Demographic transition forecasting notes

Richard W. Evans

version (20.02.a)

Put intro here.

1 Interpolation and Curve Fitting

Interpolation is a general term than signifies, in its broadest sense, the prediction of variable value

1.1 Interpolation

Interpolation

1.2 Curve fitting

Two important functions that we use in curve fitting in this project are a parameterized arctangent and a parameterized exponential. For data that monotonically increase or decrease in both value and slope from or to some horizontal asymptote, the three-parameter negative exponential is a flexible and parsimonious function.

$$(EX_3): \quad f(x|a,b,c) = e^{ax^2+bx+c} \quad \forall x,a,b,c \tag{1}$$

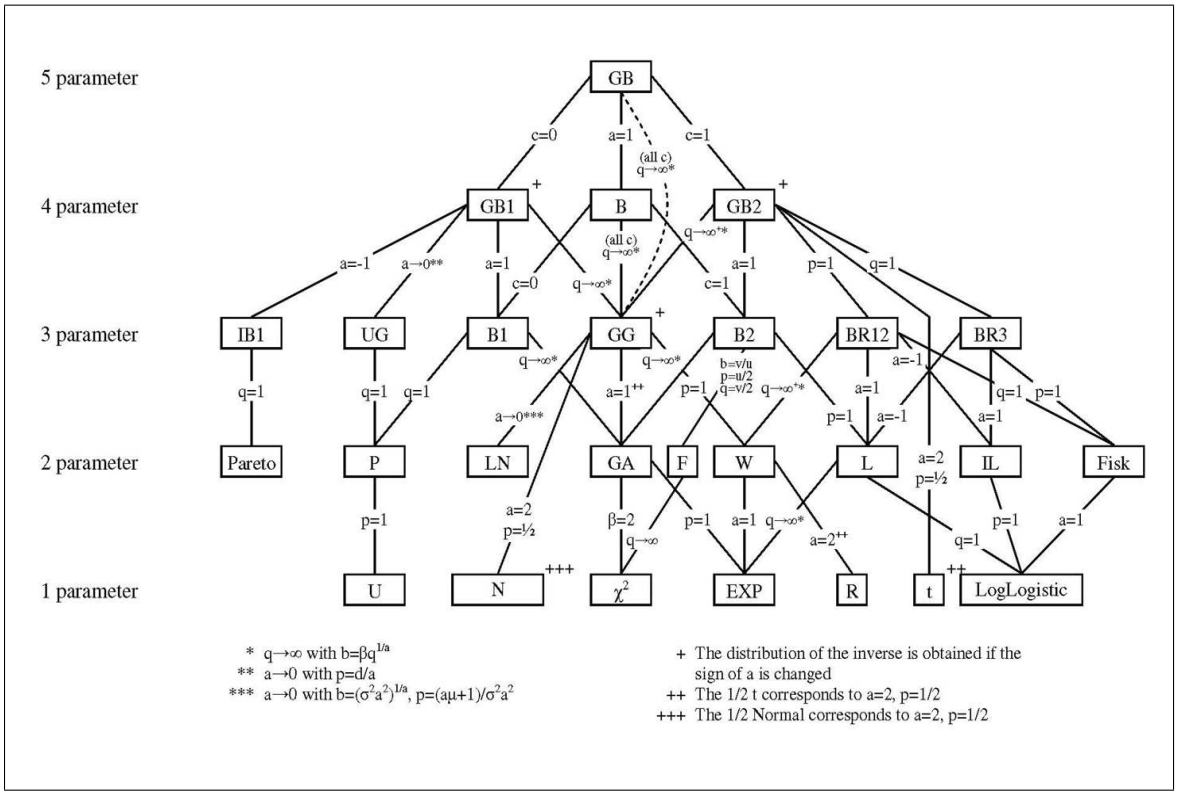
For monotonic data that start out with low slope, then increase in slope, then decrease again in slope, the parameterized arctangent function is a flexible and parsimonious function.

$$(Arctan_3): \quad f(x|a,b,c) = \arctan(ax+b)+c \quad \forall x,a,b,c \tag{2}$$

2 Estimation with Hierarchical Distribution Families

It is often valuable to fit curves or distributions with parametric functions. These functions use a limited and fixed number of parameters to capture general shapes of different families of curves. One such class of univariate probability density functions

Figure 1: Generalized beta family of distributions (McDonald and Xu, Fig. 2, 1995)



with positive support $x \geq 0$ is the generalized beta family of distributions shown in Figure 1 taken from McDonald and Xu (1995, Fig. 2).

Two of the distributions from Figure 1 that we will use for curve fitting are the two-parameter gamma $GA(\alpha, \beta)$ distribution defined in (3) and its three-parameter parent distribution the generalized gamma $GG(\alpha, \beta, m)$ defined in (4).[1]

$$(GA): \quad f(x|\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, \quad x \in [0,\infty],\ \alpha,\beta > 0$$
$$\text{where} \quad \Gamma(z) \equiv \int_0^{\infty} v^{z-1} e^{-v} dv \tag{3}$$

$$(\text{GG}): \quad f(x|\alpha,\beta,m) = \frac{m}{\beta^{\alpha}\Gamma\left(\frac{\alpha}{m}\right)} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{m}}, \quad x \in [0,\infty],\ \alpha,\beta,m > 0$$
$$\text{where} \quad \Gamma(z) \equiv \int_0^{\infty} v^{z-1} e^{-v} dv \tag{4}$$

[1]Code for the probability density functions of the GA distribution and GG distribution can be found in the `distributions.py` module as the `GA_pdf()` and `GG_pdf()` functions, respectively.

3 Exercises

Exercise 1. Define the fertility rate as $f_{s,t}$ which signifies the percent of age-s individuals that have children in year t. This fertility rate must account for both male and female population.

a. Get U.S. fertility rate data by age for as many years as possible from 1980 to the most recent year available. If the data for a given year t are less detailed than fertility rate for every age year, interpolate every age year using a cubic spline. If the data do not cover from age 21 up to age 100, choose a method to extrapolate the end points such that the extrapolation function has the same value and slope as the interpolated function at the interpolated function end points and that the function has monotonic slope and finishes at a logical endpoint. Good extrapolation function candidates are the three-parameter exponential (1) or the three-parameter arctangent (2) functions.

b. For each year of fertility-rate-by-age data $f_{s,t}$, fit a four-parameter function to the data by GMM (minimize the sum of squared errors), which function is comprised of the three-parameter generalized gamma $GG(\alpha, \beta, m)$ distribution defined in (4) and scaled by parameter $\chi > 0$.[2]

$$g(f_{s,t}|\chi_t, \alpha_t, \beta_t, m_t) = \chi_t \frac{m_t}{\beta_t^{\alpha_t}\Gamma\left(\frac{\alpha_t}{m_t}\right)} f_{s,t}^{\alpha_t - 1} e^{-\left(\frac{f_{s,t}}{\beta_t}\right)^{m_t}}, \quad \forall t,\ f_{s,t} \in [0, \infty],\ \chi_t, \alpha_t, \beta_t, m_t > 0$$
$$\text{where} \quad \Gamma(z) \equiv \int_0^\infty v^{z-1} e^{-v} dv \tag{5}$$

Plot the estimated fertility rate functions $g(f_{s,t}|\hat{\chi}_t, \hat{\alpha}_t, \hat{\beta}_t, \hat{m}_t)$ for each year of the data. Make separate time series plots for each of the four estimated parameters $\hat{\chi}_t$, $\hat{\alpha}_t$, $\hat{\beta}_t$, $\hat{m}_t$.

c. Using least squares, fit a negative arctangent (2) or exponential function (1) to each of the four estimated parameters time series. If the current period is $t = 2020$, make sure that the fitted functions asymptote to a clear value around period $t = 2100$. Plot each of the four fitted functions against the respective data for periods $1980 \leq t \leq 2100$.

d. Using your four fitted functions for the time path of parameters of the GG function $\hat{\chi}_t$, $\hat{\alpha}_t$, $\hat{\beta}_t$, $\hat{m}_t$, plot for each year $1980 \leq t \leq 2100$ the estimated fertility rate by age $f_{s,t}$.

Exercise 2. Define the mortality rate as $\rho_{s,t}$ which signifies the percent of age-s individuals that die in period t (i.e., do not live to period $t + 1$).

[2] This new function is only a probability density function for $\chi = 1$.

a. Get U.S. mortality rate data by age for as many years as possible from 1980 to the most recent year available. If the data for a given year t are less detailed than mortality rate for every age year, interpolate every age year using a cubic spline. Force the mortality rate at age 100 to be 1.0 (or 100%). If the data do not cover from age 21 up, choose a method to extrapolate the end points such that the extrapolation function has the same value and slope as the interpolated function at the interpolated function end points and that the function has monotonic slope and finishes at a logical endpoint. Good extrapolation function candidates are the three-parameter exponential (1) or the three-parameter arctangent (2) functions.

b. For each year of mortality-rate-by-age data $\rho_{s,t}$, fit a three-parameter exponential function (1) to the data by GMM (minimize the sum of squared errors).

$$g(\rho_{s,t}|a_t, b_t, c_t) = e^{a_t(\rho_{s,t})^2 + b_t\rho_{s,t} + c_t}, \quad \forall t, a, b, c, \quad \rho_{s,t} \in [0, 1] \tag{6}$$

Plot the estimated mortality rate functions $g(\rho_{s,t}|\hat{a}_t, \hat{b}_t, \hat{c}_t)$ for each year of the data. Make separate time series plots for each of the three estimated parameters $\hat{a}_t$, $\hat{b}_t$, $\hat{c}_t$.

c. Using least squares, fit a negative arctangent (2) or exponential function (1) to each of the three estimated parameters time series. If the current period is $t = 2020$, make sure that the fitted functions asymptote to a clear value around period $t = 2100$. Plot each of the three fitted functions against the respective data for periods $1980 \leq t \leq 2100$.

d. Using your three fitted functions for the time path of parameters of the three-parameter exponential EX_3 function $\hat{a}_t$, $\hat{b}_t$, $\hat{c}_t$, plot for each year $1980 \leq t \leq 2100$ the estimated mortality rate by age $\rho_{s,t}$.

Exercise 3. Define the normalized population $\tilde{\omega}_{s,t}$ as the percent of the population in period t that is age s, such that $\sum_{s=1}^{S} \tilde{\omega}_{s,t} = 1$ for all t. Define the immigration rate $i_{s,t}$ as the number of age-s immigrants in the current period as a percent of the age-s population in the previous period (i.e., the number of immigrants in the current period equals $i_{s,t}\omega_{s,t-1}$).

a. Get U.S. mortality rate data by age for as many years as possible from 1980 to the most recent year available. If the data for a given year t are less detailed than mortality rate for every age year, interpolate every age year using a cubic spline. Force the mortality rate at age 100 to be 1.0 (or 100%). If the data do not cover from age 21 up, choose a method to extrapolate the end points such that the extrapolation function has the same value and slope as the interpolated function at the interpolated function end points and that the function has monotonic slope and finishes at a logical endpoint. Good extrapolation function candidates are the three-parameter exponential (1) or the three-parameter arctangent (2) functions.

b. For each year of mortality-rate-by-age data $\rho_{s,t}$, fit a three-parameter exponential function (1) to the data by GMM (minimize the sum of squared errors).

$$g(\rho_{s,t}|a_t, b_t, c_t) = e^{a_t(\rho_{s,t})^2 + b_t\rho_{s,t} + c_t}, \quad \forall t, a, b, c, \quad \rho_{s,t} \in [0, 1] \tag{7}$$

Plot the estimated mortality rate functions $g(\rho_{s,t}|\hat{a}_t, \hat{b}_t, \hat{c}_t)$ for each year of the data. Make separate time series plots for each of the three estimated parameters $\hat{a}_t$, $\hat{b}_t$, $\hat{c}_t$.

c. Using least squares, fit a negative arctangent (2) or exponential function (1) to each of the three estimated parameters time series. If the current period is $t = 2020$, make sure that the fitted functions asymptote to a clear value around period $t = 2100$. Plot each of the three fitted functions against the respective data for periods $1980 \leq t \leq 2100$.

d. Using your three fitted functions for the time path of parameters of the three-parameter exponential EX_3 function $\hat{a}_t$, $\hat{b}_t$, $\hat{c}_t$, plot for each year $1980 \leq t \leq 2100$ the estimated mortality rate by age $\rho_{s,t}$.

References

McDonald, James B. and Yexiao Xu, "A Generalization of the Beta Distribution with Applications," *Review of Econometrics*, March-April 1995, *66* (1-2), 133–152.