Consistent Projection of Fertility for Subnational Areas Using a Logistic Curve

Extended abstract for poster by Tom McDevitt, Chief, Population Studies Branch, International Programs Center, Population Division, U.S. Census Bureau

The process of preparing assumptions for subnational projections carries with it a number of challenges not faced by the demographer involved in the process of preparing assumptions for national projections. These challenges include the requirements of consistency, not only between projected subnational and national populations but, for cohort-component projections, between assumed fertility, mortality, and migration at regional and national levels. Other challenges include inter-regional consistency, which might more accurately be referred to as the plausibility of assumptions for each region in light of trends in the determinants of, or past trends in, the components of change, and zero net internal migration for all regions taken together. Finally subnational projections require allocation of international migration across subnational regions.

This poster addresses the problem of projecting fertility for subnational regions in a developing country. Forcing consistency between national and subnational numbers of births (and, therefore, between national and subnational fertility) can be accomplished in several ways – by the addition of regional projections; by creating a residual projection for one region reflecting the difference between a national control projection and the sum of the other regional projections; or by proportionally adjusting fertility, mortality, and migration of all regions to be consistent with the national control projection levels in the projection base year and in each year of the projection. But these calculations can be made regardless of whether or not projected fertility at the regional level is plausible. This poster discusses the use of linear and logistic extrapolation of total fertility rates (TFRs), and the use of two methods based on the logistic, for helping to ensure consistency in assumed fertility across subnational areas and between subnational and national projected fertility. *A priori*, we would prefer a fertility projection methodology that does not permit projected TFR for a relatively less developed region with a recent, relatively rapid decline in fertility, from overtaking projected TFR for a more developed region (such as a national capital region) with lower but more slowly declining fertility. Ultimately, the test of the model used to project fertility should involve comparison of projected and empirical fertility across regions of a country or series of countries.

The analysis reported in this poster involves a series of comparisons between alternative functional specifications for predicted total fertility rates focusing on predictive accuracy of assumed TFR trends at the regional level. While any of a number of functional specifications have been used to project fertility (such as a straight line based on recent empirical estimates, an exponential or logarithmic function, a polynomial, and a sine curve), the logistic function has qualities that make it especially appropriate for the task. The upper and lower asymptotes of the logistic control the steepness with which projected fertility approaches some demographically plausible lower bound. The upper part of the logistic curve's characteristic S-shape accommodates the onset of fertility decline. The lower part accommodates the end of fertility decline as fertility approaches the lower bound. The middle segment is nearly linear, accommodating observed fertility for countries, and regions within countries, that are neither at the beginning nor at the end of their fertility transitions.

Comparisons are made in the poster between predicted and empirical TFR for the regions of 22 developing countries for which three or more Demographic and Health Surveys have been

conducted using a linear extrapolation of TFR as a reference fertility projection, and three variants of the logistic curve.

The specific questions asked in the poster are:

- Does the logistic curve improve predictive accuracy for assumed TFR at the subnational level for these developing countries?
- Is predictive ability of the logistic model further improved by using variants of the logistic that force greater consistency between assumed national and subnational trends in TFR?

Predictive accuracy is measured using mean absolute percent error (MAPE) between the 3-year average regional TFR from the latest DHS and the trendline-based estimate of TFR for the same date where the trendline is fitted to estimated TFRs from the first two DHSs.

Logistic Models

The first logistic variant fitted and used expresses predicted TFR as a function of defined lower and upper asymptotes and estimated slope and intercept of a regression line fitted to the logits of observed, dated TFRs.

$$\hat{TFR}_t = \left(\frac{e^{a+bt}}{1+e^{a+bt}} \right) * (UB - LB) + LB$$

where UB is the upper asymptote for the logistic function
LB is the lower asymptote for the logistic function
b is the slope of a line fitted to the logit transformations of observed TFRs
a is the intercept of a line fitted to the logit transformations of observed TFRs
t is time in years

and the logit of TFR is

$$\ln\left[\frac{(TFR - LB)/UB - LB)}{1 - (TFR - LB)/(UB - LB)} \right]$$

This variant is used to project TFR for both national population and subnational (regional) population.

The second variant, referred to in the poster as the "constrained" logistic, is defined in terms of the complement of the TFR; that is, of the difference between TFR in year t and the lower bound of the logistic. The specification ensures that proportionate changes in regional TFR track proportionate changes in (projected) national TFR and, therefore, that regional TFR tracks (projected) national TFR. In most cases, the national TFR trend used to guide the constrained logistic fitting for regions is the unconstrained logistic described above (first logistic variant). The constrained logistic is written:

$$\hat{TFR}_{t,R} = LB - (LB - TFR_{t-1,R}) * \frac{(LB - TFR_{t,N})}{(LB - TFR_{t-1,N})}$$

where $TFR_{t,N}$ is the projected TFR for country N in time t

$\hat{TFR}_{t,R}$ is the projected TFR for region R in time t

This specification is equivalent to

$$\left(\frac{LB - TFR_{t,R}}{LB - TFR_{t-1,R}}\right) = \left(\frac{LB - TFR_{t,N}}{LB - TFR_{t-1,N}}\right)$$

The third variant, referred to in the poster as the "fixed slope" variant, projects TFR for region R using the same functional form as the first variant but with the slope taken from the fitted equation from the national TFR trendline. That is,

$$\hat{TFR}_{t,R} = \left(\frac{e^{a+Bt}}{1+e^{a+Bt}}\right) * (UB - LB) + LB$$

where B is the slope of a line fitted to the logit transformations of *national* TFRs
a is the intercept of a line fitted to the logit transformations of *regional* TFRs

This specification ensures that regional TFRs decline is not only consistent with, but matches, national TFR decline in every period. This specification does not admit the possibility that national TFR will not follow a logistic trend.

Results

The results of the exercise are summarized below:

- Logistic curves tend to out-perform linear extrapolation of TFR in predicting subsequent TFR.

- None of the models compared provided the smallest mean absolute percentage error (MAPE) in every country; however, the constrained logistic and fixed-logistic models were better choices for minimizing error in projected TFR than the linear or simple logistic extrapolations. The answer to the second question posed at the outset of the poster – "Is predictive ability of the logistic model further improved by using variants of the logistic that force greater consistency between assumed national and subnational trends in TFR?" – is yes.

- For countries where both national and regional fertility are following a monotonically declining trend, where both national and regional fertility trends might be assumed to be approaching a common minimum TFR level (and, therefore, are reasonably

modeled using a logistic curve), and where there is no evidence suggesting that regional rankings are likely to change, the fixed-slope logistic model is the optimal choice for projecting subnational fertility. However, constrained modeling of regional TFR trend is recommended for countries undergoing stalled or reversals in fertility decline.

- The constrained and fixed-slope logistic models marginally out-perform the simple logistic and linear models for longer projection periods, where independently projected TFR for one or more regions may deviate substantially from the national trend if left unconstrained.