Modeling framework

Model

We assume the counts of mother to be Binomially distributed as follows,

$$y_{x,g,t}|\theta_{x,g,t} \sim \text{Binomial}(n_{x,g,t}, \theta_{x,g,t}).$$

where $\theta_{x,g,t}$ is the proportion of female being childless, and $n_{x,g,t}$ is the number of female.

$$\text{logit}(\theta_{x,g,t}) = \text{logit}(1 - f(\zeta_{g,t}, \gamma_{g,t}, \delta_{g,t})) + \sum_k B_{x,k}\alpha_{k,g,t}$$

where

$$f(\zeta_{g,t}, \gamma_{g,t}, \delta_{g,t}) = \frac{\zeta_{g,t}}{1 + e^{-\gamma_{g,t}(x-\delta_{g,t})}}$$

is the logistic function with δ the function's midpoint, ζ the supremum of the values of the function, and γ the steepness of the curve. Deviations from the parametric function are modeled with P-splines. More precisely, $\boldsymbol{B}$ is the cubic basis matrix of dimension $(A \times K)$ where $A = 44$ the upper age bound, and $K = 4$ the number of knots; $\boldsymbol{\alpha_{g,t}}$ is a cubic-splines parameter vector of dimension $(K \times 1)$ estimated in the model. We impose smoothness on the deviations by using a Random Walk 1 prior on $\boldsymbol{\alpha_{g,t}}$,

$$\alpha_{k,g,t} \sim \text{Normal}(\alpha_{k-1,g,t}, \sigma^2_{\alpha,t})$$

where the variance parameter $\sigma^2_{\alpha,t}$ is shared across all subpopulation groups in each year.

Following exploratory data analyses, we smooth the logistic parameters by assuming that

$$\zeta_{g,t} \sim \text{Normal}(2\zeta_{g,t-1} - \zeta_{g,t-2}, \sigma^2_\zeta)$$
$$\gamma_{g,t} \sim \text{Normal}(2\gamma_{g,t-1} - \gamma_{g,t-2}, \sigma^2_\gamma)$$
$$\delta_{g,t} \sim \text{Normal}(2 \cdot \delta_{g,t-1} - \delta_{g,t-2}, \sigma^2_\delta).$$

The variance parameters are shared across all subpopulation groups to enforce some pooling of information. The variance parameters take $\text{Normal}(0, 1)^+$, and $\text{Normal}(0, 0.1)^+$ priors for δ, and ζ and γ (weakly informative priors).

Discussing points

- Data issue: "[...] prior to 2012, when new editing protocols were implemented by the Census Bureau, the Current Population Survey may have overestimated childlessness somewhat." More precisely, "Prior to 2008, the CPS did not collect information on biological relatedness in the household roster portion of the survey. However, in the 2007 CPS core or "Basic" questions (which are asked in each interview), the household relationship questions were expanded to include information on the type of relationship between coresident parents and children. However, the editing procedures used in 2008 and 2010 did not utilize this information."

  - Data for correction is publicly available (momloc/poploc)? If we can perform the correction, we should focus on the years 2008, 2010, 2012, 2014, 2016, 2018, and 2020.
  - For these years, it seems ok to assume RW2 for each parameter.

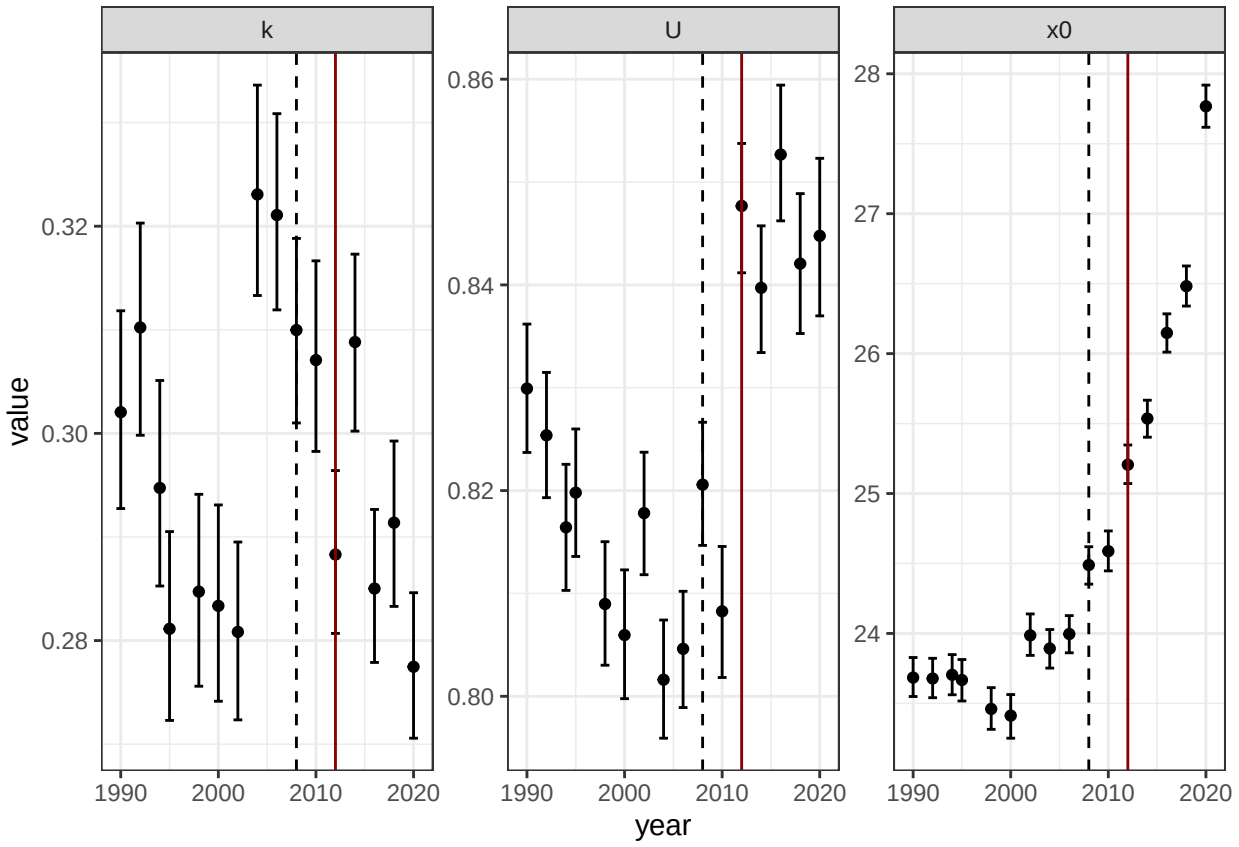

Figure 1: EDA of logistic parameters at the U.S. level

- Lots of variation in the fitted proportion of childlessness from year to year (i.e. New York, California, and Alaska in 2020). This will be even more an issue when focusing on race*state.
  - Pooling of the logistic parameters are only done through the variance parameters.
  - National logistic function in each year from which we allow deviations with P-splines for each state? But no more state-level interpretation of parameters.
  - Use linear spline to reduce flexibility and reduce the risk of over-fitting?
- Few observations at the state level (ie Wyoming has less than 5 females for ages ranging from 30 to 35 yo)