z^c \tag{8}$$

for $\nu = 4$ degrees of freedom and rates δ_j^2 determined by Raftery et al. as

$$\delta_j^2 = 15.6^2, 23.5^2, 14.5^2, 14.7^2, 3.5^2, 0.6^2$$

Figure 2: Residuals of predicted life expectancies for average life expectancies across all countries. Superimposed in black is a spline fit to the data.



The posterior distributions for each of σ_j^2 were determined simply as a product of the prior with the likelihood of the life expectancy gain, $g(\ell_{c,t}|\boldsymbol{\theta}^c)$. Prior distributions for each of the decline, shape, and scale parameters were

$$\Delta_{i,t} \sim N\left(a_i, \delta_i^2\right), \quad k_t^c \sim N\left(a_5, \delta_5^2\right), \quad \text{and} \quad z_t^c \sim N\left(a_6, \delta_6^2\right)$$

for $i = 1, 2, 3, 4$ and $\{a_1, \ldots, a_6\}$ determined from the United Nations model for life expectancy, namely, $\{15.77, 40.97, 0.21, 19.82, 2.93, 0.40\}$. The prior variances for each of Δ_i were taken to be the squared means: $\{15.77^2, 40.97^2, 0.21^2, 19.82^2, 2.93^2, 0.40^2\}$. Posterior estimates for each of these parameters were determined using Equation (5).

4.2 Computational Solutions

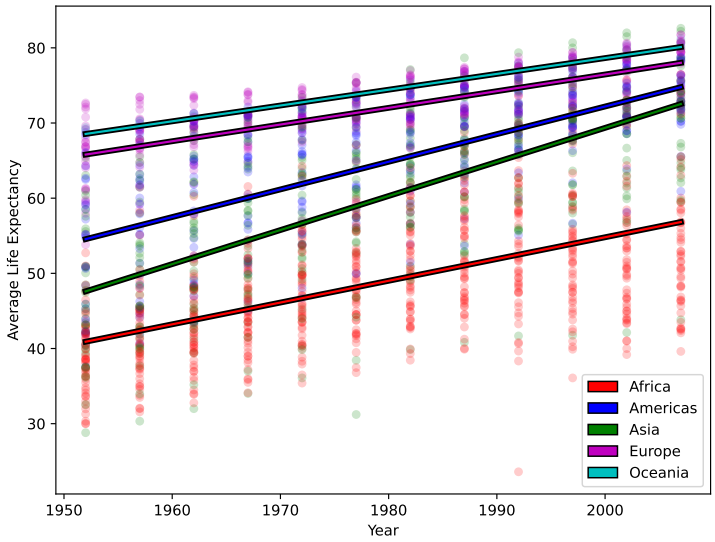
Estimation of the stationary distributions for all of $\{\Delta_1, \Delta_2, \Delta_3, \Delta_4, k, z\}$ were determined via Metropolis-Hastings sampling. Each parameter was updated over the course of 5,000 scans with a burn-in of 1,000 scans; the Rockfish cluster within the Advanced Research and Computing at Hopkins (ARCH) was used to carry out all random sampling.

5 Analysis and Discussion

A linear fit of the ordinary least squares estimate of these regressors colored by continent can be found in Figure 3. Figure 2 shows the plot of absolute residuals by life expectancy using this linear model. Nearly all residuals (barring those outliers mentioned earlier) are contained within no larger than six years: ten percent of the actual lifespans. Moreover, a spline fit to the data verifies estimation of the randomness contained within $\varepsilon_{c,t+1}$ applied to the model. The trend lines allow us to examine gains in life expectancy by continent; observations of the broad increase in life expectancy continue

to be highly relevant for national and international organizations in allotting resources and studying public health, among other endeavors.

Figure 3: Ordinary least squares result of the regression coefficients for a simple linear model estimating the temporal increase in life expectancy.



The final predictions made by our model plotted against the observed life expectancies can be found in Figure 4. Outliers within predictions, the most notable exceptions recorded in Table 1, suggest one of two issues: the model was trained on poorly collected data or the model was poorly made. The former may be attributed to a plethora or factors, the most significant of which is that the available data can only illustrate a limited portion of the living conditions within any given country. For example, though civil conflict may manifest in a country's GDP, it is impossible to quantify every facet of well-being for a populace. Such an underestimate of life expectancy gain is possibly the result of model bias towards high GDP per capita, the feature represented most uniformly among the outliers.

Another potential source of bias is illustrated across all of the Tables and Figures reported herein: popular deterministic and stochastic models alike do not anticipate societal development and national sentiment. The most significant outliers across numerous methods of life expectancy prediction occur in African countries, the majority of which are underdeveloped or notably less well-off (Osland, 2003). Hence, a probable source of discrepancies present in many of the statistical models used to predict life expectancy, especially those which rely on deterministic predictions, is an ignorance to developedness.

Acknowledgments

The authors would like to acknowldege and thank the supercomputing clusters within the Advanced Research and Computing at Hopkins (ARCH) core facility which is supported by the NSF.

Table 1: Select countries and continents which deviated from predicted life expectancy gains in some arbitrary years to the greatest extent.

Country	Continent	Predicted Life Expectancy	Actual Life Expectancy
Bahrain	Asia	-25.566	50.939
Cambodia	Asia	-336.212	31.22
Mongolia	Asia	27.117	63.625
Burkina Faso	Africa	-236.981	46.137
Cameroon	Africa	219.578	54.314
Central African Republic	Africa	-33.150	46.775
Chad	Africa	-10.194	50.525
Congo, Dem. Rep.	Africa	-29.484	47.804
Gabon	Africa	147.396	60.19
Guatemala	Africa	116.901	42.023
Madagascar	Africa	119.892	46.881
Mozambique	Africa	200.639	31.286
Namibia	Africa	-41.020	61.999
Nigeria	Africa	-7.385	46.608
Rwanda	Africa	-1018.090	23.599
South Africa	Africa	-185.880	49.339
Swaziland	Africa	-496.905	39.613

References

[1] Leontine Alkema, Adrian E. Raftery, Patrick Gerland, Samuel J. Clark, François Pelletier, Thomas Buettner & Gerhard K. Heilig (2011). Probabilistic Projections of the Total Fertility Rate for All Countries. *Demography*, 48:3, 815-839. DOI: 10.1007/s13524-011-0040-5

[2] Adrian E. Raftery, Jennifer L. Chunn, Patrick Gerland & Hana Ševčíková (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography*. 50:3, 777-801. DOI: 10.1007/s13524-012-0193-x.

[3] John Bongaart. (2005). Long-range trends in adult mortality: Models and projection methods. *Demography*, 42:1, 23–49. DOI: 10.1353/dem.2005.0003

[4] Jennifer Bryan, gapminder, (2019), GitHub repository, `https://github.com/jennybc/gapminder`

[5] Ronald D. Lee & Lawrence R. Carter (1992). Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, 87:419, 659-671. DOI: 10.1080/01621459.1992.10475265

[6] Jim Oeppen & James W. Vaupel (2002). Broken Limits to Life Expectancy. *Science*, 296:5570, 1029-1031. DOI: 10.1126/science.1069675

[7] Joyce S. Osland (2003). Broadening the Debate: The Pros and Cons of Globalization. *Journal of Management Inquiry*, 12:2, 137–154. DOI: 10.1177/1056492603012002005

Figure 4: Predicted versus actual life expectancies using the double logistic model described herein. The line for Predicted = Actual is superimposed on the graph.