CS&SS/STAT/SOC 563
Statistical Demography
Spring 2022

- Syllabus (except outline)
- Introductions
- Outline
- Timeline
- Term paper
- Twitter: @AdrianRaftery1 (mostly current scientific journal articles on statistical demography and related topics)
  - Twitter feed also available from my website, https://www.stat.washington.edu/raftery

Demography

- Demography (from Ancient Greek $\delta\hat{\eta}\mu o\varsigma$ = people) = statistical study of human populations.
- Formal demography:
  - Estimating current and past populations by age and sex, fertility, mortality and migration
  - Projecting future populations by age and sex
  - One of the most successful areas of social science
  - Used by most governments and international organizations, and by the private sector and social and health researchers
  - Relies heavily on deterministic mathematical models
- Social demography:
  - Explaining trends in fertility, mortality and migration
  - Studying consequences of changes in fertility, mortality and migration for society and individual behavior
  - Relies heavily on statistical models

Demography as a Discipline

- A bit of history:
  - Life table invented by John Graunt in 1662 in London.
  - Much 18th & 19th century work driven by insurance problems.
  - International Union for the Scientific Study of Population (IUSSP) founded in 1927 by Margaret Sanger.
  - Population Association of America (PAA) founded in 1930.
- Largely an *interdiscipline*, involving sociologists, economists, statisticians, anthropologists, geographers, historians, biologists, public health researchers, ...
  - Much demography done in national statistical agencies (e.g. U.S. Census Bureau) and international organizations (e.g. UN, UNICEF, WHO, UNAIDS, FAO, World Bank).
  - In US research universities, mostly organized in NIH-funded research centers rather than departments
  - Associations and meetings: PAA, IUSSP, national associations.
  - Journals: *Demography* (US), *Population and Development Review* (International: Population Council), *Demographic Research* (Germany), *Population Studies* (UK), *Population* (France).

Software and Data Used in the Course

- `demography` R package: Many demographic functions and data.
- `MortCast` R package: many functions for analyzing mortality rates.
- `demogR` R package: Many demographic functions and data.
- `bayesPop` R package and dependencies: Bayesian population projections
- Human Mortality Database (HMD)
- Human Fertility Database (HFD)
- `HMDHFDplus` R package: Functionality for accessing and manipulating HMD and HFD
- `wpp2019` R package: Population, mortality, fertility and migration by age and sex for all countries, 1950–2020, from the 2019 revision of the UN's *World Population Prospects*. Also `wpp2017`, `wpp2015`, `wpp2012`, `wpp2010`, `wpp2008` are sometimes useful.
- `wppExplorer` R package and online implementation: Visualization of `wpp2019` data.
- WPP 2019 data: https://population.un.org/wpp/

Review of Formal Demography: Mortality

- We start with mortality in terms of continuous age.
  - Common in medical research/biostatistics: survival analysis.
  - Engineering: reliability theory (time to malfunction of a machine or system).
- Basic primitive: instantaneous mortality rate (aka hazard rate, force of mortality, failure rate).
  - In demography, a rate is defined as the number of events per unit of something
  - Mortality rate: units = deaths per unit time.
  - For a given period of time of length h, mortality rate at age x

$$= \Pr[\text{dies in } [x, x+h) \mid \text{alive at age } x]/h.$$

  - Instantaneous mortality rate at age x

$$\mu(x) = \lim_{h \to 0} \Pr[\text{dies in } [x, x+h) \mid \text{alive at age } x]/h.$$

Simplest case: constant mortality rate

- $\mu(x) = \lambda$, a constant independent of age
- Let $X =$ age at death (a random variable).
- What probability distribution does X have? Exponential distribution.
- What is the probability density function (pdf) of X?

$$f(x) = \lambda e^{-\lambda x}.$$

- What is the cumulative distribution function (cdf) of X?

$$F(x) = \Pr[X \leq x] = \int_{u=0}^{x} f(u)du = 1 - e^{-\lambda x}.$$

- What is the survival function of X?

$$S(x) = \Pr[X > x] = 1 - F(x) = e^{-\lambda x} \quad (1)$$

- Plots
- Is this a realistic distribution in general? Ever?

General case: Mortality rate depends on age

- Mortality rate starts high at birth, declines, then rises steadily from the end of middle age.
  - There may be an "accident hump" in the middle too
- In general, let $X =$ age at death (a random variable) for a cohort
- Survival function $S(x) = \Pr[X > x]$
- probability density function (pdf): $f(x) = \frac{dF(x)}{dx} = -\frac{dS(x)}{dx}$
- $S(x) = \int_x^\infty f(u)du$
- Mortality rate

$$\begin{aligned}
\mu(x) &= \lim_{h\to 0} \frac{\Pr[x \le X < x+h | X \ge x]}{h} \\
&= \lim_{h\to 0} \frac{\Pr[x \le X < x+h]}{h \Pr[X \ge x]} \\
&= \frac{f(x)}{S(x)} \\
&= -\frac{d}{dx} \log S(x) \qquad (2)
\end{aligned}$$

- Integrating with respect to x and using $S(0) = 1$, we get

$$\begin{aligned} S(x) &= \exp\left[-\int_0^x \mu(u)du\right] \\ &= \exp[-\Lambda(x)], \end{aligned} \tag{3}$$

  where $\Lambda(x) = \int_0^x \mu(u)du$ is the *cumulative hazard function*
- By (2), any function $\mu(\cdot)$ can be a mortality rate/hazard function if:
  1. $\mu(x) \geq 0 \ \ \forall x \geq 0$
  2. $\int_0^\infty \mu(x)dx = \infty$.
  3. Why? Because $S(\infty) = 0$, and so by (3), $\Lambda(\infty) = \infty$.
- Does $\mu(x)$ have to be less than 1 for all x? No.
  - Do mortality probabilities such as $F(x)$ or $S(x)$ have to be less than 1? Yes.
  - E.g. if $\mu(x) = 2$, then $F(1) = 1 - e^{-2} = 0.86$.
  - So then $\mu(x) > 1$ but $F(1) < 1$.

Life Expectancy

- Life expectancy at age $x = e_x = E[X - x | X \geq x]$
- $e_x = \frac{\int_x^\infty (u-x) f(u) du}{S(x)}$
- By integration by parts, using the fact that $E(X) < \infty$, and so $\lim_{x\to\infty} S(x) = 0$, we get

$$e_x = \frac{\int_x^\infty S(u) du}{S(x)}. \tag{4}$$

- Setting $x = 0$ in (4) and using the fact that $S(0) = 1$ gives

$$e_0 = E(X) = \int_0^\infty S(u) du.$$

  - e_0 is called the *life expectancy at birth*.
  - What is e_0 for the exponential distribution? $1/\lambda$.
  - What is e_x for the exponential distribution? $1/\lambda$.

Demographic Data: Discrete Time, Life Tables

- Problem: demographic data usually not measured in continuous time, but in time intervals, often of 1 or 5 years.
  - A common set of intervals is ages 0, 1-4, 5-9, 10-14, ..., 95-99, 100+.
  - So we need a discrete-age version of the survival analysis theory.
- Life table for a cohort (or complete life table):
  - $\ell_x = \#$ people surviving to age x for $x = 0, 1, 2, \ldots$.
  - Note that $\frac{\ell_x}{\ell_0} = S(x)$.
  - ℓ_0 is called the *radix* of the life table.
  - Often $\ell_0 = 100,000$ (this is arbitrary). Or $\ell_0 = 1$.
  - ${}_nd_x = \ell_x - \ell_{x+n}$ is the number of deaths between age x and age $x + n$.
  - ${}_nq_x = \frac{{}_nd_x}{\ell_x} = \frac{\ell_x - \ell_{x+n}}{\ell_x} = \frac{S(x) - S(x+n)}{S(x)}$ is the probability that a person aged x dies within n years.

Age-Specific Death Rates in the Life Table

- A demographic rate (mortality, fertility, migration, marriage) is usually defined as

$$\frac{\text{Number of events}}{\text{Number of person-years at risk of the event}}.$$

  - It is in units of events per person per year (i.e. events per person-year).
  - Sometimes multiples of these units are used.
  - For example, events per five-year period rather than per year.
  - Or, events per 1,000 people.
  - It is important to get the units right. In research, getting the units wrong is a common source of mistakes.

- The age-specific mortality rate between ages x and $x + n$ is

$${}_nm_x = \frac{\#\text{deaths from age } x \text{ to age } x+n}{\#\text{person-years lived from age } x \text{ to age } x+n}.$$

- The number of person-years lived from age x to age $x+n$ is

$$ {}_nL_x = \int_x^{x+n} \ell_u \, du, $$

  treating ℓ_u for a moment as a continuous function of u.
  - ℓ_u is often well approximated by a linear function over the interval $[x, x+n)$, if n is fairly small.
  - We have $\ell_{x+n} = \ell_x - {}_nd_x$.
  - Thus ${}_nL_x \approx n\,\ell_x - \frac{1}{2}\,n\,{}_nd_x$.
- Thus the age-specific mortality rate is

$$ {}_nm_x = \frac{{}_nd_x}{{}_nL_x} \approx \frac{{}_nd_x}{n\,\ell_x - \frac{1}{2}\,n\,{}_nd_x}. $$

- What's the relationship between ${}_nm_x$ and ${}_nq_x$?
  - ${}_nm_x \approx \frac{{}_nq_x}{n(1-\frac{1}{2}\,{}_nq_x)}$.
  - So, is ${}_nm_x$ bigger or smaller than ${}_nq_x/n$? Bigger.
  - ${}_nq_x \approx \frac{n\,{}_nm_x}{1+n\,{}_nm_x/2}$.

Life Tables from Data

- It's fairly easy to estimate life table quantities for countries with good vital registration systems and regular censuses.
  - Of 201 countries over 100,000, only about 90, or 45%, satisfy this.
- Suppose we observe the mortality rate in a real population during a given period:

$$ {}_nM_x = \frac{\#\text{ deaths of people between ages } x \text{ and } x+n}{\#\text{ person-years lived between ages } x \text{ and } x+n}. $$

  - Is ${}_nM_x$ the same as ${}_nm_x$? No.
  - The difference lies in the underlying age distribution of the population.
  - In a country like the US, we can get the numerator from vital registration data.
  - We can get the denominator from a combination of censuses and vital registration data
  - ${}_nM_x$ is in what units? Deaths per person per year.
  - How can we infer ℓ_x?
  - Equivalently, how can we infer ${}_nq_x$?

- The big problem is that we want ${}_nq_x$, while data give us ${}_nM_x$.
- Suppose the distribution of ages a in the population is $p(a)$. Then

$$ {}_nM_x = \frac{\int_x^{x+n} \mu(a)p(a)da}{\int_x^{x+n} p(a)da}. \tag{5} $$

  - Problem: We typically don't know $p(a)$ within 1-year or even 5-year intervals.
- Solution 1: Assume the mortality rate is constant over time and within the age interval of interest. This is usually approximately true.
  - Then the mortality rate is constant and equal to what at all ages $a \in [x, x+n)$? ${}_nM_x$
  - Then by (3) for the constant mortality rate (exponential) distribution: $S(x) = e^{-\lambda x}$, with x replaced by n, λ replaced by ${}_nM_x$, and $S(n)$ replaced by ${}_nq_x$:

$$ {}_nq_x \approx 1 - \exp\left(-n\ {}_nM_x\right). \tag{6} $$

  - If we have ${}_nq_x$ for all age intervals $[x, x+n)$ in the life table, then we can construct the life table.

- Solution 2: Assume
  1. $p(a) \propto \ell_a$ for all $a \in [x, x+n)$;
  2. ℓ_a is a linear function of a for $a \in [x, x+n)$.
- Then by (5),

$$ {}_nM_x = \frac{\ell_x - \ell_{x+n}}{n\frac{1}{2}(\ell_x + \ell_{x+n})}. $$

- Thus

$$ \frac{\ell_{x+n}}{\ell_x} = \frac{1 - n\ {}_nM_x/2}{1 + n\ {}_nM_x/2}. $$

- It follows that

$$ \begin{aligned} {}_nq_x &= 1 - \frac{\ell_{x+n}}{\ell_x} \\ &= \frac{n\ {}_nM_x}{1 + n\ {}_nM_x/2} \end{aligned} \tag{7} $$

- Illustration: Suppose ${}_5M_x = 0.01$.
  - Then Solution 1 (6) gives

$$ {}_5q_x \approx 1 - e^{-5\times.01} = 0.04878. $$

  - Solution 2 (7) gives

$$ {}_5q_x \approx \frac{5\ {}_5M_x}{1 + 5\ {}_5M_x/2} = 0.04877. $$

  - So the two approximations are very close.
  - But this isn't always the case, especially when ${}_nM_x$ is large.

Life Tables from Data: Example

- Data (from WPP 2015):
  - Between July 1, 2010 and June 30, 2015, 927K women aged 90-94 died in the US.
  - On July 1, 2010, there were 1041K women aged 90-94 in the US.
  - On June 30, 2015, there were 1182K women aged 90-94 in the US.
  - We seek ${}_5q_{90}$ for women in 2010-2015.
- Person-years of exposure (in 000s) $\approx 5 \times \frac{1}{2}(1041 + 1182) = 5557.5$.
  - So ${}_5M_{90} = \frac{927}{5557.5} = 0.167$ units? deaths per person-year.
  - So by approximation 1 (6),

$$ {}_5q_{90} \approx 1 - e^{-5*.167} = 0.567. $$

    Units?
  - By approximation 2 (7),

$$ {}_5q_{90} \approx \frac{5\ {}_5M_{90}}{1 + 5\ {}_5M_{90}/2} = 0.589. $$

  - The two approximations more different than before, but still close.

- If we use time units equal to five-year periods, then the observed mortality rate in deaths per person per five-year period was

$$\frac{927}{\frac{1}{2}(1041 + 1182)} = 0.834.$$

  - Compare with ${}_5q_{90} \approx 0.589$. Very different.
  - So when the mortality rate is high, the mortality rate per person per period can be very different from the one-period probability of death.
  - When the mortality rate is low, they will be very similar.

Cohort and Period Life Tables

- So far we have talked about a *cohort* (or complete) life table, based on the deaths of a group of people born in the same period.
  - This is fully observed only when all the people in the cohort have died.
- Much demography is based instead on *period* life tables, and other period measures.
  - This is based on the deaths and person-years lived at each age in a population over a given period (e.g. year or 5-year period).
  - ${}_nq_x$ is calculated for each age group $[x, x+n)$, as in our example.
  - Then these calculated ${}_nq_x$ values are used to construct a *period life table*.
  - This can be used to compute period summary measures, such as life expectancy at birth, e_0, for the period of the data.
  - If age-specific mortality rates are changing over time, this life table does not reflect the mortality experience of any one group of people.
- The *Lexis diagram* is useful for computing period and cohort measures from data.

Lexis Diagram

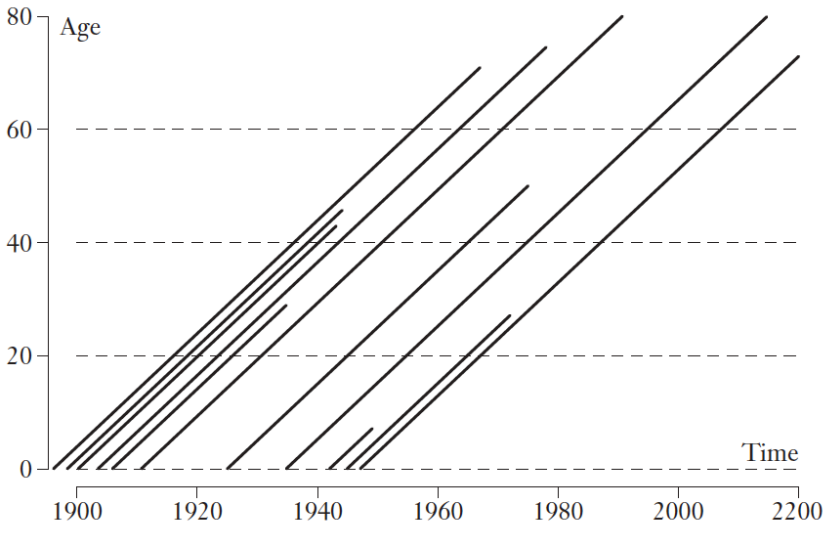


Figure 2.1 A Lexis diagram

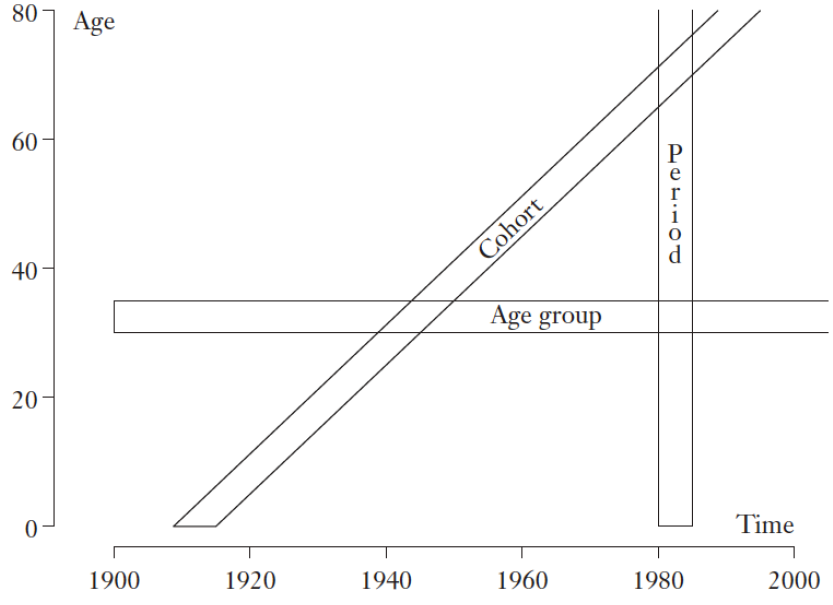


Figure 2.2 Cohort, period, and age

- Demo 1

Sampling Error Often Ignored

- Sampling error often ignored (as in example) because of large numbers.
- In the example, conditionally on the true probability of death:

$$\#\ \text{deaths} \sim \text{Binomial}(1111.5K, {}_5q_{90}).$$

- So SE $(\widehat{{}_5q_{90}}) \approx \sqrt{\frac{\widehat{{}_5q_{90}}(1-\widehat{{}_5q_{90}})}{1111.5K}} = 0.00047$.
- So a CI for ${}_5q_{90}$ is about 0.455 ± 0.001, i.e. [0.454, 0.456].
- Sampling error (0.001) is dwarfed by approximation error due to unknown age distribution within the 90–94 age interval.
  - Difference of $0.589 - 0.567 = 0.022$ between our two approximations.
  - Other sources of uncertainty, especially when estimated from a survey (not vital registration):
    non-representativeness of survey; systemmatic response errors (biases).

Sampling Variation (ctd)

- For projecting the future, sampling variation also dwarfed by uncertainty about the future mortality rate.
- Ignoring sampling variation thus makes almost no difference, but simplifies calculations greatly.
- By contrast, for small populations, it *is* essential to take account of the binomial sampling variation. Example?
  - Forecasting school enrollments
  - Forecasting populations of small cities
- Roughly, when the total population is over 100,000, it's OK to ignore sampling variation.
  - When the total population is below 20,000, you need to incorporate it.
  - In between is a grey area

Fertility Rates

- Crude birth rate = # births / # person-years in years $[T_1, T_2]$.
  - Does this reflect fertility accurately?
  - No. It depends on the proportion of women of reproductive age.
  - And on the age distribution of these women.
- Age-specific fertility rate:

$$ {}_nF_x[T_1, T_2] = $$

$$ \frac{\text{Number of births to women aged } x \text{ to } x+n \text{ in years } T_1 \text{ to } T_2}{\text{Number of person-years lived by women aged } x \text{ to } x+n \text{ in years } T_1 \text{ to } T_2} $$

Total Fertility Rate

- Total fertility rate:

$$\mathrm{TFR}[T_1, T_2] = n \times \sum_{\text{all age groups } [x, x+n]} {}_nF_x[T_1, T_2]$$

- Interpretation:
  - The number of children a woman would bear if she survived the reproductive interval (i.e. typically to age 49), and experienced at each age interval, x to $x + n$, the fertility rate ${}_nF_x[T_1, T_2]$.
  - If $T_1 = T_0 + x$ and $T_2 = T_0 + x + n$, then $\mathrm{TFR}[T_1, T_2]$ is the *cohort TFR* for women born in year T_0.
  - It is more usual to use the *period TFR*, where T_1 and T_2 are the same for all ages.
  - This gives the number of children a woman would bear if she survived the reproductive interval, and experienced at each age interval x to $x + n$ the fertility rate of women of that age in period $[T_1, T_2]$.
  - Would this reflect the fertility of any real group of women?
    No, unless fertility rates were not changing over time.
- TFR does not depend on the age distribution of the population.

Proportional Age-Specific Fertility Rates (PASFR)

- Defined as

$$ {}_n\mathrm{PASFR}_x = \frac{n \times {}_n F_x[T_1, T_2]}{\mathrm{TFR}[\mathrm{T}_1, \mathrm{T}_2]} $$

- $\sum_{\text{all age groups } [x, x+n]} {}_n\mathrm{PASFR}_x = 1.$

Migration Rates

- Definitions:
  - Internal migration: Migration from one part of a country to another.
  - International migration.
    - UN definition: A migrant moves to another country and stays there for at least 12 months.
    - Individual countries use different definitions.
    - E.g. Poland: 3 months.
  - Stock of migrants: Number of people living in a country who were born in another country.
  - Flow of migrants: Number of people who migrate in a given time period.
- What proportion of the world's population is not living in their country of birth? About 3%.
- Emigration rate: ${}_nE_x[T_1, T_2]$ = number of emigrants aged x to $x + n$ in period $[T_1, T_2]/(T_2 - T_1)$ (i.e. emigrants per year).
  - Emigration rate per person-year = Emigration rate divided by the average number of people living in the country between times T_1 and T_2.

Immigration Rate

- Immigration rate:

$$ {}_nI_x[T_1, T_2] = \frac{\text{number of immigrants aged } x \text{ to } x+n \text{ in period } [T_1, T_2]}{(T_2 - T_1)}, $$

  Units? Immigrants per year.
- Is this a demographic rate in the sense of events/person-years at risk of the event? No.
- Would it be be a true demogrphaic rate if it was divided by the number of person-years lived in the country between times T_1 and T_2?. No. Why not? Because the inhabitants of the country or area are not at risk of immigration.
- One could express it as a rate for the rest of the world (by dividing by the number of person-years lived in the rest of the world between times T_1 and T_2). Would this be useful?
- Generally, larger countries tend to have larger immigration flows on average, so there could be some argument for dividing ${}_nI_x[T_1, T_2]$ by the number of person-years in the receiving country. It would not be a demographic rate in the strictest sense, though.

Net Migration Rate

- Age-specific net migration rate:

$$
\begin{aligned}
{}_nG_x[T_1, T_2] &= {}_nI_x[T_1, T_2] - {}_nE_x[T_1, T_2] \\
&= \frac{\#\text{immigrants} - \#\text{emigrants aged } x \text{ to } x+n \text{ in } [T_1, T_2]}{(T_2 - T_1)}
\end{aligned}
$$

  Units? Net migrants per year.

- Total net migration rate: $G[T_1, T_2] = \sum_x {}_nG_x[T_1, T_2]$. Units?
- In WPP, the UN reports the net migration rate in a country over a given five-year period as
  - $G[T_1, T_2]/((T_2 - T_1) \times$ # person-years lived in the country in $[T_1, T_2]/1,000)$.
  - = net number of migrants per year per 1,000 population in the receiving country. Units?

Population Projection: Demographic Balancing Equation

For the population projection section, we will simplify notation initially by giving results for one-year age groups and one-year time periods.

- Thus the subscript n in ${}_nq_x$ and ${}_nF_x$ will be taken to be $n = 1$ and suppressed.
- This will work with no change if projections are in n-year periods for n-year age groups.
- Then everything is in terms of time periods.
- For example, the UN produces projections for 5-year periods and 5-year age groups (i.e. time unit = 5 years).
- We will denote time periods by t.
- Some complication for the common age-group set 0, 1-4, 5-9, etc.

Demographic Balancing Equation (ctd)

- Demographic balancing equation:

$$N_{t+1} = N_t + B_t - D_t + I_t - E_t, \quad \text{where}$$

  - N_t = population at time t
  - B_t = births in time interval $(t, t+1]$
  - D_t = deaths in time interval $(t, t+1]$
  - I_t = immigration in time interval $(t, t+1]$
  - E_t = emigration in time interval $(t, t+1]$

Age-Structured Demographic Balancing Equation

- Let $N_{x,t}$ = population of age $x = 0, 1, \ldots, (A-1)+$ in year t
- $\boldsymbol{N}_t = (N_{0,t}, N_{1,t}, \ldots, N_{A-1,t})^T$
- Consider only one sex (female), and assume the population is closed to migration. Then

$$N_{0,t+1} \approx \sum_{x=1}^{A-1} \tilde{F}_x N_{x,t},$$

  where $\tilde{F}_x$ = expected number of female births to a woman aged x last birthday, who survive to time $t+1$.
- Are $\tilde{F}_x$ and F_x the same? No. Why not? Three reasons:
  - $\tilde{F}_x$ refers to female births only, while F_x refers to all births.
  - Some women aged x may die before time $t+1$, and so not contribute a full person-year of exposure.
  - Some babies may die before time t_1.

Projecting the Number of Female Births

- We use

$$\tilde{F}_x = F_x \times \frac{1}{1+\mathrm{SRB}} \times \frac{1}{2}\left(1 + s_{x-1}\frac{N_{x-1,t}}{N_{x,t}}\right) \times (1 - q_0/2),$$

  - where SRB = sex ratio at birth = number of males born per female
  - The 1st factor is the age-specific fertility rate, F_x.
  - The 2nd factor adjusts for the fact we're only looking at female births.
  - The 3rd factor adjusts for female mortality during the interval.
  - The 4th factor, $(1 - q_0/2)$, is included to approximate the proportion of babies born during the year who die before time $t + 1$.
  - More exact expression in Preston et al, Section 6.3.1.
- What is the normal value of SRB? 1.05
  - What is the normal range of SRB? 1.04-1.06.

Projecting the Number of Women Aged x at the Next Time Period

- Let ${}_ns_x = 1 - {}_nq_x$, which is the probability that a woman aged x survives n years.
  - We abbreviate ${}_1s_x$ to s_x, which is the probability of a woman aged x now still being alive in one year, at age $x + 1$.
- Then

$$N_{x+1,t+1} = s_x N_{x,t}.$$

- For the highest age-group (ages $(A-1)+$), the approximation is often made that this has a constant survival rate, s_{A-1}. So

$$N_{A-1,t+1} = s_{A-2} N_{A-2,t} + s_{A-1} N_{A-1,t}.$$

  - Sometimes s_{A-1} is taken to be zero, meaning that everyone dies before reaching age A.
  - Usually A is taken high enough that it doesn't matter much (e.g. 100+ or more for humans).

Leslie Matrix

- These equations can be written in matrix form as

$$\boldsymbol{N}_{t+1} = L\,\boldsymbol{N}_t, \quad \text{where}$$

$$L = \begin{bmatrix} 0 & \tilde{F}_1 & \tilde{F}_2 & \cdots & \tilde{F}_{A-2} & \tilde{F}_{A-1} \\ s_0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & s_1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & s_{A-2} & s_{A-1} \end{bmatrix}$$

- Sometimes s_{A-1} is taken to be zero.
- L is the Leslie (or projection) matrix (Leslie 1945, *Biometrika*).
- Easy to construct by hand.
  - There is also software to do it, e.g. the *make.leslie.matrix* function in the `popReconstruct` R package.
  - The `demogR` R package also does it

Population Projection

- Population projection:
  - Projecting one period ahead:

$$\boldsymbol{N}_{t+1} = L\boldsymbol{N}_t.$$

  - Projecting k periods ahead:

$$\boldsymbol{N}_{t+k} = L^k\boldsymbol{N}_t.$$

- This is called the *cohort-component method of population projection* (CCMPP).

Rate of Increase and Stable Age Distribution

- Theorem: Asymptotically, as $t \to \infty$, $\boldsymbol{N}_t$ converges to

$$\lambda^t \boldsymbol{u}, \quad \text{where}$$

  - λ is the dominant right eigenvalue of L.
  - λ exists, is unique, and is real (by the Perron-Froebenius theorem).
  - $\log(\lambda)$ is the annual (instantaneous) rate of increase of the population
  - $\boldsymbol{u}$ is the dominant right eigenvector of L, called the *stable age distribution*.
- Remark: It's more accurate to write $(\boldsymbol{N}_t/\lambda^t) \longrightarrow \boldsymbol{u}$ as $t \longrightarrow \infty$.

Reproductive Value and Leslie Matrix

- Reproductive value v_a of someone aged a
  - = expected number of future offspring of an individual aged a (R.A. Fisher 1930)
  - Important in population genetics
  - Typically low at birth, increases to a peak soon after the start of reproduction, and then declines.
- Theorem: The vector $\boldsymbol{v} = (v_0, \ldots, v_{A-1})$ of reproductive values is the dominant left eigenvector of the Leslie matrix L.

Population Projection with Migration

- Migration is a continuous process over the time interval
- Some migrants won't survive to the end of the interval
- Some migrants may bear children who will survive to the end of the interval.
- Hard to model exactly
  - It can be done using Integral Projection Models (Easterling et al 2000).
  - R package IPMpack.
- We consider four discrete-time approximations.

Population Projection with Migration: Discrete-Time Approximations

- Approximation 1: Assume all the migration happens at the beginning of the period. Let G_t^F be the vector of age-specific net numbers of female migrants in the period $(t, t+1]$. Then

$$\boldsymbol{N}_{t+1} = L(\boldsymbol{N}_t + G_t^F).$$

- Approximation 2: Assume all the migration happens at the end of the period. Let G_t^F be the vector of age-specific net numbers of female migrants in the period $(t, t+1]$. Then

$$\boldsymbol{N}_{t+1} = L\boldsymbol{N}_t + G_t^F.$$

- Approximation 3: Assume half the net migration happens at the beginning of the interval, and half at the end. Then

$$\boldsymbol{N}_{t+1} = L\left(\boldsymbol{N}_t + \frac{G_t^F}{2}\right) + \frac{G_t^F}{2}.$$

- Approximation 4: Assume all the migration happens half-way through the interval. Then

$$\begin{aligned} \boldsymbol{N}_{t+1} &= L^{\frac{1}{2}}\left(L^{\frac{1}{2}}\boldsymbol{N}_t + G_t^F\right) \\ &= L\boldsymbol{N}_t + L^{\frac{1}{2}}G_t^F \end{aligned}$$

  where $L^{\frac{1}{2}}$ is a matrix square root of L such that $L^{\frac{1}{2}}L^{\frac{1}{2}} = L$.
  - However, $L^{\frac{1}{2}}$ does not seem to be a real matrix in general.
- $L^{\frac{1}{2}}$ could be approximated by $\frac{1}{2}(I + L)$.
  - This leads to the same result as Approximation 3.

Some Issues

- There are two sexes (!)
  - Complicates formalism, but basic approach still works
  - Sex ratio at birth important
  - See Preston et al, Section 6.3.2.
- Leslie matrix model is deterministic, but reality is stochastic
  - Numbers of births and deaths have (at least) binomial variation
  - Leslie matrix model ignores this
  - But human populations are usually large, so not a big problem
- Fertility and mortality rates vary with time
  - For projection, future fertility and mortality rates are unknown and uncertain.
  - Cohort-component method ignores this.
  - If future fertility and mortality rates vary but are known, then L becomes L_t, and so

$$\boldsymbol{N}_{t+k} = (L_{t+k-1}L_{t+k-2}L_{t+1}\ldots L_t)\boldsymbol{N}_t = \left(\prod_{i=k-1}^{0} L_{t+i}\right)\boldsymbol{N}_t.$$

Issues (ctd.)

- Numbers of past and present births and deaths unknown in many countries
- The HIV/AIDS epidemic has changed the *pattern* of age-specific mortality in many countries in recent decades.

Modeling Age-Specific Rates: Purposes

- To fill in a whole table from partial information
  - E.g. in many developing countries without good vital registration systems, much of what we know is from surveys,
  - in the form of child mortality (${}_5q_0$) from surveys of mothers,
  - and adult mortality ($\sim {}_{45}q_{15}$), from the sibling method.
- to generate age-specific rates for population projections
- to identify suspicious data
- to reconstruct past rates from fragmentary data

Modeling Age-Specific Rates: Approaches

- Parametric models
- Model life tables and relational models
- Lee-Carter model and extensions

Parametric Models of Mortality: Gompertz and Makeham Models

- $\mu(x) = \alpha e^{\beta x}$ (Gompertz 1825).
- or log-linear model: $\log \mu(x) = \log(\alpha) + \beta x$.
- Improvement (Makeham 1860): $\mu(x) = \gamma + \alpha e^{\beta x}$
- Work very well for ages 50+.
  What two major features do they miss?
  1. Infant and child mortality
  2. Young adult mortality hump

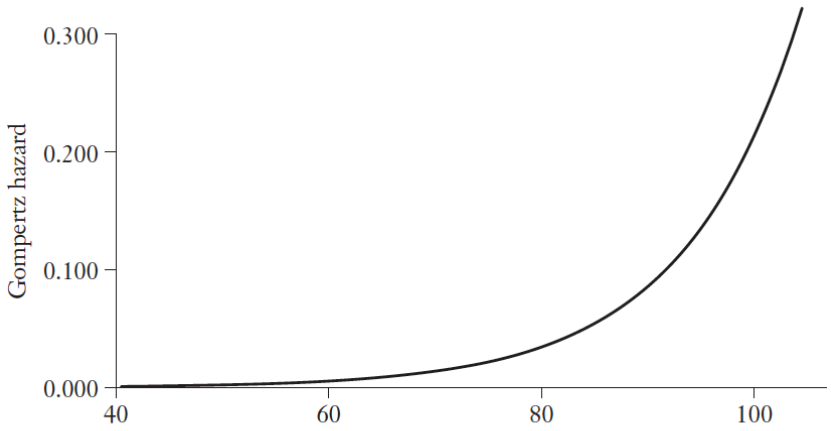


Figure 3.5 A Gompertz hazard function

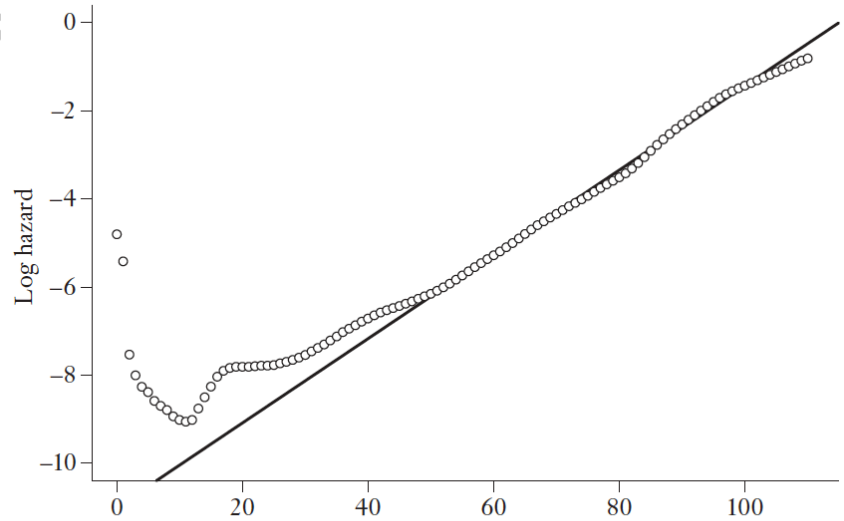


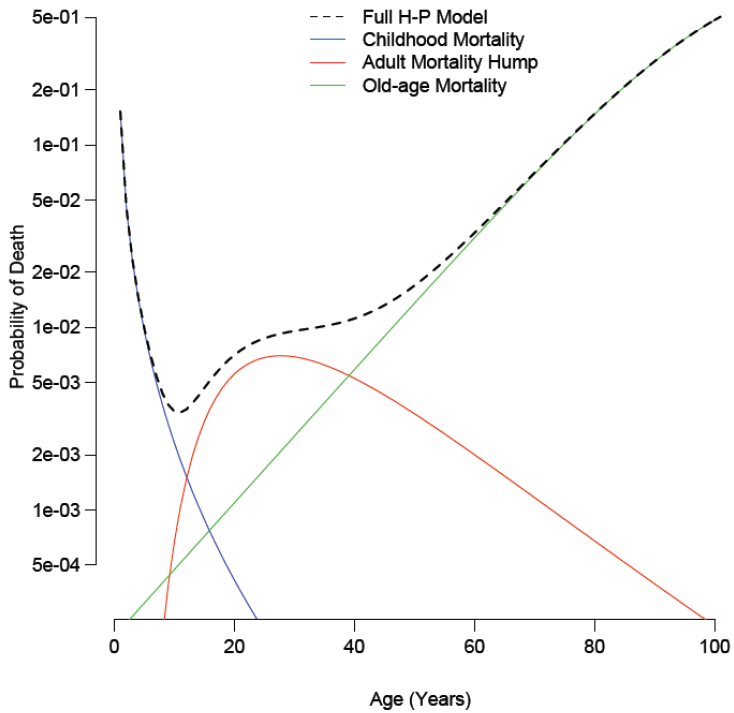
Figure 3.6 Logarithm of the hazard function

Heligman-Pollard Model

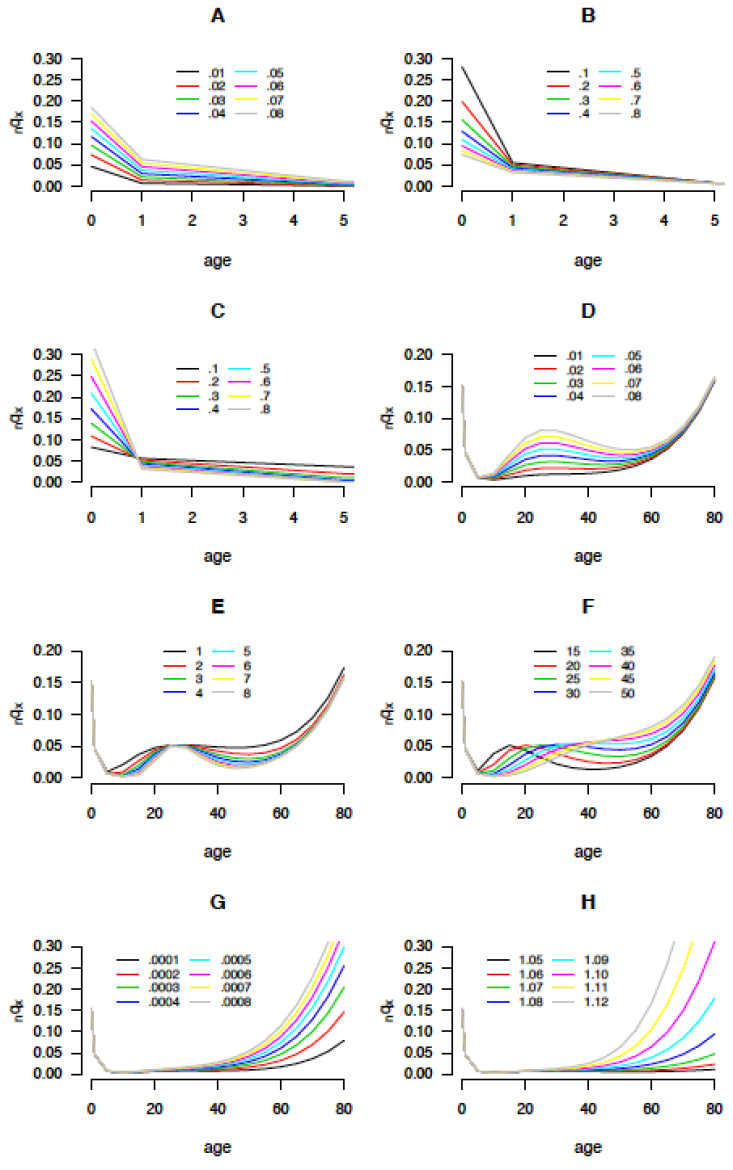
- Realistic 8-parameter model:

$$\frac{q_x}{1-q_x} = A^{(x+B)^{-C}} + De^{-E\,(\log(x)-\log(F))^2} + GH^x,$$

  where A, B, C, D, E, F, G, H are positive constants.
  - 1st term = child mortality
  - 2nd term = adult accident hump
  - 3rd term = Old-age mortality
- Decomposition of H-P curve (Sharrow et al 2013):

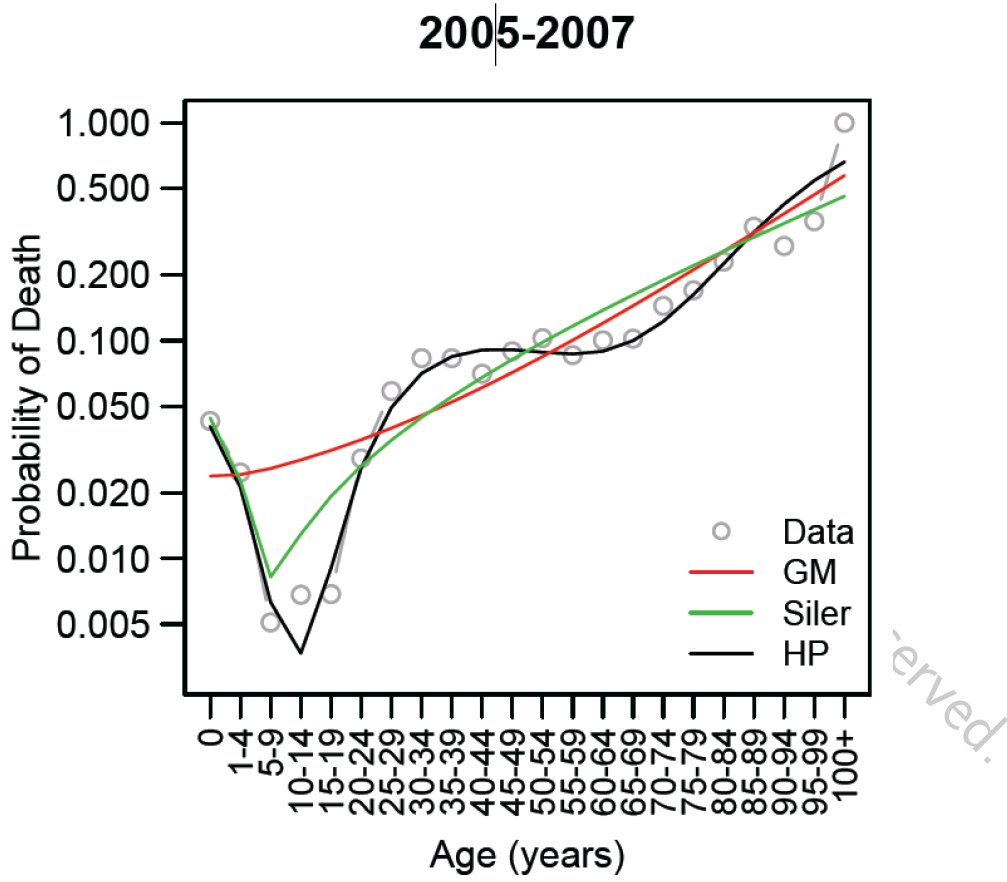


Heligman-Pollard Parameters: Variation Plots



Heligman-Pollard Model: Fit to Data

- Used by Sharrow et al (2013) to model HIV/AIDS mortality in South Africa.
- Maximum likelihood and Bayesian estimation methods developed.
- R package: `MortalityLaws`
- Performed better than other parametric models.



Model Life Tables

- Coale-Demeny (1966, 1983) life tables:
- Based on 326 empirical mortality schedules, mostly European.
- They were divided into 4 subsets: North, South, East, West (Europe).
  - Now the Coale-Demeny West model life table is the main one used.
- Model schedules derived for each subset.
- Implemented in the `demogR` R package (functions `cdmlt*`)

Brass Relational Model

(Brass 1971)

- Brass (1971) relational model:
- Starting from a model life table with survival probabilities $S^*(x)$:

$$\text{logit}(S_{\alpha,\beta}(x)) = \alpha + \beta \, \text{logit}(S^*(x)).$$

- Yields a family of life tables.
- α and β can be estimated from data.
- R code in Applied Demography Toolbox

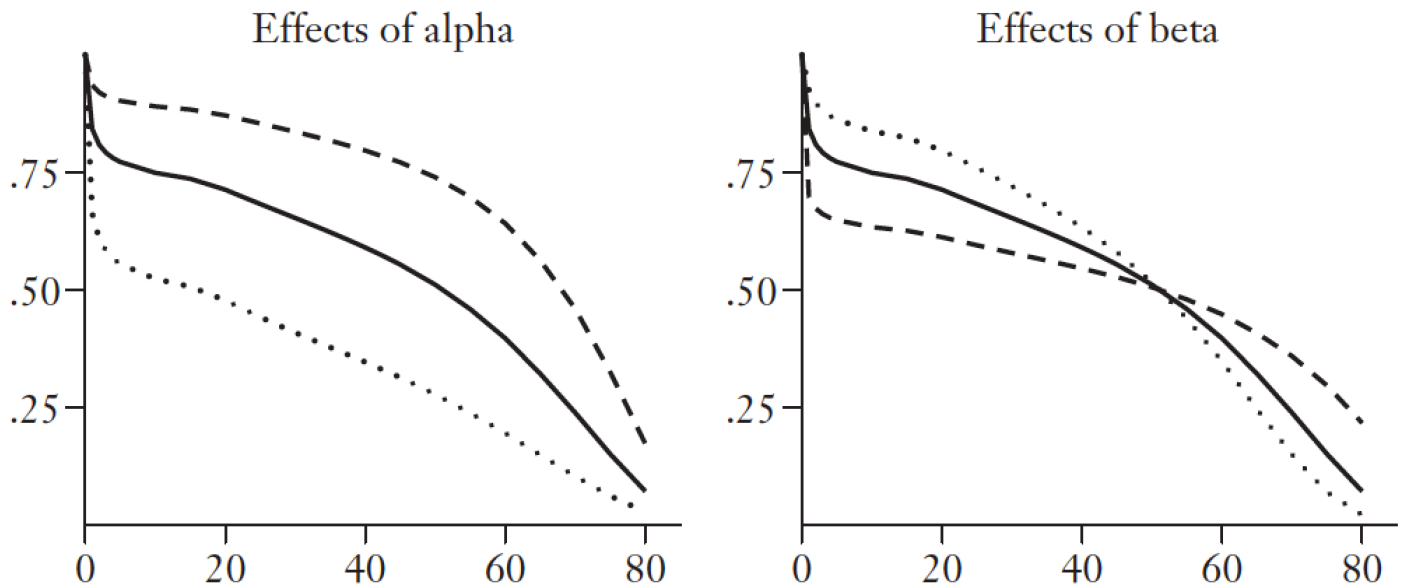


Figure 7.2 Shapes of Brass model lifetables

Lee-Carter Model

- Needs age-specific mortality rates for at least 3 time points
- Model:

$$\begin{aligned} \log(m_{x,t}) &= a_x + k_t b_x + \varepsilon_{x,t} \\ \varepsilon_{x,t} &\overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2) \\ k_t &= k_{t-1} - \eta + \delta_t \\ \delta_t &\overset{\text{iid}}{\sim} N(0, \sigma_\delta^2) \end{aligned} \tag{8}$$

- Interpretation:
  - a_x = baseline log-mortality rate
  - k_t = scalar for period t, called the mortality index
  - b_x reflects the pattern of change in mortality at age x
  - k_t modeled as a random walk with negative drift, $-\eta$.
  - So, in expectation, k_t is linear:

$$E[k_t|k_1] = k_1 - (t-1)\eta.$$

Lee-Carter Model: Identification

- Identification:
  - $\log(m_{x,t}) = a_x + k_t b_x + \varepsilon_{x,t}$
  - If b_x is replaced by $b_x c$ for all x, and k_t is replaced by k_t/c for all t, for any scalar c, then the model is unchanged. Any other versions?
  - Also if a_x is replaced by $a_x - b_x c$ for each x, and k_t is replaced by $k_t + c$, for any scalar c, then the model is unchanged.
- Identifying constraints:
  - $\sum_{x=0}^{A-1} b_x = 1$. item $\sum_{t=1}^{T} k_t = 0$.
  - (Not the only possible set of constraints).

Lee-Carter Model: Approximate Least Squares Estimation

- $\log(m_{x,t}) = a_x + k_t b_x + \varepsilon_{x,t}$
- Let us sum this (i.e. Eq. (8)) over t. Then the identifying constraints suggest

$$\hat{a}_x = \frac{1}{T} \sum_{t=1}^{T} \log(m_{x,t}), \quad \forall x.$$

- Similarly, summing over x and taking expectations, we get

$$\hat{k}_t = \sum_{x=0}^{A-1} [\log(m_{x,t}) - a_x] \quad \forall t.$$

- $\hat{b}_x$ is then obtained as the least squares regression coefficient in the regression of $[\log(m_{x,t}) - a_x]$ on k_t with zero intercept, for each x.

Estimating Parameters of the Random Walk Model for k_t

- The parameters of the random walk model for k_t, i.e. the drift η and the variance σ_δ^2, are estimated by standard time series methods, given the estimated values of k_t:

$$\begin{aligned} \hat{\eta} &= \frac{k_1 - k_T}{T-1} \\ \hat{\sigma}_\delta^2 &= \frac{1}{T-1}\sum_{t=2}^{T}(k_t - k_{t-1} + \hat{\eta})^2 \end{aligned}$$

  - Or

$$\hat{\sigma}_\delta^2 = \frac{1}{T-2}\sum_{t=2}^{T}(k_t - k_{t-1} + \hat{\eta})^2$$

    can be used.
  - (With denominator $T-2$ it's unbiased; with $T-1$ it minimizes mean squared error.)
- Implemented in the `MortCast` and `demography` R packages.

Singular Value Decomposition

- Singular Value Decomposition (SVD):
  - Let C be an $A \times T$ matrix of rank r. (If there are no linear dependencies, then $r = \min\{A, T\}$.)
  - Then, in general,
$$C = U\Lambda V^T, \tag{9}$$
  - where U is an $A \times r$ matrix, V is a $T \times r$ matrix, and
  - both U and V are *orthonormal*, i.e. $U^T U = I$ and $V^T V = I$.
  - Λ is a diagonal $r \times r$ matrix with positive elements.
- (9) can also be written
$$C = \sum_{j=1}^{r} \lambda_j u_j v_j^T, \tag{10}$$

  where u_j is the j-th column of U (j-th left singular vector of C), and v_j is the j-th column of V (j-th right singular vector of C).
  - If the columns of C are centered, then u_j is the normalized j-th principal component of the data in the matrix C (where the columns of C correspond to variables).

Least Squares Estimation by SVD

- Let the (x, t) element of C be $[\log(m_{x-1,t}) - \hat{a}_{x-1}]$.
  - Ignoring the error term, the Lee-Carter model (8) can then be written

$$C = b \cdot k^T,$$

    where $b = (b_0, \ldots, b_{A-1})^T$ and $k = (k_1, \ldots, k_T)^T$.
  - This is a rank-? SVD of C? Rank-one.
  - So b and k can be estimated by the first (dominant) left and right singular vectors of an SVD of the matrix C.
  - The Lee-Carter model can be viewed as a reduced-rank (rank one) approximation of the matrix of log-mortality rates.
- b is a normalized version of the first principal component of the dataset of mortality patterns, after removal of the mean mortality rate.
  - Nice explanation of SVD for mortality by Clark (2019)
- Can be directly implemented using the `svd()` R function.
  - Also implemented in the `demography` R package.

Lee-Carter Model: Other Estimation Methods

- Maximum likelihood estimation with Gaussian and Poisson errors (Renshaw & Haberman 2003, JRSS C).
  - Implemented in `ilc` R package.
  - Caution needed with use of Poisson errors (default in `ilc`), because they don't account for overdispersion.
- Bayesian estimation: Pedroza (2006); Wisniowski et al (2015).
- Lee-Carter is for one sex.
  - Could be used for each sex separately
  - Big problem?
    - In reality male $>$ female mortality at (almost) all ages
    - But Lee-Carter separately gives cross-overs
    - Solution: "coherent Lee-Carter" (Li and Lee 2005)
    - Implemented in `MortCast`

Lee-Carter Model: Pros

- Fitted 20th century US mortality data very well
- Provides probabilistic forecasts of mortality
- Focuses uncertainty on a one-number summary of mortality for each year (a major intellectual contribution).
- Now the dominant framework
  - (although modifications used in practice).

Lee-Carter Model: Cons

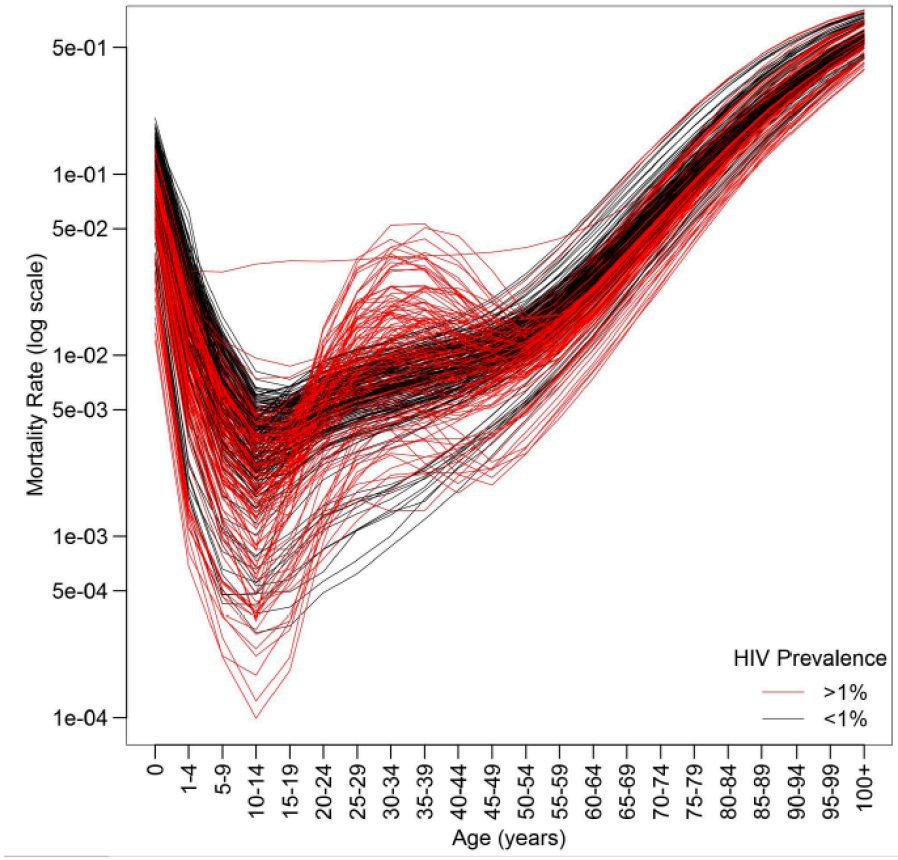
- It requires full age-specific mortality data for at least 3 time points. For many countries, this is not available.
- It is linear in time on the log-mortality rate scale:
  - $\log(m_{x,t}) = a_x + k_t b_x + \varepsilon_{x,t}$
  - $E[k_t|k_1] = k_1 - (t-1)\eta$
  - However, empirically progress in mortality has been more linear on the e_0 scale than on the scale of k_t (Lee & Miller 2001).
  - Lee & Miller (2001) proposed a modified method in which $e_{0,t}$ is forecast linearly and k_t is then fit to the forecasted $e_{0,t}$ numerically.
  - An implication is that as time goes by and mortality declines, progress (defined as percentage decline per year in mortality, or equivalently absolute decline on the log scale) accelerates at older ages relative to younger ages,
  - while Lee-Carter implies that progress accelerates/stays the same/ decelerates at older compared to younger ages.? The same.
  - Li, Lee & Gerland (2013) proposed a modified Lee-Carter method in which the b_x vector rotates over time, to address this.
  - Implemented in MortCast

- The Lee-Carter method assumes that at each age x, progress in overall mortality is constant over time in expectation.
  - However, in practice mortality improves slowly for countries at low levels of e_0, fast for countries at middle levels of e_0 (around 60), and more slowly for countries at the highest levels (over 80).
  - This can be seen only in data from multiple countries, including developing ones.
  - The Lee-Carter model specification reflects the fact it was developed for one country, the US.
  - Addressed by UN Bayesian model (Raftery, Chunn et al 2013). See later.
- Demo 2

Lee-Carter for HIV/AIDS Epidemics

(Sharrow et al 2014, PLoS One)

- Issue: Age pattern different with HIV epidemic
  - Large mortality hump at ages 25-50; size changes with HIV prevalence
  - Female 5-year mortality rate schedules for 40 countries with generalized HIV epidemics, 1970–2010:



Lee-Carter for HIV/AIDS Epidemics: Model

- Data on 320 five-year life tables for each sex for the 40 countries with generalized HIV/AIDS epidemics, from WPP.
- Use 3 singular vectors instead of 1 as in original Lee-Carter:

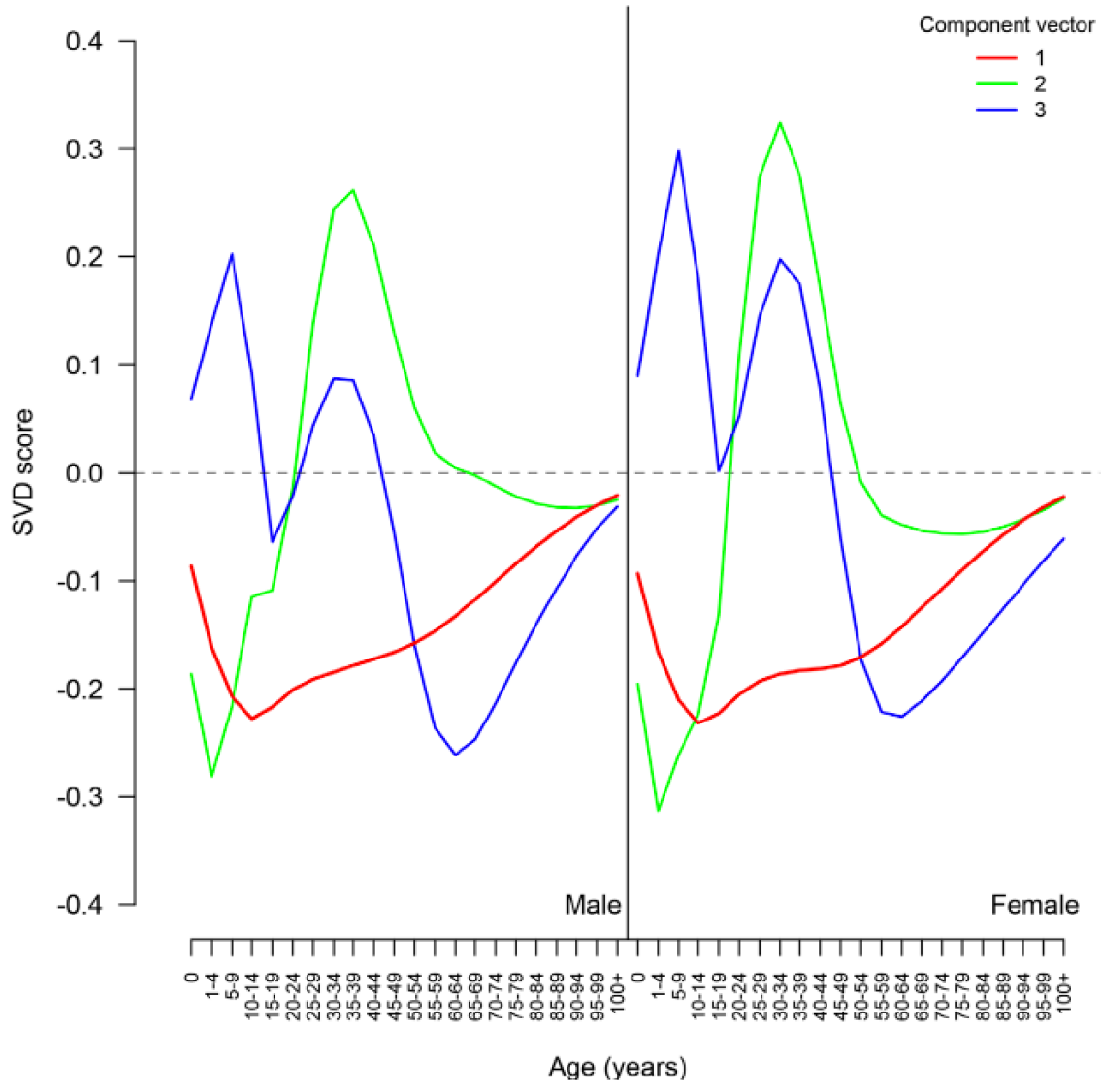
$$\log(m_{x,\ell}) = c_\ell + \sum_{j=1}^{3} k_{j,\ell} b_{j,x} + \varepsilon_{x,\ell},$$

  where ℓ indexes the life tables (not just time), and x indexes age-group *and* sex.
- Weight, $k_{j,\ell}$, of the j-th singular vector for the ℓ-th table modeled as a function of HIV prevalence and e_0:

$$k_{j,\ell} = \alpha_{0,j} + \alpha_{1,j}\mathrm{HIV}_\ell + \alpha_{2,j} e_{0,\ell}.$$

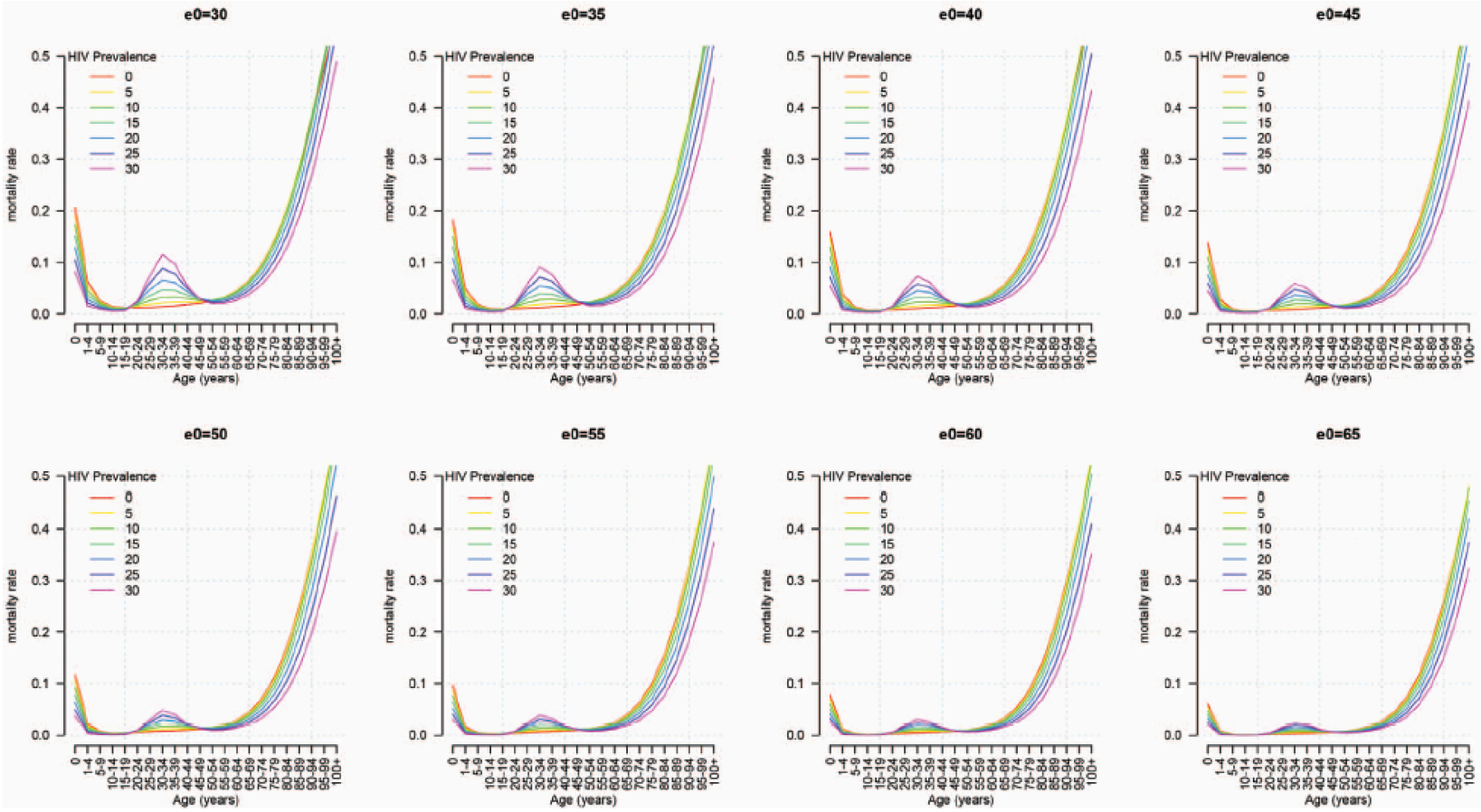
Lee-Carter for HIV/AIDS Epidemics: 3 Components

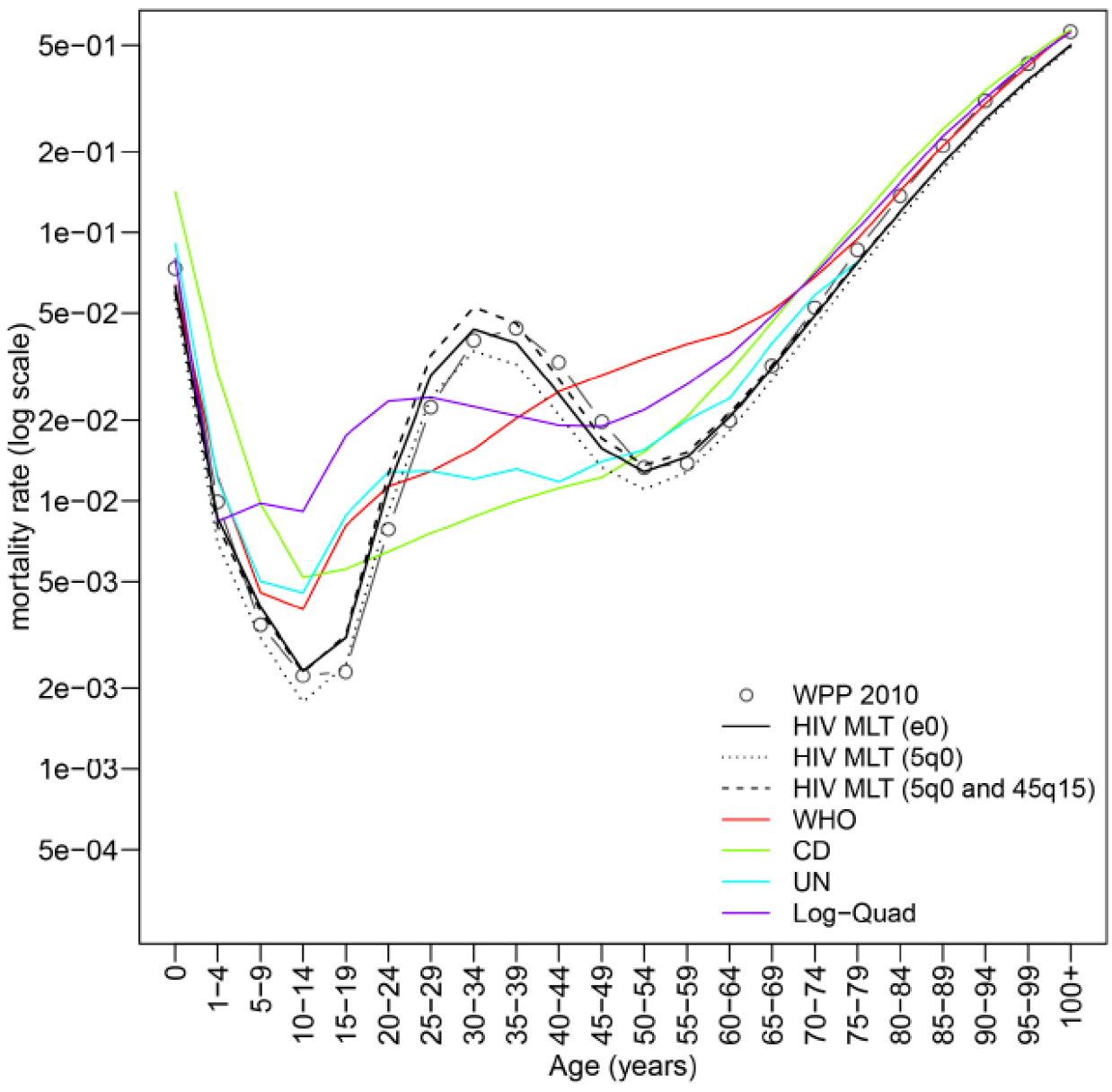
(left singular vectors)



Lee-Carter for HIV/AIDS Epidemics: Fitted Mortality Rates

- Fitted mortality rates at varying HIV prevalence and e_0:




Lesotho 2005−2010
mortality rate (log scale)
5e−01
2e−01
1e−01
5e−02
2e−02
1e−02
5e−03
2e−03
1e−03
5e−04
0
1−4
5−9
10−14
15−19
20−24
25−29
30−34
35−39
40−44
45−49
50−54
55−59
60−64
65−69
70−74
75−79
80−84
85−89
90−94
95−99
100+
Age (years)
WPP 2010
HIV MLT (e0)
HIV MLT (5q0)
HIV MLT (5q0 and 45q15)
WHO
CD
UN
Log−Quad

Lee-Carter Model: Functional Data Analytic Extensions

(Hyndman & Ullah 2007)

- Continuous age: b_x replaced by continuous function of age
- More than one singular vector used.
- Extended to fertility

Application to Fertility

- Lee & Tuljapurkar (1994): Lee-Carter model doesn't work so well for fertility.
- Application to fertility using 3 singular vectors: (Pantazis & Clark 2018, PLoS One)

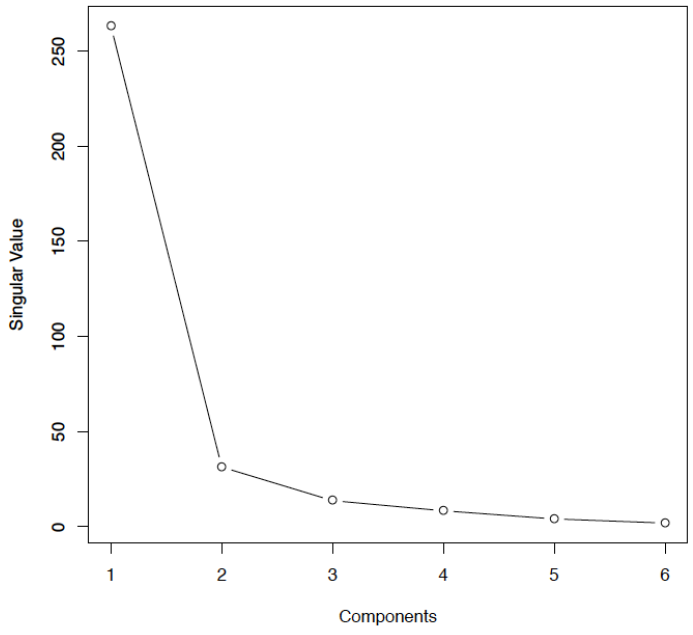


Figure 2: Scree plot of singular values from SVD factorization of age-specific fertility schedules.

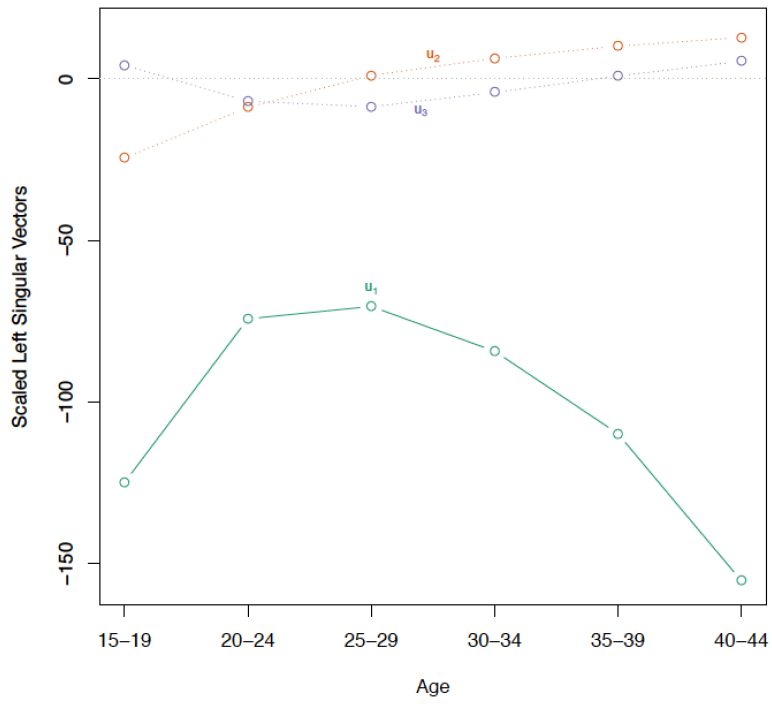


Figure 3: First three components (scaled left singular vectors) from SVD factorization of age-specific fertility schedules.

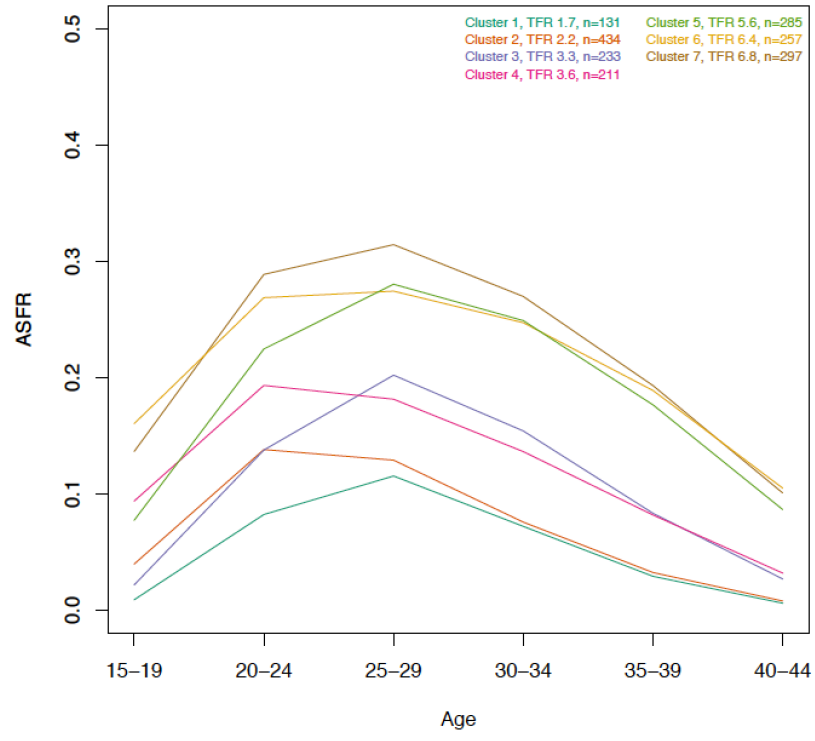


Figure 10: Median age-specific fertility schedules by cluster