CS&SS/STAT/SOC 563 — Statistical Demography
Spring 2022 — Lectures II

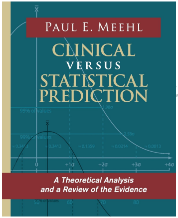
Population Projections in Practice

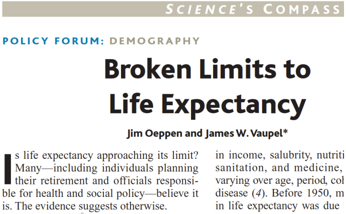
- Population projections: used by governments at all levels for planning, private sector, researchers.
- UN Population Division publishes projections of age- and sex-specific population counts and vital rates for all countries by 5-year age groups in 5-year periods to 2100, every two years in *World Population Prospects* (WPP)
  - used throughout UN system
  - input for development planning, monitoring (e.g. SDGs 2030) and global modeling (e.g. food security, climate)
  - last Revision: WPP 2019 (published mid-2019)
  - Based on probabilistic methods developed at UW Stat since 2015
- Population can be projected well into the future using current and recent population and vital rates:
  - Governments project to 2050 (Ireland), 2060 (USA), 2070 (EU), 2095 (US SSA), 2100 (Japan)
  - UN projects to 2100 for global trends, climate modeling, etc.
  - From 1958 to 2000, world population more than doubled, but ...
  - UN's 1958 projection of 2000 world population was accurate to within 4%

Traditional Methods for Population Projections

- Traditional method (UN to 2008; most national offices currently) starts from current population estimates and projects population forward deterministically using the cohort component method.
- Requires assumptions about future fertility, mortality, migration:
  - Often produced by in-house experts or expert panels
- Uncertainty communicated by scenarios:
  - e.g. UN traditionally has published High, Medium, Low variants
  - High, Low: Medium fertility $\pm$ half a child per woman
  - No probabilistic basis
  - Can be implausible over multiple projection periods

Issues with Current Methods




SCIENCE'S COMPASS

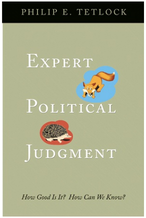
POLICY FORUM: DEMOGRAPHY

Broken Limits to Life Expectancy

Jim Oeppen and James W. Vaupel*

Is life expectancy approaching its limit? Many—including individuals planning their retirement and officials responsible for health and social policy—believe it is. The evidence suggests otherwise.

in income, salubrity, nutriti sanitation, and medicine, varying over age, period, col disease (4). Before 1950, m in life expectancy was due

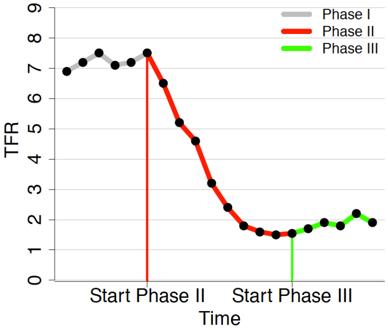


- Methods from 1940s. Successful overall. But ...
- Expert panels good at assessing science and data, designing models
- But not so good at producing forecasts from scratch:
  - Meehl (1954): Statistical models beat expert forecasts
  - Oeppen and Vaupel (2002): Demographic expert forecasts poor historically.
  - Tetlock (2005): Dart-throwing chimpanzees beat expert forecasts
- Not probabilistic. Need:
  - general assessment of accuracy
  - assessing changes and differences between outcomes and expectations
  - making decisions that avoid *risks*

Probabilistic Population Projections: Overview

- Probabilistic projections of each of the 3 components of population change: fertility, mortality, migration
- Probabilistic projections of TFR from a statistical model $\longrightarrow$ sample of trajectories from the joint predictive distribution of future TFR for each country and period
  - Convert each trajectory to age-specific fertility rates
- Similar approach for mortality rates
- Apply projection model to each sample
  - Yields many possible population futures of the world
- Method assessed by out-of-sample prediction for 5, 10, . . . , 30 years:
  - Predictions better than UN's previous method
  - Projection intervals reasonably well calibrated

Probabilistic TFR Projections



(*Source:* Alkema et al, 2011, *Demography*)

- 3 phases:
  - Phase I: high fertility pre-transition
  - Phase II: fertility transition/decline to below replacement level
  - Phase III: low fertility post-transition turnaround and fluctuations
- Fertility transition has started in all countries $\Longrightarrow$ Phase I not modelled

Phase II model: Fertility Transition

- Fertility decline:
  - starts slowly
  - accelerates
  - decelerates
  - stops below replacement level
- Expected 5-year declines in TFR modelled by a double logistic function (sum of two logistic functions) for each country
  - Flexible 5-parameter function
  - General form:

$$g(f) = \frac{d}{1 + \exp\left[-\frac{(f - a_2)}{a_1}\right]} - \frac{d}{1 + \exp\left[-\frac{(f - a_4)}{a_3}\right]}.$$

  - Interpretation:
    - d = upper bound
    - a_1 represents time taken for upswing
    - a_2 = middle of upswing
    - a_3 represents time taken for downswing
    - a_4 = middle of downswing
- Demo 3

Parameterization of Double Logistic Function for TFR

- If we write a single logistic function as

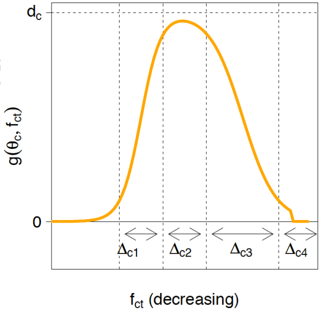
$$L(f) = \frac{d}{1 + \exp\left[-\frac{2\log(p)}{\Delta}(f - f_{50\%})\right]},$$

  - the function increases from 0 to d
  - the midpoint of the increase is at $f_{50\%}$, such that $L(f_{50\%}) = \frac{d}{2}$
  - Δ is the length of the interval in which $L(\cdot)$ increases from $\frac{1}{p+1}d$ to $\frac{p}{p+1}d$.
  - Thus setting $p = 9$ gives $\Delta = f_{90\%} - f_{10\%}$,
  - called the 80% range of the logistic function
- We write the expected decline in TFR as a function of current TFR, f:
  - $g(f;\theta) =$ Expected decline

$$= \frac{d}{1 + \exp\left(-\frac{2\ln(9)}{\triangle_1}(f - \sum_{i=2}^{4}\triangle_i + 0.5\triangle_1)\right)} - \frac{d}{1 + \exp\left(-\frac{2\ln(9)}{\triangle_3}(f - \triangle_4 - 0.5\triangle_3)\right)},$$

  where $\theta = (d, \triangle_1, \triangle_2, \triangle_3, \triangle_4)$.

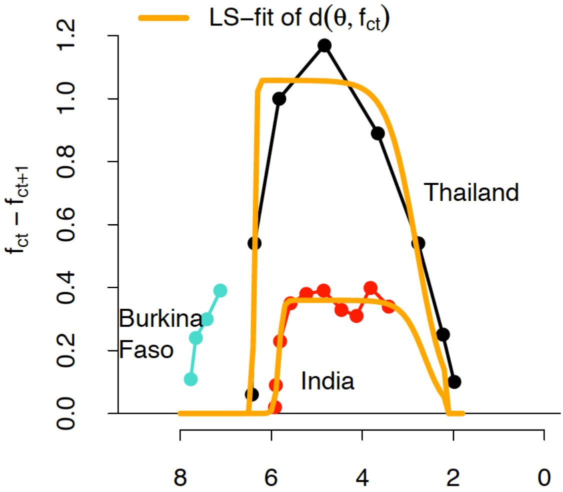
Double Logistic Function for TFR



- Interpretation:
  - d is the upper bound of the expected decline
  - Midpoint of the first (right-hand) logistic function is $0.5\Delta_3 + \Delta_4$.
  - Δ_3 is the 80% range of the first logistic function
  - Midpoint of the 2nd logistic function is $0.5\Delta_1 + \Delta_2 + \Delta_3 + \Delta_4$
  - Δ_1 is the 80% range of the second (left-hand) logistic function.

Probabilistic Model for TFR Decline

- Made probabilistic by adding a random error term
- $\Longrightarrow$ random walk with non-constant drift



Estimation: Bayesian Hierarchical Model

- Separate estimation for each country not feasible because of few data and only part of the evolution observed
- Solution: For each country, draw on information from other countries
- Hierarchical model:
  - Double logistic parameters for a country distributed about "world average" $\Longrightarrow$ "prior" $\approx$ range of possible curves for the country
  - World parameters estimated
  - "Prior" refined by country's historical experience
  - Result: Estimate for a specific country $\approx$ weighted average of world average and estimate based on its data only
- Bayesian estimation using Markov chain Monte Carlo (MCMC)
- Gives a sample of many possible future trajectories of TFR in all countries and periods
- Between-country correlation in forecast errors included in projection model (Fosdick & Raftery 2014)

Bayesian Hierarchical Models: Review of Basic Ideas

- Introduction to Bayesian statistics: Hoff, P.D. (2009). *An Introduction to Bayesian Statistics*.
- Data y_i for observations i (e.g. school students), where each observation belongs to a group $j[i]$.
  - Basic Analysis of Variance (ANOVA) model:

$$\begin{aligned} y_i &= \alpha_{j[i]} + \varepsilon_i,\ i = 1, \dots n, \\ \varepsilon_i &\overset{\text{iid}}{\sim} N(0, \sigma_y^2). \end{aligned}$$

- (Non-Bayesian) random intercept model/ANOVA random effects model:

$$\begin{aligned} y_i &= \alpha_{j[i]} + \varepsilon_i,\ i = 1, \dots n, \\ \varepsilon_i &\overset{\text{iid}}{\sim} N(0, \sigma_y^2), \\ \alpha_j &\overset{\text{iid}}{\sim} N(\mu_\alpha, \sigma_\alpha^2),\ j = 1, \dots, J. \end{aligned}$$

- What are the hyperparameters?

Random Effects ANOVA: Estimating Hyperparameters by MLE

- Hyperparameters? $\sigma_y^2, \sigma_\alpha^2, \mu_\alpha$.
- These can be estimated by MLE, e.g. using `lme4` R package.
- To get the likelihood, integrate out the random effects α_j analytically, and maximize over the hyperparameters.
- MLE ignores uncertainty about them.
- MLE hard for complex models
- Can have convergence problems for optimization in degenerate situations.
- Doesn't answer questions like, is School 3 performing less well than School 30?

Priors for Bayesian Random Intercept Model

- Conventional prior distributions for hyperparameters:

$$
\begin{aligned}
1/\sigma_y^2 &\sim \mathrm{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \\
1/\sigma_\alpha^2 &\sim \mathrm{Gamma}(\eta_0/2, \eta_0\tau_0^2/2) \\
\mu_\alpha|\sigma_\alpha^2 &\sim N(\mu_0, \sigma_\alpha^2/\kappa_0)
\end{aligned}
$$

- One possible choice: Unit information prior:
  - $\nu_0 = \eta_0 = 1$
  - $\sigma_0^2 = \hat{\mathrm{Var}}(y_i)$
  - $\tau_0^2 =$ empirical variance of the group means, $\bar{y}_j$.
  - $\mu_0 =$ empirical mean of the group means, $\bar{y}_j$..
  - $\kappa_0 = 1$.

Priors for Bayesian Random Intercept Model (ctd)

- Other priors are possible also. E.g. Gelman & Hill (2007) use

$$\begin{aligned} \sigma_y &\sim \text{Uniform}(0, 100) \\ \sigma_\alpha &\sim \text{Uniform}(0, 100) \\ \mu_\alpha &\sim N(0, 100^2) \end{aligned}$$

  - But the choices of 100 may not be appropriate
- Figure of Bayesian hierarchical model

Inference for Hierarchical Models

- For most Bayesian hierarchical models, the posterior distribution is not available in analytic form.
- Direct Monte Carlo inference is also hard.
- So instead we use a different form of Monte Carlo inference: Markov chain Monte Carlo (MCMC)
  - Instead of simulating independent samples from the posterior distribution (difficult)
  - We simulate a Markov chain that converges to the posterior distribution
  - This gives a *dependent* sequence of simulations that are approximately drawn from the posterior
  - The simplest form of MCMC is the Gibbs sampler

Gibbs Sampler

- Suppose we want to simulate from a joint posterior distribution of a vector of parameters $\phi = (\phi_1, \ldots, \phi_p)$.
- Suppose that we can simulate from the conditional distribution of each ϕ_j given all the other parameters and the data.
- Then the Gibbs sampler goes as follows:
  1. Pick a starting point $\phi^{(0)} = \{\phi_1^{(0)}, \ldots, \phi_p^{(0)}\}$
  2. Generate $\phi^{(s)}$ from $\phi^{(s-1)}$ as follows:
     - 2.1 sample $\phi_1^{(s)} \sim p(\phi_1 | \phi_2^{(s-1)}, \phi_3^{(s-1)}, \ldots, \phi_p^{(s-1)})$
     - 2.2 sample $\phi_2^{(s)} \sim p(\phi_2 | \phi_1^{(s)}, \phi_3^{(s-1)}, \ldots, \phi_p^{(s-1)})$

       $\vdots$
     - p. sample $\phi_p^{(s)} \sim p(\phi_p | \phi_1^{(s)}, \phi_2^{(s)}, \ldots, \phi_{p-1}^{(s)})$
- This algorithm generates a *dependent* sequence of vectors:

$$\begin{aligned} \phi^{(1)} &= \{\phi_1^{(1)}, \ldots, \phi_p^{(1)}\} \\ \phi^{(2)} &= \{\phi_1^{(2)}, \ldots, \phi_p^{(2)}\} \\ &\vdots \\ \phi^{(S)} &= \{\phi_1^{(S)}, \ldots, \phi_p^{(S)}\} \end{aligned}$$

Convergence of Gibbs Sampler

- Under certain conditions, no matter what $\phi^{(0)}$ is,

$$\Pr(\phi^{(s)} \in A) \to \int_A p(\phi)\, d\phi \quad \text{as } s \to \infty.$$

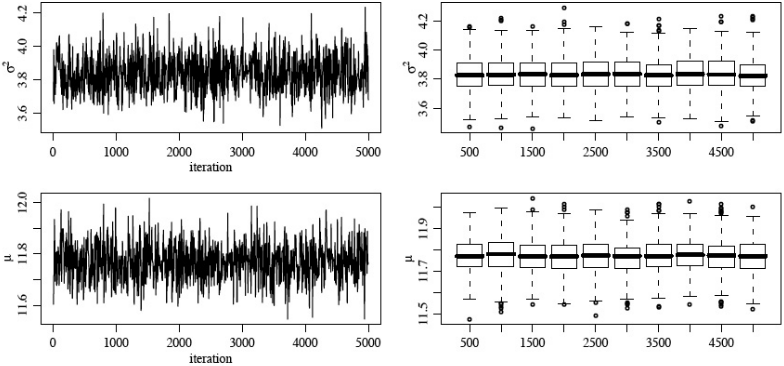
  - This means that the probability that the s-th iteration is in the set A gets close to the true posterior probability that ϕ is in the set A as s gets large.
  - Thus the *sampling distribution* of $\phi^{(s)}$ approaches the *true posterior distribution* of ϕ as $s \to \infty$.

- For most functions g of interest,

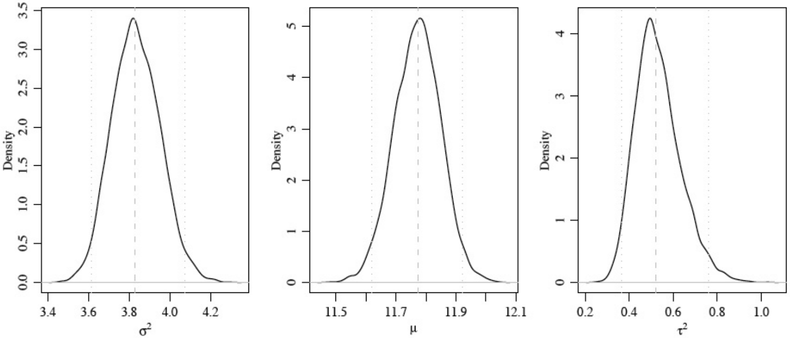
$$\frac{1}{S}\sum_{s=1}^{S} g(\phi^{(s)}) \to E(g(\phi)) = \int g(\phi)p(\phi)\, d\phi \quad \text{as } S \to \infty$$

  - This means we can approximate $E(g(\phi))$ with the sample average of $\{g(\phi^{(1)}),\ldots,g(\phi^{(S)})\}$

Gibbs Sampling Results for Netherlands Schools



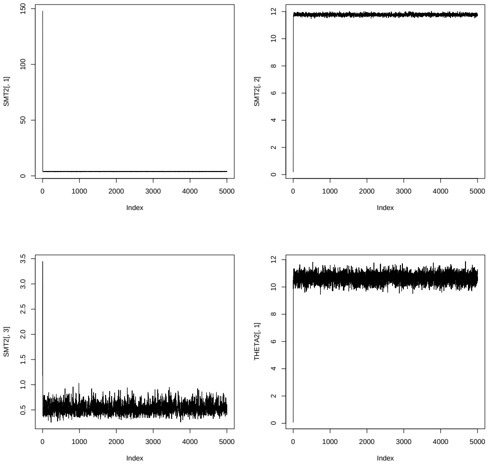
Marginal Posterior Distributions



Convergence and Mixing

- How many iterations are needed in the Gibbs sampler?
  - This is really two different questions.
  - Convergence: Has the sampler converged to the part of parameter space with high posterior probability?
  - Mixing: Have there been enough iterations to estimate quantities of interest with enough accuracy? (or to "explore the posterior distribution")
  - Software: R coda package
- Convergence:
  - Visual inspection of trace plots
  - Gelman-Rubin diagnostic
  - Remove the initial "burn-in" period

Burn-in by visual inspection



- burn-in = 6 iterations

Convergence: Gelman-Rubin Diagnostic

- This is based on parallel chains from multiple starting values
  - It is a normalized version of the ratio of the between chain variance to the average within-chain variance
  - If it is much above 1 (conventionally above 1.1), the chains may not have converged.
  - The width of the posterior interval for the mean of the chain (i.e. the parameter estimate) can be reduced by this factor
  - It can be computed by `gelman.diag` in the `coda` R package.
- For the Netherlands school example, for one run with 2 chains and 5000 iterations (one with a good and one with a bad starting value):
  - For all the 132 α_j's, the Gelman-Rubin diagnostic was less than 1.02, so there was no problem with these.
  - For σ_y^2, μ_α and σ_α^2, it was 1.29, 1.23, 1.01.
  - So there may not have been full convergence

Mixing: Raftery-Lewis Diagnostic

- How many iterations are needed to estimate a posterior quantile of a parameter to a desired accuracy?
- The Raftery-Lewis diagnostic answers this using Markov chain theory
- It finds how many iterations are needed to estimate quantile q to within $\pm r$ with probability s.
- The default is $q = .025$, $r = .005$, $s = .95$.
- But I now think that $r = .0125$ is small enough.
  - For a perfectly independent chain (the ideal), this gives the answer 600 iterations
- I suggest running it for the hyperparameters for both $q = .025$ and $q = .975$, and also for some of the random effects.
- The required number of iterations is the maximum of these.
- It can be computed by `raftery.diag` in the `coda` R package.

Software for Bayes and MCMC

- WinBUGS (and OpenBUGS, BUGS called from R, etc): good for hierarchical models
- JAGS: a more recent version of WinBUGS
- MCMCpack R function: Good for regression-type models and extensions
- MCMCglmm R function: Bayesian generalized linear multilevel (mixed) models
- MLwin: For multilevel models
- STAN: MCMC for general Bayesian models (not necessarily hierarchical), but can be used for Bayesian hierarchical models.
- NIMBLE: Fast MCMC inference for Bayesian hierarchical models
- INLA: Analytic approximation instead of MCMC
- TMB: Similar to INLA
- Direct programming in R: More labor-intensive but may be necessary for more complex models
- Demo 4