Microtheses and Nanotheses provide space in the Newsletter for current and recent research students to communicate their research findings with the community. We welcome submissions for this section from current and recent research students. See newsletter.lms.ac.uk for preparation and submission guidance.

Microthesis: Combining data sources to develop a fertility projection model

Joanne Ellison

Fertility projections are a key determinant of population projections, which are widely used by government policymakers and planners. They are also vital to anticipate demand for maternity services and school places. My thesis presents a fertility projection model that combines individual-level and population-level data sources to exploit their opposing strengths.

Individual-level fertility data

Individual-level fertility data is often extracted from surveys collecting retrospective fertility histories and additional information from women in the population. The date of birth tends to be recorded imprecisely to the nearest month or year. A key feature is the presence of the five clocks, namely age, period, cohort, time since last birth (TSLB) and parity. Parity is the number of children previously borne, and we take the reproductive age range to be 15-44.

We illustrate three of the clocks in Table 1 (rows 2-4) for a woman aged 41 who had single births at ages 25 and 27, and twins at age 31. As the woman is currently 41, this year of age is not fully observed, and so she contributes 26 sets of observations (or person-year records) out of 30, from ages 15-40. Each person-year record has a corresponding binary response variable B equal to 1 if the woman had a birth at that age and 0 otherwise (see row 1 of Table 1). The person-year record then also contains the values of any extra variables included in the model.

B	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
Age	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Parity	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2
TSLB	-	-	-	-	-	-	-	-	-	-	-	1	2	1	2

B	0	1	0	0	0	0	0	0	0	0	0				
Age	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
Parity	2	2	4	4	4	4	4	4	4	4	4				
TSLB	3	4	1	2	3	4	5	6	7	8	9				

Table 1. Binary response variable B and three of the five clocks for a hypothetical woman aged 41.

Population-level fertility data

Whereas individual-level data only covers a sample of the population, population-level data includes everyone. The Office for National Statistics (ONS) publishes fertility rates by parity for England and Wales (tinyurl.com/rpz2f44x). This dataset uses mid-year population estimates and birth registration data. The rates are indexed by age, period, cohort and parity only, unlike the individual-level data where we have time since last birth and a range of survey variables. In this way the two levels of data complement each other, with a shortcoming of one (e.g. sample size, detail) being a strength of the other.

Modelling individual-level data

Individual-level data informs the base of our model, namely responses from Wave 1 (2009-11) of a survey called the UK Household Longitudinal Study or UKHLS (tinyurl.com/3kf2n6t3). Our sample of 18,218 women (357,287 person-years) were born in 1945-1992 and resided in England or Wales when interviewed. For parities 0, 1, 2 and 3+, we learn about the smooth dependence of B on age, cohort and time since last birth, as well as the effect of highest educational qualification. We do this by fitting logistic generalized additive models (GAMs) (see "GAMs", and [1] for an excellent introduction).

We henceforth focus on parity 0 and model the probability of a first birth. Our chosen model includes smooth effects of age, cohort, and their interaction; it also includes the categorical highest qualification variable and its smooth interaction with age.

GAMs

Just as a generalized linear model (GLM) generalizes linear regression to allow the response variable distribution to be non-normal, a GAM generalizes a GLM to allow the response to depend on smooth functions of covariates in some way.

Let Y_i be the ith response variable, following an exponential family distribution and with $\mu_i \equiv E(Y_i)$. A GAM has the following form

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{1i}, x_{3i}) + \dots$$

where g is a monotonic link function, $\mathbf{X}_i$ is the ith row of the fixed effects model matrix, $\boldsymbol{\beta}$ are the corresponding parameters, and the f_j are smooth functions of covariates x_k. These can be one-dimensional curves or higher-dimensional surfaces.

Integrating population-level data

To incorporate the ONS data into our individual-level model, we marginalize over the single additional covariate, i.e. qualification (Q). For each age and cohort, this amounts to taking a weighted average of the Q-specific probabilities, where the weights are the probabilities of a woman belonging to each Q category given her age and cohort. We model these probabilities using multinomial logistic regression. Lastly, we weight the contributions of the two data sources according to our prior beliefs about their relative importance. Terming these 'integrated models', we fit several of them, varying the weights.

Results

In Figure 1 we plot the mean probabilities of a first birth by age and cohort for all women and two of the qualification categories (GCSE, Degree). The model in the first row only uses UKHLS data (100% UKHLS, 0% ONS), while the second balances the information in the data sources roughly equally (50/50).

Incorporating the ONS data closely aligns the "All" probabilities to the observed ONS rates. In particular, we see that the quickly increasing trends forecast by the 100/0 model completely reverse to slow declines.

The Q-specific forecasts show us how changes are shared across population subgroups. The recent teenage fertility declines are strongly felt by the GCSE category, whose bimodal curves change to unimodal for the youngest cohorts, approaching those of the Degree category. Such additional insights demonstrate the value and richness of integrating different levels of data for demographic forecasting. For further details on this work, see [2].

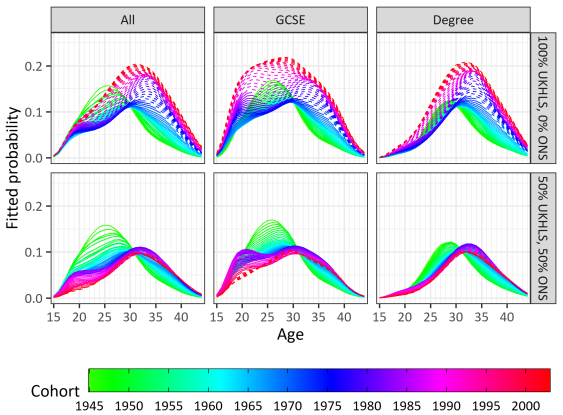

Figure 1. Mean conditional probabilities of a first birth for various integrated models; dashed lines indicate forecasts.

Acknowledgements

This work was funded by EPSRC (award 1801045), and partly supported by the ESRC FertilityTrends project (grant ES/S009477/1) and the ESRC Centre for Population Change - phase II (grant ES/K007394/1).

FURTHER READING

[1] S. N. Wood. *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall/CRC Press, 2017.

[2] J. Ellison. Stochastic modelling and projection of age-specific fertility rates. PhD Thesis, *University of Southampton*, 2021. eprints.soton.ac.uk/450468/

Joanne Ellison

Joanne is a Research Fellow at the University of Southampton, and is currently working on the ESRC-funded FertilityTrends project. Her main research interests are in developing statistical models for forecasting demographic processes. Despite being a big tennis fan, she is yet to pick up a racket herself.